Preface

The enormous progress over the last decades in our understanding of the mechanisms behind the complex system "Earth" is to a large extent based on the availability of enlarged data sets and sophisticated methods for their analysis. Univariate as well as multivariate time series are a particular class of such data which are of special importance for studying the dynamical processes in complex systems. Time series analysis theory and applications in geo- and astrophysics have always been mutually stimulating, starting with classical (linear) problems like the proper estimation of power spectra, which has been put forward by Udny Yule (studying the features of sunspot activity) and, later, by John Tukey.

In the second half of the 20th century, more and more evidence has been accumulated that most processes in nature are intrinsically non-linear and thus cannot be sufficiently studied by linear statistical methods. With mathematical developments in the fields of dynamic system's theory, exemplified by Edward Lorenz's pioneering work, and fractal theory, starting with the early fractal concepts inferred by Harold Edwin Hurst from the analysis of geophysical time series, nonlinear methods became available for time series analysis as well. Over the last decades, these methods have attracted an increasing interest in various branches of the earth sciences. The world's leading associations of geoscientists, the American Geophysical Union (AGU) and the European Geosciences Union (EGU) have reacted to these trends with the formation of special nonlinear focus groups and topical sections, which are actively present at the corresponding annual assemblies.

Surprisingly, although nonlinear methods have meanwhile become an established, but still developing toolbox for the analysis of geoscientific time series, so far there has not been a book giving an overview over corresponding applications of these methods. The aim of this volume is therefore to close this apparent gap between the numerous excellent books on (i) geostatistics and the "traditional" (linear) analysis of geoscientific time series, (ii) the nonlinear modelling of geophysical processes, and (iii) the theory of nonlinear time series analysis.

This volume contains a collection of papers that were presented in a topical session on "Applications of Nonlinear Time Series Analysis in the Geosciences" at the General Assembly of the European Geosciences Union in Vienna from April 15–20, 2007. More than 30 colleagues from various countries used this opportunity to present and discuss their most recent results on the analysis of time series from problems originated in the fields of climatology, atmospheric sciences, hydrology, seismology, geodesy, and solar-terrestrial physics. Oral and poster sessions included a total of 38 presentations, which attracted the interest of many colleagues working both theoretically on and practically with nonlinear methods of time series analysis in the geosciences. The feedback from both presenters and audience has encouraged us to prepare this volume, which is dedicated to both experts in nonlinear time series analysis and practitioners in the various geoscientific disciplines who are in need of novel and advanced analysis tools for their time series. In this volume, presentations shown at the conference are complemented by invited contributions written by some of the most distinguished colleagues in the field.

In order to allow the interested reader to easily find methods that are suitable for his particular problems or questions, we have decided to arrange this book in three parts that comprise typical applications from the fields of climatology, geodynamics, and solar-terrestrial physics, respectively. However, especially in the latter case, the assignment of the different subjects has not always been unique, as there are obvious and rather strong links to the two other fields. Moreover, we would like to note that there are methods whose application has already become very common for studying problems from either of these fields.

The first 7 chapters deal with problems from climatology and the atmospheric sciences. A. Gluhovsky discusses the potential of subsampling for the analysis of atmospheric time series, which usually cannot be described by a simple linear stochastic model. In such cases, traditional estimates of already very simple statistics can be significantly biased, a problem that can be solved by using subsampling methods. J. Mikšovský, P. Pišoft, and A. Raidl report results on the spatial patterns of nonlinearity in simulations of global circulation models as well as reanalysis data. S. Hallerberg, J. Bröcker, and H. Kantz discuss different methods for the prediction of extreme events, a challenging problem of contemporary interest in various geoscientific disciplines. D.B. Percival presents an overview about the use of the discrete wavelet transform for the analysis of climatological time series, with a special consideration of ice thickness and oxygen isotope data. G.S. Duane and J.P. Hacker describe a framework for automatic parameter estimation in atmospheric models based on the theory of synchronisation. W.W. Hsieh and A.J. Cannon report on recent improvements on nonlinear generalisations of traditional multivariate methods like principal component analysis and canonical correlation analysis, which are based on the application of neural networks and allow the extraction of nonlinear, dynamically relevant components. R. Donner, T. Sakamoto, and N. Tanizuka discuss methods for quantifying the complexity of multivariate

time series, and how such concepts can be used to study variations and spatio-temporal dependences of climatological observables. As a particular example, the case of Japanese air temperature records is considered.

The next 5 chapters describe the analysis of time series in the fields of oceanography and seismology. S.M. Barbosa, M.E. Silva, and M.J. Fernandes discuss the issue of characterising the long-term variability of sea-level records in the presence of nonstationarities, trends, or long-term memory. A. Ardalan and H. Hashemi describe a framework for the empirical modelling of global ocean tide and sea level variability using time series from satellite altimetry. J.A. Hawkins, A. Warn-Varnas, and I. Christov use different linear as well as nonlinear Fourier-type techniques for the analysis of internal gravity waves from oceanographic time series. M.E. Ramirez, M. Berrocoso, M.J. González, and A. Fernández describe a time-frequency analysis of GPS data from the Deception Island Volcano (Southern Shetland Islands) for the estimation of local crustal deformation. A. Jiménez, A.M. Posadas, and K.F. Tiampo use a cellular automaton approach to derive a simple statistical model for the spatio-temporal variability of seismic activity in different tectonically active regions.

The final 4 chapters discuss problems related to dynamical processes on the Sun and their relationship to the complex system "Earth". I.M. Moroz uses a topological method, the so-called template analysis, to study the internal structure of chaos in the Hide-Skeldon-Acheson dynamo, and compares her results with those for the well-known Lorenz model. N.G. Mazur, V.A. Pilipenko, and K.-H. Glassmeier describe a framework for the analysis of solitary wave signals in geophysical time series, particularly satellite observations of electromagnetic disturbances in the near-Earth environment. M. Paluš and D. Novotná introduce a nonlinear generalisation of singular spectrum analysis that can be used to derive dynamically meaningful oscillatory components from atmospheric, geomagnetic, and solar variability signals. Finally, R. Donner demonstrates the use of phase coherence analysis for understanding the long-term dynamics of the north-south asymmetry of sunspot activity.

We would like to express our sincerest thanks to those people who made the idea of this book becoming reality: the authors, who prepared their excellent results for publication in this book and the numerous referees, who helped us evaluating the scientific quality of all contributions and making them being ready for publication. We also acknowledge the support of Springer at all stages during the preparation of this book. We do very much hope that it will inspire many readers in their own scientific research.

Dresden / Porto,
January 2008

Reik Donner
Susana Barbosa

Contents

List of Contributors

AU: Please provide e-mail id for all the missing contributions.

Alireza A. Ardalan
Department of Surveying and Geomatics Engineering,
Center of Excellence in Surveying Engineering and Disaster Prevention,
Faculty of Engineering,
University of Tehran,
Tehran-Iran,
e-mail: ardalan@ut.ac.ir

Susana M. Barbosa
Universidade do Porto,
Faculdade de Ciências,
e-mail: susana.barbosa@fc.up.pt

Manuel Berrocoso
Laboratorio de Astronomía,
Geodesia y Cartografía.
Departamento de Matemáticas.
Facultad de Ciencias.
Universidad de Cádiz.

Jochen Bröcker
Max Planck Institute for the Physics of Complex Systems,
Nöthnitzer Str. 38, D 01187 Dresden,
Germany

Alex J. Cannon
Meteorological Service of Canada,
Environment Canada,
201-401 Burrard Street,
Vancouver, BC,
Canada V6C 3S5,
e-mail: alex.cannon@ec.gc.ca

Ivan Christov
Naval Research Laboratory,
Stennis Space Center,
MS 39529, USA;
Northwestern University,
Evanston, IL 60208, USA

Reik Donner
Institute for Transport and Economics,
Dresden University of Technology,
Andreas-Schubert-Str. 23, 01062 Dresden, Germany,
e-mail: donner@vwi.tu-dresden.de;
Graduate School of Science,
Osaka Prefecture University,
1-1 Gakuen-cho, Naka-ku,
Sakai-shi, 599-8531 Japan

Gregory S. Duane
National Center for Atmospheric Research, Boulder, CO,
e-mail: gduane@ucar.edu

AU: Please provide full affiliation.

Maria Joana Fernandes
Universidade do Porto,
Faculdade de Ciências

A. Fernández
Laboratorio de Astronomía,
Geodesia y Cartografía.
Departamento de Matemáticas.
Facultad de Ciencias.
Universidad de Cádiz.

Karl-Heinz Glassmeier
Institut für Geophysik und extraterrestrische Physik,
Technische Universität
Braunschweig, Germany,
e-mail: kh.glassmeier@tu-bs.de

Alexander Gluhovsky
Department of Earth & Atmospheric Sciences,
Department of Statistics, and Purdue Climate Change Research Center (PCCRC), Purdue University,
West Lafayette,
Indiana 47907, USA,
e-mail: aglu@purdue.edu

María José González
Laboratorio de Astronomía,
Geodesia y Cartografía.
Departamento de Matemáticas.
Facultad de Ciencias.
Universidad de Cádiz.

Joshua P. Hacker
National Center for Atmospheric Research, Boulder, CO

Sarah Hallerberg
Max Planck Institute for the Physics of Complex Systems,
Nöthnitzer Str. 38,
D 01187 Dresden,
Germany

Hassan Hashemi
Department of Surveying and Geomatics Engineering,
Center of Excellence in Surveying Engineering and Disaster Prevention,
Faculty of Engineering,
University of Tehran,
Tehran-Iran,
e-mail: hashemih@ut.ac.ir

Jim A. Hawkins
Planning Systems Inc.,
Slidell, LA 70458, USA,
e-mail: jhawkins@psislidell.com

William W. Hsieh
Department of Earth and Ocean Sciences,
University of British Columbia 6339 Stores road,
Vancouver, BC, Canada V6T 1Z4,
e-mail: whsieh@eos.ubc.ca

Abigail Jiménez
Department of Earth Sciences
Biological and Geological Sciences,
University of Western Ontario,
London, Canada,
e-mail: ajimene@uwo.ca

Holger Kantz
Max Planck Institute for the Physics of Complex Systems,
Nöthnitzer Str. 38,
D 01187 Dresden, Germany,
e-mail: kantz@pks.mpg.de

Nikolay G. Mazur
Institute of the Physics of the Earth,
Russian Academy of Sciences,
Moscow, Russia,
e-mail: n.g.mazur@mtu-net.ru

Jií Mikšovský
Department of Meteorology and Environment Protection,
Faculty of Mathematics and Physics,
Charles University,
Prague, Czech Republic,
e-mail:
jiri.miksovsky@mff.cuni.cz

Irene M. Moroz
Mathematical Institute,
24-29 St Giles',
Oxford OX1 3LB, UK
e-mail: moroz@maths.ox.ac.uk

Dagmar Novotná
Institute of Atmospheric Physics,
Academy of Sciences of the Czech Republic,
Boční II/1401,
141 31 Prague 4, Czech Republic,
e-mail: nov@ufa.cas.cz

Milan Paluš
Institute of Computer Science,
Academy of Sciences of the Czech Republic,
Pod vodárenskou věží 2,
182 07 Prague 8, Czech Republic,
e-mail: mp@cs.cas.cz

Donald B. Percival
Applied Physics Laboratory,
University of Washington,
Box 355640, Seattle, WA,
98195–5640, USA,
e-mail: dbp@apl.washington.edu

Viacheslav A. Pilipenko
Space Research Institute,
Russian Academy of Sciences,
Moscow, Russia,
e-mail: pilipenko_va@mail.ru

Petr Pišoft
Department of Meteorology and Environment Protection,
Faculty of Mathematics and Physics,
Charles University,
Prague, Czech Republic

Antonio M. Posadas
Department of Applied Physics,
University of Almería,
Spain,
e-mail: aposadas@ual.es

Aleš Raidl
Department of Meteorology and Environment Protection,
Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic

María Eva Ramírez
Laboratorio de Astronom'ıa,
Geodesia y Cartograf'ıa.
Departamento de Matemáticas.
Facultad de Ciencias.
Universidad de C'adiz.,
e-mail: mariaeva.ramirez@uca.es

Takahide Sakamoto
Electrical Engineering Course,
Osaka Municipal Ikuno Technical High School,
2-3-66 Ikuno-higashi, Ikuno-ku,
Osaka-shi, 544-0025 Japan,
e-mail:
taka_sakamoto@mem.iee.or.jp;
Graduate School of Science,
Osaka Prefecture University,
1-1 Gakuen-cho, Naka-ku, Sakai-shi,
599-8531 Japan

Maria Eduarda Silva
Universidade do Porto,
Faculdade de Ciências

AU: Please provide full affiliation.

Noboru Tanizuka
Graduate School of Science,
Osaka Prefecture University,
1-1 Gakuen-cho,
Naka-ku,
Sakai-shi,
599-8531 Japan,
e-mail:
tanizuka@mi.s.osakafu-u.ac.jp

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

Alex Warn-Varnas
Naval Research Laboratory,
Stennis Space Center,
MS 39529, USA

Part I

Applications in Climatology and Atmospheric Sciences

Subsampling Methodology for the Analysis of Nonlinear Atmospheric Time Series

Alexander Gluhovsky

Department of Earth & Atmospheric Sciences, Department of Statistics, and Purdue Climate Change Research Center (PCCRC), Purdue University, West Lafayette, Indiana 47907, USA, aglu@purdue.edu

Abstract. This contribution addresses the problem of obtaining reliable statistical inference from meteorological and climatological records. The common practice is to choose a *linear* model for the time series, then compute confidence intervals (CIs) for its parameters based on the estimated model. It is demonstrated that such CIs may become misleading when the underlying data generating mechanism is nonlinear, while the computer intensive subsampling method provides an attractive alternative (including situations when linear models are entirely out of place, e.g., when constructing CIs for the skewness).

Keywords: Nonlinear time series, Confidence intervals, Subsampling, Skewness

1 Introduction

Conventional statistical methods are commonly based on strong assumptions that are rarely met in atmospheric data sets (e.g., [1]). These include the assumption that observations follow a normal (Gaussian) distribution, or the assumption of a linear model for the observed time series. In fact, distributions of many meteorological and climatological variables are not normal, as the velocity field in a turbulent flow [2, 3], the precipitation amount or economic damage from extreme weather events [4, 5]. At the same time, it has become clear that while departures from normality, nonlinearities in real data generating mechanisms (DGMs) may render conventional statistical inference misleading, on the positive side, computer intensive resampling methods [6, 7, 8, 9] could bring about valid results without imposing questionable assumptions on DGMs (e.g., [10, 11]). This work describes how one such

method, subsampling [8], can be used in practice to obtain reliable inference from observed and modeled time series.

As a motivating example, consider the time series of the vertical velocity of wind recorded from an aircraft (50 m above Lake Michigan, 70 m/s flight speed, 20 Hz sampling rate) during an outbreak of a polar air mass over the Great Lakes region [12]. Figure 1 shows a record of 4096 data points (corresponding to about 14.3 km) that has passed the test for stationarity [13].

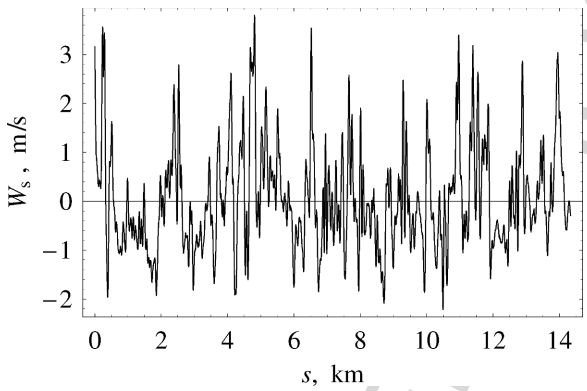


Fig. 1: Record of 20-Hz aircraft vertical velocity measurements.

The sample mean, variance, and skewness of the vertical velocity computed from the record are, respectively, 0.03, 1.11, and 0.84. Sample characteristics like these (routinely obtained in field programs as well as in laboratory experiments and computer simulations) are just point estimates (our "best guesses") of the true values of the parameters, so confidence intervals (CIs) are duly employed to learn how much importance is reasonable to attach to such numbers. However, since the DGM is usually unknown, the common practice is to assume a linear parametric model for it, then estimate the model from the observed record, and compute CIs for parameters of the underlying time series from the estimated model.

In Sect. 2, it is demonstrated via Monte Carlo simulations of a model nonlinear time series that nonlinearities in the *real* DGM may render useless the inference (90% CIs for the variance) based on the estimated linear parametric model. Then in Sect. 3, we show how the subsampling method [8] allows one to avoid time series analysis anchored in parametric models with imposed perceived physical assumptions. Previously [11, 14], we have considered subsampling CIs for the mean and variance of explicit nonlinear time series as well as the related enhancements of the subsampling methodology. Completely new in this paper are the analysis of the coverage of conventional CIs in case of an implicit nonlinear time series (those retaining the general

ARMA form but excited by nonnormal white noise [15]) and the construction of CIs for the skewness from a single record of limited length. The latter is of particular importance since CIs for the skewness cannot be obtained from linear models, while nonzero skewness and limited record lengths are frequent attributes of atmospheric time series, especially of those relevant to extreme weather and climate events.

2 Inadequacy of Linear Models

2.1 The Model

The data were generated using variations of the following nonlinear model [16],

$$Y_t = X_t + a(X_t^2 - 1), \tag{1}$$

where X_t is a first order autoregressive process (AR(1)),

$$X_t = \phi X_{t-1} + \epsilon_t, \tag{2}$$

$0 < \phi < 1$ is a constant, and ϵ_t is a white noise process (a sequence of uncorrelated random variables with mean 0 and variance σ_ϵ^2).

AR(1) with a Gaussian white noise is widely employed in studies of climate as a default model for correlated time series (e.g., [17, 18]). When the white noise in model (2) is not Gaussian, the model may exhibit nonlinear behavior and is referred to as an *implicit* nonlinear model [15], as opposed to an *explicit* nonlinear model (1).

If a record of length n is generated from model (2) with a Gaussian white noise (i.e., called a *linear* model here), then, say, a 90% CI for the variance of X_t is given by

$$\hat{\sigma}_X^2 \pm 1.645\sigma_X^2 \left(\frac{2}{n} \cdot \frac{1+\phi^2}{1-\phi^2}\right)^{1/2}, \tag{3}$$

where $\hat{\sigma}_X^2$ is the sample variance, an estimate of the *true* variance of X_t,

$$\sigma_X^2 = \sigma_\epsilon^2 / \left(1 - \phi^2\right). \tag{4}$$

Equation (3) follows from the fact that $\hat{\sigma}_X^2$ is asymptotically normal with mean σ_X^2 and variance $(2/n)\sigma_X^4 \left(1 + \phi^2\right) / \left(1 - \phi^2\right)$ [19]. Parameters ϕ and σ_ϵ are generally estimated from the data. Using our model example, it will be demonstrated in Sects. 2.2 and 2.3 that common practice of fitting a linear model to data that are generated (unknown to us) by an explicit or implicit nonlinear model, may result in invalid CIs, even though customary postfitting diagnostic checking indicates that the model provides an adequate description of the data.

Also, a linear model may match the first two moments (mean and variance) of the observed time series, but such model has a zero skewness, while a

nonlinear model may be capable of matching all three moments. Note that at $a = 0.14$, the mean, variance, and skewness of Y_t (nonlinear model (1)) are, respectively, 0, $1+2a^2 \approx 1.04$, and $6a+8a^3 \approx 0.86$, i.e., close to corresponding sample characteristics (0.04, 1.11, and 0.84) of the vertical velocity time series discussed in Sect. 1. Thus, Y_t might provide a better description for that series than linear models. At the same time, the subsampling method (see Sect. 3) may not require that *any* model, linear or nonlinear, be fitted to the time series under study to obtain reasonable CIs, though an approximating model is often desirable to make subsampling CIs more accurate.

2.2 Monte Carlo Simulations and Actual Coverage of Conventional CIs for the Variance of an Explicit Nonlinear Model

A 90% CI is the range of numbers that traps an unknown parameter with probability 0.90 called the *coverage probability*. This implies that if instead of one time series record commonly available in practice, the records could be generated over and over, and from each record a 90% CI was computed, then 90% of the resulting CIs would contain the parameter. Such coverage probability (often referred to as a *nominal* or *target* coverage probability, e.g., [7]) is attained only if all assumptions underlying the method for the CI construction are met. This is the case of CIs (3) when the data are generated from linear model (2). In geosciences, however, such assumptions are rarely met, so that the *actual* coverage probability may differ from the target level (sometimes considerably as is demonstrated below). Intervals with confidence levels other than 90% (e.g., 95% or 99%) are also often used in various applications (the higher the confidence level the wider the CI).

Using the above probabilistic interpretation, the actual coverage probability could be determined through Monte Carlo simulations when the DGM is known. Although this assumption is unrealistic, still Monte Carlo simulations of models possessing statistical properties shared by real processes may provide valuable information on what can be expected in situations of practical interest. With all this in mind, Monte Carlo simulations of the nonlinear model (1) were implemented as follows. First, 1000 records, each of 1024 observations, were generated from the model with $\phi = 0.67$ and Gaussian white noise with zero mean and variance $\sigma_\epsilon^2 = 1 - \phi^2 = 0.5511$ (which makes $\sigma_X^2 = 1$). The choice of 1024 observations was motivated by the practical implementation of subsampling in Sect. 3.3 requiring a record length to be a power of 2 ($1024 = 2^{10}$), and because at the chosen value of ϕ, about 1000 data points from model (1) allow the same accuracy in the estimation of variance as 400 independent normal observations (see, e.g., [19]). Note also that while the variance and skewness of Y_t grow with a (see the last paragraph in Sect. 2.1), the effective sample size does not increase due to nonlinearities: linearly filtering time series data to remove correlation, results in the white noise that nevertheless retains the nonlinear structure of the original time series [20].

Next, pretending that, as in reality, the data generating mechanism was unknown, an AR(1) model was fitted to each realization, the goodness of fit confirmed by commonly employed diagnostic checking procedures, and the 90% CIs were computed using (3). After that, from the resulting set of 1000 CIs, the actual coverage probability was determined by counting the fraction of times the variance of Y_t, $V = 1 + 2a^2$, was covered by the CIs. Finally, the procedure was repeated for various values of a: 0.00, 0.05, 0.1, 0.15, 0.20, 0.25, 0.30, 0.35.

The results are shown in Fig. 2. Not surprisingly, at $a = 0$ (when the data generating model is linear) the actual coverage probability of CI (3) is about the nominal value, 0.90. In contrast, with growing nonlinearity (characterized by increasing a) the actual coverage rapidly decreases from the target value, making such CIs misleading [11].

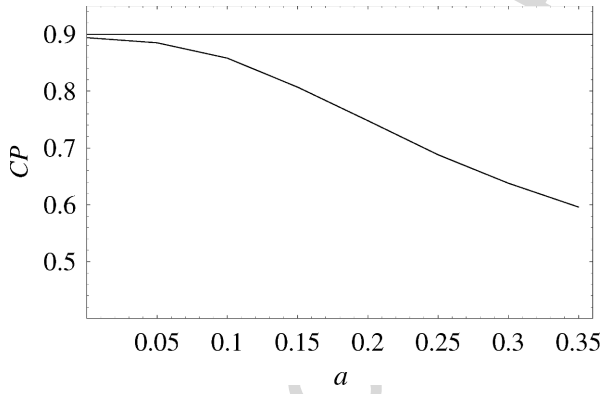


Fig. 2: Actual coverage probabilities (CP) of 90% CIs (3) for the variance of time series (1) at various values of nonlinearity constant a. The horizontal line indicates the target 0.90 coverage.

2.3 Actual Coverage of Conventional CIs for the Variance of Implicit Nonlinear Model

Consider again model (1), this time assuming that the white noise is no longer Gaussian, but follows a Student's t distribution (thus introducing nonlinearity implicitly [15]). Figure 3 shows PDFs of t distribution with three (short-dashed) and six (long-dashed) degrees of freedom, as well as that of standard normal distribution (solid). Similar to the standard normal curve, the t curves are symmetric about zero, and they become practically indistinguishable from the standard normal at larger degrees of freedom. But of importance for the problem discussed here is that the tails of the t curves lie above the tails of the normal.

As revealed by Monte Carlo simulations of model (1) with $a = 0$ and a Student's t white noise (i.e., implicit nonlinear model (2)), such *heavier* tails result in poor performance of CIs (3). Namely, their actual coverage in case of the t white noise with six and three degrees of freedom is about 0.83 and 0.43, respectively, as opposed to nearly 0.90 for model (2) with the Gaussian white noise (see Sect. 2.2) – also a warning against aptness of AR models for fitting heavy tailed data from extreme events.

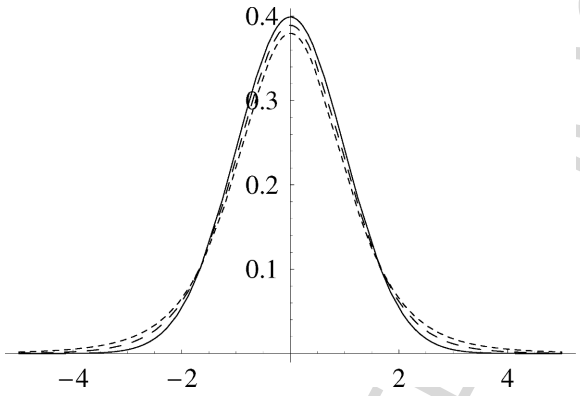


Fig. 3: Probability density functions for standard normal (*solid*) and Student's t distributions with 3 (*short-dashed*) and 6 (*long-dashed*) degrees of freedom.

In addition to heavy tails, another source of trouble for the validity of CIs (3) is a nonzero skewness, cf. [10]. The same experiment with model (2), now with the white noise following a *skewed* distribution (lognormal (0,1)), has resulted in the actual coverage of CIs (3) of just 0.35.

Figure 4 shows that nonnormal noise exacerbates the effect of explicit nonlinearity on the actual coverage of CIs (3) when data are generated from the general model (1). The lower curve in Fig. 4 (corresponding to the t white noise with six degrees of freedom) differs markedly from the upper curve (corresponding to the Gaussian white noise) that was taken from Fig. 2.

3 Subsampling Confidence Intervals

As an alternative, consider subsampling [8], a computer-intensive method that works under even weaker assumptions than other bootstrap techniques (it only requires a nondegenerate limiting distribution for the properly normalized statistic of interest), thus delivering us from having to rely on questionable assumptions about data. Subsampling is based on the values of the statistic of interest recomputed over *subsamples* of the record of the time series, Y_t, i.e., blocks of consecutive observations of the same length b (*block size*) sufficient

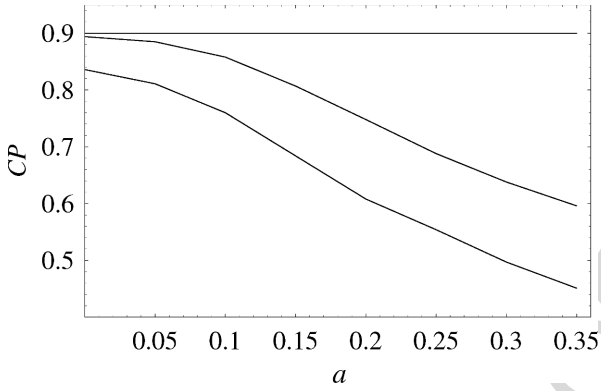


Fig. 4: Same as in Fig. 2 with the *lower solid curve* added that corresponds to Student's t white noise with 6 degrees of freedom in Eq. (2).

to retain the dependence structure of the time series. Three blocks of size b (the first, the ith, and the last) are underscored in a record below containing n observations and, therefore, $n - b + 1$ blocks:

$$\{\underbrace{Y_1, ..., Y_b}_{b}, ..., \underbrace{Y_i, ..., Y_{i+b-1}}_{b}, ..., \underbrace{Y_{n-b+1}, ..., Y_n}_{b}\}. \tag{5}$$

When the DGM (the model) is known, one could (following the probabilistic interpretation of CIs in the beginning of Sect. 2.2) generate a very large number of realizations, compute the sample variance from each realization, estimate the 0.05 and 0.95 quantiles of its distribution and use them as the lower and upper confidence limits of a 90% (percentile) CI. In practice, i.e., when the model is unknown and usually only one record of the observed time series is available, subsampling comes to the rescue by replacing computer generated realizations from the known model with subsamples of the single existing record.

In Monte Carlo simulations to determine the actual coverage probabilities of *subsampling* CIs, there is no need to fit a model to the data. In other respects, the simulations below were carried out as described in Sect. 2.2: from each of 1000 realizations of Y_t, a subsampling CI for the variance or the skewness of Y_t was computed, then the actual coverage probability was determined by counting the fraction of times the *known* value of the parameter was covered by the CIs.

The results of Monte Carlo simulations demonstrate the superiority of subsampling CIs over conventional CIs in estimating the variance of a nonlinear time series (Sect. 3.1), and well as in estimating the skewness (Sect. 3.2), where linear models are inapplicable. The choice of block size b is treated in Sect. 3.3.

3.1 Actual Coverage of Subsampling CIs for the Variance

Coverage probabilities of 90% subsampling CIs for the variance of Y_t in model (1) are presented in Fig. 5 by a long-dashed curve. The plummeting solid curves are taken from Fig. 4; they show the coverage of conventional CIs (3) for the variance of Y_t.

The diminishing actual coverage of CIs (3) is due to the fact that they fail to grow noticeably with a. In contrast, subsampling CIs enlarge with increasing a, so that their coverage remains practically the same (and close to the target of 0.90) for all values of a. Using calibration [8], this allows to achieve even better coverage. That is, one might replace the nominal 90% CIs providing the actual coverage of 0.86 (at $a = 0$) with nominal 95% CIs providing the actual coverage of 0.90 at $a = 0$ and 0.87 at $a = 0.35$, as seen from the short-dashed line in Fig. 5. In practice, calibration can be carried out using a model time series that shares certain statistical properties with the one under study (e.g., model (1) with $a = 0.14$ for the vertical velocity time series).

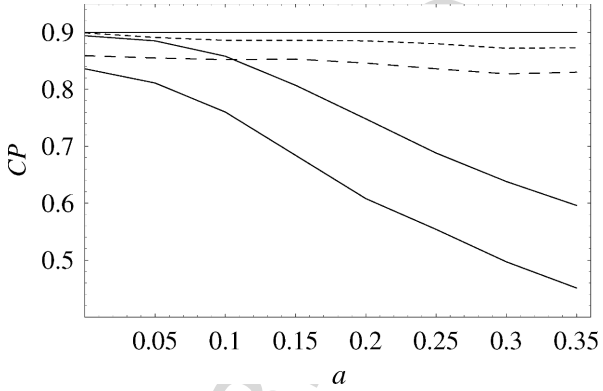


Fig. 5: Actual coverage probabilities (CP) of 90% conventional (*solid curves* from Fig. 4) and subsampling (*long-dashed curve*) CIs for the variance of time series (1) at various values of nonlinearity constant a. The *short-dashed curve* shows the result of calibration, the *solid horizontal line* indicates the target 0.90 coverage.

3.2 Actual Coverage of Subsampling CIs for the Skewness

Nonzero skewness is a frequent attribute of atmospheric and climatic time series, but CIs for the skewness cannot be obtained from linear models, which imply zero skewness. Yet subsampling works here as well.

In Fig. 6, the long-dashed curve shows the actual coverage of subsampling CIs for the skewness of time series generated from model (1). Not as good

as its counterpart in Fig. 5 for the variance, since estimating the skewness requires much longer records [13, 16] (in our simulations, records of length $n = 1024$ were used).

When feasible, a simple way to improve the coverage is to increase the record length. The solid curve in Fig. 6 shows a better coverage, thanks to longer records with $n = 4096$. Otherwise, a calibration can be used: the short-dashed line demonstrates improved (due to calibration) coverage for the original records of $n = 1024$ observations (as in Sect. 3.1, nominal 90% CIs in the subsampling procedure were replaced by nominal 95% CIs).

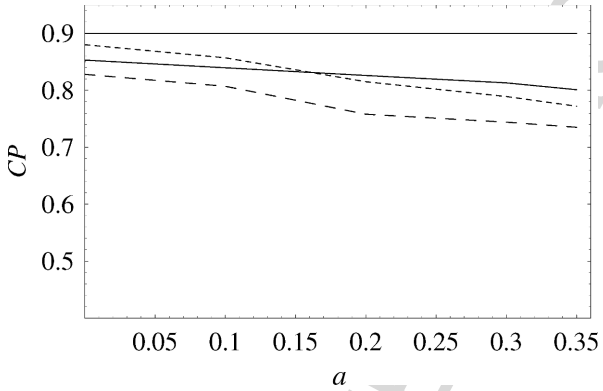


Fig. 6: Actual coverage probabilities of 90% subsampling CIs for the skewness of time series (1) at various values of nonlinearity constant a: original (*long-dashed*) and calibrated (*short-dashed*). The *solid curve* shows the coverage for records four times longer (4096 observations, not calibrated).

3.3 Block Size Selection

Figure 7 demonstrates that the actual coverage of subsampling CIs depends considerably on the block size b. The two curves in Fig. 7, one (dashed) for the variance and the other (solid) for the skewness, were obtained by Monte Carlo simulations of model (1), similar to those in Sects. 3.1 and 3.2, but with fixed $a = 0.14$ and varying b. For each curve, there exists a range of block sizes (around its maximum) that would be appropriate for subsampling. Accordingly, subsampling CIs in Sects. 3.1 and 3.2 were computed with thus determined optimal block sizes for all a (at $a = 0.14$, as seen from 7, they were $b = 80$ for the variance and $b = 140$ for the skewness).

In practice, where the model is unknown and typically only one record is available, the choice of the block size turns out to be the most difficult problem in subsampling, shared by all blocking methods. The asymptotic result [8],

$$b \to \infty \qquad \text{and} \qquad b/n \to 0 \qquad \text{as} \qquad n \to \infty, \tag{6}$$

that the block size needs to tend to infinity with the sample size but slower, does not help to choose the block size for relatively short atmospheric and climatic records.

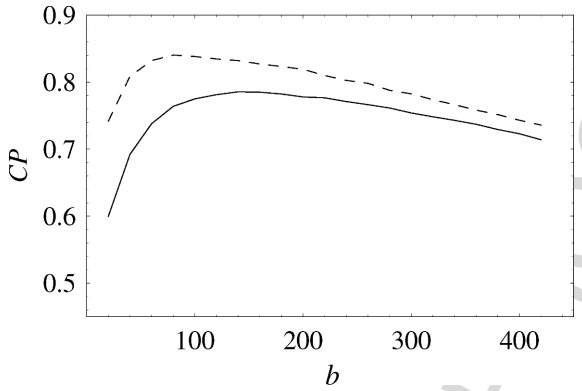


Fig. 7: Actual coverage probabilities of 90% subsampling CIs at various block sizes for the variance (*dashed curve*) and skewness (*solid curve*) from Monte Carlo simulations of model (1) at $a = 0.14$.

We developed, however, another resampling technique that permits determining the optimal block size from one record [14]. Obviously, we had to modify the first step in previous simulations, where 1000 independent realizations, each of $n = 1024$ data points, were generated. These are now replaced by pseudo realizations from the single available record via the following version of the circular bootstrap [21]. The record of $n = 1024$ data points is "wrapped" around a circle, then $p = 2^k < n$ points (say, $p = 32$) on the circle are chosen at random (following a uniform distribution on the circle) as starting points for p consecutive segments of a pseudo realization. The length of each segment is n/p, so the pseudo realization is again of length n. In the current implementation of the technique it was convenient to choose both n and p to be powers of 2. The procedure is repeated to generate N such pseudo realizations, that substitute 1000 independent realizations of a model time series. In [14] this technique was tested on subsampling CIs for the mean of X_t (linear model (2)).

The actual coverage of subsampling CIs for the skewness of Y_t (nonlinear model (1)) obtained from 1000 independent realizations of Y_t at $a = 0.14$ is shown in Fig. 8 by the the solid curve (taken from Fig. 7). In practice, however, with only one realization available, such a curve (that permits to choose the appropriate block size) would not be available. What then can be obtained from its substitute resulting from pseudo realizations?

Each dashed curve in Fig. 8 was computed using a different record of length $n = 1024$ generated from model (1). As described above, $N = 10000$

(this results in smooth curves) pseudo realizations were obtained from the record, then the actual coverage was found by counting the fraction of times the skewness of Y_t, $S = 6a+8a^3$, was covered by the 10000 CIs. The maxima of dashed curves vary wildly (depending on the initial record used), so that each dashed curve typically fails to provide the correct coverage. Nevertheless, the dashed curves essentially retain the shape of the solid curve, thus indicating a suitable block size to be used in subsampling.

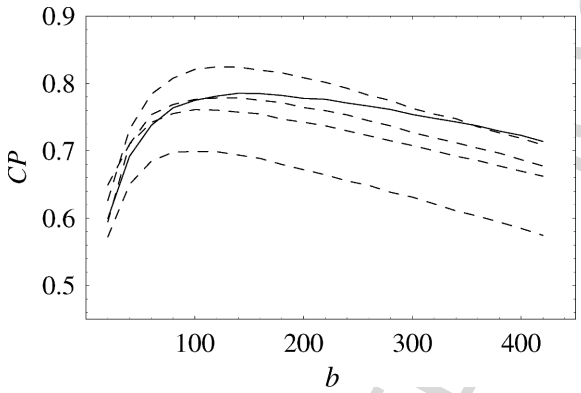


Fig. 8: Selection of block size for subsampling in case of one available record. Each *dashed curve* results from pseudo realizations generated from one different record of Y_t (nonlinear model (1)). *Solid curve* (from Fig. 7) shows actual coverage probabilities of 90% subsampling CIs for the skewness computed from independent realizations of Y_t.

3.4 Vertical Velocity Skewness

Return now to a real life example: the vertical velocity of the wind shown in Fig. 1. From this single record, the curve for the skewness similar to dashed curves in Fig. 8 was obtained (see Fig. 9) using the technique described in Sect. 3.3.

This indicated $b = 100$ as a suitable block size. Then subsampling with $b = 100$ has resulted in the following 90% subsampling CI for the skewness of the vertical velocity time series,

$$(0.66, 1.02), \tag{7}$$

which reasonably confirms its positive skewness. Calibration could slightly modify CI (7), while making its coverage closer to the target.

How should one proceed with a calibration? High coverage exhibited by the curve in Fig. 9 at $b = 100$ is too good to be true: another record similar to that in Fig. 1 may result in a considerably less impressive curve, as seen

from Fig. 8. Since the actual coverage of the subsampling CI, on which the calibration in Sect. 3.2 was based, remains unknown in real situations, an approximating nonlinear model may be used.

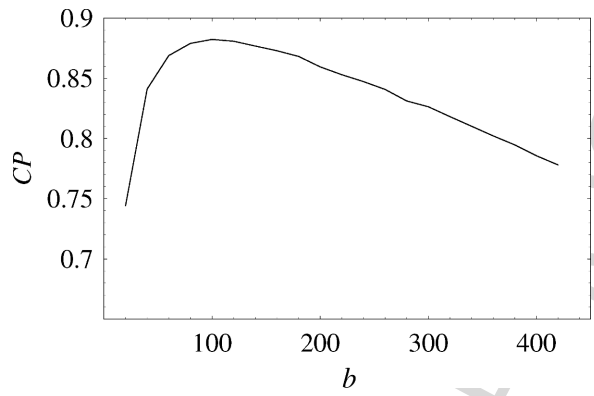


Fig. 9: An analog of *dashed curves* in Fig. 8 computed from the record in Fig. 1.

4 Conclusion

This study has addressed the problem of obtaining reliable statistical inference from atmospheric time series. Since these originate from an inherently nonlinear system, statistical inference based on linear models may be questionable. To investigate how nonlinearities may affect CIs commonly computed from estimated linear models, an AR(1) process driven by a Gaussian white noise (typically used as a default model for correlated time series in climate studies) was altered with (i) a nonlinear component, (ii) a Student's t white noise replacing a Gaussian one.

It was demonstrated by Monte Carlo simulations that the actual coverage probabilities of such common CIs for the variance of our model nonlinear time series become inadequate. In contrast, CIs for the variance obtained via subsampling method proved valid for both linear and nonlinear versions of the model.

Many atmospheric time series are nonstationary, and the subsampling method is by no means restricted to stationary series [8]. However, to emphasize the issues central to this work, only stationary time series are treated here. Besides, atmospheric time series are often considered trend stationary [22], i.e., modeled as the sum, $Y_t = \mu_t + X_t$, where μ_t is a deterministic trend and X_t is a stationary process, commonly a linear one. For example, fitting a linear trend μ_t to a temperature time series, Bloomfield [23] selected an AR(4)

model for X_t. In another study, Y_t was a polynomial trend plus a Gaussian fractionally differenced noise [24].

For the same reason, problems with the linear models approach to constructing CIs for parameters of potentially nonlinear time series are illustrated using only the simplest linear model, AR(1). Although a higher order ARMA model may provide a better fit to an observed time series, no linear model is capable to produce CIs for the skewness (the focus of this work). In contrast, employing subsampling has resulted in reasonable CIs for the skewness of the model nonlinear time series. Subsampling may also be helpful in statistical analyses of extreme events (though not without difficulties preventing easy applications [25]), since for extreme value distributions with nonzero skewness and heavy tails, common CIs based on asymptotic maximum likelihood fail to capture the real variability [26].

Acknowledgements. I am grateful to D. N. Politis and E. M. Agee for useful discussions. This work was supported by National Science Foundation Grants ATM-0514674 and ATM-0541491.

References

1. M. Ghil, M.R. Allen, M.D. Dettinger, et al., Advanced spectral methods for climate time series. *Rev. Geophys.*, 40, Art. No. 1003 (2002)
2. M. Lesieur, *Turbulence in Fluids*, 3rd ed., Kluwer, Dordrecht (1997)
3. A. Maurizi, On the dependence of third- and fourth moments on stability in the turbulent boundary layer. *Nonlin. Processes Geophys.*, 13, 119–123 (2006)
4. R.W. Katz, Techniques for estimating uncertainty in climate change scenarious and impact studies. *Clim. Res.*, 20, 167–185 (2002)
5. R.W. Katz, M.B. Parlange, P. Naveau, Statistics of extremes in hydrology. *Adv. Water Resour.*, 25, 1287–1304 (2002)
6. B. Efron, R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, London (1993)
7. A.C. Davison, D.V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, Cambridge (1997)
8. D.N. Politis, J.P. Romano, M. Wolf, *Subsampling*, Springer, New York (1999)
9. S.N. Lahiri, *Resampling Methods for Dependent Data*, Springer, New York (2003)
10. R.R. Wilcox, *Applying Contemporary Statistical Techniques*, Academic Press, San Diego (2003)
11. A. Gluhovsky, E. Agee, On the analysis of atmospheric and climatic time series. *J. Appl. Meteorol. Climatol.*, 46, 1125–1129 (2007)
12. E.M. Agee, S.R. Gilbert, An aircraft investigation of mesoscale convection over Lake Michigan during the 10 January 1984 cold air outbreak. *J. Atmos. Sci.*, 46, 1877–1897 (1989)
13. A. Gluhovsky, E. Agee, A definitive approach to turbulence statistical studies in planetary boundary layers. *J. Atmos. Sci.*, 51, 1682–1690 (1994)

14. A. Gluhovsky, M. Zihlbauer, D.N. Politis, Subsampling confidence intervals for parameters of atmospheric time series: block size choice and calibration. *J. Stat. Comput. Simul.*, 75, 381–389 (2005)
15. J. Fan, Q. Yao, *Nonlinear Time Series*, Springer, New York (2003)
16. D.H. Lenschow, J. Mann, L. Kristensen, How long is long enough when measuring fluxes and other turbulence statistics. *J. Atmos. Oceanic Tech.*, 11, 661–673 (1994)
17. H. von Storch, F.W. Zwiers, *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge (1999)
18. D.B. Percival, J.E. Overland, H.O. Mofjeld, Modeling North Pacific climate time series. In: D.R. Brillinger, E.A. Robinson, F.P. Schoenberg (eds.), *Time Series Analysis and Applications to Geophysical Systems*, Springer, New York, 151–167 (2004)
19. M.B. Priestley, *Spectral Analysis and Time Series*, Academic Press, San Diego (1981)
20. J. Theiler, S. Eubank, Don't bleach chaotic data. *Chaos*, 3, 771–782 (1993)
21. D.N. Politis, J.P. Romano, A circular block-resampling procedure for stationary data. In: R. LePage, L. Billard (eds.), *Exploring the Limits of Bootstrap*, Wiley, New York, 263–270 (1992)
22. R.H. Shumway, D.S. Stoffer, *Time Series Analysis and Its Applications*, Springer, New York (2000)
23. P. Bloomfield, Trends in global temperature. *Climatic Change*, 21, 1–16 (1992)
24. P.F. Craigmile, P.Guttorp, D.B, Percival, Trend assessment in a long memory dependence model using the discrete wavelet transform. *Environmetrics*, 15, 313–335 (2004)
25. S.I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer, New York (2007)
26. M. Kallache, H.W. Rust, J. Kropp, Trend assessment: applications for hydrology and climate research. *Nonlin. Processes Geophys.*, 12, 201–210 (2005)

Global Patterns of Nonlinearity in Real and GCM-Simulated Atmospheric Data

Jiří Mikšovský, Petr Pišoft, and Aleš Raidl

Department of Meteorology and Environment Protection, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic,
jiri.miksovsky@mff.cuni.cz

Abstract. We employed selected methods of time series analysis to investigate the spatial and seasonal variations of nonlinearity in the NCEP/NCAR reanalysis data and in the outputs of the global climate model HadCM3 of the Hadley Center. The applied nonlinearity detection techniques were based on a direct comparison of the results of prediction by multiple linear regression and by the method of local linear models, complemented by tests using surrogate data. Series of daily values of relative topography and geopotential height were analyzed. Although some differences of the detected patterns of nonlinearity were found, their basic features seem to be identical for both the reanalysis and the model outputs. Most prominently, the distinct contrast between weak nonlinearity in the equatorial area and stronger nonlinearity in higher latitudes was well reproduced by the HadCM3 model. Nonlinearity tends to be slightly stronger in the model outputs than in the reanalysis data. Nonlinear behavior was generally stronger in the colder part of the year in the mid-latitudes of both hemispheres, for both analyzed datasets.

Keywords: Nonlinearity, Reanalysis, Global climate model, Surrogates

1 Introduction

The Earth's climate system, as well as its atmospheric component, is an intrinsically nonlinear physical system. This nonlinearity is generally reflected in many series of climatic variables such as atmospheric pressure or temperature, but whether it is detectable and how strong it is depends on the type of the variable [1, 2, 3], geographic area of its origin [2, 4, 5, 6] or length of the signal [3]. The manifestations of nonlinearity in time series can be studied in numerous ways, using different statistics or criteria of the presence of nonlinear behavior. The techniques applied so far to meteorological data involve the calculation of the mutual information or persistence [1, 7, 8], statistics based on the performance of a nonlinear predictive method [3, 4, 9], nonlinear correlations [10] or the examination of the character of the prediction

residuals [2, 4, 5]. Tests using some form of surrogate data are frequently employed [1, 2, 3, 4, 7, 9, 10]. The presence of nonlinearity can also be assessed by comparing the performance of a linear and a nonlinear time series analysis method. In the atmospheric sciences, such studies are frequently associated with the application of statistical methods for prediction [6, 11, 12], or downscaling and postprocessing tasks [6, 13, 14, 15, 16]. Alongside with a wide spectrum of techniques for the detection of nonlinearity, different authors studied diverse types of signals, ranging from various variables related to the local temperature [1, 3, 6, 7, 10, 13, 14, 15, 16] or pressure [1, 2, 3, 4, 5, 7] to characteristics of larger-scale dynamics such as the mean hemispheric available potential energy [8]. Heterogeneity of the methods and datasets applied by different researchers makes it difficult to directly compare the results and use them to create a consistent global picture of the geographic variations of nonlinearity. However, it also seems that there are some systematic regularities in the spatial distribution of nonlinearity or of the related characteristics [2, 5, 6, 10, 17]. Here, we investigate this matter further, using a comparison of the results of linear and nonlinear prediction and tests based on the surrogate data.

A significant portion of the existing studies dealing with the issue of nonlinearity in time series focus on the analysis of individual scalar signals, typically employing time delayed values for the construction of the space of predictors or phase space reconstruction. Due to the complex behavior the atmosphere exhibits, and the relatively small size of the available records, the information content in a single series is limited and often insufficient for an effective application of nonlinear techniques. But meteorological measurements are frequently available for more than one variable, and they are carried out for multiple locations. When a multivariate system is used instead of a single scalar series, more information about the local state of the climate system can be obtained. It also seems that multivariate systems exhibit a generally stronger detectable nonlinear behavior [3]. For these reasons, and because using multiple input variables is common in many tasks of statistical meteorology and climatology, we focused on settings with multivariate predictors in this study. We restricted our attention to just a few of the available variables, defining the temperature and pressure structure of the atmosphere. The two illustrative cases presented here are based on forecasts of daily values of the relative topography 850-500 hPa (which is closely related to the temperature of the lower troposphere) and of the geopotential height of the 850 hPa level (one of the variables characterizing the structure of the field of atmospheric pressure). Along with investigating the character of the series derived from actual measurements (NCEP/NCAR reanalysis), attention was paid to the potential of the global climate model HadCM3 to reproduce the structures detected in the observed data. This should help to assess whether such simulation is able to capture not just the basic characteristics of the Earth's climate, but also the eventual nonlinear features of the respective time series. The utilized datasets are presented in Sect. 2, the techniques applied to quantitatively evaluate

nonlinearity are described in Sect. 3. Section 4 is devoted to the study of the spatial variations of nonlinearity. Section 5 focusses on the influence of the presence of the annual cycle in the series and seasonal changes of the detected patterns. Finally, in Sect. 6, the results are discussed with regard to their possible physical cause and practical implications. Color versions of the presented maps of the geographical distribution of nonlinearity (Figs. 3, 5, 6, 8 and 9) can be accessed at http://www.miksovsky.info/springer2008.htm.

2 Data

Direct atmospheric observations and measurements suffer from a number of potential problems. Their locations are typically unevenly spaced and coverage of some areas of the Earth is limited. Data from different sources are often incompatible and sometimes flawed. This restricts the usability of raw measurements for an analysis such as ours, the goal of which is to derive globally comparable results. To avoid or reduce the aforementioned problems, we used a gridded dataset in this study instead of direct measurements – the NCEP/NCAR reanalysis [18, 19] (hereinafter NCEP/NCAR). The reanalysis is a dataset derived from measurements at weather stations, as well as inputs from rawinsondes, meteorological satellites and other sources. The input observations are processed by a fixed data assimilation system, including a numerical forecast model, and the resulting series are available in a regular horizontal grid of 2.5° by 2.5°. Here, daily values of the geopotential height of the 850 hPa level (hereinafter H850) and 500 hPa level have been employed in a reduced 5° by 5° horizontal resolution, for the period between 1961 and 2000. From the values of the geopotential heights, the relative topography 850-500 hPa (RT850-500) has been computed. This quantity describes the thickness of the layer between the 850 hPa and 500 hPa levels and it is proportional to its mean virtual temperature. According to the classification used by Kalnay et al. [18], geopotential heights fall into the A–category of variables, thus reflecting the character of actual measurements rather than the specific properties of the model applied to create the reanalysis. A typical example of the analyzed series of RT850-500 in the equatorial area and in the mid-latitudes is shown in Fig. 1.

The recently increased interest in climate change instigated an intensive development of the models of the global climate. These simulations, to be reasonably realistic, must describe all key components of the climate system as well as the connections among them. As a result, the models are very complex and demanding with respect to the required computational resources. But despite their sophistication, no model is able to mimic the observed climate with absolute accuracy. A very important task in climate modeling is therefore validating the models, i.e., assessing their ability to reproduce the real climate. The common validation procedures are usually based on the basic statistical characteristics of the model outputs; here, we focus on the ability of a climate

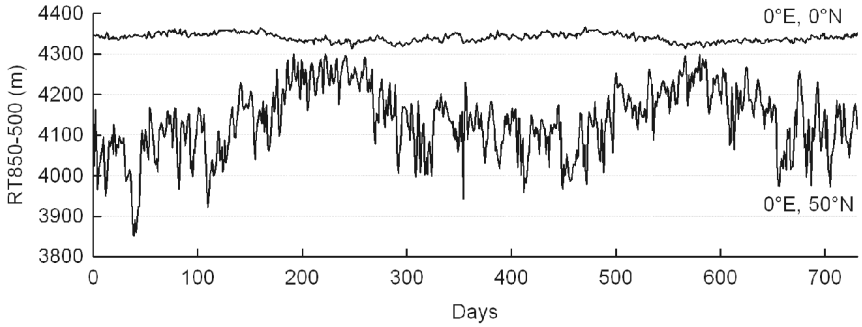


Fig. 1: A section of the analyzed data: Time series of daily values of the relative topography 850-500 hPa in the equatorial area (0°E, 0°N) and in the mid-latitudes (0°E, 50°N), for the years 1991 and 1992.

simulation to produce time series with the same nonlinear qualities as the real climate. For this task, we chose one of the major global climate models, HadCM3 of the Hadley Centre [20, 21]. The model outputs were used in a reduced horizontal resolution of 3.75° (longitude) by 5° (latitude). The model integration employed here was based on the observed concentrations of the greenhouse gasses and estimates of past changes in ozone concentration and sulfur emissions prior to the year 1990, and the emission scenario SRES B2 afterwards [21]. Since we only used the period from 1961 to 2000 for our analysis, and there is just very little difference among the SRES scenarios in the 1990s, the specific scenario choice should not be crucial.

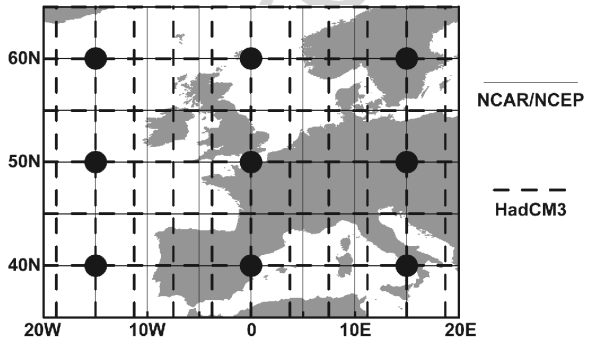


Fig. 2: An example of the structure of the pattern of predictors, displayed for the predictand series located at 0°E, 50°N. Black circles mark the positions of the predictors, the grid illustrates the reduced horizontal resolution of the NCEP/NCAR and HadCM3 data, used in this study.

3 Methods

3.1 General Settings

One of the key issues of the multivariate approach to the construction of the space of predictors is the selection of a suitable set of input variables. Unlike for some simple low-dimensional dynamical systems, a perfect phase-space reconstruction is impossible from climatic time series, due to the complexity of the underlying system. In the case of practical time series analysis tasks, a finite-dimensional local approximation of the phase space may suffice. To predict values of a scalar series in some grid point, we used a pre-set pattern of predictors, centered on the location of the predictand and spanning 30° in longitude and 20° in latitude (Fig. 2). A different configuration of predictors was chosen for each of the two tasks presented: In the case of the RT850-500 forecast, the dimension of the predictor space was $N = 18$, with 9 values of RT850-500 and 9 values of H850 in a configuration shown in Fig. 2. For the forecast of H850, 9 predictors were used, all of which were of the H850 type. Note that, despite the different spatial resolution of the NCEP/NCAR reanalysis and the HadCM3 model, the selected pattern of predictors could be applied for both of them directly, without interpolating the data to a common grid.

All predictors $x_i(t), i = 1, \ldots, N$, were transformed to have zero mean and standard deviation equal to $\sqrt{\cos\varphi}$, using the linear transformation $x_i(t) \rightarrow \sqrt{cos(\varphi)}(x_i(t) - \overline{x}_i)/\sigma_i$ (φ being latitude of the respective grid point, $\overline{x}_i$ mean value of the predictor series and σ_i its standard deviation). Hence, the predictor's variance was proportional to the size of the area characterized by the corresponding grid point. The presented results were derived from the outcomes of prediction one day ahead, carried out for grid points located between 70°N and 70°S (the areas closest to the poles were excluded from the analysis, due to the severe deformation of the applied spatial pattern of predictors in high latitudes).

3.2 Direct Comparison-Based Approach

Our primary technique of quantification of nonlinearity was based on a direct comparison of the root mean square errors (RMSEs) of prediction by a linear reference method, multiple linear regression, and by its nonlinear counterpart, the method of local linear models. In the case of linear regression, the value of the scalar predictand y at time $t+1$ was computed as a linear combination of the values of individual predictors $x_i, i = 1, \ldots, N$, in the previous time step

$$\hat{y}(t+1) = a_0 + \sum_{i=1}^{N} a_i x_i(t)\,, \tag{1}$$

where the coefficients $a_j, j = 0, \ldots, N$, were calculated to minimize the sum of the squared values of the residuals $\hat{y}(t) - y(t)$.

Even a nonlinear system can be described rather well when the linear model is applied locally for smaller portions of the phase space instead of a global linear approximation. This concept has been successfully utilized for the construction of forecast models for many different types of time series. Several related studies are reprinted in [22] and the basic principles of the method of local models are also described, e.g., in [23]. The dynamics in the individual regions of the input space is approximated by linear mappings based on (1), but an individual linear predictive model (or a set of coefficients a_i, respectively) is constructed for each value of t. To create such a local model, only a certain number M of the predictors-predictand pairs, representing the states of the system most similar to the one at time t, is employed to compute the coefficients. The similarity of individual states was quantified by the distance of the respective N-dimensional vectors of predictors $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_N(t))$ here, using the Euclidian norm.

To calculate the out-of-sample root mean square error of the prediction, the analyzed series were divided into two subintervals. The years 1961–1990 were used as a calibration set, i.e., for the computation of the coefficients of the above described models. These were then tested for the years 1991–2000. The values of RMSE we obtained for the prediction by multiple linear regression ($RMSE_{\mathrm{MLR}}$) and local linear models ($RMSE_{\mathrm{LM}}$) were compared by computing

$$SS_{\mathrm{LM}} = 1 - (RMSE_{\mathrm{LM}}/RMSE_{\mathrm{MLR}})^2 \,, \tag{2}$$

which will be referred to as the local models' skill score. Its definition is based on the commonly used concept of a skill score, described, e.g., in [24]. SS_{LM} vanishes when both methods perform equally well in terms of RMSE and it equals to one for a perfect forecast by local models (presuming that $RMSE_{\mathrm{MLR}} \neq 0$). The number M of predictors-predictand pairs used for the computation of the coefficients of the local models is one of the adjustable parameters of the method of local models. Depending on the specific structure of the local climate system, different values of M may be suitable to minimize RMSE. Here, local models constructed with $M = 250$, 500 and 1000 were tested for each grid point; the variant giving the lowest RMSE was then used in the subsequent analysis.

3.3 Surrogate Data-Based Approach

The above described approach yields results which are interesting from a practical perspective, but, strictly speaking, it only refers to a relation of two particular techniques, both of which may have their specifics. Another method, which does not rely on comparing different mappings, exists. It uses modified series (so-called surrogate series or surrogates), which preserve selected properties of the original signal, but are consistent with some general null hypothesis. Here, the hypothesis is that the data originates from a linear

Gaussian process, the output of which may have been modified by a static monotonic nonlinear filter. The values of a nonlinearity-sensitive statistic are then compared for the original series and multiple surrogates, and if a statistically significant difference is detected, the null hypothesis is rejected. It should be noted that the formal rejection does not necessarily prove the presence of nonlinearity in the signal, as it can be caused by other reasons, such as nonstationarity of the series or imperfection of the surrogate-generating procedure. For details see, e.g., [9], where the principles of the surrogate data-based tests are presented in depth, or [25], where the usability of several methods of generating surrogates is discussed for various geophysical data.

For each grid point, 10 surrogates were created from the respective multivariate system of time series. Prediction by the method of local linear models was carried out for each of the surrogates and an arithmetic average $RMSE_{\mathrm{SURR}}$ of the resulting RMSEs was computed. A skill score-based variable, analogous to (2), was then calculated using RMSE for the original series $RMSE_{\mathrm{LM}}$ and $RMSE_{\mathrm{SURR}}$:

$$SS_{\mathrm{SURR}} = 1 - (RMSE_{\mathrm{LM}}/RMSE_{\mathrm{SURR}})^2. \tag{3}$$

In order to keep the computational demands at a reasonable level, the surrogate data-based analysis was performed just for $M = 250$. Also, the years 1991–2000 were used for both calibration and testing of the mappings. The surrogate series were generated by the iterative amplitude adjusted Fourier transform [26] in its multivariate form [9]; the program package TISEAN by Hegger et al. [27] was applied for this task.

4 Spatial Patterns of Nonlinearity

Figure 3a shows the geographical distribution of the local models' skill score SS_{LM}, obtained for the NCEP/NCAR RT850-500 forecast. The most prominent feature of the detected pattern is the strong latitudinal variance of nonlinearity. Near the equator, just very small and mostly statistically insignificant difference between the performance of purely linear regression and local linear models was found. Nonlinear behavior becomes visibly stronger in the midlatitudes, and it is more pronounced on average in the northern hemisphere, where major nonlinearity was detected for all grid points north of circa 25°N (Fig. 4). In the southern hemisphere, the strongest nonlinearity was located in a band approximately between 25°S and 50°S. This structure seems to be well reproduced by the HadCM3 model (Fig. 3b), although the nonlinear behavior is slightly stronger in the model data in the northern hemisphere – see Table 1, columns 1 and 2. The spatial correlation of the SS_{LM} fields for the NCEP/NCAR and HadCM3 data was evaluated by computing the Pearson correlation coefficient, after linear interpolation of the HadCM3 data-based values of SS_{LM} to the 5° by 5° grid of NCEP/NCAR. For the entire area

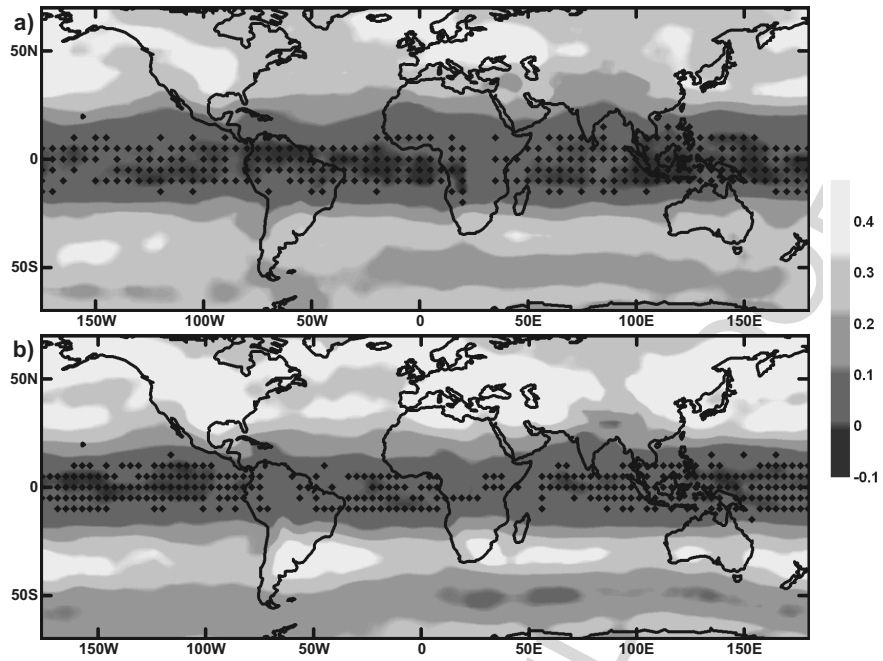


Fig. 3: Geographical distribution of the local models' skill score SS_{LM}, obtained for the RT850-500 prediction, using the NCEP/NCAR (**a**) and HadCM3 (**b**) data. Diamonds mark the positions of the grid points where daily errors of prediction by the method of local models were not statistically significantly lower than for linear regression at the 95% confidence level, according to the one-sided paired sign test.

between 70°N and 70°S, the correlation was 0.91. When just extratropical areas were taken into account, the resemblance of the SS_{LM} patterns was stronger in the northern hemisphere than in the southern one (Table 1, column 3). Similar values of correlation were also obtained when the Spearman rank-order correlation coefficient was used instead of the Pearson one.

Aside from the dominant latitudinal dependence, the detected nonlinearity patterns also exhibited a distinct finer structure. As can be seen in Fig. 3a for the NCEP/NCAR reanalysis data, local maxima of nonlinearity were found over Europe, North America, East Asia and the northern part of the Pacific Ocean, and east of the landmasses of the southern hemisphere. The HadCM3 data yielded a very similar pattern (Fig. 3b). After the average latitudinal structure was filtered out by subtracting the respective latitudinal averages from the values of SS_{LM} in every grid point, the spatial correlation of the NCEP/NCAR and HadCM3 SS_{LM} patterns was still rather high, though the resemblance of both fields was clearly stronger in the northern hemisphere (Table 1, column 4).

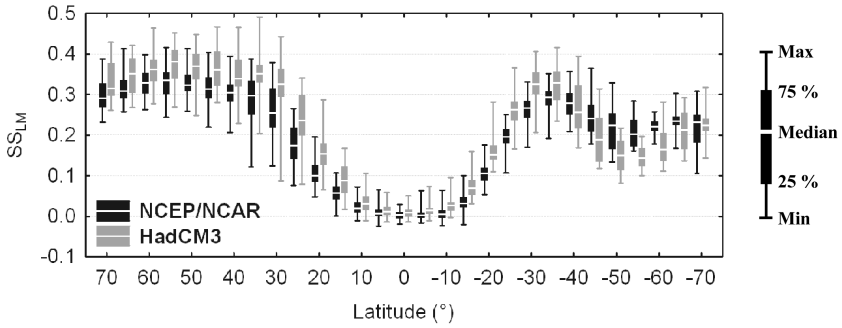


Fig. 4: Distribution of SS_{LM} in different latitudes, obtained for the RT850-500 prediction (latitude values are positive north of the equator).

Table 1: Regional averages of SS_{LM}, obtained for the RT850-500 prediction in the case of the NCEP/NCAR (column 1) and HadCM3 (column 2) data and spatial correlations of the NCEP/NCAR and HadCM3 SS_{LM} patterns for the original values of SS_{LM} (column 3) and after the average latitudinal dependence has been filtered out (column 4).

	SS_{LM}		Correlation	
Region	NCEP/NCAR	HadCM3	Original	Filtered
25°N–70°N	0.29	0.34	0.75	0.60
20°S–20°N	0.04	0.06	0.89	0.45
70°S–25°S	0.24	0.23	0.55	0.48

When the results of the H850 prediction were applied as a basis for a nonlinearity detection, a somewhat different pattern emerged (Fig. 5). The basic latitudinal structure with very weak nonlinearity in the equatorial area was still present, but other details of the detected structure differed from the ones found for the RT850-500 prediction. In the northern hemisphere, maximum values of SS_{LM} were located over the northwestern part of the Atlantic Ocean and the adjacent part of North America, as well as over the northern part of the Pacific Ocean. Both these maxima were rather well expressed, while the rest of the northern hemisphere exhibited weaker nonlinearity. In the southern hemisphere, the maxima of SS_{LM} were less localized. The overall degree of nonlinearity was lower than for the RT850-500 prediction (Table 2, columns 1 and 2). The similarity of the patterns obtained from the NCEP/NCAR and HadCM3 data was again very strong, with a value of global spatial correlation of 0.9. The nonlinearity was stronger on average in the HadCM3 outputs than in the NCEP/NCAR reanalysis. As for the match of the patterns of SS_{LM} with filtered-out latitudinal dependence, there was still a high positive correlation, stronger in the northern hemisphere (Table 2, column 4).

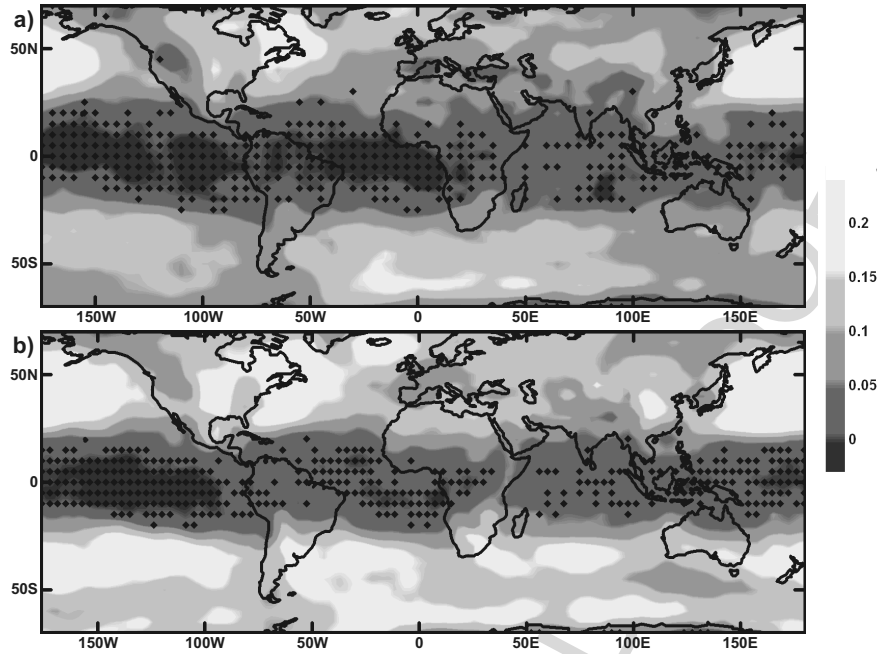


Fig. 5: Same as Fig. 3, for the prediction of H850 instead of RT850-500.

Table 2: Same as Table 1, for the prediction of H850 instead of RT850-500.

	SS_{LM}		Correlation	
Region	NCEP/NCAR	HadCM3	Original	Filtered
25°N–70°N	0.10	0.13	0.84	0.85
20°S–20°N	0.02	0.03	0.71	0.52
70°S–25°S	0.09	0.14	0.75	0.71

As can be seen from Fig. 6, the pattern of nonlinearity obtained for the RT850-500 prediction by means of surrogate data and expressed through SS_{SURR} is very similar to the one presented above for the direct comparison technique (Fig. 3a). To illustrate the distribution of RMSE in the ensemble of surrogates, a more detailed example of the outcomes is shown in Fig. 7 for the grid points along the 0° meridian. The results for the HadCM3 data are not shown, but they also confirm the outcomes of the direct comparison of multiple linear regression and local linear models. Similarly, surrogate data-based verification of the results derived from the H850 prediction showed no major differences either.

It should be mentioned that when an identical setting is used for direct comparison-based and surrogate data-based tests, including an equal size of the calibration set, SS_{LM} is systematically smaller than SS_{SURR}. The reason

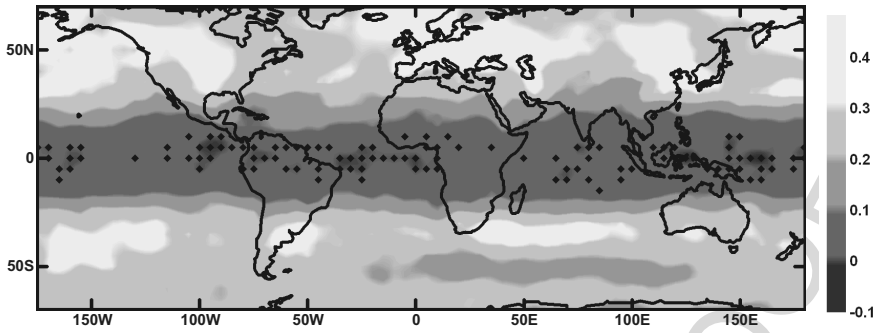


Fig. 6: Geographical distribution of SS_{SURR}, obtained for the RT850-500 prediction, using the NCEP/NCAR data. Diamonds mark positions of the grid points, where the value of RMSE for the original series was not smaller than for all 10 surrogates. This is equivalent to the non-rejection of the hypothesis of a linear Gaussian generating process at the confidence level of about 91%, according to the usually applied one-sided rank-order test, described, e.g., in [9]. Testing at a higher confidence level would require more surrogates, but even then, the results would be almost identical, as additional tests have shown for selected individual grid points.

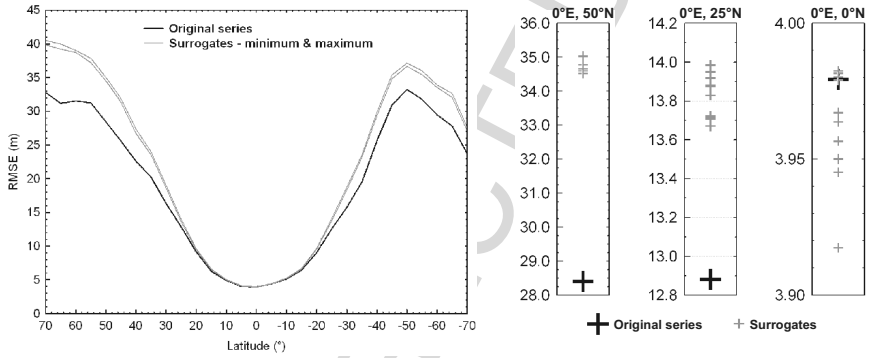


Fig. 7: *Left panel*: RMSE of the RT850-500 prediction by the local linear models method and its range for the respective surrogate series, for grid points at 0°E (NCEP/NCAR data). *Right panels*: Values of RMSE (m) obtained for the original series and individual surrogates in the three selected grid points.

for this difference is related to the behavior of the method of local models for purely linear series. When the processed signal contains no deterministic nonlinear component (like surrogates do) and M is smaller than the size of the calibration set, the method of local models performs slightly worse than linear regression. Our choice of a shorter calibration set for the surrogate data-based tests (Sect. 3.3) has actually partly compensated for this shift, because the magnitude of detected nonlinearity generally decreases with the reduction of the size of the calibration set.

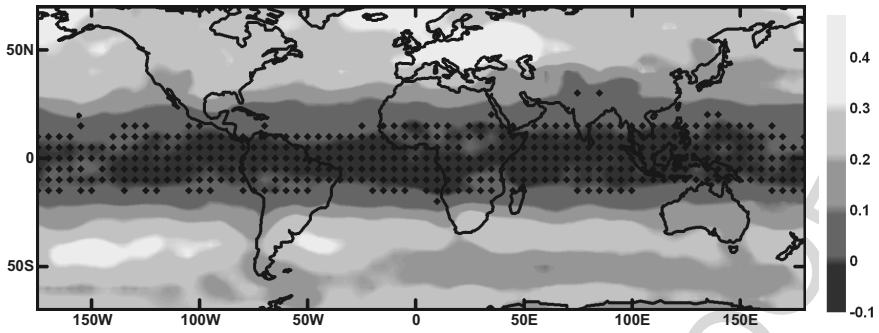


Fig. 8: Same as Fig. 3a, for series with removed annual cycle.

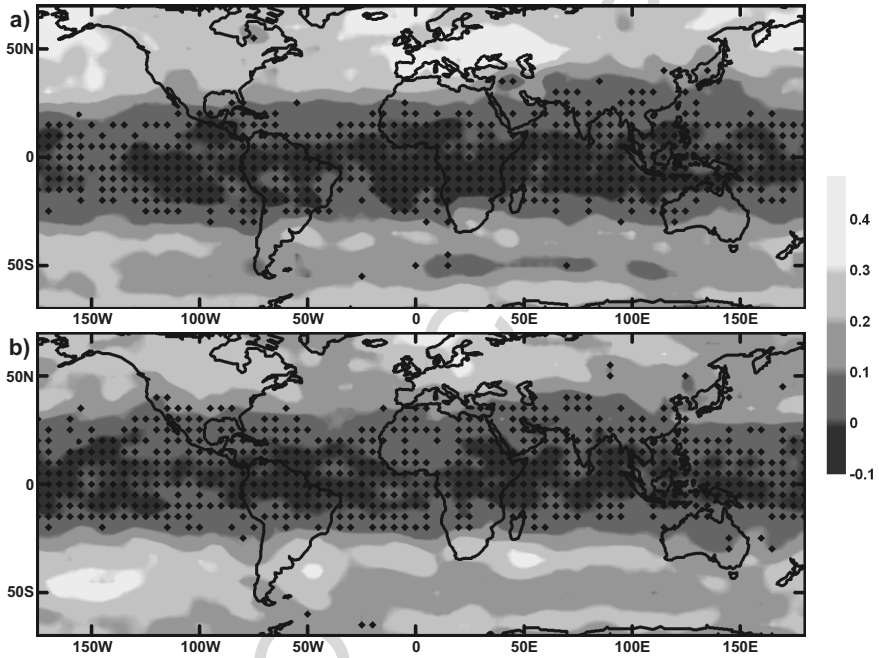


Fig. 9: Same as Fig. 3a, for the DJF (**a**) and JJA (**b**) seasons (winter and summer in the northern hemisphere) instead of the entire year.

5 Seasonal Variations of Nonlinearity

The annual cycle is among the strongest oscillations in the climate system. It dominates series of many climatic variables, especially in higher latitudes (see example in Fig. 1). This also means that the geographical areas with a well-defined annual cycle coincide to some degree with the regions, where strong nonlinearity was detected. To assess the possible relationship, we repeated some of the tests for the series with removed annual cycle. The removal was carried out by subtracting the mean climatological annual cycle of the respective variable, computed for the years 1961–2000 and smoothed by an 11–day moving average. An example of the results is shown in Fig. 8, for the RT850-500 prediction. As a comparison to Fig. 3a reveals, the values of SS_{LM} generally decreased after the annual cycle removal. Although this change was relatively small on average (e.g., the average value of SS_{LM} decreased from 0.29 to 0.23 in the area north of 25°N, and from 0.24 to 0.21 south of 25°S), it was profound in the regions with the highest amplitude of the annual cycle of RT850-500. For instance, the maximum of SS_{LM}, originally detected over East Asia and the adjacent part of the Pacific Ocean, disappeared almost completely. In the southern hemisphere, the changes associated with the annual cycle removal were generally smaller. In the case of the H850 prediction, the shape of the pattern of SS_{LM} remained practically identical for the annual cycle-free series, though the average degree of nonlinearity also slightly decreased.

In many situations, the annual cycle cannot be treated as simply an oscillation superposed to the variations at other time scales. Different seasons are associated with different atmospheric dynamics in many regions, and properties of the analyzed time series, including their eventual nonlinearity, may thus periodically vary throughout the year. Because of this, the analysis of climatic data is often performed separately for different parts of the year, typically seasons or months. We used this approach to investigate the seasonal variations of SS_{LM}. The results below are shown for the parts of the year corresponding to climatological winter (December, January and February – DJF) and summer (June, July and August – JJA) of the northern hemisphere. When the analysis was carried out for separate seasons, the RMSE of the prediction by linear regression decreased for most grid points in the annual average. The performance of the method of local models usually became worse, primarily due to the reduction of the amount of data available for the calibration of the mappings. As a result, the average magnitude of nonlinearity decreased somewhat, compared to the situation when the series were analyzed as the whole. Despite this change, the basic features of the patterns of SS_{LM} were still the same, as can be seen from an example of the results based on the RT850-500 forecast (Fig. 9). In the equatorial area, the nonlinearity remained very weak or undetectable in all seasons. In higher latitudes, the patterns retained some of the basic shape, detected for the year as the whole, but their magnitude visibly varied with the season. The overall nonlinearity was stronger in the

DJF season than in JJA in the northern hemisphere, while in higher latitudes of the southern hemisphere, this variation was reversed and JJA exhibited stronger nonlinearity than DJF on average (Fig. 10, Table 3, columns 1 and 2). The seasonal changes were stronger expressed in the northern hemisphere. The seasonal variation was well simulated by the HadCM3 model (Table 3, columns 3 and 4) and it was also detectable in the results based on the forecast of H850, for both the NCEP/NCAR and HadCM3 data (not shown).

6 Discussion

All performed analyses revealed a common basic latitudinal structure with just negligible nonlinearity in the equatorial regions, but generally stronger nonlinear behavior in the mid-latitudes of both hemispheres. A detailed analysis of the factors behind the observed patterns might be problematic, because they do not seem to be a result of a single driving force, but rather their complex combination. There are, however, some possible links worth mentioning. In the case of the results based on the RT850-500 prediction, there may be a connection between more pronounced nonlinearity in the mid-latitudes and the activity of the polar front. The strongest nonlinear behavior over Europe and North America seems to coincide with the position of the zones where air masses of different origin often interact. In the southern hemisphere, where the landmasses are less extensive, areas of the strongest nonlinearity are typically located rather east of the continents, possibly because of the interaction of the landmass with the prevailing westerlies. Between approximately 50°S and 60°S, where the amount of land is very small, nonlinearity is weaker on average. A removal of the annual cycle from the series slightly decreases the magnitude of detected nonlinearity, but except for the regions where the annual variation is very strong (East Asia), the effect of the annual cycle presence does not dominate the results. For the H850 forecast, there appears to be a certain connection of the areas with strong nonlinearity to the zones of high horizontal gradient of H850. In the northern hemisphere, such areas are typically associated with deep stationary cyclones, which are usually present over the North Atlantic and North Pacific during winter. The match is not perfect though, and there may be some other factors involved. Altogether, it seems that nonlinearity tends to be stronger in the regions with more complex dynamics, where strong driving or perturbing factors are in effect. This hypothesis is supported by the fact that nonlinearity is generally more pronounced during the colder season in the mid-latitudes of both hemispheres, i.e., in situations when the temperature gradient between the equatorial area and the polar region is strongest. The fact that the seasonal variations are more distinct in the northern hemisphere is probably an effect of the uneven distribution of the continents, resulting in a larger influence of the continental climate in the northern mid-latitudes.

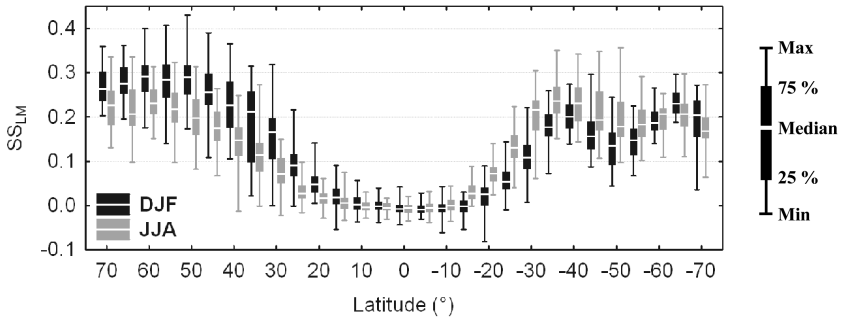


Fig. 10: Distribution of SS_{LM}, obtained for the prediction of RT850-500 for the DJF and JJA seasons in different latitudes (NCEP/NCAR data).

Table 3: Seasonal variations of nonlinearity (expressed by SS_{LM}) in the NCEP/NCAR and HadCM3 data, for the RT850-500 prediction.

	NCEP/NCAR		HadCM3	
Area	DJF	JJA	DJF	JJA
25°N–70°N	0.23	0.16	0.27	0.20
20°S–20°N	0.01	0.01	0.02	0.02
70°S–25°S	0.16	0.20	0.16	0.18

The two presented cases, based on the prediction of geopotential height and relative topography one day ahead, represent just a fraction of possible settings. From additional tests, carried out for different predictand-predictors combinations, it seems that the basic structure with weak nonlinearity in the equatorial area is typical for most situations. On the other hand, the finer details of the detected patterns vary, especially with the type of the predictand. The exact number and geographical configuration of predictors seem to be less important, as long as they sufficiently characterize the local state of the atmosphere. Beside the type of the studied variables, we also paid attention to the sensitivity of the results to the specific details of the tests. It appears that the results are rather robust to the changes of the size of the source area of predictors, although a use of a too big or too small area leads to a general increase of the prediction error and a weakening of detected nonlinearity. The outcomes remain very similar when the input data are pre-processed by principal component analysis, instead of using the point-wise predictors directly. The method of eventual normalization of the predictors also does not appear to be of major importance. The observed patterns of nonlinearity seem to be rather stable in time, i.e., the specific choice of the analyzed period does not have any major effect on the outcomes of the tests. The relatively most distinct changes compared to the presented results were detected in the

NCEP/NCAR data when the 1960s were chosen as the testing set instead of the years 1991–2000, especially in the southern hemisphere. This difference can probably be contributed to the variations in the amount of observational data, entering the reanalysis, as discussed below. The applied tests were all based on prediction one day ahead – with an increase of the lead time, nonlinearity quickly weakened and it became undetectable for predictions more than approximately five days ahead, even in the regions where the nonlinear behavior was originally strongest. This is in good agreement with the fact that a deterministic weather prediction is impossible for too long lead times, regardless of the method.

Most of the patterns of nonlinearity identified in the NCEP/NCAR reanalysis data were also found in the outputs of the HadCM3 model. From the perspective of applied nonlinear time series analysis tasks (such as statistical downscaling carried out by nonlinear methods), the fact that a climate model is able to reproduce the character of the observed data is encouraging. Still, from the results obtained for a single representative of global climate models, it is not possible to infer whether all existing climate simulations do behave in a similar fashion. It is interesting that the correspondence of the structures found in the NCEP/NCAR and HadCM3 data tends to be better in the northern hemisphere. Although this fact can at least partially be a consequence of the specifics of the model's physics, it might also be contributed to the character of the reanalysis data. To assess the possible influence of the specific properties of the NCEP/NCAR reanalysis, we repeated some of the tests for another commonly used gridded dataset based on observations, the ERA-40 reanalysis [28]. Although some differences were found, the resemblance of the results from the NCEP/NCAR and ERA-40 data was generally strong in the northern hemisphere, but somewhat weaker in the southern one. This implies that caution is needed in interpretation of the model-reanalysis differences, particularly in the southern latitudes, as they may be a result of a limited amount of observational data used by the reanalysis (and possibly some other specifics of the NCEP/NCAR dataset), not just imperfections of the climate model. This especially applies to the period preceding the era of meteorological satellites – e.g., the amount of data entering the NCEP/NCAR reanalysis is very low before the year 1979 south of approximately 40°S [19].

We have shown that the direct comparison of prediction by linear regression and by local linear models yields nonlinearity patterns very similar to the approach based on the application of local linear models for surrogate data. A practical advantage of the direct comparison lies in its speed, as there is no need for multiple realizations of a nonlinear model. This is especially convenient in the case of an analysis like ours, carried out for thousands of grid points and repeated for numerous settings. Another benefit of the direct comparison is that it provides specific information about the potential gain from employing a nonlinear method; its fundamental drawback is that such information may only be valid for the combination of the methods applied.

7 Conclusions

By analyzing the series of selected atmospheric variables, we were able to confirm the presence of systematic geographical and seasonal variations of nonlinearity. Simple and unequivocal physical explanation of the results beyond the basic tropics/mid-latitudes and summer/winter contrast may be problematic, because the finer details of the detected patterns are probably a product of multiple influences and they are subject to the type of the predictand variable and some other factors. To find out whether any other general regularities exist would require a systematic analysis performed for a large number of variables and pressure levels. Regardless of the exact cause of the detected structures, their character was simulated fairly well by the HadCM3 model. From the practical perspective, this finding is rather promising, as it confirms that data produced by the current generation of global climate models can be utilized for the study of nonlinear properties of the climate system.

Acknowledgements. This study was supported by the Czech Science Foundation (grant 205/06/P181) and by the Ministry of Education of the Czech Republic (research plan MSM0021620860). The presented work would not be possible without the utilized datasets: NCEP/NCAR reanalysis (obtained from NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at http://www.cdc.noaa.gov), HadCM3 model outputs (provided by the Met Office Hadley Centre) and ERA-40 reanalysis (obtained from the Data Server of the European Centre for Medium-Range Weather Forecasts). The authors would also like to express their gratitude to the two anonymous reviewers of the manuscript and to R. Donner for their valuable comments and suggestions.

References

1. M. Paluš, D. Novotná: Testing for nonlinearity in weather records, Phys. Lett. A **193**, 67 (1994)
2. D. A. S. Patil, B. R. Hunt, J. A. Carton: Identifying low-dimensional nonlinear behavior in atmospheric data, Mon. Weather Rev. **129**, 2116 (2001)
3. J. Miksovsky, A. Raidl: Testing for nonlinearity in European climatic time series by the method of surrogate data, Theor. Appl. Climatol. **83**, 21 (2006)
4. M. C. Casdagli: Characterizing Nonlinearity in Weather and Epilepsy Data: A Personal View. In:*Nonlinear Dynamics and Time Series*, ed by C. D. Cutler, D. T. Kaplan (American Mathematical Society, Providence, Rhode Island 1997) pp 201–222
5. G. Sugihara, M. Casdagli, E. Habjan, et al.: Residual delay maps unveil global patterns of atmospheric nonlinearity and produce improved local forecasts, P. Natl. Acad. Sci. USA **96**, 14210 (1999)
6. J. Miksovsky, A. Raidl: Testing the performance of three nonlinear methods of time series analysis for prediction and downscaling of European daily temperatures, Nonlinear Proc. Geoph. **12**, 979 (2005)
7. M. Paluš: Detecting nonlinearity in multivariate time series, Phys. Lett. A **213**, 138 (1996)

8. A. A. Tsonis: Probing the linearity and nonlinearity in the transitions of the atmospheric circulation, Nonlinear Proc. Geoph. **8**, 341 (2001)
9. T. Schreiber, A. Schmitz: Surrogate time series, Physica D **142**, 346 (2000)
10. I. Bartos, I. M. Jánosi: Nonlinear correlations of daily temperature records over land, Nonlinear Proc. Geoph. **13**, 571 (2006)
11. V. Pérez-Muñuzuri, I. R. Gelpi: Application of nonlinear forecasting techniques for meteorological modeling, Ann. Geophysicae **18**, 1349 (2000)
12. B. Tang, W. W. Hsieh, A. H. Monahan, F. T. Tangang: Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial Pacific sea surface temperatures, J. Climate **13**, 287 (2000)
13. A. Weichert, G. Bürger: Linear versus nonlinear techniques in downscaling, Climate Res. **10**, 83–93 (1998)
14. J. T. Schoof, S. C. Pryor: Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks, Int. J. Climatol. **21**, 773 (2001)
15. M. Casaioli, R. Mantovani, F. P. Scorzoni, et al.: Linear and nonlinear post-processing of numerically forecasted surface temperature, Nonlinear Proc. Geoph. **10**, 373 (2003)
16. E. Eccel, L. Ghielmi, P. Granitto, et al.: Prediction of minimum temperatures in an alpine region by linear and non-linear post-processing of meteorological models, Nonlinear Proc. Geoph. **14**, 211 (2007)
17. W. von Bloh, M. C. Romano, M. Thiel: Long-term predictability of mean daily temperature data, Nonlinear Proc. Geoph. **12**, 471 (2005)
18. E. Kalnay, M. Kanamitsu, R. Kistler, et al.: The NCEP/NCAR 40-year reanalysis project, Bull. Amer. Meteor. Soc. **77**, 437 (1996)
19. R. Kistler, E. Kalnay, W. Collins, et al.: The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation, Bull. Amer. Meteor. Soc. **82**, 247 (2001)
20. C. Gordon, C. Cooper, C. A. Senior, et al.: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments, Clim. Dynam. **16**, 147 (2000)
21. T. C. Johns, J. M. Gregory, W. J. Ingram, et al.: Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios, Clim. Dynam. **20**, 583 (2003)
22. E. Ott, T. Sauer, J. A. Yorke (eds.): *Coping with Chaos: Analysis of Chaotic Data and The Exploitation of Chaotic Systems* (John Wiley & Sons, New York 1994)
23. H. Kantz, T. Schreiber: *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge 1997)
24. D. S. Wilks: *Statistical Methods in the Atmospheric Sciences*, 2nd edn (Elsevier, Amsterdam 2006)
25. V. Venema, S. Bachner, H.W. Rust, C. Simmer: Statistical characteristics of surrogate data based on geophysical measurements, Nonlinear Proc. Geoph. **13**, 449 (2006)
26. T. Schreiber, A. Schmitz: Improved surrogate data for nonlinearity tests, Phys. Rev. Lett. **77**, 635 (1996)
27. R. Hegger, H. Kantz, T. Schreiber: Practical implementation of nonlinear time series methods: The TISEAN package, CHAOS **9**, 413 (1999)
28. S. M. Uppala, P. W. Kållberg, A. J. Simmons, et al.: The ERA-40 re-analysis, Quart. J. R. Meteorol. Soc. **131**, 2961 (2005)

Prediction of Extreme Events

Sarah Hallerberg, Jochen Bröcker, and Holger Kantz

Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38,
D 01187 Dresden, Germany, kantz@pks.mpg.de

Abstract. We discuss concepts for the prediction of extreme events, based on time series data. We consider both probabilistic forecasts and predictions by precursors. Probabilistic forecasts employ estimates of the probability for the event to follow, whereas precursors are temporal patterns in the data typically preceeding events. Theoretical considerations lead to the construction of schemes that are optimal with respect to several scoring rules. We discuss scenarios in which, in contrast to intuition, events with larger magnitude are better predictable than events with smaller magnitude.

AU: Please provide keywords.

1 Prediction of Events

Geophysical processes are characterized by complicated time evolutions, which are generally aperiodic on top of potential seasonal oscillations and exhibit large fluctuations. This applies to all processes related to or caused by the atmosphere, but also is true for geological processes. The prediction of extreme events is of particular interest due to their usually large impact on human life, as exemplified by earthquakes, storms, or floods. For many of such processes no detailed physical models and also no useful observations to put into such models are available, such that their prediction is very often a time series task. But even in much more favorable situations where sophisticated models, sophisticated observations, and hence model based forecasts exist, extreme events pose challenges. Due to its immense relevance in all aspects of daily life, the weather has been subject to forecasts for centuries, on various levels of sophistication. Weather predictions are nowadays generated on a daily basis with the involvement of an enormous body of scientific results and computational resources. This type of prediction is different though from the prediction of extreme events in one specific aspect: Weather predictions are designed to perform well under a large variety of "typical" situations, most of which are not extreme in any sense. Hence, a prediction scheme which works excellently on average might completely fail in rare but extreme situations. Indeed, there

have been situations in recent years where public warnings of extreme weather situations turned out to be inadequate, such as with the Great Strom of October 15/16, 1986 in South England [1, 2] or the extreme precipitation event in Saxony on 12/13 August, 2002, leading to floodings of the river Elbe. Both events were misplaced or overlooked by medium range weather forecasts [3]. In this chapter, we discuss the predictability and prediction schemes for extreme events, based exclusively on time series analysis. We are not employing any prior knowledge about physical processes or models for the phenomenon under study, but rely only on recordings of past data. For weather prediction over lead times larger than a few hours, this approach would not be too reasonable, as the atmospheric phenomena governing the evolution of weather are fairly well understood, and physical models have been demonstrated to yield useful forecasts over a large spectrum of spatial and temporal scales. In a variety of other circumstances though, a time series approach is the only possibility for predictions at all. A frequent reason is that good models for the specific situation are not available. But even if, to generate a useful forecast, some state variables have to be fed into such models, which in turn have to be estimated from (often rather incomplete and noisy) measurements, which often requires more than reasonable effort.[1]

In the time series setting, we will assume that the unknown dynamics is a nonlinear and stochastic process, which can generally be described in a nonparametric and data driven way by all its joint probabilities. An even stronger assumption is that the dynamics is a generalized Markov process, which can be described by a finite set of transition probabilities (rather than the infinite set of *all* joint probabilities). For the present work, this assumption is only important in so far as the envisaged prediction methods are suboptimal if the process is not Markovian. The stochastic character of the processes under concern suggest that the forecasts are probabilistic. In other words, our schemes will provide us with probabilities of an event to come. We will argue that probabilistic predictions will require performance measures which are different from standard ones such as the root mean squared prediction error. As a possible alternative, the concept of scores provides measures to quantify the success of probabilistic forecasts. We will discuss and employ two popular examples, the Brier score and the Ignorance score. A problem with scores in connection with extreme events is that the average score over many forecast instances is taken as a quality measure for the forecasting system. Since by its very nature the base rate of an extreme event is very small, there are very few instances where having a good forecast actually makes a difference, whence the score of an excellent forecast is on average only marginally better than the score of a just mediocre forecast.

A probably much more suitable scoring scheme for our purpose is the Receiver Operating Characteristic (ROC). An important observation is that both

[1] This problem, known as data assimilation, takes about 50% of the total CPU time required to generate a medium range (≅10 days) weather forecast [4].

the scores and the ROC encourage essentially similar forecasting strategies, namely to use the probability of the event given the previously observed time series. A simple but very effective approximation of this conditional probability leads to a scheme using precursors of extreme events, an approach which can be motivated independently. We discuss the performance of our prediction schemes for simple model processes, and verify these findings by predictions of turbulent wind gusts and large fluctuations in laboratory turbulence. As a striking result, we find that under certain circumstances, events are the better predictable the more extreme they are.

2 Time Series Data and Conditional Probabilities

Suppose we are given a time series $\{x_1, \ldots, x_N\}$ of N data points, which are evenly distributed in time, where N is called the sample size. The time series can be vector valued, but often is scalar. Regardless of what process and what measurement function creates the data, we will interprete this sequence as being generated by a stochastic process. As it is well known [5, 6], a stochastic process is fully characterized by all its joint cumulative probabilities, $\mathbb{P}(x_{i_1} < \theta_1, x_{i_2} < \theta_2, \ldots, x_{i_k} < \theta_k)$ for all k, and all possible sequences of indices $i_j; j = 1 \ldots k$. For simplicity we will assume the process to be stationary, which implies that all joint probabilities depend only on times relative to the time of the first argument, such that we can set $i_1 = 0$ always. Moreover, in the following, we will order the time indices in descending order, the indices further right refer to times further in the past.

Stationarity is a property which almost never applies to realistic processes such as atmospheric turbulence. Applying concepts from stationary processes to data which might originate from a non-stationary process could result in reduced performance of our prediction algorithms, but is not a fundamental problem in the examples of non-stationary data,[2] which we study in this contribution. Moreover, the methods proposed in this contribution should also be suitable in the special case of non-stationarity due to slowly varying system parameters, as it was argued in [7]. In terms of the prediction of wind-speeds in high frequency wind speed data those slowly varying system parameters are related to changing weather conditions or change of the time of the day.

From a joint probability one easily computes conditional probabilities, which denote the probability to find a specific value for the variable x_Δ, if the values for the past variables $x_0, x_{-1}, \ldots, x_{1-\tau}$ are given. Joint and conditional probabilities are connected by the well known Bayes rule

$$p(x_\Delta | x_0, x_{-1}, \ldots, x_{1-\tau}) := p(x_\Delta, x_0, x_{-1}, \ldots, x_{1-\tau}) / p(x_0, x_{-1}, \ldots, x_{1-\tau}). \tag{1}$$

[2] Data are called non-stationary, if the null-hypothesis of stationarity can easily be rejected.

Note that Bayes theorem and many other expressions in this contribution could also be formulated in terms of probability densities. Since the distinction between probabilities and probability densities is not of relevance, as far as numerical estimates of probability densities are involved, we formulate the correspondent expressions mostly in terms of probabilities and only refer to probability densities, when analytical considerations are involved.

Conditional probabilities provide the information needed for (probabilistic) predictions: Knowing $p(x_\Delta|x_0, x_{-1}, \ldots, x_{1-\tau})$ as a function of $x_0, x_{-1}, \ldots, x_{1-\tau}$, and given specific values for the $x_0, x_{-1}, \ldots, x_{1-\tau}$'s, one can calculate the probability that the observation Δ time steps in the future will fall into a given interval. Generally, the probability density function or probability mass function of x_Δ will be the sharper, the further into the past the conditioning extends. Ideally, the entire past of the process would be observed and the conditional PDF for infinite conditioning would be known, thus yielding optimal knowledge of the future (which does not mean that this conditional probability becomes necessarily sharp peaked like a δ–function). In practice this is absolutely out of reach. The practical difficulty here is to estimate the conditional probability from the sample of N data points, as this estimate gets the worse the larger τ. If the observed time series were governed by a generalized Markov process of order τ_0, then the τ_0-step conditioning would be optimal and any additional conditioning would not improve (or in fact change) the forecast. In general, although the process is not Markovian, finite conditioning still provide a rather good approximation to infinite conditioning, or in more colloquial terms, there is nothing wrong about basing one's predictions on finite τ-conditioning, the worst to happen is that this is sub-optimal.

So far, we have been discussing $p(x_\Delta|x_0, x_{-1}, \ldots, x_{1-\tau})$. When we want to predict the occurrence of events, one could in principle make a prediction of x_Δ, and then derive from the value of x_Δ whether this value fulfills the criterion for an event to follow or not. However, simpler, faster, and more general is the following approach: We assume to have a second time series $\{\chi_1, \chi_2, \ldots, \chi_N\}$, $\chi_i \in \{0, 1\}$, which is the event time series, where $\chi_i = 1$ generally means that an event takes place. In many applications this series is derived from the original time series $\{x_1, \ldots, x_N\}$, for example by defining

$$\chi_i = \begin{cases} 1 & x_i \geq \eta, \\ 0 & x_i < \eta, \end{cases} \tag{2}$$

if the event under study is defined as a crossing of a given threshold η, or

$$\chi_i = \begin{cases} 1 & x_{i+1} - x_i \geq \eta, \\ 0 & x_{i+1} - x_i < \eta, \end{cases} \tag{3}$$

if the event is defined as an increment $x_{i+1} - x_i$ larger or equal to η. However, the events could also be defined using the observation of some other quantity. In fact, the only important requirements are that at the time when the forecast has to be made, χ_Δ is unknown, and that its actual value is revealed later.

Since the relative time distance Δ between the last observed value x_o and the event is irrelevant for the following discussion, we shift the time indices of the event series in such a way that the event with time index i is to be predicted from the observation sequence terminating with the time index i, regardless of how far into the future the prediction is made.

In the following we will concentrate on events, which are extreme. For our purposes we specify that we understand as an extreme event an event which is rare, which is recurrent, and to which we can assign a magnitude η, which assumes a large value if the event takes place.[3]

Using then both time series, one can construct the joint probabilities $p(\chi_0, x_0, x_{-1}, \ldots, x_{1-\tau})$ which contain all dependencies between the sequence of observations down to τ temporal steps into the past. By the stationarity assumption this joint probability is equal to $p(\chi_i, x_i, x_{i-1}, \ldots, x_{i+1-\tau})$ for any i. The prediction schemes to be discussed in the following will exploit such joint probabilities, which themselves will be estimated from the data records. We will abbreviate a vector of τ successive time series elements $(x_i, x_{i-1}, \ldots, x_{i-\tau+1}) =: \mathbf{x}_i$. The explicit value of τ is suppressed in this notation. Also note that strictly speaking, $p(\chi_i, \mathbf{x}_i)$ is a probability in the argument χ_i, but a density in the argument $\mathbf{x}_i$. The interpretation is that for any volume V in $\mathbb{R}^k$,

$$\mathbb{P}(\chi_i = 1, \mathbf{x}_i \in V) = \int_V p(\chi_i, \mathbf{x}_i) d\mathbf{x}_i. \tag{4}$$

3 Probabilistic Forecasts and Prediction Through Precursors

Assuming the process which generates the observations and the events to be stochastic calls for probabilistic predictions. Such predictions consist of random variables $\hat{p}_i$, called forecast probabilities, which are issued at time i. If the τ-dimensional vector $\mathbf{x}_i$ is used to represent the current state of the process, then the forecast probabilities are a function $\hat{p}(\mathbf{x}_i)$ of $\mathbf{x}_i$ with values between zero and one. The function $\hat{p}(\mathbf{x}_i)$ is called a probabilistic predictor. Intuitively, one would hope that $\hat{p}(\mathbf{x}_i)$ gives the probability of $\chi_i = 1$ given $\mathbf{x}_i$, or

$$\hat{p}(\mathbf{x}_i) = p(\chi_i = 1 | \mathbf{x}_i). \tag{5}$$

We will see in Sect. 4 that many reasonable measures of forecast success support this intuition, that is, they give maximum possible scores if $\hat{p}(\mathbf{x}_i)$ indeed agrees with the probability of $\chi_i = 1$ given $\mathbf{x}_i$.

A seemingly different way to motivate $p(\chi_i = 1 | \mathbf{x}_i)$ as a good forecast probability is through reliability. Reliability means that on condition that the

[3] Note that this is not a general definition, and that other people might understand the term *extreme event* in a different way.

forecast (approximately) equals z, the event should occur with a relative frequency (approximately) equal to z, too. As an optimality criterion, reliability is not sufficient to single out a particular forecasting scheme, since *any* conditional probability of the form $p(\chi_i = 1|I)$ is reliable, independent of what I is. In particular, the unconditional probability $\hat{p}(\mathbf{x}) = \text{ const. } = p(\chi_i = 1)$ is reliable as well. Hence, in addition to reliability, the forecast should feature a high correlation with the actual event. This property is known as sharpness. It can be demonstrated that $p(\chi_i = 1|\mathbf{x}_i)$ is indeed the reliable forecast which features maximum sharpness among all functions of $\mathbf{x}_i$. As we will briefly discuss in Sect. 4, it is in fact for the same reason that $\hat{p}(\mathbf{x}_i) = p(\chi_i = 1|\mathbf{x}_i)$ achieves optimal scores.

If experimental data is to be investigated, the exact shape of $p(\chi_i = 1|\mathbf{x}_i)$ (or any other probability, for that matter) is of course unknown and has to be estimated from data. There exist many sophisticated algorithms to approximate conditional probabilities. These algorithms, although of great value in the analysis of time series, often result in exceedingly complex models for $p(\chi_i = 1|\mathbf{x}_i)$ and are therefore of limited use in real time implementations, where simple and fast algorithms are required.

An intuitive and indeed widespread approach to the prediction of extremes is to search for precursors (e.g., [8, 9]). A precursor is a pattern in the time series, i.e., a sequence of τ values $(x_0, \ldots, x_{1-\tau})$ which "typically" precedes an event. In the following, precursors will be denoted by $\mathbf{u} := (u_0, \ldots, u_{1-\tau})$. The assumed stochastic nature of the process implies that there are events which are not preceeded by a sequence of observations which are similar to the precursory pattern, but that there are also incidents where the sequence of observations is very similar to the precursor, though no event follows. The prediction by precursors requires first to choose one or more precursory patterns (after fixing the parameter τ). We are going to address the issue of how to identify such precursors at the end of this section. Suppose for the moment that we had already chosen a precursor $\mathbf{u}$ in one way or another. We can then define an *alarm volume* $V(\delta, \mathbf{u})$ around each precursor as the set of all $\mathbf{x}_i$ for which $||\mathbf{x}_i - \mathbf{u}|| \leq \delta$, where $||\cdot||$ denotes a norm which can be, for example, the Euclidean norm or the maximum norm. When using the maximum norm, then the alarm volume consists of all time series segments which fall into a δ-tube around the precursory pattern $\mathbf{u}$. The challenge in this approach is to determine good precursory structures, since they are the core of this prediction scheme. Two approaches have been studied, both of some intuitive appeal:

Strategy I: After having collected all events $\chi_i = 1$ from the recorded data, one studies what typically happens before these events. This leads one to study the conditional probability $p(\mathbf{x}_i|\chi_i = 1)$. A reasonable way to extract a distinguished pattern $\mathbf{u}$ from these is to ask for $\mathbf{u}$ to maximize this probability. Then, the precursor represents the time series pattern which is most probably observed before an event takes place.

Strategy II: Alternatively, one can look through all possible precursory patterns $\mathbf{x}$ and define $\mathbf{u}$ to be the pattern for which $p(\chi_i = 1|\mathbf{x}_i)$ is maximal, that is, the pattern which has the largest probability to be followed by an event.

Note that the conditional probability used for strategy II is the conditional probability which was suggested for probabilistic forecasting at the beginning of this section.

Strategy I might seem a very intuitive approach to look for precursory structures. However, the considerations in Sect. 4 and the results in Sect. 5 will show that $p(\chi_i = 1|\mathbf{x}_i)$ is in some sense the optimal prediction scheme. As strategy II essentially approximates this conditional probability, the performance obtained by identifying precursors with strategy I is expected to be worse or equal, but not better than with strategy II.

4 Scoring Schemes

In this section, the question of how to quantify the performance of forecasts is addressed. Performance measures are important not only in order to rank existing forecast schemes but also in the design of such schemes, for example the tuning of free parameters. Measuring the success of predictions in terms of how "close" they eventually come to the truth is a paradigm which presumably requires no further motivation. The (root) mean squared error, briefly revisited in Sect. 4.1, is just one among many possible variants of this paradigm, albeit a very important and popular one. If we envisage to formulate our forecasts in terms of probabilities though, the paradigm cannot be applied readily without modification, as the notion of "distance" between forecast and truth ceases to be meaningful. But probability forecasts essentially quantify how likely a given potential event will come true, thus already providing a sort of self rating. Hence it seems reasonable to value the success of a probability forecast in terms of how confident the forecast was of the event which eventually occurred, in relation to other events which did not. This idea is implemented in the concept of *scores*, explained in Sect. 4.2.

A third approach to measuring the quality of probabilistic forecast is the Receiver Operating Characteristic (ROC), presented in Sect. 4.3. Different from scores, the ROC, albeit taking the probabilistic character of the forecast into account, is insensitive to the reliability of the forecast.

4.1 RMS Error

When predicting future values of some time series, a commonly used criterion to quantify the success is the root mean squared (RMS) prediction error: Let $\hat{x}_i$ be a prediction generated by some algorithm and x_i the observation, then one defines the RMS error as

$$\bar{e} = \frac{1}{N} \sum_{i=1}^{N} (\hat{x}_i - x_i)^2 \ . \tag{6}$$

The same quantity can be computed for predictions $\hat{\chi}_i$ of the event variable χ_i, where both χ_i and $\hat{\chi}_i$ can only assume the values 0 or 1. In our context, where $\chi_i = 1$ is rare, such a scoring will favor predictions schemes which predominantly ignore the occurrence of events at all. Assume that χ_i denotes the occurrence of an earthquake of large amplitude on day i. It is evident that a predictor which does never predict the earthquake to come will fail only once during very many years, thus yielding an almost zero RMS error. In the uniform average over time, mis-prediction of rare events is simply averaged out. Also, the RMS error relies on the norm of the difference between prediction and future value and is therefore symmetric, i.e., the "costs" implied by a false alarm are identical to missing a hit. When discussing extreme events, such costs are usually very different, but their quantification usually is an art in its own, so that no values for these costs are available. This suggests that a better scoring should consider these two types of errors, namely false alarm and missed hit, separately.

4.2 Brier Score and Ignorance

A *scoring rule* is a function $S(\hat{p}, z)$ where $\hat{p} \in [0, 1]$ and z is either zero or one. For our purpose we specify that $z = \chi_i$. A scoring rule [10, 11, 12, 13] effectively defines two functions $S(\hat{p}, 1)$, quantifying the score in case the forecast is $\hat{p}$ and the event happens, and $S(\hat{p}, 0)$, quantifying the score in case the forecast probability is $\hat{p}$ but the event does not happen. Two important examples are the ignorance score, given by the scoring rule

$$S(\hat{p}, \chi_i) := -\log(\hat{p}) \cdot \chi_i - \log(1 - \hat{p}) \cdot (1 - \chi_i), \tag{7}$$

and the Brier score, given by the scoring rule

$$S(\hat{p}, \chi_i) := (\chi_i - \hat{p})^2 = (1 - \hat{p})^2 \cdot \chi_i + \hat{p}^2 \cdot (1 - \chi_i). \tag{8}$$

These definitions imply the convention that a smaller score indicates a better forecast.

A score is a "point–wise" (evaluated at every single time instance) measure of performance. It quantifies the success of individual forecast instances by comparing the random variables $\hat{p}$ and χ pointwise. The general quality of a forecasting system (as given here by the random variable $\hat{p}$) is commonly measured by the average score $E\left[S(\hat{p}, \chi_i)\right]$, which can be estimated from the empirical mean

$$E\left[S(\hat{p}, \chi_i)\right] \cong \frac{1}{N} \sum_{i=1}^{N} S(\hat{p}_i, \chi_i) \tag{9}$$

over a sufficiently large data set $(\hat{p}_i, \chi_i)$.

The rationale behind the two mentioned scoring rules, the ignorance and the Brier score, is rather obvious. If the event occurs, the score should become better (i.e., decrease) with increasing $\hat{p}$, while if it does not occur, the score should become worse (i.e. *in*crease) with increasing $\hat{p}$. But why then not taking just $1-\hat{p}$ if the event occurs, and $\hat{p}$ if it does not? To see the problem with this "linear" scoring rule, define the *scoring function*

$$s(\hat{p}, q) := S(\hat{p}, 1) \cdot q + S(\hat{p}, 0) \cdot (1 - q) \tag{10}$$

where q is another probability, i. e. , a number in the unit interval. Note that the scoring function is the score averaged over cases where the forecast is $\hat{p}$ but in fact q is the true distribution of χ. In view of the interpretation of the scoring function, it seems reasonable to require that the average score of the forecast $\hat{p}$ should be best (i.e. minimal) if and only if $\hat{p}$ in fact coincides with the true distribution of χ. This means that the *divergence function* (or loss function)

$$d(\hat{p}, q) := s(\hat{p}, q) - s(q, q) \tag{11}$$

has to be positive definite, i. e. , it has to be nonnegative, and zero only if $\hat{p} = q$. A scoring rule with the corresponding divergence function having this property is called *strictly proper* [12, 14]. The divergence function of the Brier score for example is $d(\hat{p}, q) := (\hat{p}-q)^2$, demonstrating that this score is strictly proper. While the ignorance is proper as well, the linear score though is easily shown to be *improper*.

4.3 The Receiver Operating Characteristic

The ROC [15] is a concept originating in signal detection, but it is applicable to any problem in which, based on some evidence, we have to decide whether a certain event will happen or not, for example if it will rain tomorrow or if an extreme wind gust will occur within the next minute. We assume the evidence to be a (rather general) random variable $\mathbf{x}$. To stay with the wind example, the evidence used in this case could be the delay vector $\mathbf{x}_i = (x_i, x_{i-1}, \ldots, x_{i-\tau+1})$ of previous wind measurements. Suppose that $r(\mathbf{x}_i)$ is a real-valued function of the evidence with the idea that a large r is indicative of an event, while a small r is indicative of a non-event. An $r(\mathbf{x}_i)$ exceeding a certain threshold δ could be interpreted as signaling an impending event. The variable $r(\mathbf{x}_i)$ will henceforth be referred to as the decision variable. Referring back to Sect. 3, if we use the precursor technique to forecast an event, the decision variable could be the (negative of) the Euclidean distance of the delay vector $\mathbf{x}_i$ to the (pre-defined) precursor $\mathbf{u}$

$$r(\mathbf{x}) := -\|\mathbf{x}_i - \mathbf{u}\|. \tag{12}$$

Giving an alarm if $r(\mathbf{x}_i) \geq \delta$ is equivalent to giving an alarm if $\mathbf{x}_i$ falls into the alarm volume V_δ. Alternatively, (an approximation to) the conditional probability $r(\mathbf{x}_i) = p(\chi_i = 1|\mathbf{x}_i)$ could be used as decision variable.

The ROC curve for a certain decision variable r comprises a plot of the *hit rate*

$$H(\delta) := p(r \geq \delta | \chi_i = 1) \tag{13}$$

versus the *false–alarm rate*

$$F(\delta) := p(r \geq \delta | \chi_i = 0), \tag{14}$$

with δ acting as a parameter along the curve. Alternative names for the hit rate are rate of true positives or the power of the test ($r \geq \delta$). Alternative names for the false–alarm rate are rate of false positives or the size of the test ($r \geq \delta$). It follows readily from the definitions that both H and F are monotonously decreasing functions of δ with limits 0 for increasing δ and 1 for decreasing δ. Hence, the ROC curve is a monotonously *in*creasing arc connecting the points $(0, 0)$ and $(1, 1)$. Furthermore, note that monotonically increasing transformations of the decision variable do not change the ROC at all, as is easily seen using the definitions of the hit rate and the false alarm rate. This is exactly the reason why, when using the precursor approach to predict events, it is already sufficient to specify the level sets V_δ as a function of δ, but not necessary to assign a probability value to each level set. A typical ROC curve is shown in Fig. 1.

The obvious question is of course as to when a ROC curve should be considered good. Arguably, a decision variable r_1 should be taken as superior to another decision variable r_2, if for any fixed false–alarm rate F, the hit rate H_1 of r_1 is equal or larger than the hit rate H_2 of r_2. If this is the case, we will refer to r_1 as being *uniformly superior* to r_2. It can be demonstrated that the decision variable $p(\chi_i = 1|\mathbf{x}_i)$ is uniformly superior to any decision variable of the form $r(\mathbf{x}_i)$ (this follows from the Neyman–Pearson–Lemma [16] and the fact that $p(\chi_i = 1|\mathbf{x}_i)$ is a monotonically increasing function of the likelihood ratio, see Eq. (22)).

As a consequence, $p(\chi_i = 1|r)$ is uniformly superior to any transformation $\phi(r)$ (in particular r itself). But if r is a function of $\mathbf{x}_i$, then $p(\chi_i = 1|\mathbf{x}_i)$ is still uniformly superior to $p(\chi_i = 1|r(\mathbf{x}_i))$. An easy calculation will reveal that the slope of the ROC curve is a monotonically increasing function of $p(\chi_i = 1|r)$, so replacing the decision variable r with the slope of the ROC curve at r is an alternative way of getting the optimal transformation of r. If the ROC curve is concave though, then the slope (as a function of r) is monotonically increasing, thus using it as a new decision variable does not alter the ROC plot. We can conclude that a concave ROC is optimal in that it cannot be any further improved by a transformation of the decision variable.

If we have to compare two arbitrary decision variables r_1 and r_2, then the notion of "uniformly superior" is not so useful, as the two ROC curves might cross. This is a problem if a criterion is required in order to optimize a prediction algorithm, in particular, when we search for optimal precursors. There is no reason why the ROC curves corresponding to any two predictors should not cross. Hence, summary statistics of ROC curves are needed, for example the following:

Proximity to (0,1): A good ROC should be close to the point $(0, 1)$, that is where the false–alarm rate is zero while the hit rate is 1. The point closest to $(0, 1)$ would simultaneously define an operation point for the algorithm.

Area under ROC curve: The area under the ROC curve (AUC) is a well established summary index for ROC curves, which should be maximal. It can be shown that this quantity gives the probability that on an instance when the event takes place, the decision variable is actually larger than on an independent instance when the event does not take place. It is a global quantity, averaging over all alarm rates.

Maximal hit rate for fixed alarm volume: Optimizing ROC for precursors by asking for a maximal hit rate without any further constraints is not a useful criterion, since all decision variables have a maximum hit rate of 1, achievable by just giving always alarms. Fixing the alarm volume, this criterion leads to precursors according to strategy I of Sect. 3. Note that the false alarm rate is not considered at all in such an optimization, so that the optimal hit rate for fixed alarm rate might be achieved at the cost of an unreasonably large false alarm rate. Inverting the criterion to minimizing the false alarm rate for fixed alarm volume leads to the same precursor.

Ratio of hit rate and false alarm rate: A maximum ratio of hit rate versus false alarm rate in the limit of small false alarm rates yields another well established summary index, the slope of the ROC curve at the origin. If the ROC is concave, this is the same as the *overall* maximum ratio of hit rate versus false alarm rate. Maximizing this summary index leads to the prediction scheme called *strategy II* in Sect. 3.

It should be noted that an uniformly superior decision variable is superior with respect to any mentioned summary index, but not vice versa. A decision variable might for example have a larger AUC than another, but still their ROC curves might cross, in which case neither of the two is uniformly superior to the other.

4.4 Applying the Scoring Schemes for the Prediction of Extreme Events

As has been argued in the context of the RMS error, the Brier score might have its shortcomings if applied to forecasts of very unlikely events. If the overall probability of the event is very small, a forecast which successfully separates events from non-event gets little credit over a forecast which plainly states that the event will not happen at all. A more formal analysis shows that this is due to the loss function of the Brier score depending only on the difference between the forecast and the correct probability.

The ignorance in contrast severely punishes erroneous forecasts close to both one or zero. In fact, forecasting zero probability for an event which actually occurs, or holding an event for certain which then fails to materialize

yields a score of infinity. Hence the ignorance might be a more appropriate score for extreme event forecasts, albeit a rather harsh one.

The ROC avoids a direct dependence on the overall probability of events and non-events by definition.

$$H(\delta) := p(r \geq \delta | \chi_i = 1) = \frac{p(r \geq \delta, \chi_i = 1)}{p(\chi_i = 1)} \tag{15}$$

$$F(\delta) := p(r \geq \delta | \chi_i = 0) = \frac{p(r \geq \delta, \chi_i = 0)}{p(\chi_i = 0)} \tag{16}$$

Since the cumulative probabilities $p(r \geq \delta | \chi_i = 1)$ and $p(r \geq \delta | \chi_i = 0)$ of giving an alarm and observing an event (non-event) are normalized with the total probability to find events (non-events) the rates do not depend explicitely on the total probabilities. However one cannot exclude an implicit dependence which is given through the relation between precursor and event or the definition of the events.

5 Performance of the Prediction Schemes

For the purpose of precise understanding, the two different prediction strategies of the precursor based prediction schemes were evaluated for an extremely simple time series model, an AR(1) process, $x_{i+1} = a x_n + \xi_i$, with the correlation coefficient a and the sequence of normal distributed i. i. d. random numbers ξ_i, in [17] and [18]. As events χ_i, we considered both threshold crossing, $x_i \geq \eta$, and large increments, $x_{i+1} - x_i \geq \eta$. Due to the short–range correlation of the process, we used only the last value x_i as evidence, i.e., $\tau = 1$ and the vector $\mathbf{x}_i$ reduces to x_i. Correspondingly the precursor consists of the special value u, which x_i can obtain and the alarm volume $V(\delta, \mathbf{u})$ becomes an alarm interval $I(\delta, u)$. This setting allows us to compute all relevant expressions, analytically for the example of the AR(1) process. We can then compare the different prediction strategies by creating ROC-curves. This was done by expressing the hit rate $H(\delta)$ and the rate of false alarms $F(\delta)$ in terms of probability densities,

$$\begin{aligned} H(\delta) &= \int_{I(\delta,u)} \rho(x_i | \chi_i = 1) dx \\ F(\delta) &= \int_{I(\delta,u)} \rho(x_i | \chi_i = 0) dx. \end{aligned} \tag{17}$$

The values of the precursor u were taken to be either the maximum of $p(x_i|\chi_i = 1)$ (strategy I) or of $p(\chi_i = 1|x_i)$ (strategy II).

5.1 Generation of ROC-Curves

One can obtain numerical estimates for the conditional probabilities $p(x|\chi_i = 1)$ and $p(\chi_i = 1|x_i)$ by using a kernel estimator or by "binning and counting".

Since for all stochastic processes studied in this section sufficiently many data points were available, we used the latter numerical method in order to compute the ROC-curves shown in Figs. 1 and 2.[4] In the next step we identified the values of x_i in which $p(x_i|\chi_i = 1)$ and $p(\chi_i = 1|x_i)$ are maximal and used them as precursors u_I and u_{II}. We can then define sets of alarm intervals $I(u_I, \delta)$ and $I(u_{II}, \delta)$ and count for each value of δ, how many values of our process are within this interval and how many of these alarms actually were followed by an event. Both, the analytical results and the numerical results for the AR(1) process were in good agreement and are displayed in Fig. 1.

5.2 Comparing Different Strategies of Identifying the Precursor

The first essential finding is that consistent with our theoretical investigations, strategy II is uniformly superior to strategy I, for arbitrary parameters of the AR(1) process and for all event sizes. Figure 1 illustrates this result, which is in good agreement with the theoretical considerations in Sect. 4.3 about optimal ROC curves obtained by using $p(\chi_i = 1|\mathbf{x}_i)$ as decision variable.

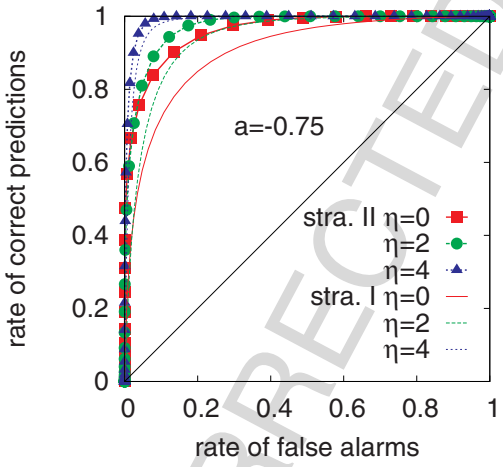


Fig. 1: Performance of strategies I and II for the prediction of increments generated by an AR(1) process with correlation coefficient $a = -0.75$. Different event magnitudes $\eta = (x_{i+1} - x_i)/\sigma$, with σ being the variance of the process under study, are compared.

We will now try to obtain an intuitive understanding of the superiority of strategy II (finding a precursor which maximizes $p(\chi_i = 1|\mathbf{x_i})$) over strategy I (maximizing $p(\mathbf{x}_i|\chi_i = 1)$) by investigating the slope of the ROC plot in the

[4] In Sect. 6.1 we applied a box-kernel estimator for the same purpose. The details of the box-kernel estimator will be explained in Sect. 6.1.

vicinity of the origin. We therefore assume again the more general case of a multi-dimensional decision variable $\mathbf{x}$ and a more dimensional precursor $\mathbf{u}$. As shown above, the hit rate $H(\delta)$ and the false alarm rate $F(\delta)$ (for arbitrary precursor $\mathbf{u}$) can be expressed by conditional probability densities $\rho(\mathbf{x}_i|\chi_i)$

$$\begin{aligned} H(\delta) &= \int_{V(\delta,\mathbf{u})} \rho(\mathbf{x}_i|\chi_i = 1) d\mathbf{x}_i \\ F(\delta) &= \int_{V(\delta,\mathbf{u})} \rho(\mathbf{x}_i|\chi_i = 0) d\mathbf{x}_i \end{aligned} \tag{18}$$

and the slope m of the ROC curve is given by

$$m = \frac{dH(\delta)}{dF(\delta)}. \tag{19}$$

For small alarm volumes $V(\delta, \mathbf{u})$

$$H(\delta) \approx c\delta^\tau \rho(\mathbf{u}|\chi_i = 1), \quad F(\delta) \approx c\delta^\tau \rho(\mathbf{u}|\chi_i = 0). \tag{20}$$

The geometry parameter c defines how the alarm volume scales with δ^τ. Inserting Eq. (20) in Eq. (19), this factor cancels out. Hence the slope of the ROC curve in the vicinity of the origin is given by

$$m \approx \frac{\rho(\mathbf{u}|\chi_i = 1)}{\rho(\mathbf{u}|\chi_i = 0)}. \tag{21}$$

The right hand side is known as the likelihood ratio. Using Bayes' Theorem, the right side of Eq. (21) can be written in terms of the conditional probability $p(\chi_i = 1|\mathbf{x})$ and the total probability $p(\chi_i = 1)$ to find events:

$$m \approx \frac{p(\chi_i = 1|\mathbf{u})}{(1 - p(\chi_i = 1|\mathbf{u}))} \frac{p(\chi_i = 1)}{(1 - p(\chi_i = 1))}. \tag{22}$$

Note that the total probability to find events is given by the process under study and does not influence the choice of the precursor. The specific precursor $\mathbf{u}$ which maximizes m is given by setting the derivative of m with respect to $\mathbf{u}$ equal to zero. One easily finds that this requires $\partial p(\chi_i = 1|\mathbf{u})/\partial \mathbf{u} = 0$, a condition which is fulfilled by the $\mathbf{u}$ which maximizes $p(\chi_i = 1|\mathbf{u})$. This is exactly what we called strategy II before. Strategy I aims at maximizing $\rho(\mathbf{u}|\chi_i = 1)$, but does not take the denominator of the likelihood ratio (see Eq. (21)) into account. Hence we have shown that in the vicinity of the origin strategy II is always superior or equal to strategy I. This corresponds to the considerations in Sect. 4.3 concerning the uniform superiority of predicting through the conditional probability $p(\chi_i = 1|\mathbf{x}_i)$.

In the limit of small alarm volumes, the probabilistic prediction and the precursor based prediction according to strategy II are equivalent. However, for larger alarm volumes the influence of the specific structure of $p(\chi_i = 1|\mathbf{x}_i)$ leads to slightly different predictions, especially, when $p(\chi_i = 1|\mathbf{x}_i)$ is not symmetric around its maximum or exhibits multiple maxima. Then the alarm volume $V(\delta, \mathbf{u})$ in general does not match any level set of $p(\chi_i = 1|\mathbf{x}_i)$.

5.3 The Influence of the Event Magnitude

As the second relevant finding we quote here results from detailed studies which show that for a large variety of processes, the larger the events to be predicted, the better is the corresponding ROC curve [17, 19]. The magnitude η of an event is measured in units of the standard deviation σ, e.g., $\eta = (x_{i+1} - x_i)/\sigma$ for the prediction of increments and $\eta = x_{i+1}/\sigma$ for the prediction of threshold crossings.

The mentioned studies provide a deeper understanding of empirical observations reported for the prediction of avalanches in models which display self organized criticality [20] and in multi-agent games [21]. Our investigations also led to a criterion as to whether larger events are better predictable than smaller or not, depending on the joint distribution of precursor and event [19]. With some restrictions imposed on the length of the data set, the criterion can be evaluated numerically for arbitrary time series data.

We found some especially interesting results for the prediction of large increments. Analytical studies show that large increments are better to predict in terms of the ROC, if the probability distribution of the process under study is Gaussian. For data following a symmetric exponential distribution, $\rho(x) \propto \exp(-\gamma|x|)$, there is no significant dependence of the prediction skill on the magnitude of the increment, while for data whose distribution has a power–law tail, $\rho(x) \propto x^{-\alpha}$, for $x > 0$ and $\alpha > 2$, larger increments are harder to predict than smaller increments. This is illustrated in Fig. 2. It is intuitively clear that for short term predictions, only the short range correlation structure is relevant. Hence results for short range correlated processes can be qualitatively transfered to processes of arbitrary correlation, as long as they exhibit the same joint distribution of event and precursor. We studied a class of long range correlated Gaussian data numerically and confirmed the improved predictability for larger events also in this situation [19].

However, in the analogous study for the prediction of threshold crossings in AR(1) processes with Gaussian, approximately exponential and approximately power-law distribution, we obtained qualitatively different results [18]. In contrast to the results for increments, threshold crossings were for all tested distributions the better predictable, the larger the specified threshold was.

6 Application to Experimental Data

6.1 The Influence of the Event Rate on the Brier Score

To illustrate the concepts introduced in the previous sections, we will perform predictions of wind speeds, with the aim of forecasting the occurrence of particularly strong wind gusts, i.e., of sudden increases of the wind speed. We used recordings of horizontal wind speed, sampled with 8 Hz at 30 m above ground at the Lammefjord measurement site [22]. We fixed a gust strength

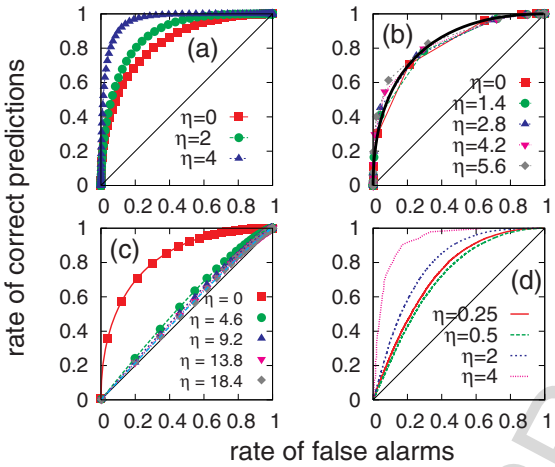


Fig. 2: Typical ROC curves for the prediction of large increments via strategy II, for different event magnitudes $\eta = x_{i+1} - x_i/\sigma$, with σ being the standard deviation of the process under study. *Panel a*: predictions within normal i. i. d. random variables, *panel b*: predictions within symmetric exponential i. i. d. random variables, *panel c*: predictions within power law i. i. d. random variables, with exponent $\alpha = 4$, *panel d*: predictions within a long range correlated Gaussian process. Each data set contained about 10^7 data points.

g and define $\chi_i = 1$ if $x_{i+\Delta} - x_i > g$, where $x_i, i = 1 \ldots N$ are wind speed measurements. The time horizon Δ between the prediction and the event corresponds to the time through which the increment is defined. Since the data is strongly correlated we choose a time horizon of $\Delta = 32$ in order to observe sufficiently many large increments. The time horizon $\Delta = 32$ corresponds to an increase of the wind speeds 4 s ahead in time. Various gust strengths g were considered, and we also compared different values for the conditioning τ, that is, conditional probabilities $p(\chi_i = 1|x_i, x_{i-1}, \ldots, x_{i-\tau+1})$ for various values of τ were considered as basis for our predictions.

The following results are based on 10^6 prediction trials at equidistant times, where for every sequence $\mathbf{x}_i = (x_i, x_{i-1}, \ldots, x_{i-\tau+1})$ of τ successive observations, the prediction was compared to the known value of χ_i. All information needed for the prediction is extracted from the same time series.

The conditional probabilities $p(\chi_i|\mathbf{x}_i)$ can be either estimated by binning and counting or by a kernel estimator. The latter is more time consuming but less memory consuming and slightly more accurate, if the dimension τ of the condition $\mathbf{x}_i$ is large. We use a box-kernel of width ϵ, hence

$$\hat{p}_i = \frac{\sum_{j=1}^{N} \chi_j \theta(\epsilon - |\mathbf{x}_i - \mathbf{x}_j|)\theta(100 - |i-j|)}{\sum_{j=1}^{N} \theta(\epsilon - |\mathbf{x}_i - \mathbf{x}_j|)\theta(100 - |i-j|)} \,. \tag{23}$$

The right hand side denotes the relative number of events following those vectors $\mathbf{x}_j$ which are in the ϵ-neighborhood of $\mathbf{x}_i$. The second θ-function acting on the time indices simply excludes all those time series elements from the estimation which are too close in time to the actual observation and hence might be correlated with it. This guarantees that we perform true out-of-sample predictions, since thus all sample points with time indices $j: \ |j-i| < 100$ were ignored. We expect those sample points to be highly correlated with x_i, allowing to form good predictions. But since these sample points are not available in a real forecast situation, including them here would lead to overoptimistic performance assessments. For every trial, we numerically estimate the conditional probability $\hat{p}(\mathbf{x}_i) = p(\chi_i = 1|\mathbf{x}_i)$ with an adaptive kernel size, thereby ensuring a local sample size of at least 20 points. For these predicted prob-

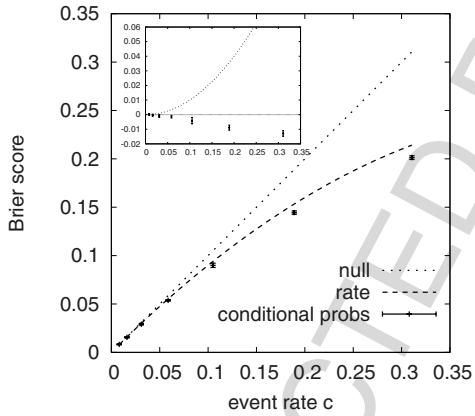
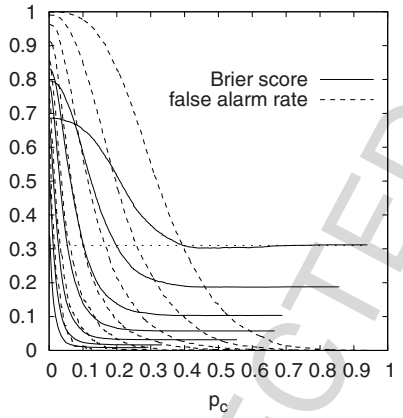


Fig. 3: The Brier scores for wind gust prediction using as predictor $\hat{p} = p(\chi_i|\mathbf{x}_i)$, compared to the constant predictions $\hat{p} = c$ and $\hat{p} = 0$. The larger the magnitude g of the events, the smaller their rates c. The confidence intervals of the Brier scores are derived from ten disjoint samples of 10^5 prediction trials each. The inset shows the scores again but reduced by the score $c(1-c)$ of the predictor $\hat{p} = c$.

abilities we compute the average Brier score as performance indicator. This indicator can be compared to the score of the constant prediction $\hat{p} = c$, where c is the rate of events, $c = \sum_i \chi_i / N$. This rate depends on the gust magnitude g and is very small if g is large. This simple predictor has an average Brier score of $c(1-c)$. For small event rates c, it is also reasonable to use the trivial constant prediction $\hat{p} = 0$, that is, to predict the event never to happen. This even simpler predictor has an average Brier score of c (only missed events get counted). In Fig. 3, the Brier scores of these three prediction schemes are shown as functions of the event rate. The skills of the two constant predictors $\hat{p} = 0$ and $\hat{p} = c$ are fully determined by the event rate, that is, they are independent of any information about the future, as stored in the preceding

time series elements, and are therefore independent of any temporal correlations in the data sequence. The dependence of these two predictors on the rate demonstrates that the absolute value of the Brier score (and hence its interpretation) depends on the event rate, which renders a comparison between prediction schemes for different event rates difficult. The remarkable finding is that the Brier score of the predictor $\hat{p} = p(\chi_i = 1|\mathbf{x}_i)$, which at least in theory should not be inferior to the constant predictor $\hat{p} = c$ is only insignificantly better than the latter when the event rate is very small. In this situation, the score of the "null" predictor $\hat{p} = 0$ is also almost as good. This demonstrates that the Brier score is not very useful when the event rate is small. The inset of Fig. 3 shows the same results but presented in a different way: We plot the difference between the scores of the predictors and the score of the constant rate predictor $\hat{p} = c$.



AU: Please provide text citation for figure 4.

Fig. 4: The Brier score and the rate of false alarms for gust prediction of wind speeds. The predicted probability $\hat{p}$ is converted into warnings whenever it exceeds the threshold value p_c. For large values of p_c, the scores for different gust strengths saturate at the respective event rates c. For small p_c and small event rates c, the Brier score of these "filtered" probabilities is dominated by the false alarms. The horizontal line at $y = 0.31$ just shows that for the event rate $c \approx 0.31$ the Brier score has a nontrivial minimum around $p_c \approx 0.45$.

In order to convert probabilistic predictions into warnings, one would introduce a threshold p_c and predict a "filtered" probability $\tilde{p} = \theta(\hat{p} - p_c)$ so that whenever the predicted probability is below p_c, a filtered probability of 0 is issued, and a probability of unity (or a warning) otherwise. Inspecting Eqs. (8) and (9) for these special probabilities $\tilde{p}$ shows that the Brier score is the number of false alarms plus the number of unpredicted events, normalized by the number of prediction trials. If the warnings were given randomly with constant rate, then the Brier score would be a linear interpolation between

the event rate (warning rate being zero: no hits) and one minus the event rate (warning rate unity: maximal number of false alarms). Numerical results of the score of $\tilde{p}$ as a function of p_c obtained for a set of g-values ranging from 0.5 to 3.5 are shown in Fig. 5. As suspected earlier, for small event rates, the Brier score is best if no alarms are ever given ($p_c = 1$), since then no false alarms are made, at the cost of missing the large but rather few events. For low rates and low thresholds $p_c \approx 0$, the Brier score is almost identical to the false alarm rate, which is also shown. Only for the largest event rate $c \approx 0.31$ (gust strength $g = 0.5$), we find a nontrivial minimum of the Brier score with this scheme.

Without using any additional scoring scheme, the findings discussed so far might suggest that large wind gusts are not predictable. The ROC plot obtained from the prediction scheme $\hat{p} = p(\chi_i = 1|\mathbf{x}_i)$ for the same range of threshold gust strengths is shown in Fig. 5. This plot clearly demonstrates predictive skill of this algorithm. We note two interesting observations: First, the larger the magnitude of the event (the larger g), the better is the ROC curve. Second, when we increase the conditioning from $\tau = 1$ to $\tau = 8$, the predictive skill in the range of small false alarm rates improves. Whereas the first finding is understood fairly well theoretically (see Sect. 5.3), the second one suggests nontrivial temporal correlations in the wind speed data.

Although in this study we focused on the Brier score, we argue that other scores such as the ignorance would suffer from similar difficulties, in particular if predicted probabilities are to be converted into actual warnings. The reason for this is that the interpretation of forecast skill in terms of scores is difficult if they depend in a nontrivial way on the event rate of the process. The need to convert probabilistic predictions into alarms for practical purposes turns the prediction of rate extreme events into a classification problem. Classification problems though can be conveniently evaluated by the ROC statistics. For the remainder of the paper, we will therefore restrict ourselves to ROC analysis of prediction schemes.

6.2 Prediction of Increments in a Free Jet Flow

In a last step, we study the influence of the event magnitude in experimental data. This is done by applying the prediction scheme called strategy II to experimental data with the aim of predicting increments. The potential complications are that stationarity is violated to a smaller or larger extent, that the correlation structure is more complicated, and that the distribution is only approximately one of the above studied classes.

We start with data from a well controlled laboratory experiment, namely from free jet turbulence [23]. Using hot wire anemometry, the velocity of the air in front of a nozzle is measured at a sampling rate of 8 Hz and at a position where the flow can in good approximation be considered as being isotropically turbulent. Taking increments a_i of such a sequence v_i over short time intervals, $a_i = v_{i+k} - v_i$, for k small, yields approximately a symmetric exponential

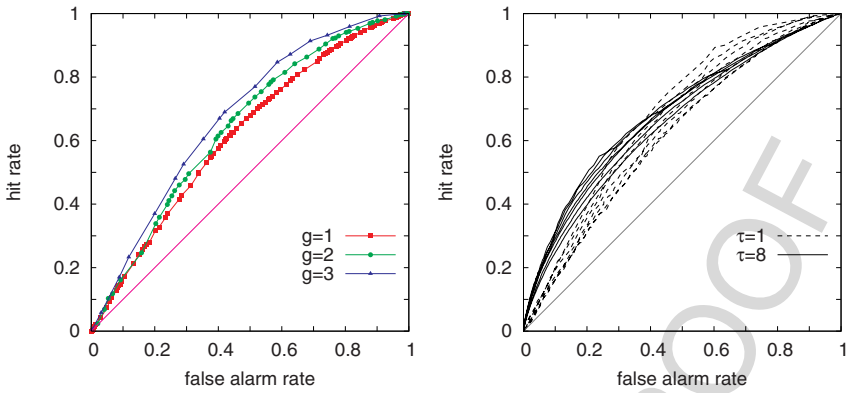


Fig. 5: The ROC statistics for the prediction of large increments of wind speeds (wind gusts) for different gust strength using the estimated values $\hat{p} = p(\chi_i|\mathbf{x}_i)$. On the right the *lower bundle* of curves refers to $\tau = 1$, i.e., $\mathbf{x}_i$ has only one component (the most recent observation), whereas the *upper bundle* represents predictions where the condition $\mathbf{x}_i$ is the vector of the $\tau = 8$ last observations. In this example, the predictive skill improves when the condition is extended. Every bundle of curves represents gusts strengths ranging from $g = 0.5$ to $g = 3.5$, from *bottom to top*. On the left the influence of the event size is studied for $\tau = 1$. One can notice the better predictability of larger and thus rarer events. In both ROCs the overall performance is rather poor (all lines fall close to the diagonal), but the larger the gust strength, the better the predictability. Predictability can be further improved by redefining wind gusts to be large increments occurring in future time intervals rather than at a specific instance in time.

distribution for a_i, whereas for long time intervals, i.e., large k, the distribution of the increments is approximately Gaussian[5] [24, 25], see Fig. 6. The time horizon Δ between the prediction and the event corresponds to the time through which the increment is defined. Since the data is strongly correlated we choose a time horizon $\Delta = 285$, which corresponds to an increment 35.625 s ahead in time, in order to observe sufficiently many large increments. Note that the increment size is defined again in units of the standard deviation σ of the process under study, i.e., $\eta = (a_{i+\Delta} - a_i)/\sigma$.

In Fig. 7, we show the ROC statistics for the prediction of large increments $a_{i+\Delta} - a_i$ in the increment time series $\{a_i\}$. The ROC statistics were generated according to the algorithm described in Sect. 5.1.

As predicted by the theoretical considerations for symmetric exponentially distributed i. i. d. random numbers, the quality of the prediction does not significantly depend on the magnitude of the increment, if the distribution of

[5] For longitudinal velocity increments, one wing of the distributions is higher than the other. This effect can be understood via Kolmogorovs four-fifths law, which demands a non-zero skewness of the velocity increment [26].

the data is approximately exponential which corresponds to small values of k. In contrast, in the case of large k the probability distribution follows approximately a Gaussian distribution and larger increments are better predictable. Both predictions were made by determining precursors in the first part of the data set ($7 \cdot 10^6$ data points) and then predicting increments in the second part of $\{a_i\}$ (also $7 \cdot 10^6$ data points).

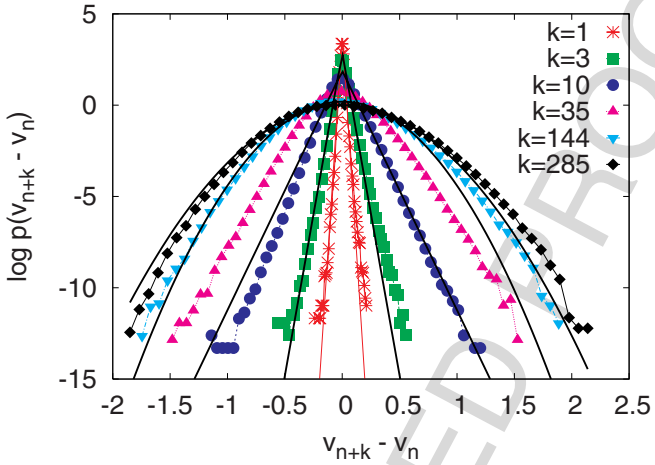


Fig. 6: Histograms for velocity increments $a_i = v_{i+k} - v_i$ in a free jet flow. For small increments, e.g., $k = 3$, the pdfs of the increment time series $\{a_i\}$ follow approximately an exponential distribution, for larger k the pdfs of the increments are approximately Gaussian distributed.

6.3 Prediction of Increments in Wind Speeds

As a second example, we study wind speeds measured from a measurement site about 66 m above ground at a sampling rate of 1 Hz. These data reflect the full complications of field measurements, including non-stationarity and inaccuracies, but also represent a much more complicated turbulent state, namely boundary layer turbulence which is strongly affected by the interaction of the air flow with the earth's surface. Hence the deviations from the asymptotic distributions are larger than in the laboratory experiment, expecially in the tails of the distributions.

We predict large increments in the acceleration of the wind, so called turbulent gusts which are of relevance for controlling wind turbines or scheduling aircraft take-off and landing. The time horizon Δ between the prediction and the event corresponds to the time through which the increment is defined. Since the data is strongly correlated we choose a time horizon $\Delta = 35$, which corresponds to an increment 35 s ahead in time, in order to observe sufficiently

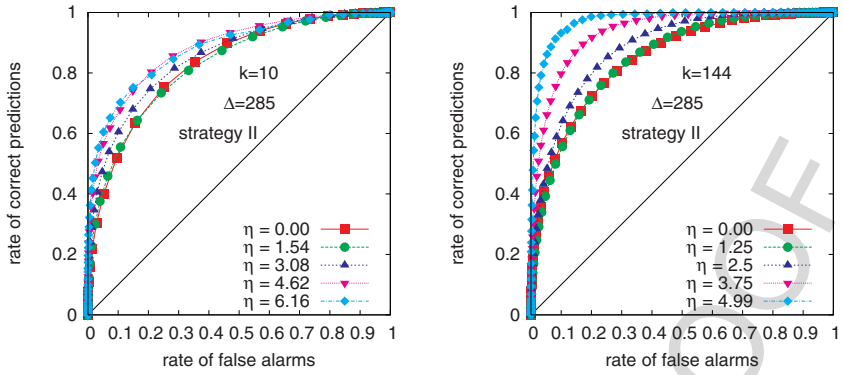


Fig. 7: Two ROC curve for the prediction of increments $\Delta = 285$ time steps ahead of the increment time series a_i of isotropic turbulence: For increments defined by short time intervals (*left panel*), the predictability is almost independent of the event magnitude, whereas for increments defined by large time intervals (*right panel*), larger events are better predictable.

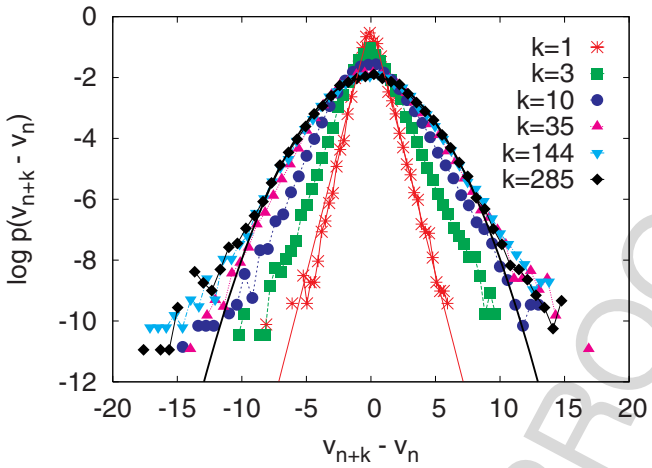
many large increments. Note, that the increment size is defined again in units of the standard deviation σ of the process under study, i.e., $\eta = (a_{i+\Delta} - a_i)/\sigma$. Again, the predictions were made via identifying the precursors in the first part of the data set and then predicting in the second part. The ROC curves (Fig. 9) show again a better predictability of larger events if the data set $\{a_i\}$ is asymptotically Gaussian distributed and a much weaker dependence on the event size in the asymptotically exponential case.

7 Conclusion

We presented an overview over some aspects of the predictions of rare events χ_i in time series and showed how to make use of the properties of the process to construct predictors. As a general result, the prediction scheme should exploit the conditional probability $p(\chi_i = 1|\mathbf{x}_i)$ and not $p(\mathbf{x}_i|\chi_i = 1)$. In practice prediction from $p(\chi_i = 1|\mathbf{x}_i)$ can be either drawn by using the probability itself or by using the values of $\mathbf{x}_i$ for which $p(\chi_i = 1|\mathbf{x}_i)$ is maximal as precursory structures.

Furthermore we discussed the role of several scoring schemes with respect to their performance for the prediction of extreme and thus rare events. We find that the RMS and the Brier Score are not the optimal scoring schemes for the prediction of rare events, since they involve an averaging over the whole number of prediction trials. Hence, the influence of the correctly predicted rare events, is suppressed due to the influence of the large number of non-events.

The ignorance tries to avoid this effect by taking the logarithm of the forecast probabilities. Also the ROC statistics is in particular suitable for the



AU: Please provide text citation for figure 8.

Fig. 8: Histograms of velocity increments in wind speed. Again, we find that on short scales the pdfs of the increments follow approximately an exponential distribution, whereas the increments are approximately Gaussian distributed for larger k. However, the deviations from the asymptotic distributions are larger than in the laboratory experiment, expecially in the tails of the distributions.

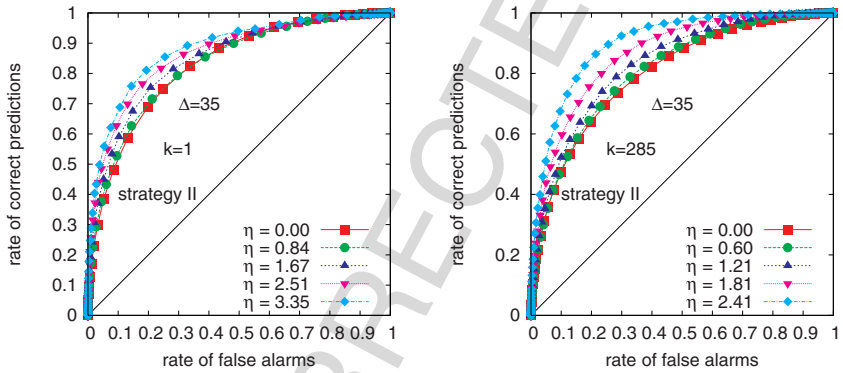


Fig. 9: ROC statistics for the prediction of large increments of wind speeds for different event magnitudes, predicting $\Delta = 35$ time steps ahead.

prediction of rare events, since is does not explicitly depend on the rate of the event, provided that there are sufficiently many events to create ROC statistics.

Another aspect, which we addressed is the dependence on the magnitude of the events under study. One can show that for stochastic processes, this dependence is mainly given by the underlying probability distribution of the process under study. This leads to particularly interesting results for the prediction

of increments. If the probability distribution of the process under study is Gaussian, larger increments are the better predictable, the larger they are. If the probability distribution is exponential, there is no significant dependence on the magnitude of the increment and for power–law distributed processes we find that larger increments are the harder to predict, the larger they are.

The corresponding results for the prediction of threshold crossings are qualitatively different. We find for processes with Gaussian, exponential and power–law probability distribution that larger threshold crossings are the better predictable, the larger they are.

By applying the previously developed concepts for the prediction of increments in the acceleration of a free jet flow and in wind speed measurements, we showed that the dependence on the increment magnitude, which was previously described and theoretically understood for stochastic processes, can be also found in experimental data, which exhibits non-stationarity and long-range correlations.

Acknowledgements. We are grateful to Joachim Peinke and his group from the University of Oldenburg for providing both their excellent free jet and wind speed data sets.

This contribution has benefited from measurements downloaded from the internet database: *Database of Wind Characteristics* located at DTU, Denmark, `http://www.winddata.com`. A wind field time series from the following site has been studied: Vinderby (Riso National Laboratories, Denmark).

References

1. *1987 Great Storm: Terrible blow, not a knockout*, Daily Telegraph, 13 October 2007.
2. *Great Storm of 1987*, English wikipedia, `http://en.wikipedia.org/wiki/Great_Storm_of_1987`
3. *Starkniederschläge in Sachsen im August 2002*, Publication of the German Weather Service DWD October 2002
4. User Guide to ECMWF Forecast Products, `http://www.ecmwf.com/products/forecasts/guide/index.html`
5. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, *Time Series Analysis*, Prentice-Hall Inc. (1994)
6. P. J. Brockwell, R. A. Davis, *Time Series: Theory and Methods*, Springer (1998)
7. H. Kantz, M. Ragwitz, Phase space reconstruction and nonlinear predictions for stationary and nonstationary Markovian processes, *Int. J. Bifurcation Chaos,* **14** 1935–1945 (2004)
8. D. J. Jackson, Hypothesis testing and earthquake prediction, *Proc. Natl. Acad. Sci. USA,* **93**, 3772–3775 (1996)
9. C. E. Elger, K. Lehnertz, Seizure prediction by non-linear time series analysis of brain electrical activity, *Eur. J. Neurosci.,* **10**, 786–789 (1998)
10. I. J. Good, Rational decisions, *J. Royal Statist. Soc.,* **XIV(1)** B, 107–114 (1952)

11. Jr. J. L. Kelly, A new interpretation of information rate, *Bell System Techn. J.,* **35**, 917–926 (1956)
12. T. A. Brown, *Probabilistic Forecasts and Reproducing Scoring Systems*, RAND Corporation **RM − −6299 − −ARPA** (1970)
13. L. J. Savage, Elicitation of personal probabilities and expectation, *J. Amer. Statist. Ass.,* **66**, 783–801 (1971)
14. J. Bröcker, L. A. Smith, Scoring probabilistic forecasts: the importance of being proper, *Weather and Forecasting,* **22**, 382–388 (2007)
15. J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press (1975)
16. A. M. Mood, F. A. Graybill, D. C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill (1974)
17. S. Hallerberg, E. G. Altmann, D. Holstein, H. Kantz, Precursors of extreme increments, *Phys. Rev. E,* **75**, 016706 (2007)
18. S. Hallerberg, H. Kantz, How does the quality of a prediction depend on the magnitude of the events under study?, *Nonlin. Proc. Geophys.,* subm.
19. S. Hallerberg, H. Kantz, Influence of the event magnitude on the predictability of extreme events, *Phys. Rev. E,* **77**, 011108 (2008)
20. A. B. Shapoval, M. G. Shrirman, How size of target avalanches influence prediction efficiency, *Int. J. Mod. Phys. C,* **17** 1777–1790 (2006)
21. D. Lamper, S. D. Howison, N. F. Johnson, Predictability of large future changes in a competitive evolving population, *Phys. Rev. Lett.,* **88**, 017902 (2002)
22. The wind-speed data were recorded at the Riso National Laboratory, Technical University of Denmark, `http://www.risoe.dk/vea`, see also `http://winddata.com`
23. C. Renner, J. Peinke, R. Friedrich, Experimental indications for Markov properties of small-scale turbulence, *J. Fluid. Mech.,* **433** 383–409 (2001)
24. C. W. Van Atta, J. Park, Statistical self-similarity and initial subrange turbulence, In: *Statistical Models and Turbulence, Lect. Notes in Phys.* **12**, pp 402–426, eds. M. Rosenblatt and C. W. Van Atta, Springer Berlin (1972)
25. Y. Gagne, E. Hopfinger, U. Frisch, A new universal scaling for fully developed turbulence: the distribution of velocity increments. In: *New Trends in Nonlinear Dynamics and Pattern-Forming Phenomena*, NATO ASI **237**, pp 315–319, eds. P. Coullet and P. Huerre, Plenum Press, New York (1990)
26. U. Frisch, *Turbulence*, Cambridge University Press, Cambridge (1995)

AU: Please update the volume and page numbers for Ref. [18].

Analysis of Geophysical Time Series Using Discrete Wavelet Transforms: An Overview

Donald B. Percival

Applied Physics Laboratory, University of Washington, Box 355640, Seattle, WA, 98195–5640, USA, dbp@apl.washington.edu

Abstract. Discrete wavelet transforms (DWTs) are mathematical tools that are useful for analyzing geophysical time series. The basic idea is to transform a time series into coefficients describing how the series varies over particular scales. One version of the DWT is the maximal overlap DWT (MODWT). The MODWT leads to two basic decompositions. The first is a scale-based analysis of variance known as the wavelet variance, and the second is a multiresolution analysis that reexpresses a time series as the sum of several new series, each of which is associated with a particular scale. Both decompositions are illustrated through examples involving Arctic sea ice and an Antarctic ice core. A second version of the DWT is the orthonormal DWT (ODWT), which can be extracted from the MODWT by subsampling. The relative strengths and weaknesses of the MODWT, the ODWT and the continuous wavelet transform are discussed.

AU: Please provide Keywords.

1 Introduction

The wide-spread use of wavelets to analyze data in the geosciences can be traced back to work by Morlet and coworkers [1, 2] in the early 1980s. Their efforts were motivated by signal analysis in oil and gas exploration and resulted in the continuous wavelet transform (CWT). Work in the late 1980s by Daubechies, Mallat and others [3, 4, 5, 6] led to various discrete wavelet transforms (DWTs), which are the focus of this article. While CWTs and DWTs are closely related, DWTs are more amenable to certain types of statistical analysis, making them the transform of choice for tackling certain – but not all – problems of interest in geophysical data analysis. The intent of this article is to give an overview of how DWTs can be used in the analysis of geophysical time series, i.e., a sequence of observations recorded over time (usually at regularly spaced intervals such as once per second).

The remainder of this article is structured as follows. In Sect. 2 we review the important notion of scale and the basic ideas behind the maximal overlap DWT (MODWT). The MODWT leads to two basic decompositions. The first (the subject of Sect. 3) is a scale-based analysis of variance known as the

wavelet variance (or wavelet spectrum). The second (Sect. 4) is an additive decomposition known as a multiresolution analysis, in which a time series is reexpressed as the sum of several new series, each associated with a particular physical scale. In Sect. 5 we discuss another form of the DWT known as the orthonormal DWT (ODWT) that can be extracted from the MODWT and that has certain strengths and weaknesses in comparison to the MODWT. Our overview concentrates on the so-called Haar wavelet, but we note the existence of other wavelets in Sect. 6 and discuss why they might be preferred over the Haar wavelet for certain types of analyses. Finally we make some concluding comments in Sect. 7, including a comparison of the strengths and weaknesses of DWTs and CWTs.

2 Maximal Overlap Discrete Wavelet Transform

Let X_n, $n = 0, 1, \ldots, N-1$, represent the nth value of a time series that has N values in all. We assume that, for all n, the time at which X_n was observed can be expressed as $t_0 + n\,\Delta$, where t_0 is the time associated with X_0, and Δ is the sampling interval between any two adjacently recorded values X_n and X_{n+1}. Given $\tau_j = 2^{j-1}$ for some positive integer j, consider

$$A_{j,n} = \frac{1}{\tau_j} \sum_{l=0}^{\tau_j - 1} X_{n-l}\,, \tag{1}$$

which is the average of τ_j adjacent values of the series starting with $X_{n-\tau_j+1}$ and ending at X_n. We refer to the above as a scale τ_j average. The variable τ_j is sometimes called a dyadic scale since its values are restricted to be powers of two. It is a dimensionless scale that is associated with a physical scale of $\tau_j\,\Delta$. Since $\tau_1 = 1$ and hence $A_{1,n} = X_n$, we can think of the original series as being unit scale 'averages'.

The definition for $A_{j,n}$ makes sense as long as $\tau_j - 1 \le n \le N-1$; however, $A_{j,n}$ is ill-defined when $\tau_j \ge 2$ and $0 \le n \le \tau_j - 2$ because (1) would then involve X_{-1} and possibly other values of the time series we don't have access to. To force $A_{j,n}$ to be well defined for the full range $0 \le n \le N-1$, we assume that the time series is periodic with a period of N; i.e., $X_n = X_{n+N}$ for all integers n. With this definition, $X_{-1} = X_{N-1}$, $X_{-2} = X_{N-2}$ and so forth. This assumption introduces some 'boundary' averages such as $A_{2,0} = (X_0 + X_{N-1})/2$, which combine nonadjacent values from the original series when $N > 2$. For $\tau_2 = 2$, the only possible boundary average is $A_{2,0}$, while the other $N-1$ averages $A_{2,1}, \ldots, A_{2,N-1}$ involve adjacent values from the time series.

If we let $a_{j,l} = 1/\tau_j$ for $0 \le l \le \tau_j - 1$, we can reexpress (1) in filtering notation as

$$A_{j,n} = \sum_{l=0}^{\tau_j - 1} a_{j,l} X_{n-l}\,, \quad n = 0, 1, \ldots, N-1\,.$$

The left-hand portion of Fig. 1 shows the filters $a_{j,l}$ for the dyadic scales indexed by $j = 1, 2, 3$ and 4 (we define $a_{j,l}$ to be zero when $l < 0$ or $l \geq \tau_j$).

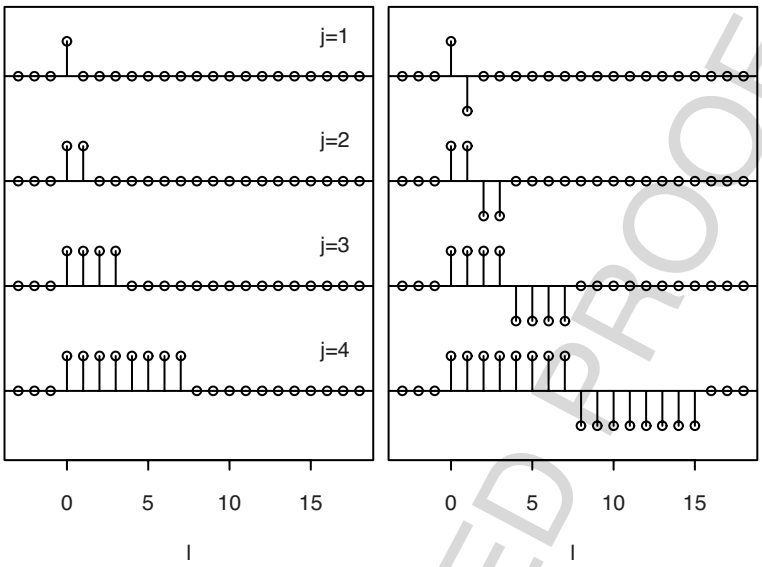


Fig. 1: Averaging filters $a_{j,l}$ (*left-hand panel*) for dyadic scales $\tau_j = 2^{j-1}$, $j = 1, 2, 3$ and 4, and related differencing filters $d_{j,l}$ (*right*). The averaging filters are proportional to Haar scaling filters, and the differencing filters, to Haar wavelet filters.

While averages of time series over various scales are of interest in their own right, what is often of more interest is how these averages change over time. For example, a key question about various indicators of climate is whether their average values over certain time scales have changed significantly with time. The wavelet transform is a mechanism that allows us to quantify how averages of a time series over particular scales change from one interval of time to the next. These changes are quantified in wavelet coefficients, which form the bulk of any DWT.

Wavelet coefficients in a DWT are organized into sets. There is one set for each dyadic scale τ_j, and each coefficient in this set is proportional to the difference between two adjacent nonoverlapping averages. Mathematically, these differences are given by

$$D_{j,n} = A_{j,n} - A_{j,n-\tau_j} = \sum_{l=0}^{2\tau_j - 1} d_{j,l} X_{n-l} \,, \quad n = 0, 1, \ldots, N-1 \,, \qquad (2)$$

where

$$d_{j,l} = \begin{cases} 1/\tau_j, & l = 0, \ldots, \tau_j - 1 \; ; \\ -1/\tau_j, & l = \tau_j, \ldots, 2\tau_j - 1 \; ; \\ 0, & \text{otherwise.} \end{cases}$$

The right-hand portion of Fig. 1 shows the differencing filters $d_{j,l}$ associated with the averaging filters $a_{j,l}$. If $D_{j,n}$ is close to zero, then $A_{j,n-\tau_j}$ and $A_{j,n}$ are close to each other, indicating that there is not much change in these adjacent nonoverlapping averages of scale τ_j; on the other hand, if $D_{j,n}$ has a large magnitude, then the two scale τ_j averages differ considerably.

We can now define the Haar maximal overlap discrete wavelet transform (MODWT) of maximum level J_0, where J_0 is a positive integer that we are free to select. This transform consists of $J_0 + 1$ sets of N coefficients, for a total of $(J_0+1) \times N$ coefficients in all. There are J_0 sets of wavelet coefficients, and the remaining set consists of the so-called scaling coefficients. For $j = 1, \ldots, J_0$, the wavelet coefficients are given by $\widetilde{W}_{j,n} = D_{j,n}/2$, while the single set of scaling coefficients is given by $\widetilde{V}_{J_0,n} = A_{J_0+1,n}$, where $n = 0, 1, \ldots, N-1$ in both cases. Let $\mathbf{X}$ be an N dimensional column vector containing the time series X_n, and let $\widetilde{\mathbf{W}}_j$ be a similar vector containing the level j MODWT wavelet coefficients $\widetilde{W}_{j,n}$. We can then write

$$\widetilde{\mathbf{W}}_j = \widetilde{\mathcal{W}}_j \mathbf{X} \; , \tag{3}$$

where $\widetilde{\mathcal{W}}_j$ is an $N \times N$ matrix whose rows can be deduced by studying (2). For example, if $N = 7$ and $j = 2$ so that $\tau_2 = 2$, we find that

$$\widetilde{\mathcal{W}}_2 = \begin{bmatrix} 1/4 & 0 & 0 & 0 & -1/4 & -1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 0 & 0 & -1/4 & -1/4 \\ -1/4 & 1/4 & 1/4 & 0 & 0 & 0 & -1/4 \\ -1/4 & -1/4 & 1/4 & 1/4 & 0 & 0 & 0 \\ 0 & -1/4 & -1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & -1/4 & -1/4 & 1/4 & 1/4 & 0 \\ 0 & 0 & 0 & -1/4 & -1/4 & 1/4 & 1/4 \end{bmatrix} .$$

Note that any of the bottom six rows in $\widetilde{\mathcal{W}}_2$ can be obtained by circularly shifting the row above it to the right by one, a pattern that holds for all $\widetilde{\mathcal{W}}_j$. Note also that the first three rows yield boundary wavelet coefficients since they combine together values of the time series that are not contiguous in time (in general, there are $\min\{2\tau_j - 1, N\}$ boundary coefficients). In a similar manner, if $\widetilde{\mathbf{V}}_{J_0}$ is an N dimensional column vector containing the scaling coefficients $\widetilde{V}_{J_0,n}$, then we can write

$$\widetilde{\mathbf{V}}_{J_0} = \widetilde{\mathcal{V}}_{J_0} \mathbf{X} \; , \tag{4}$$

where $\widetilde{\mathcal{V}}_{J_0}$ is an $N \times N$ matrix whose rows are dictated by (1).

In practice the MODWT wavelet and scaling coefficients are not computed directly via (3) and (4), but rather via an efficient recursive procedure known as the pyramid algorithm (for pseudo-code describing this algorithm, see pp. 177–178 of [7]).

3 Analysis of Variance via the Wavelet Variance

The MODWT leads to two basic decompositions for a time series X_n. The first is an analysis of variance (ANOVA) that is based on a decomposition of the 'energy' in X_n (the second is discussed in Sect. 4). By definition the energy in a time series is just the sum of its squared values:

$$\sum_{n=0}^{N-1} X_n^2 = \mathbf{X}^T\mathbf{X} = \|\mathbf{X}\|^2 ,$$

where 'T' denotes the transpose operation, and $\|\mathbf{X}\|$ is the Euclidian norm of $\mathbf{X}$. This decomposition states that

$$\|\mathbf{X}\|^2 = \sum_{j=1}^{J_0} \|\widetilde{\mathbf{W}}_j\|^2 + \|\widetilde{\mathbf{V}}_{J_0}\|^2 , \tag{5}$$

so the energy in the series is preserved in its MODWT wavelet and scaling coefficients.

Let σ_X^2 be the sample variance for the time series:

$$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{n=0}^{N-1} \big(X_n - \overline{X}\big)^2 = \frac{1}{N}\sum_{n=0}^{N-1} X_n^2 - \overline{X}^2 , \text{ where } \overline{X} = \frac{1}{N}\sum_{n=0}^{N-1} X_n .$$

It follows from (5) that

$$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{j=1}^{J_0} \|\widetilde{\mathbf{W}}_j\|^2 + \frac{1}{N}\|\widetilde{\mathbf{V}}_{J_0}\|^2 - \overline{X}^2 . \tag{6}$$

In the above, we refer to $\|\widetilde{\mathbf{W}}_j\|^2/N = \hat{\nu}_j^2$ as the empirical wavelet variance. We can regard $\hat{\nu}_j^2$ as an appropriate definition for the sample variance of the level j wavelet coefficients. This assumes that the mean value of $\widetilde{\mathbf{W}}_j$ can be taken to be zero, which is reasonable for certain X_n because of the differencing operation inherent in the filters used in (2). On the other hand, $\frac{1}{N}\|\widetilde{\mathbf{V}}_{J_0}\|^2 - \overline{X}^2$ is the sample variance of the scaling coefficients because $\widetilde{\mathbf{V}}_{J_0}$ is a running average of $\mathbf{X}$ and hence has a sample mean of $\overline{X}$. Equation (6) thus gives us a scale-based ANOVA, in that we are breaking $\hat{\sigma}_X^2$ up into J_0+1 pieces, each of which can be interpreted in terms of sample variances of either differences in averages over the dyadic scales $\tau_1, \ldots, \tau_{J_0}$ or averages over a scale of $2\tau_{J_0} = \tau_{J_0+1}$. If $N = 2^J$ for some positive integer J and if we set $J_0 = J$, the contribution to $\hat{\sigma}_X^2$ due to the scaling coefficients drops out because $\widetilde{\mathbf{V}}_{J_0}$ becomes a vector whose elements are all equal to $\overline{X}$, and we then have

$$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{j=1}^{J_0} \|\widetilde{\mathbf{W}}_j\|^2 = \sum_{j=1}^{J_0} \hat{\nu}_j^2 . \tag{7}$$

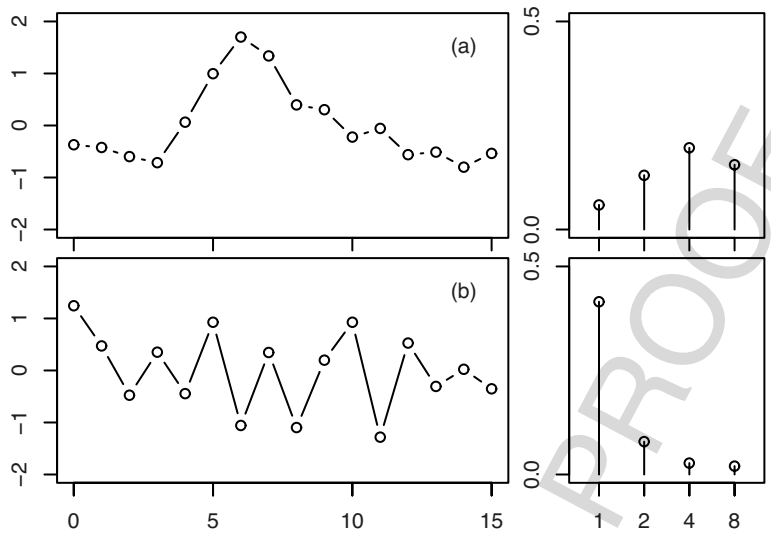


Fig. 2: Two time series (*left-hand plots*), each with $N = 16$ values. Both series have the same sample means and variances. The *right-hand plots* show the corresponding the Haar MODWT wavelet variances over the dyadic scales 1, 2, 4 and 8.

Even if these stipulations on N and J_0 are dropped, the above is still a good approximation as along as J_0 is large enough so that τ_{J_0+1} is close to N.

To see how the wavelet variance can help characterize time series, consider the two artificial series shown in the left-hand column of plots in Fig. 2. By construction both series have exactly the same sample mean and variance, but their appearances are quite different. Series (a) varies more slowly than series (b), which tends to fluctuate back and forth from one time point to the next. The right-hand plots show the corresponding empirical wavelet variances versus the dyadic scales $\tau_1, \ldots, \tau_4$. The wavelet variances for the two series have their largest values at different scales, namely, scale $\tau_3 = 4$ for (a) and $\tau_1 = 1$ for (b). Small-scale fluctuations are thus an important part of the overall variability of series (b), but less so for (a), where larger scale fluctuations are more prominent. Although the sample mean and variance are here incapable of distinguishing between the two series, the scale-based ANOVA given by the wavelet variance can in a manner that is intuitively reasonable.

The next two subsections consider 'real world' examples, both involving Arctic sea ice. Other examples of the use of the wavelet variance in geophysics include the study of the El Niño–Southern Oscillation [8], surface albedo and temperature in desert grasslands [9], soil variations [10], the relationship between rainfall and runoff [11], ocean surface waves [12], solar coronal

activity [13], North Atlantic sea levels [14], atmospheric turbulence [15] and the impact of large multi-purpose dams on water temperature variability [16].

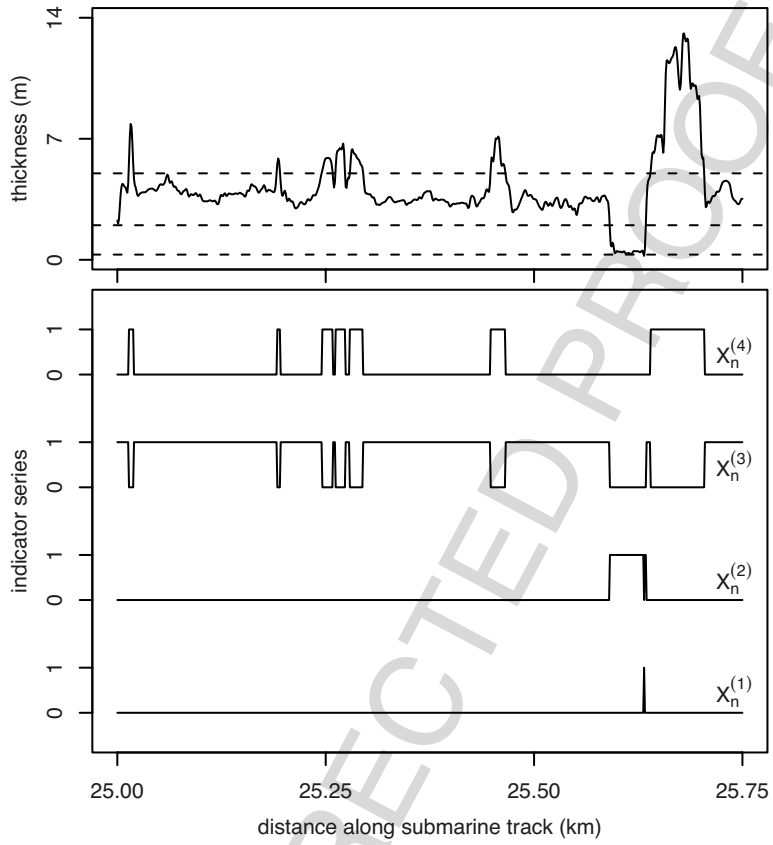


Fig. 3: Portion of a series of Arctic ice thickness measurements X_n versus distance along a submarine track (*upper plot*), along with four binary-valued series (*lower*) indicating the absence/presence (0/1) of four ice types: leads and new ice (defined as $X_n < 0.3$ m and denoted as $X_n^{(1)}$), first year ice ($0.3 \leq X_n < 2$, $X_n^{(2)}$), medium multiyear ice ($2 \leq X_n < 5$, $X_n^{(3)}$) and ridged ice ($X_n \geq 5$, $X_n^{(4)}$). The sampling interval is $\Delta = 0.001$ km. The horizontal dashed lines in the upper plot depict the defining boundaries for the ice types. These data were taken near the North Pole in April of 1991 and are archived at the National Snow and Ice Data Center (`http://nsidc.org/`).

3.1 Wavelet Variance Analysis of Arctic Ice Types

Naval submarines with upward-looking sonars have collected data on sea-ice thickness in the Arctic Ocean since 1958. Currently data from 34 cruises conducted by the U.S. Navy between 1975 and 2000 have been publically archived. These data provide a unique direct look at the climatology of Arctic ice thickness as a function of space and time. The upper plot of Fig. 3 shows a 0.75 km portion of one such series of ice thickness measurements X_n (in meters) taken near the North Pole in April of 1991 (the entire set of measurements extends over 50 km). We can regard X_n as a time series with $\Delta = 0.001$ km, where here 'time' is considered as a surrogate for distance along the submarine track under the ice (the observations were recorded at regular intervals of time, but the submarine was moving at a constant speed).

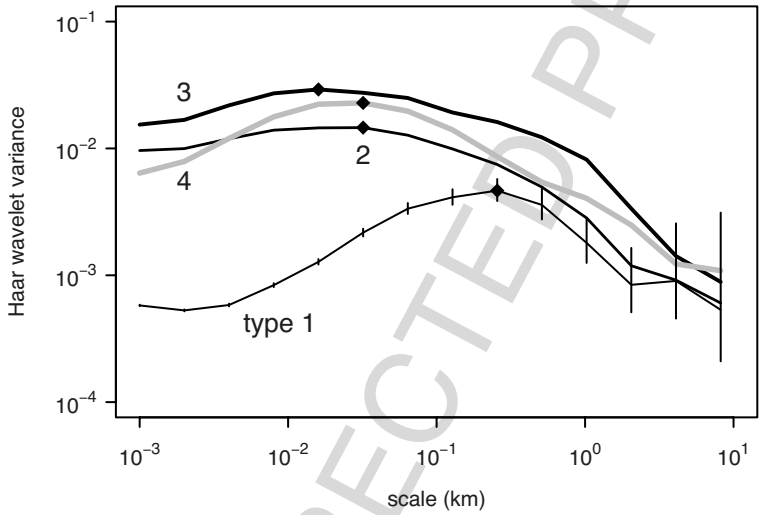


Fig. 4: Empirical Haar wavelet variances $\hat{\nu}_j^2$ versus physical scales $\tau_j \Delta$, $j = 1, 2, \ldots, 14$, for the four binary-valued ice type series $X_n^{(i)}$ shown in Fig. 3. The vertical lines emanating from each $\hat{\nu}_j^2$ for $X_n^{(1)}$ represent 95% confidence intervals for a hypothesized theoretical wavelet variance. The largest wavelet variances for each ice type are indicated by a solid diamond – their locations define wavelet-based characteristic scales.

Ice thickness can be classified into four types, which are driven by different physical processes [17, 18]. The first type consists of leads and new ice and has a thickness below 0.3 m; the second is first year ice and ranges from 0.3 to 2 m; the third is medium multiyear ice, from 2 to 5 m; and the fourth is ridged ice, anything above 5 m. The divisions between the four types are marked on

Fig. 3 by horizontal dashed lines. Let $X_n^{(i)}$ be a binary-valued series indicating the absence or presence (using 0 or 1) of ice type i at measurement X_n. These four indicator series are plotted in the bottom of Fig. 3.

Figure 4 shows empirical Haar wavelet variances for the four indicator series $X_n^{(i)}$ plotted versus $\tau_j \Delta$ for j ranging from 1 to 14 (i.e., physical scales from 0.001 up to 8.192 km). If we regard $\hat{\nu}_j^2$ as an estimate of a hypothesized theoretical wavelet variance, we can determine how far our estimates are likely to be off from the true wavelet variances (for details, see Chap. 8 of [7]). The vertical lines in Fig. 4 indicate 95% confidence intervals (CIs) for the true wavelet variances for ice type 1 (the three other ices types would have CIs with similar widths). Note that the widths of the CIs increase as τ_j increases.

All four wavelet variance curves in Fig. 4 have a single broad peak. The largest $\hat{\nu}_j^2$ for each ice type is marked with a solid diamond. While the scale at which the largest value occurs is similar for types 2, 3 and 4 (either 16 or 32 m), the one for type 1 is an order of magnitude larger (256 m). We can consider the location of these peak values as defining a characteristic scale for each ice type. A question of geophysical interest is how stable these characteristic scales are both spatially and temporally. This question can be addressed by using $\hat{\nu}_j^2$ to determine these scales from data taken at other locations and times across the Arctic basin. For this application, the wavelet variance thus extracts a summary statistic that picks out the largest scale-based contributor to the sample variance of an ice-type indicator series, and this statistic can be studied across space and time to deduce possible changes in the climatology of Arctic ice thickness.

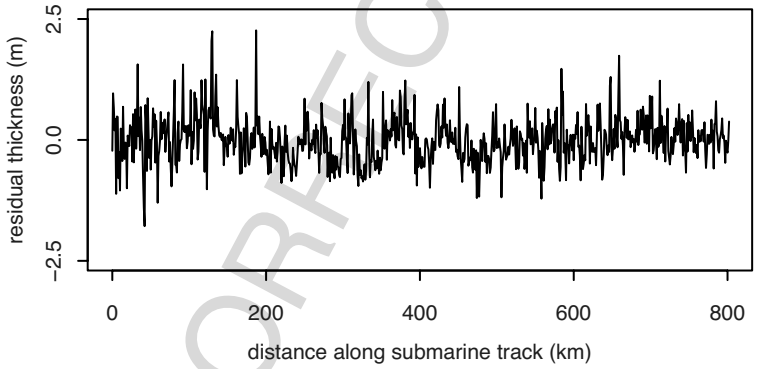


Fig. 5: Arctic ice thickness residuals X_n versus distance along a submarine track. The residuals are the deviations from a least squares fit of a line to a series of 1 km average thicknesses (the sampling interval Δ is also 1 km). There are $N = 803$ thickness measurements, and these were collected from a SCientific ICe EXpedition (SCICEX) cruise within the Arctic Ocean in September of 1997 and are archived at the National Snow and Ice Data Center (`http://nsidc.org/`).

3.2 Wavelet Variance Analysis of Averaged Ice Thickness

As a second example, let us consider another series of ice thickness measurements, but now consisting of one kilometer averages that have been detrended by subtracting off a line fit via least squares (the sampling interval is $\Delta = 1$ km). The residuals from this fit are plotted in Fig. 5.

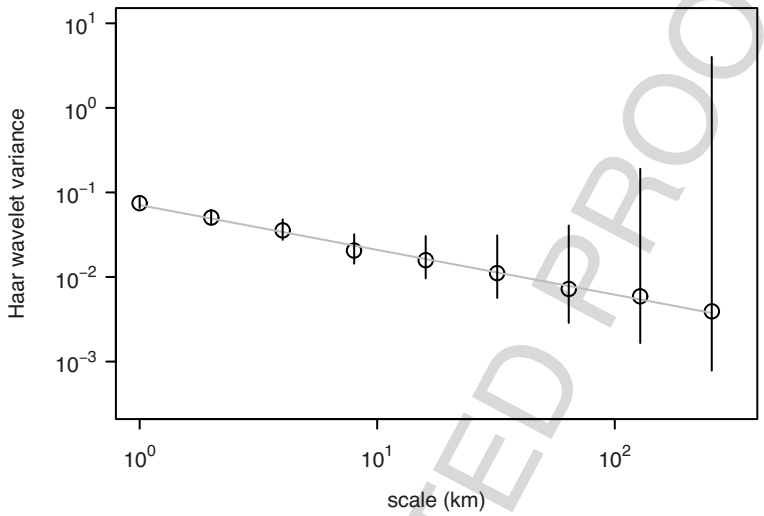


Fig. 6: Empirical Haar wavelet variances $\hat{\nu}_j^2$ versus physical scales $\tau_j \Delta$, $j = 1, 2, \ldots, 9$, for the residual thickness series shown in Fig. 5. The vertical lines emanating from each $\hat{\nu}_j^2$ represent 95% confidence intervals for a hypothesized theoretical wavelet variance. The line through the variances is a least squares fit of $\log_{10}(\hat{\nu}_j^2)$ versus $\log_{10}(\tau_j \Delta)$ and has a slope of -0.53.

The empirical wavelet variances $\hat{\nu}_j^2$ for $j = 1, \ldots, 9$ for the residual thicknesses are shown in Fig. 6, along with 95% confidence intervals for a hypothesized theoretical wavelet variance (the vertical lines). A linear least squares fit of $\log_{10}(\hat{\nu}_j^2)$ versus $\log_{10}(\tau_j \Delta)$ is also shown (the line with a slope of -0.53). With 1 km averaging, the largest wavelet variance occurs at the smallest scale, but what is of more interest is the rate of decay of $\hat{\nu}_j^2$ with increasing scale. This decay is very close to linear on a log/log scale. This form of decay is indicative of a stationary process whose spectral density function (SDF) $S(f)$ is approximately proportional to a power law $|f|^\alpha$, where f is a Fourier frequency. For such a process, it can be argued that the theoretical wavelet variance should be approximately proportional to $\tau_j^{-\alpha-1}$, which implies that a log/log plot of $\hat{\nu}_j^2$ versus $\tau_j \Delta$ should be approximately linear,

with a slope given by $-\alpha - 1$. The observed slope of -0.53 thus maps into a power-law exponent of $\alpha = -0.47$. A process whose SDF is proportional to $|f|^{-0.47}$ exhibits long-range dependence, which is characterized by an autocovariance function that decays at a slower rate than standard models such as autoregressive and/or moving average processes. This slower rate of decay has implications in assessing the sampling variability in various statistics derived from ice thickness measurements (for details, see [19, 20]).

4 Multiresolution Analysis

We now turn to the second basic decomposition afforded by the MODWT, which is an additive decomposition known in the wavelet literature as a multiresolution analysis (MRA). This decomposition says that we can reexpress $\mathbf{X}$ as the sum of $J_0 + 1$ new time series, each of which has a scale-based interpretation. In particular, define

$$\widetilde{\mathbf{D}}_j = \widetilde{\mathcal{W}}_j^T \widetilde{\mathbf{W}}_j \text{ and } \widetilde{\mathbf{S}}_{J_0} = \widetilde{\mathcal{V}}_{J_0}^T \widetilde{\mathbf{V}}_{J_0}, \tag{8}$$

where $\widetilde{\mathbf{D}}_j$ and $\widetilde{\mathbf{S}}_{J_0}$ are N dimensional vectors known as, respectively, the jth level detail and the J_0th level smooth. We can now write

$$\mathbf{X} = \sum_{j=1}^{J_0} \widetilde{\mathbf{D}}_j + \widetilde{\mathbf{S}}_{J_0}, \tag{9}$$

where $\widetilde{\mathbf{D}}_j$ is a time series reflecting variations in averages over a scale of τ_j in $\mathbf{X}$, whereas $\widetilde{\mathbf{S}}_{J_0}$ is a series reflecting averages over a scale of τ_{J_0+1}. Note that we can recover our original time series $\mathbf{X}$ from its MODWT, which tells us that no information about the series has been lost in transforming it and that (8) constitutes the pieces of an inverse MODWT. Thus, if we know how a time series varies at the dyadic scales $\tau_1, \ldots, \tau_{J_0}$ and if we know its averages over a scale of τ_{J_0+1}, then we can reconstruct the series perfectly. If we compare (9) to a level $J_0 + 1$ decomposition, namely,

$$\mathbf{X} = \sum_{j=1}^{J_0+1} \widetilde{\mathbf{D}}_j + \widetilde{\mathbf{S}}_{J_0+1},$$

we can deduce that, for all j,

$$\widetilde{\mathbf{S}}_j = \widetilde{\mathbf{S}}_{j+1} + \widetilde{\mathbf{D}}_{j+1}\ , \tag{10}$$

and hence the details can be interpreted as the differences between successive smooths. If $N = 2^J$ and if we again set $J_0 = J$, then (9) becomes

$$\mathbf{X} = \sum_{j=1}^{J_0} \widetilde{\mathbf{D}}_j + \overline{X}\mathbf{1}, \tag{11}$$

where $\mathbf{1}$ is an N dimensional vector, all of whose elements are ones.

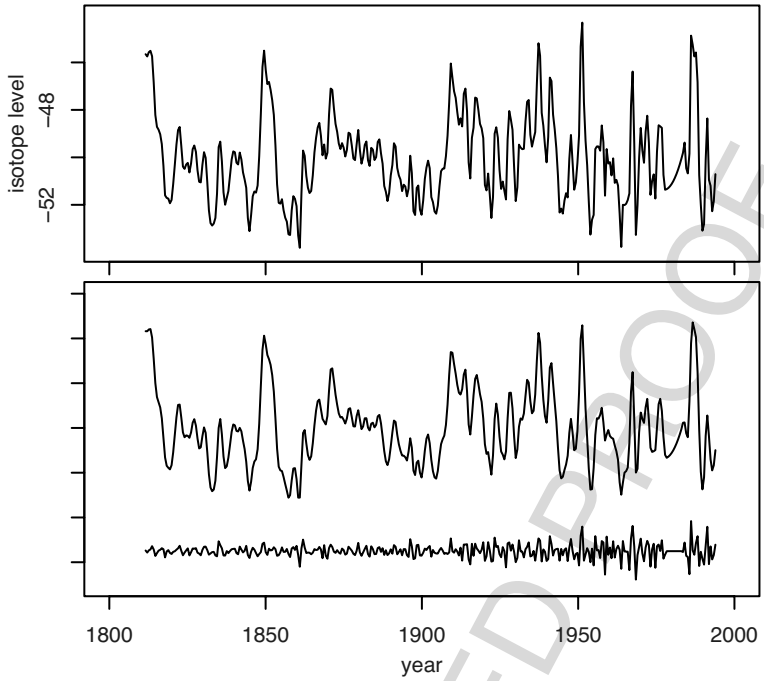


Fig. 7: Oxygen isotope measurements from an Antarctic ice core (*top panel*), along with a Haar MODWT-based multiresolution analysis of level $J_0 = 1$ consisting of the smooth series $\widetilde{\mathbf{S}}_1$ and a single detail series $\widetilde{\mathbf{D}}_1$ (*bottom panel*, *upper* and *lower plots*, respectively). Due to the large difference between the beginning and end of the series, the MODWT was computed using so-called reflection boundary conditions rather than the periodic conditions described in Sect. 2 (for details, see p. 140 of [7]). Data is courtesy of Lars Karlöf, Norwegian Polar Institute, Polar Environmental Centre, Tromsø, Norway.

As a simple example of an MRA, let us consider a series $\mathbf{X}$ of $N = 352$ oxygen isotope measurements from an ice core taken at one location on a spatial array with 3.5 to 7 km spacing in Dronning Maud Land, Antarctica. Here the spacing between observations is taken to be $\Delta = 0.5$ years (the raw measurements are indexed by distance along the core, but these are then mapped to values at half-year intervals). The series is plotted in the upper panel of Fig. 7 and has a temporal span of 176 years. For each of the cores in the array, an MRA was conducted in order to compare details with similar scales to ascertain which scales are dominated by environmental noise and which might contain a common signal (see [21] for details). Here we demonstrate that the simplest possible MRA for the core shown in Fig. 7 reveals some interesting properties not readily apparent in a plot of the data itself.

The lower panel of Fig. 7 depicts a level $J_0 = 1$ Haar MODWT-based MRA, consisting of a smooth series $\widetilde{\mathbf{S}}_1$ and a single detail series $\widetilde{\mathbf{D}}_1$, which, upon being added together, yield $\mathbf{X}$. The smooth series is the portion of $\mathbf{X}$ that can be attributed to averages over a scale of a year, whereas the detail series represents variations over a half-year scale. What is interesting is that the local variability in $\widetilde{\mathbf{D}}_1$ increases gradually with the passage of time. This increase is not readily apparent in the plot of $\mathbf{X}$ itself, but the MRA pulls it out clearly. The physical mechanism behind this increase is not fully understood, but is thought to be due to diffusion.

In addition the MRA reveals an artifact in $\widetilde{\mathbf{D}}_1$ centered at 1981, around which the detail series is flat for a stretch of 5.5 years. This is due to a linear interpolation scheme used to fill in a break in the ice core. Differencing a series of the form $X_n = a + bn$ leads to wavelet coefficients that are proportional to the slope b and a detail series that is flat. For the questions that this MRA and those for other cores in the spatial array were used to address, filling in a small number of short gaps by linear interpolation is acceptable. Had we been interested in estimating the wavelet variance for a series with many gaps, linear interpolation could bias the estimates unacceptably towards zero. In this case it is advisable to use either a wavelet variance estimator that is specifically designed to work with gappy time series [22] or a stochastic interpolation scheme that preserves the small scale properties of the time series based upon a nominal stochastic model (see [19], Appendix B).

Other examples of the use of MRAs in geophysics include the analysis of subtidal sea level fluctuations [23], magnetic storm activity [24], the Lisbon and Gibraltar North Atlantic Oscillation winter indices [25], spatial variation of microflora abundance in agricultural soil [26], the December 26th 2004 tsunami as recorded along the southeastern coast of Brazil [27] and large-scale coherent structures in turbulent separation bubbles [28].

5 Orthonormal Discrete Wavelet Transform

While a level J_0 MODWT of a time series of length N consists of a total of $(J_0 + 1) \times N$ values, it is also possible to define a discrete wavelet transform that consists of just N values. This transform is orthonormal, which means that the transpose of $N \times N$ matrix $\mathcal{W}$ relating the time series to the transform coefficients is the inverse of $\mathcal{W}$. We hence use the acronym 'ODWT' to denote this transform. We can readily define the ODWT in terms of the MODWT if N happens to be an integer multiple of 2^{J_0} (if N is not of this form, an ODWT can still be defined, but not as easily – see pp. 141–145 of [7] for details). The ODWT wavelet coefficients are given by

$$W_{j,n} = 2^{j/2}\widetilde{W}_{j,2^j(n+1)-1}\ , \quad n = 0, 1, \ldots, \frac{N}{2^j} - 1\ , \quad j = 1, 2, \ldots, J_0\ ;$$

i.e., the ODWT coefficients are obtained by subsampling and rescaling the MODWT coefficients. For example, at level $j = 1$, the ODWT coefficients are

formed by taking the MODWT coefficients with odd indices and multiplying them by $\sqrt{2}$, whereas, at level $j = 2$, we subsample every fourth coefficient and multiply them by 2:

$$\sqrt{2}\begin{bmatrix} \widetilde{W}_{1,1} \\ \widetilde{W}_{1,3} \\ \widetilde{W}_{1,5} \\ \vdots \\ \widetilde{W}_{1,N-3} \\ \widetilde{W}_{1,N-1} \end{bmatrix} = \begin{bmatrix} W_{1,0} \\ W_{1,1} \\ W_{1,2} \\ \vdots \\ W_{1,\frac{N}{2}-2} \\ W_{1,\frac{N}{2}-1} \end{bmatrix} = \mathbf{W}_1 \;\&\; 2\begin{bmatrix} \widetilde{W}_{2,3} \\ \widetilde{W}_{2,7} \\ \widetilde{W}_{2,11} \\ \vdots \\ \widetilde{W}_{2,N-5} \\ \widetilde{W}_{2,N-1} \end{bmatrix} = \begin{bmatrix} W_{2,0} \\ W_{2,1} \\ W_{2,2} \\ \vdots \\ W_{2,\frac{N}{4}-2} \\ W_{2,\frac{N}{4}-1} \end{bmatrix} = \mathbf{W}_2 \,.$$

As the level j increases, we need to subsample fewer and fewer MODWT wavelet coefficients in order to create the corresponding ODWT coefficients in $\mathbf{W}_j$. In a similar manner the ODWT scaling coefficients are defined by

$$V_{J_0,n} = 2^{J_0/2}\widetilde{V}_{J_0,2^{J_0}(n+1)-1}, \quad n = 0, 1, \ldots, N_{J_0} - 1 \,,$$

and can be placed in an N_{J_0} dimensional vector denoted as $\mathbf{V}_{J_0}$. The ODWT of level J_0 consists of the collection of vectors $\mathbf{W}_1$, $\mathbf{W}_2$, …, $\mathbf{W}_{J_0}$ and $\mathbf{V}_{J_0}$, whose dimensions are, respectively, $N/2$, $N/4$, …, $N/2^{J_0}$ and $N/2^{J_0}$, which collectively sum to N.

As was true for the MODWT, the ODWT leads to a scale-based ANOVA and an MRA. We start by considering the analogs of (3) and (4):

$$\mathbf{W}_j = \mathcal{W}_j\mathbf{X} \;\text{ and }\; \mathbf{V}_{J_0} = \mathcal{V}_{J_0}\mathbf{X} \,, \tag{12}$$

where $\mathcal{W}_j$ is an $\frac{N}{2^j} \times N$ matrix whose rows are selected rescaled rows from $\widetilde{\mathcal{W}}_j$, while $\mathcal{V}_{J_0}$ is an $\frac{N}{2^{J_0}} \times N$ matrix whose rows are selected rescaled rows from $\widetilde{\mathcal{V}}_{J_0}$. The ODWT-based ANOVAs and MRAs are easy to state: just remove all the tildes from Eqs. (5) through (11)! In practice the ODWT wavelet and scaling coefficients are not computed by subsampling the corresponding MODWT coefficients, but rather via an efficient pyramid algorithm (pseudo-code for this algorithm is given on pp. 100–101 of [7]).

In general, MODWT-based ANOVAs and MRAs tend to perform better than their ODWT equivalents because of the deleterious effect that subsampling can have on the ODWT (for details, see Sects. 5.1, 5.6 and 8.3 of [7]); however, the ODWT is the transform of choice for certain other types of analyses. For example, if a time series can be modeled as a signal plus Gaussian white noise, then its ODWT consists of a transformed signal plus Gaussian white noise. Certain types of signals are more easily recognized in the ODWT domain than in their original time domain representation, which makes it possible to design effective data-adaptive procedures for extracting signals buried in white noise. This fact is exploited in the large body of literature devoted to wavelet shrinkage; see [29] for a recent review article that emphasize this

use of the ODWT. (There is a device called 'cycle spinning' in which ODWT-based signal extraction is applied to a time series and all its possible circular shifts, followed by an averaging of the N extracted signals. This procedure is equivalent to a signal estimation procedure based upon the MODWT; for details, see pp. 429–431 of [7]).

As a second example, the ODWT transforms certain – but not all – time series into a collection of wavelet coefficients that are approximately uncorrelated within and between levels, but that have possibly level-dependent variances. Time series with long-range dependence are examples of ones that are effectively decorrelated by the ODWT. This decorrelating property can be put to good use in formulating wavelet-based approximate maximum likelihood estimators of parameters associated with processes with long-range dependence, in simulating series with long-range dependence and in formulating bootstrap procedures for assessing the sampling variability in certain statistics (for details, see [7, 30]).

6 Beyond the Haar Wavelet

Our discussion so far has focused on the Haar MODWT and corresponding ODWT, but there are other versions of both transforms. For a selected maximum level J_0, these transforms can be formulated in terms of wavelet filters of levels $j = 1, \ldots, J_0$ and a scaling filter of level J_0. Figure 8 shows the level $j = 3$ wavelet filters for the Haar transform and LA(8) transform, where 'LA(8)' stands for the member of the Daubechies 'least asymmetric' family whose level $j = 1$ filter has width $L = 8$ [31]. The shape of the Haar filter tells us that the corresponding wavelet coefficients are proportional to differences of adjacent simple averages of scale 4. The shape of the LA(8) filter says that the wavelet coefficients can be interpreted as the difference between a centrally located weighted average and weighted averages occurring before and after it. Once the wavelet and scaling filters have been used to properly formulate the matrices $\widetilde{\mathcal{W}}_j$, $\widetilde{\mathcal{V}}_{J_0}$, $\mathcal{W}_j$, and $\mathcal{V}_{J_0}$ of (3), (4) and (12), all of the equations involving the Haar MODWT and ODWT presented in Sects. 2–5 also hold for the corresponding LA(8) transforms.

The LA(8) transform can yield a more informative ANOVA and MRA than the Haar for certain time series because the latter can suffer from 'leakage' effects in which the wavelet coefficients for a particular scale are locked into patterns driven by a nearby dominant scale. The fact that the wavelet coefficients are highly correlated between different levels is undesirable because the transform is then not successfully partitioning out different aspects of a time series into different coefficients. In many geophysical applications, including the ones used as examples in Sects. 3 and 4, an analysis based upon the Haar wavelet is entirely adequate, and there is no need to consider other wavelets. An effective procedure for deciding if the Haar wavelet is adequate or not is to compare analyses based upon the Haar wavelet with those based

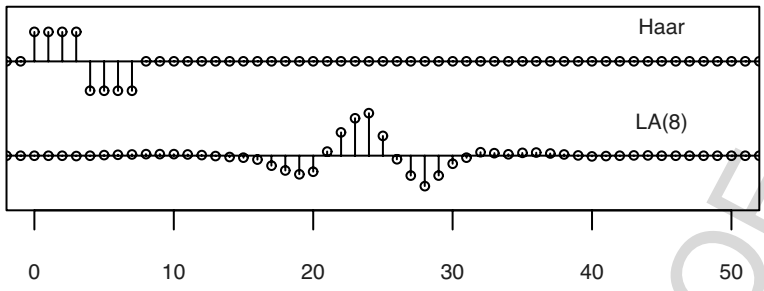


Fig. 8: Filters used to produce scale τ_3 wavelet coefficients based upon the Haar wavelet filter (*top*) and the Daubechies least asymmetric wavelet filter of width 8 (*bottom*).

upon other wavelets. If the analyses are basically the same, there is no need to use anything other than the Haar; if not, an analysis using something other than the Haar wavelet might be called for. Use of non-Haar MODWTs and ODWTs produces more boundary coefficients, so there is a price to pay in abandoning a Haar-based analysis.

7 Concluding Comments

Hopefully the overview presented here has given the reader some idea of the potential uses of DWTs in analyzing geophysical time series. There are many aspects of wavelet analysis that we have not touched upon, including the fact that all of the procedures we have discussed can be applied to time series whose statistical properties are evolving over time. The ability of DWTs to handle this case, which is eluded to briefly in the MRA for the oxygen series presented in Fig. 7, is tied up with the fact that the wavelet coefficients extract information not only across different scales, but also across time. For example, a wavelet variance estimator in which the squared wavelet coefficients are averaged locally rather than globally (as in the construction of $\hat{\nu}_j^2$) is an effective way of studying time-varying properties in a time series. The reader should consult [7] for details on this and other aspects of wavelet analysis not covered in this brief overview.

In Sect. 5 we discussed some of the relative strengths and weaknesses of the MODWT and the ODWT. These DWTs are closely related to corresponding continuous wavelet transforms (CWTs), which also are quite commonly used to analyze geophysical time series. A CWT might be called an 'anti-statistic' in the sense that, rather than summarizing the information in a time series, it converts it into a two-dimensional field. As a result, there is a considerable amount of redundant information in a CWT, which is both a strength and a weakness. One example where this redundancy is a strength is in the analysis of certain types of singularities ('cusps'), where the nature of the singularity

can be deduced by tracing the wavelet transform modulus maxima across a fine grid of scales (see, e.g., [32], Fig. 6.5). The dyadic scales used in DWTs are typically too coarsely spaced to make this type of singularity analysis feasible.

The redundancy in the CWT, however, can make proper interpretation of 'heat' plots of the CWT problematic, i.e., scale versus time plots in which the magnitudes of the CWT coefficients are color-coded. These plots often have rather striking structures that our eyes are drawn toward, but that can be largely attributed to the fact that CWT coefficients are typically highly correlated both spatially and temporarily. Proper statistical assessment of the significance of these structures involves some subtle issues [33], particularly if they are picked out by eye prior to being assessed. Subsampling to the dyadic scales in the MODWT and ODWT essentially breaks this correlation structure spatially, and subsampling the MODWT to get the ODWT does the same temporarily. The fact that collections of coefficients from these DWTs are approximately uncorrelated makes it easier to devise statistical tests and to implement bootstrapping procedures (the latter are not feasible with CWTs).

Finally we note that the CWT does not formally involve components in a time series that are handled in DWTs by the scaling coefficients. These are often useful for extracting large-scale trends that are an important part of some geophysical time series and that are a key component in wavelet-based signal extraction.

Acknowledgements. The author gratefully acknowledges partial support for preparation of this article from the U.S. National Science Foundation under award number 0529955.

References

1. Goupillaud, P., Grossmann, A. and Morlet, J. (1984). Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, **23**, 85–102.
2. Grossmann, A. and Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, **15**, 723–736.
3. Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, **41**, 909–996.
4. Mallat, S.G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Transactions of the American Mathematical Society*, **315**, 69–87.
5. Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.
6. Mallat, S.G. (1989). Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 2091–2110.
7. Percival, D.B. and Walden, A.T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, UK.

8. Torrence, C. and Compo, G.P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, **79**, 61–78.
9. Pelgrum, H., Schmugge, T., Rango, A., Ritchie, J. and Kustas, B. (2000). Length-scale analysis of surface albedo, temperature, and normalized difference vegetation index in desert grassland. *Water Resources Research*, **36**, 1757–1766.
10. Lark, R.M. and Webster, R. (2001). Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform. *European Journal of Soil Science*, **52**, 547–562.
11. Labat, D., Ababou, R. and Mangin, A. (2001). Introduction of wavelet analyses to rainfall/runoffs relationship for a karstic basin: the case of Licq–Atherey karstic system. *Ground Water*, **39**, 605–615.
12. Massel, S.R. (2001). Wavelet analysis for processing of ocean surface wave records. *Ocean Engineering*, **28**, 957–987.
13. Rybák, J. and Dorotovič, I. (2002). Temporal variability of the coronal green-line index (1947–1998). *Solar Physics*, **205**, 177–187.
14. Barbosa, S.M., Fernandes, M.J. and Silva, M.E. (2006). Long range dependence in North Atlantic sea level. *Physica A: Statistical Mechanics and its Applications*, **371**, 725–731.
15. Cornish, C.R., Bretherton, C.S. and Percival, D.B. (2006). Maximal overlap wavelet statistical analysis with application to atmospheric turbulence. *Boundary-Layer Meteorology*, **119**, 339–374.
16. Steel, E.A. and Lange, I.A. (2007). Using wavelet analysis to detect changes in water temperature regimes at multiple scales: effects of multi-purpose dams in the Willamette River basin. *River Research and Applications*, **23**, 351–359.
17. Flato, G.M. (1995). Spatial and temporal variability of Arctic ice thickness. *Annals of Glaciology*, **21**, 323–329.

AU: Please provide publisher details for Reference [18].

18. World Meteorological Organization (2007). *Sea-Ice Information Services in the World*. (3rd ed.) Publication 574.

AU: Please update the volume and page numbers for references [19] and [20].

19. Percival, D.B., Rothrock, D.A., Thorndike, A.S. and Gneiting, T. (2008). The variance of mean sea-ice thickness: effect of long-range dependence. *Journal of Geophysical Research – Oceans*, in press.
20. Rothrock, D.A., Percival, D.B. and Wensnahan, M. (2008). The decline in arctic sea-ice thickness: separating the spatial, annual, and interannual variability in a quarter century of submarine data. *Journal of Geophysical Research – Oceans*, in press.
21. Karlöf, L., Winebrenner, D.P. and Percival, D.B. (2006). How representative is a time series derived from a firn core? A study at a low accumulation site on the Antarctic plateau. *Journal of Geophysical Research – Earth Surface*, **111**, F04001, doi:10.1029/2006JF000552.

AU: Please update the reference [22].

22. Mondal, D. and Percival, D.B. (2008). Wavelet variance analysis for gappy time series. *Annals of the Institute of Statistical Mathematics*, under review.
23. Percival, D.B. and Mofjeld, H.O. (1997). Analysis of subtidal coastal sea level fluctuations using wavelets. *Journal of the American Statistical Association*, **92**, 868–880.
24. Jach, A., Kokoszka, P., Sojka, J. and Zhu, L. (2006). Wavelet-based index of magnetic storm activity. *Journal of Geophysical Research – Space Physics*, **111**, A09215, doi:10.1029/2006JA011635.
25. Barbosa, S., Silva, M.E. and Fernandes, M.J. (2006). Wavelet analysis of the Lisbon and Gibraltar North Atlantic Oscillation winter indices. *International Journal of Climatology*, **26**, 581–593.

26. Barnes, R.J., Baxter, S.J. and Lark, R.M. (2007). Spatial covariation of *Azotobacter* abundance and soil properties: a case study using the wavelet transform. *Soil Biology and Biochemistry*, **39**, 295–310.
27. França, C.A.S. and De Mesquita, A.R. (2007). The December 26th 2004 tsunami recorded along the southeastern coast of Brazil. *Natural Hazards*, **40**, 209–222.
28. Chun, S.J., Liu, Y.Z. and Sung, H.J. (2007). Multi-resolution analysis of the large-scale coherent structure in a turbulent separation bubble affected by an unsteady wake. *Journal of Fluids and Structures*, **23**, 85–100.
29. Wang, Y. (2006). Selected review on wavelets in statistics. In *Frontiers in Statistics: Dedicated to Peter John Bickel in Honor of His 65th Birthday* (J. Fan and H.L. Koul, eds.). Imperial College Press, London, pp. 163–182.
30. Percival, D.B., Sardy, S. and Davison, A.C. (2001). Wavestrapping time series: adaptive wavelet-based bootstrapping. In *Nonlinear and Nonstationary Signal Processing* (W.J. Fitzgerald, R.L. Smith, A.T. Walden and P.C. Young, eds.). Cambridge University Press, Cambridge, UK, pp. 442–470.
31. Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.
32. Mallat, S.G. (1999). *A Wavelet Tour of Signal Processing* (2nd ed.). Academic Press, San Diego.
33. Maraun, D., Kurths, J. and Holschneider, M. (2007). Nonstationary Gaussian processes in wavelet domain: synthesis, estimation, and significance testing. *Physical Review E* **75**, 016707.

Automatic Parameter Estimation in a Mesoscale Model Without Ensembles

Gregory S. Duane and Joshua P. Hacker

National Center for Atmospheric Research, Boulder, CO, gduane@ucar.edu

Abstract. In numerical forecasting, unknown model parameters have been estimated from a time series of observations by regarding them as extra state variables, and applying standard data assimilation methods that use ensembles to represent background error. In many situations, however, the use of ensembles is prohibitively expensive and/or impracticable because of the inability to properly account for model error in the initialization scheme. If one is seeking to estimate model parameters as data is assimilated, it is possible to take advantage of the assumed relative constancy of such parameters over large regions of time and space to derive an estimate from a single realization. The approach follows from a general result on synchronously coupled dynamical systems, where one system here represents "truth" and the other "model": If two such systems can be made to synchronize when their corresponding parameters are identical, for any coupling scheme (such as might be used in conventional data assimilation) a parameter estimation law can generally be added that will dynamically reduce a total cost (Lyapunov) function including parameter mismatch terms as well as state mismatch terms.

The approach is used to estimate a parameter that quantifies the effect of soil moisture in a single-column version of the Weather Research and Forecasting (WRF) model. The scheme can be extended to infer a 2D map of soil parameter values for a 3D model, using the fact that the parameter is slowly varying almost everywhere. Discontinuities are represented as additional degrees of freedom, and the Lyapunov function is augmented so as to penalize for horizontal variations in the soil parameter value except at locations of such discontinuities. The constrained optimization approach that is proposed should be useful for a variety of parameter estimation problems in numerical weather prediction (NWP), and will extend the power of ensemble methods.

AU: Please provide Keywords.

1 Introduction

Any scheme for meteorological data assimilation has the goal of synchronizing a computational model with the real climate system, based on a limited set of noisy observations. Where the purpose is prediction of the state of the real system in the not-too-distant future, this synchronization view contrasts *a priori* with the usual view, in which observations are combined with the current model state to form the best possible estimate of the *current* state of the real system.

In this chapter, we extend previous work on synchronization-based data assimilation, summarized in the next two sections, to show that parameters can be readily synchronized, as well as states. The synchronization view lends itself particularly well to the estimation of slowly varying parameters, a point made with a simplified, single-column version of an actual weather prediction model in Sect. 4. It is argued that the synchronization view also lends itself to the estimation of local parameter values that are slowly varying in space almost everywhere. In the concluding section, it is argued that the approach to parameter estimation can be extended to a more general scheme for machine learning.

2 Background: Data Assimilation and Synchronized Chaos

The phenomenon of chaos synchronization [1] was first brought to light by Fujisaka and Yamada [2] and independently by Afraimovich et al. [3], but extensive research on the subject in the '90s was spurred by the seminal work of Pecora and Carroll [4], who considered two chaotic systems in a master-slave relationship defined by a shared subsystem. Pecora and Carroll considered configurations such as the following combination of Lorenz systems:

$$\begin{aligned} \dot{X} &= \sigma(Y - X) \\ \dot{Y} &= \rho X - Y - XZ \qquad & \dot{Y}_1 &= \rho X - Y_1 - XZ_1 \\ \dot{Z} &= -\beta Z + XY & \dot{Z}_1 &= -\beta Z_1 + XY_1 \end{aligned} \tag{1}$$

which synchronizes rapidly, slaving the Y_1, Z_1-subsystem to the master X, Y, Z-subsystem, as seen in Fig. 1, despites differing initial conditions and despite sensitive dependence on initial conditions.

If we imagine that the first Lorenz system represents the world, and that the second Lorenz system is a predictive model, then synchronization effects data assimilation of observed variables into the running model. The only observed variable in the foregoing example is X, but that is sufficient to cause the desired convergence of model to truth. Synchronization is known to be tolerant of reasonable levels of noise, as might arise in the observation channel, and occurs with partial coupling schemes that do not completely replace a model variable with a variable of the observed system.

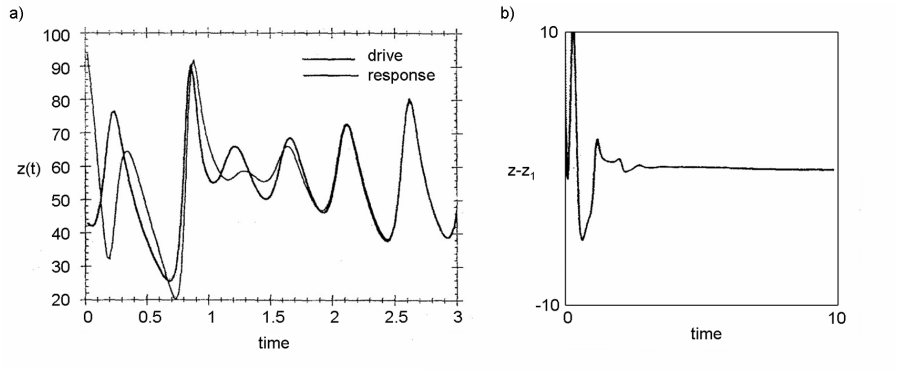


Fig. 1: The trajectories of the synchronously coupled Lorenz systems in the Pecora-Carroll complete replacement scheme (1) rapidly converge (**a**). Differences between corresponding variables approach zero (**b**).

Specifically, systems can also synchronize when coupled *diffusively*, as with a pair of directionally coupled Rossler systems:

$$\begin{aligned} \dot{X} &= -Y - Z + \alpha(X_1 - X) & \dot{X}_1 &= -Y_1 - Z_1 + \alpha(X - X_1) \\ \dot{Y} &= X + aY & \dot{Y}_1 &= X_1 + aY_1 \\ \dot{Z} &= b + Z(X - c) & \dot{Z}_1 &= b + Z_1(X_1 - c) \end{aligned} \tag{2}$$

where α parametrizes the coupling strength. The diffusive coupling scheme can be seen to resemble the "nudging" approach to data assimilation [5]. For judicious choice of nudging coefficient, it can be seen to resemble 3DVar, and for time-varying coefficient, Kalman filtering (as defined in e.g. [6]).

It is commonly not the existence, but the stability of the synchronization manifold that distinguishes coupled systems exhibiting synchronization from those that do not (such as (2) for different values of α). N Lyapunov exponents can be defined for perturbations in the N-dimensional space that is transverse to the synchronization manifold $\mathcal{M}$. If the largest of these, $h^{\perp}_{max}$, is negative, then motion in the synchronization manifold is stable against transverse perturbations. In that case, the coupled systems will synchronize for some range of differing initial conditions. However, since $h^{\perp}_{max}$ only determines *local* stability properties, the size of the basin of attraction for the synchronized regime remains unknown. As $h^{\perp}_{max}$ is increased through zero, the system undergoes a *blowout bifurcation*. For small positive values of $h^{\perp}_{max}$, on-off synchronization occurs (a special case of on-off intermittency), as illustrated in Fig. 2b, where degradation results from a time lag in the coupling. The other panels of Fig. 2 show an increasing rate of bursting as the time-lag increases. Vestiges of synchronization are discernible even far from the blowout bifurcation point (Fig. 2c), a phenomenon that was used to predict new teleconnection patterns [7].

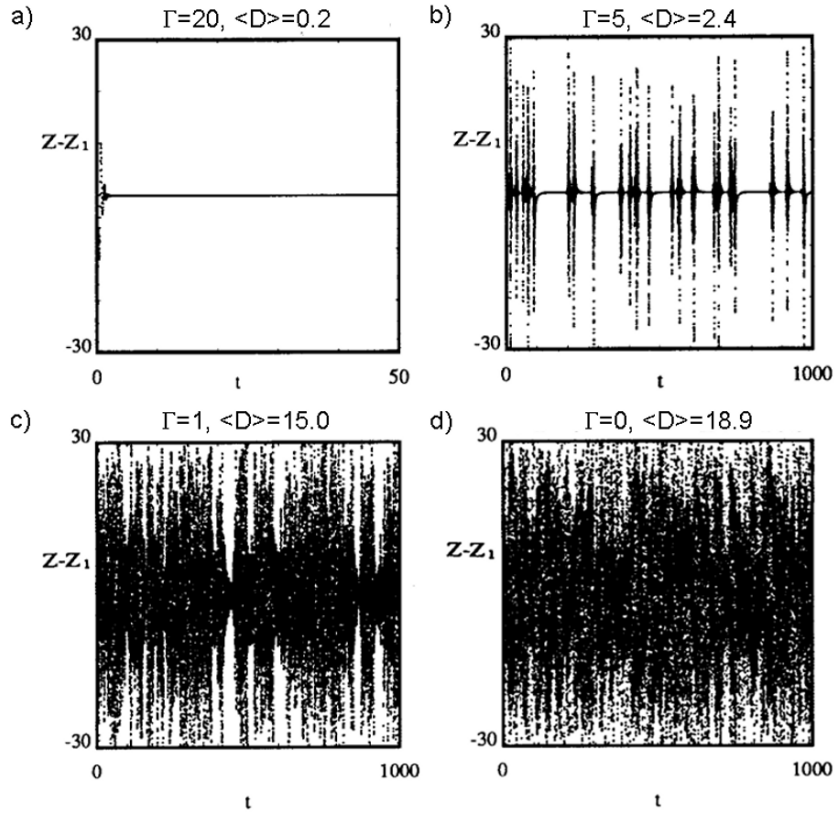


Fig. 2: The difference between the simultaneous states of two Lorenz systems with time-lagged coupling, represented by $Z(t) - Z_1(t)$ vs. t for various values of the inverse time-lag Γ illustrating complete synchronization (**a**), intermittent or "on-off" synchronization (**b**), partial synchronization (**c**), and de-coupled systems (**d**). Average euclidean distance $\langle D \rangle$ between the states of the two systems in X, Y, Z-space is also shown. The trajectories are generated by adaptive Runge-Kutta numerical integrations with $\sigma = 10.$, $\rho = 28.$, and $\beta = 8/3$.

The early work on synchronized chaos was spurred by an intended application to secure communications, since the signal connecting the two synchronized systems can be difficult to distinguish from noise. Practical applications to cryptography have not emerged, largely because system parameters can be extracted from the coupling signal, with some effort, for low-dimensional systems. But the essence of the phenomenon is that two systems that are effectively unpredictable, connected by a signal that may be almost indecipherable, can still exhibit significant correlations. It is argued that this phenomenon makes weather prediction possible, and will be more generally useful for real-time computational modeling of natural systems.

3 Background: Synchronization in Geophysical Fluid Dynamics

Synchronization in geophysical fluid models was demonstrated by Duane and Tribbia [8], originally with a view toward predicting and explaining new families of long-range teleconnections [9]. The uncoupled single-system model in this work was derived from one described by Vautard et al. [10].

The model is given by the quasigeostrophic equation for potential vorticity q in a two-layer reentrant channel on a β-plane:

$$\frac{Dq_i}{Dt} \equiv \frac{\partial q_i}{\partial t} + J(\psi_i, q_i) = F_i + D_i \tag{3}$$

where the layer $i = 1, 2$, ψ is streamfunction, and the Jacobian $J(\psi, \cdot) = \frac{\partial \psi}{\partial x}\frac{\partial \cdot}{\partial y} - \frac{\partial \psi}{\partial y}\frac{\partial \cdot}{\partial x}$ gives the advective contribution to the Lagrangian derivative D/Dt. Equation (3) states that potential vorticity is conserved on a moving parcel, except for forcing F_i and dissipation D_i. The discretized potential vorticity is

$$q_i = \mathrm{f}_0 + \beta \mathrm{y} + \nabla^2 \psi_\mathrm{i} + \mathrm{R}_\mathrm{i}^{-2}(\psi_1 - \psi_2)(-1)^\mathrm{i} \tag{4}$$

where $f(x, y)$ is the vorticity due to the Earth's rotation at each point (x, y), f_0 is the average f in the channel, β is the constant df/dy and R_i is the Rossby radius of deformation in each layer. The forcing F is a relaxation term designed to induce a jet-like flow near the beginning of the channel: $F_i = f_0(q_i^* - q_i)$ for q_i^* corresponding to the choice of ψ^* shown in Fig. 3a. The dissipation terms D_i, boundary conditions, and other parameter values are given in Ref. [9].

Two models of the form (3), $Dq^A/Dt = F^A + D^A$ and $Dq^B/Dt = F^B + D^B$ were coupled diffusively in one direction by modifying one of the forcing terms:

$$F_\mathbf{k}^B = f^B[a_\mathbf{k}(q_\mathbf{k}^* - q_\mathbf{k}^B) + b_\mathbf{k}(q_\mathbf{k}^A - q_\mathbf{k}^B)] \tag{5}$$

where the flow has been decomposed spectrally and the subscript $\mathbf{k}$ on each quantity indicates the wave number $\mathbf{k}$ spectral component. (The layer index i has been suppressed.) The two sets of coefficients $a_\mathbf{k}$ and $b_\mathbf{k}$ were chosen to couple the two channels in some medium range of wavenumbers and to force each channel only with the low wavenumber components of the background flow:

$$a_\mathbf{k} = \begin{cases} 0 & \text{if } |k_x| \le k_{x0} \text{ and } |k_y| \le k_{y0} \\ (k_n/|\mathbf{k}|)^4 & \text{if } |\mathbf{k}| > k_n \\ 1 - (k_0/|\mathbf{k}|)^4 & \text{otherwise} \end{cases}$$

$$b_\mathbf{k} = \begin{cases} 1 - a_\mathbf{k} \text{ if } |\mathbf{k}| \le k_n \\ 0 \qquad\quad \text{if } |\mathbf{k}| > k_n \end{cases}$$

as in Ref. [9], where the constants k_0, k_{x0}, k_{y0} and k_n are defined.

It was found that the two channels thus coupled rapidly synchronize (Fig. 3), starting from initial flow patterns that are arbitarily set equal to the forcing in one channel, and to a different pattern in the other channel.

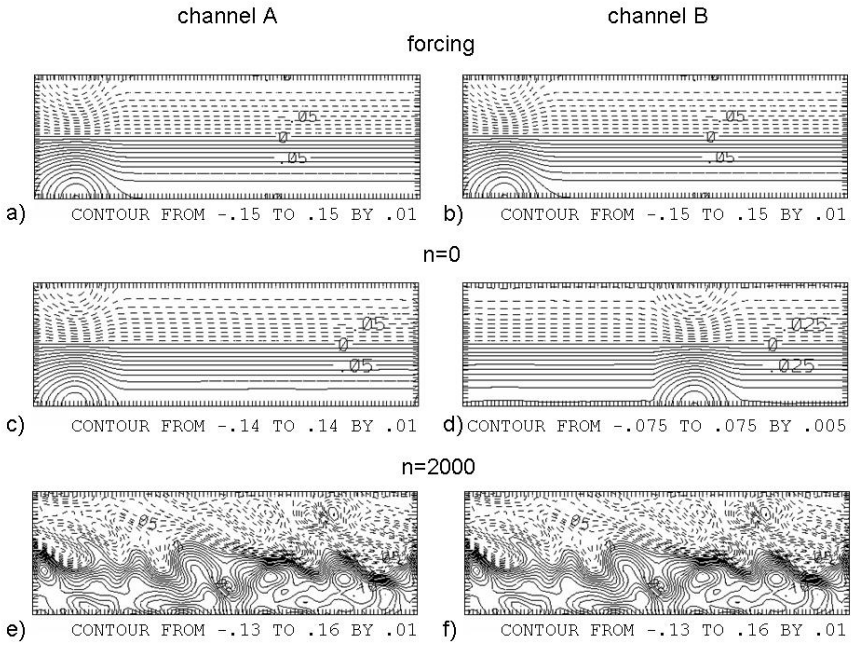


Fig. 3: Streamfunction (in units of $1.48 \times 10^9 m^2 s^{-1}$) describing the forcing ψ^* (**a,b**), and the evolving flow ψ (**c**–**f**), in a parallel channel model with bidirectional coupling of medium scale modes for which $|k_x| > k_{x0} = 3$ or $|k_y| > k_{y0} = 2$, and $|k| \leq 15$, for the indicated numbers n of time steps in a numerical integration. Parameters are as in Ref. [9]. An average streamfunction for the two vertical layers $i = 1, 2$ is shown. Synchronization occurs by the last time shown (**e,f**), despite differing initial conditions.

(Results are shown for bidirectional coupling defined by adding an equation for $F^A_{\mathbf{k}}$ analogous to (5). The synchronization behavior for coupling in just one direction is very similar.) With unidirectional coupling, the synchronization effects data assimilation from the A channel into the B channel.

4 Parameter Adaptation in a Mesoscale Model

Machine learning might also be realized in the synchronization context, so as to correct for deterministic model error in the resolved degrees of freedom. By allowing model parameters to vary slowly, generalized synchronization that is defined by a complex non-identical correspondence between variables in the two models would be transformed to more nearly identical synchronization. Indeed, parameter adaptation laws can be added to a synchronously coupled pair of systems so as to synchronize the parameters as well as the states. Parlitz

[11] showed for example that two unidirectionally coupled Lorenz systems with different parameters:

$$\begin{aligned} \dot{X} &= \sigma(Y - X) & \dot{X}_1 &= \sigma(Y - X_1) \\ \dot{Y} &= \rho X - Y - XZ & \dot{Y}_1 &= \rho_1 X_1 - \nu Y_1 - X_1 Z_1 \\ \dot{Z} &= -\beta Z + XY & \dot{Z}_1 &= -\beta Z_1 + X_1 Y_1 \end{aligned} \tag{6}$$

could be augmented with parameter adaptation rules:

$$\begin{aligned} \dot{\rho}_1 &= (Y - Y_1)X_1 \\ \dot{\nu} &= (Y_1 - Y)Y_1 \\ \dot{\mu} &= Y - Y_1 \end{aligned} \tag{7}$$

so that the Lorenz systems would synchronize, and additionally $\rho_1 \to \rho$, $\nu \to 1$, and $\mu \to 0$.

Equations for a synchronously coupled pair of systems can in fact always be augmented to allow parameter adaptation as well, provided that relevant dynamical variables are observed, as shown by Duane, Yu, and Kocarev [12]. Consider, for example, two quasigeostrophic channel models of the form (3) coupled according to (5), which are known to synchronize, as discussed in the background section. The model parameter to be estimated, f^B, is an overall coefficient in the forcing term

$$F^B_{\mathbf{k}} = f^B a_{\mathbf{k}}(q^*_{\mathbf{k}} - q^B_{\mathbf{k}}) + f^B b_{\mathbf{k}}(q^A_{\mathbf{k}} - q^B_{\mathbf{k}}) \tag{8}$$

where the layer index l is suppressed and the coefficients $a_{\mathbf{k}}, b_{\mathbf{k}}$ are slightly smoothed step functions of $\mathbf{k}$, as before, so that each spectral component is either coupled to the corresponding component in the A system or to the background flow q^* or neither. The coefficients are chosen as before so as to couple only the medium-scale components. (The forcing for the A system is correspondingly: $F^A_{\mathbf{k}} = f^A a_{\mathbf{k}}(q^*_{\mathbf{k}} - q^A_{\mathbf{k}})$).

The parameter estimation rule in spectral space:

$$\dot{f}^B = \sum_{\mathbf{k} \in S} (q^A_{\mathbf{k}} - q^B_{\mathbf{k}})[a_{\mathbf{k}}(q^*_{\mathbf{k}} - q^B_{\mathbf{k}}) + b_{\mathbf{k}}(q^A_{\mathbf{k}} - q^B_{\mathbf{k}})] \tag{9}$$

for a restricted range of wavenumbers in S, as in the figure, causes f^B to converge to f^A, as shown in Fig. 4.

The derivation of the rule (9) is instructive. One chooses the parameter adaptation rule so that a Lyapunov function that quantifies both state error and parameter error is monotonically decreasing, using the fact that a Lyapunov function for the identical-parameter situation is known to be monotonically decreasing, since the identical systems synchronize. The latter, "core" Lyapunov function, $L_o(q^A, q^B)|_{f^A = f^B} \equiv \int d^2x \ (q^A - q^B)^2$, is known to be decreasing at any point (q^A, q^B) in a large region of the coupled-system state space. Consider the more general Lyapunov function for the case of parameter mismatch:

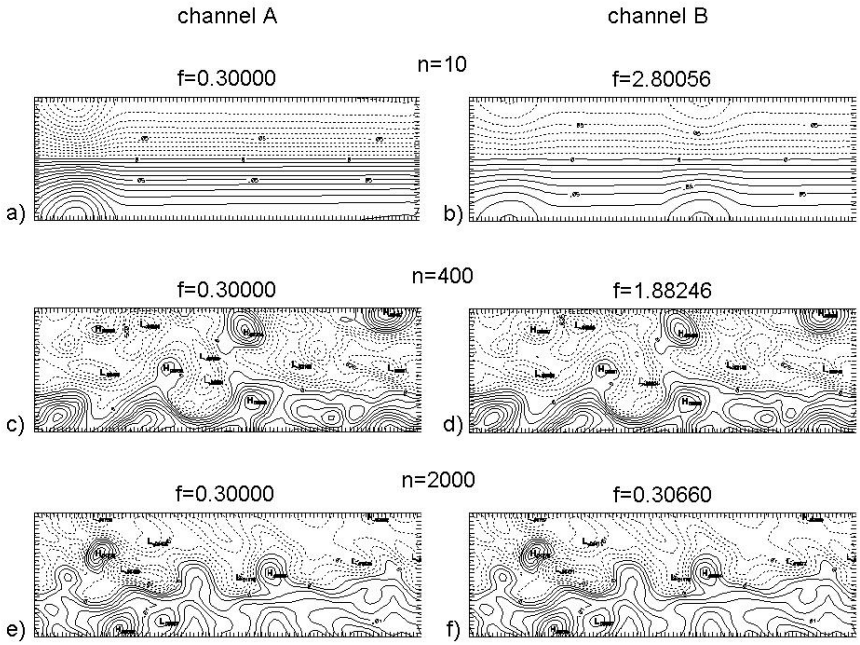


Fig. 4: The evolving flow ψ (a-f) for two quasigeostrophic channel models that are synchronously coupled as in (Duane and Tribbia 2001; Duane and Tribbia 2004) (but in one direction only), and with the forcing parameter f^B for the second channel (denoted μ_o in the reference) allowed to vary according to the truncated parameter adaptation rule (9) with $S = \{\mathbf{k} : \mathbf{k}_x, \mathbf{k}_y \leq 12\}$ (in waves per channel-length). Starting from the initial value $f^B = 3.0$ at time step $n = 0$ (not shown), f^B converges to the value of the corresponding parameter $f^A = 0.3$ in the first channel, as the flows synchronize. (An average of the two layers $l = 1, 2$ is shown.)

$$L(q^A, q^B, f^A, f^B) \equiv (f^A - f^B)^2 + \int d^2x \, (q^A - q^B)^2 \tag{10}$$

and compute the time derivative

$$\dot{L} = -2\dot{f}^B(f^A - f^B) + \frac{d}{dt}\int d^2x \, (q^A - q^B)^2 \tag{11}$$

The second term on the right-hand side can be expanded using the dynamical equation (3), with the forcing term as defined in (8), to compute the extra contribution to $\dot{q}^B$ from the time-varying coefficient f^B

$$\begin{aligned}\frac{d}{dt}\int d^2x\,(q^A-q^B)^2 &= 2\int d^2x\,(q^A-q^B)(\dot{q}^A-\dot{q}^B)\\ &= \dot{L}_o|_{f^B=f^A} + 2(f^A-f^B)\\ &\qquad\times\int d^2x\,(q^A-q^B)\sum_{\mathbf{k}}\left[a_{\mathbf{k}}(q^*_{\mathbf{k}}-q^B_{\mathbf{k}})+b_{\mathbf{k}}(q^A_{\mathbf{k}}-q^B_{\mathbf{k}})\right]e^{i\mathbf{k}\cdot\mathbf{x}}\\ &= \dot{L}_o|_{f^B=f^A} + 2(f^A-f^B)\sum_{\mathbf{k}}(q^A_{\mathbf{k}}-q^B_{\mathbf{k}})\left[a_{\mathbf{k}}(q^*_{\mathbf{k}}-q^B_{\mathbf{k}})+b_{\mathbf{k}}(q^A_{\mathbf{k}}-q^B_{\mathbf{k}})\right]\end{aligned} \tag{12}$$

where the sum on the second line, multiplied by $f^A - f^B$, is the extra contribution to F in (3). From (11) and (12), it is seen that if we choose the adaptation law (9), with S universal, we will have

$$\dot{L} = \dot{L}_o|_{f^B=f^A} \leq 0 \tag{13}$$

as desired. The adaptation law for restricted S can be derived by using a different, correspondingly truncated Lyapunov function.

It is also instructive to consider the effect of a simpler parameter adaptation law. If one ignores the occurence of the parameter f^B in the coupling of the two channels and only retains the first term in the forcing (8), then the parameter adaptation law that guarantees (13) is:

$$\dot{f}^B = \int d^2x\,(q^*-q^B)(q^A-q^B) \tag{14}$$

Under the truncated adaptation rule (14) the monotonic convergence of f^B to the correct value f^A (Fig. 5a) is replaced by oscillatory convergence, as plotted in Fig. 5b. The robustness of the general approach to parameter estimation is apparent in this example.

The general approach that we have illustrated was formalized by Duane et al. [12]. Consider a "real system" given by ODE's:

$$\dot{\mathbf{x}} = \boldsymbol{f}(\mathbf{x},\mathbf{p}), \tag{15}$$

$$\dot{\mathbf{p}} = 0, \tag{16}$$

where $\mathbf{x} \in I\!R^N$, $\boldsymbol{f} : I\!R^N \to I\!R^N$, and $\mathbf{p} \in I\!R^m$ is the vector of (unknown, constant) parameters of the system. We further assume that $\mathbf{s} = \mathbf{s}(\mathbf{x})$, where $\mathbf{s} : I\!R^N \to I\!R^n$, $n \leq N$, is an n dimensional vector representing the experimental measurement output of the system. A "computational model" of the system (15) is given by:

$$\dot{\mathbf{y}} = \boldsymbol{f}(\mathbf{y},\mathbf{q}) + \upsilon(\mathbf{y},\mathbf{s}), \tag{17}$$

$$\dot{\mathbf{q}} = \boldsymbol{N}(\mathbf{y},\mathbf{x}-\mathbf{y}), \tag{18}$$

where $\boldsymbol{N}(\mathbf{y},0) = 0$, and υ is the control signal. Let $\mathbf{e} \equiv \mathbf{y} - \mathbf{x}$ and $\mathbf{r} \equiv \mathbf{q} - \mathbf{p}$. Choose a positive definite Lyapunov function $L_o(\mathbf{e})|_{\mathbf{q}=\mathbf{p}}$. Assume that

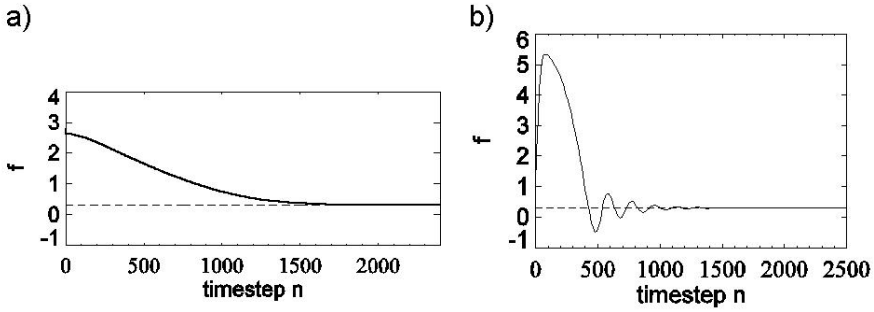


Fig. 5: (**a**) Convergence of f^B to $f^A = 0.3$ (*dashed line*) for the synchronously coupled quasigeostrophic channels, displayed in Fig. 4, and (**b**) convergence for the simplified parameter adaptation rule (14). Monotonic convergence is replaced by oscillatory convergence.

the control signal v is designed such that there is some time t_0 for which $\dot{L}_o(\mathbf{e}(t))|_{\mathbf{q}=\mathbf{p}} < 0$ when $\mathbf{e}(t) \neq 0$ and $\dot{L}_o(\mathbf{e}(t))|_{\mathbf{q}=\mathbf{p}} = 0$ when $\mathbf{e}(t) = 0$, for all $t > t_0$. That is, after time t_0, the system proceeds monotonically toward synchronization. Let $\boldsymbol{h} \equiv \boldsymbol{f}(\mathbf{y}, \mathbf{r}+\mathbf{p}) - \boldsymbol{f}(\mathbf{y}-\mathbf{e}, \mathbf{p})$. Duane et al. [12] established the following theorem:

Theorem 1. *Assume that (i) the control law v in (17) is designed such that the synchronization manifold $\mathbf{x} = \mathbf{y}$ is globally asymptotically stable, (ii) $\boldsymbol{f}$ is linear in the parameters $\mathbf{p}$, and (iii) the parameter estimation law (18) is designed such that*

$$N_j = -\delta_j \sum_i \left(\frac{\partial L_o}{\partial e_i}\right)\left(\frac{\partial h_i}{\partial r_j}\right),$$

where δ_j are positive constants. Then the synchronization manifold $\mathbf{y} = \mathbf{x}$, $\mathbf{p} = \mathbf{q}$ is globally asymptotically stable.

The theorem ensures the stability of the synchronization manifold $\mathbf{y} = \mathbf{x}$, $\mathbf{p} = \mathbf{q}$. It says that if the two systems synchronize for the case of identical parameters, then the parameters of the "real system" can be estimated when they are not known *a priori*, provided that each partial derivative $\partial L_o/\partial e_i$ is known for which the vector $\partial h_i/\partial r_j$ $(j = 1, ...)$ is not zero. For the usual form $L_o \equiv \sum_i (e_i)^2$, the requirement is that x_i be known if the equation for $\dot{y}_i$ contains parameters that one seeks to estimate. By considering a more general Lyapunov function that is defined in terms of some subset S of the state variables, or their indices, $L_o \equiv \sum_{i \in S} c_i (e_i)^2$ for positive coefficients c_i, one obtains the looser requirement for each desired parameter, that x_i be known for at least some i for which the $\dot{y}_i$ equation contains that parameter. (Convergence may be slower if fewer x_i are known.)

As a more realistic example than the quasigeostrophic channel model, the Weather Research and Forecasting (WRF) model was considered, as adapted for weather prediction over military test ranges for the Army Test and Evaluation Command (ATEC). The ATEC application is based on observations that are so frequent that they can be assumed to occur at every numerical time step, so that a continuously coupled differential equation system can be taken to reflect the actual data assimilation scenario.

At a relevant level of model detail, the prognostic equation for humidity (water vapor mixing ratio) q is:

$$\frac{\partial q}{\partial t} = \frac{\partial}{\partial z}\left\{K\left(\frac{\partial q}{\partial z} - Mf(u_0, T_0, ...)\right)\right\} \tag{19}$$

where K is a moisture diffusivity, and $M = M(x, y)$ quantifies the impact of soil moisture at each location (x, y), which is a function f of state variables such as the zonal wind u_0, temperature T_0, etc. at the surface. To study the estimation of M using the synchronization method, attention is restricted to a single vertical column $(x, y) = (x_0, y_0)$ and a model is introduced that is diffusively coupled to ("nudged" by) the true state. The model humidity q_m, for instance, is governed by:

$$\frac{\partial q_m}{\partial t} = \frac{\partial}{\partial z}\left\{K\left(\frac{\partial q_m}{\partial z} - Mf(u_{m0}, T_{m0}, ...)\right)\right\} + c(q_{obs} - q_m) \tag{20}$$

where q_{obs} is the observed humidity (at any level z where an observation is taken) that is the sum of the true q and observational noise. c is a coupling ("nudging") coefficient. Similar equations govern the evolution of temperature T, wind speed u, and other model variables, but the parameter M is thought to enter only the humidity equation (20).

In accordance with the theorem, extended to PDEs in a straightforward way, M for the model was made to vary with observational input as:

$$\dot{M} \sim -\frac{\partial K}{\partial z} f(u_{m0}, T_{m0}, ...)(q_{obs}(z) - q_m(z)) \tag{21}$$

for the case of observations taken at just one level z. For observations at multiple levels, (21) is simply averaged over several values of z.

Repeated convergence of M to its true value, each time followed by a burst away from synchronization, is seen in Fig. 6. The behavior differs from the smooth convergence in the channel model example because the state variables do not converge in the time interval under consideration (Fig. 6e) – in contrast to the nearly complete synchronization of the two channel models. Importantly, in contrast to the example of the channel model, the prognostic equation (19) is an idealization of the behavior of the adapted WRF model, as implemented in software, the details of which were not completely known to the authors.

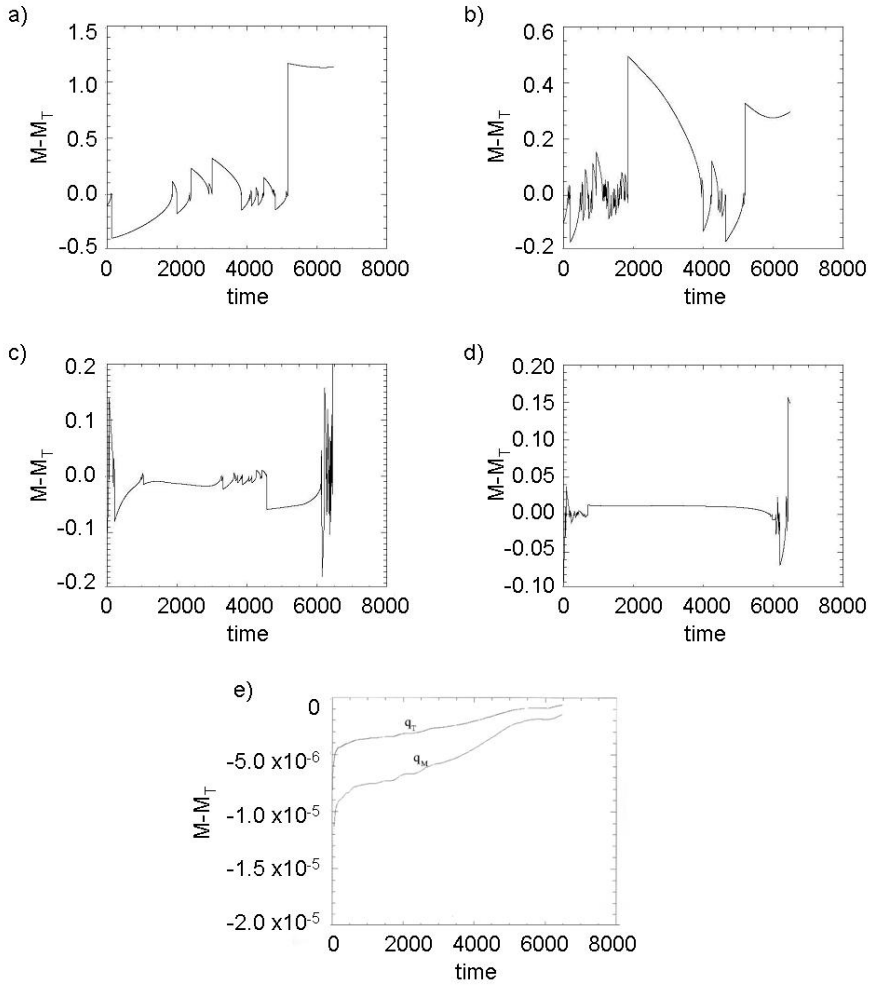


Fig. 6: The variable model parameter M converges to the true value M_T with repeated "bursting", in the dynamical parameter adaptation scheme that only requires a single realization $M - M_T$ is plotted for several cases: (**a**) observations at 7 points in the column, but nudging only at surface level, with nudging coefficient $c = 0.01$ in (20), (**b**) observations and nudging as in (**a**), but with $c = 0.15$, (**c**) observations and nudging at 7 points, with $c = 0.0025$, (**d**) observations and nudging at 4 points, with $c = 0.015$. Results are unstable, but for all cases the correct value of the parameter can be identified. State variables, plotted in (**e**) for observations and nudging at all levels, also do not converge completely over the time interval shown.

In a 3-dimensional model, soil moisture availability $M(x, y)$ is a slowly varying function of ground position (x, y) almost everywhere. It is not slowly varying only for a set of positions of measure zero, at which land cover abruptly changes. It is suggested that the Lyapunov function approach can be readily extended so as to estimate such a parameter field that is slowly varying almost everywhere.

One simply introduces a Lyapunov function of the form:

$$L = L_{\text{single-column}} + \sum_{x,y} \sum_{(x',y') \in N(x,y)} \big[A(M_{x,y} - M_{x',y'})^2 (1 - l_{x,y,x',y'})^2 + B(1 - l_{x,y,x',y'})^2 (l_{x,y,x',y'})^2 \big] + C\Phi(l) \tag{22}$$

and derives dynamical equations such that L is monotonically decreasing. The term in (22) with coefficient A tends to force smooth spatial variation in M except at locations where the new field l has a value near unity. The variable $l_{x,y,x',y'}$ is conceptually located at a position between (x, y) and (x', y'), and is intended to represent a linear discontinuity in land cover. $N(x, y)$ denotes the local neighborhood of point (x, y). The term with coefficent B tends to binarize the values of l, so that either $l \approx 0$ or $l \approx 1$. The expression $\Phi(l)$ (multiplied by an arbitrary coefficient C) denotes a collection of terms that tends to make the discontinuites along which $l \approx 1$ one-dimensional in $(x, y)-$ space, by inhibiting neighboring parallel "edges", and favoring neighboring contiguous edges.

While the suggested extension of the single-column approach has not yet been tested, it can be expected to succeed on theoretical and empirical grounds. The use of multiple neighboring columns to estimate a local soil parameter promises improvement in principle. The treatment of discontinuities resembles methods that have been effectively applied to image segmentation [13].

5 Concluding Remarks: From Parameter Estimation to Model Learning

The extension of the parameter estimation method to 3D, suggests a further extension to qualitative model learning. For problems of qualitative model optimization, as for the estimation of a 2D parameter field, the requisite Lyapunov function has multiple local optima, as does (22). The optimization problem contrasts with those described by a quadratic Lyapunov function that possess a single basin of attraction. Unlike the quadratic case, a stochastic component in the adaptation procedure might play an essential role. The stochastic component would allow jumps among the basins of attraction of the different local optima defined by the deterministic scheme, as in Fig. 7.

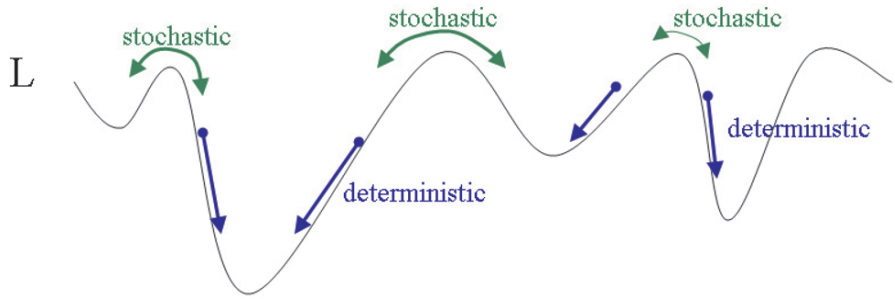


Fig. 7: Deterministic parameter estimation rules cause parameters to reach local optima of the Lyapunov function L. Stochasticity (e.g. in a simulated annealing algorithm) allows jumps among different basins of attraction.

The resulting approach would resemble that of a genetic algorithm, with a "mutation rate" proportional to synchronization error.

The synchronization approach to data assimilation and model learning stands in contrast to the use of ensembles to effectively estimate background error [14]. In an anthropomorphic view of machine learning, the synchronization approach appears more natural – one learns "on the fly", rather than forming multiple copies of oneself to test alternative possibilities. In the case of estimating slowly varying parameters, one is effectively using ergodicity to replace an ensemble average by a time average, computed dynamically, as in (9) and (21). For the full three-dimensional WRF model, the single-column version of which was discussed in the last section, such a replacement is essential, since the dimensionality of mesoscale models precludes the use of a large enough ensemble, with currently available computational resources.

Acknowledgements. The author thanks Joe Tribbia, Jeff Weiss, Dongchuan Yu, and Ljupco Kocarev for useful discussions. This work was partially supported under NSF Grant 0327929. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

References

1. S. Boccaletti, J. Kurths, G. Osipov, D.L. Valladares, C.S. Zhou, The synchronization of chaotic systems, *Phys. Rep.,* **366,** 1–101 (2002)
2. H. Fujisaka, T. Yamada, Stability theory of synchronized motion in coupled-oscillator systems, *Prog. Theor. Phys.,* **69,** 32–47 (1983)
3. V.S. Afraimovich, N.N. Verichev, M.I. Rabinovich, Stochastic synchronization of oscillation in dissipative systems, *Radiophys. Quantum Electron.,* **29,** 795–803 (1986)
4. L.M. Pecora, T.L. Carroll, Synchronization in chaotic systems, *Phys. Rev. Lett.,* **64,** 821–824 (1991)

5. R.A. Anthes, Data assimilation and initialization of hurricane prediction models, *J. Atmos. Sci.,* **31,** 702–719 (1974)
6. E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability,* Cambridge University Press, Cambridge (2003)
7. G.S. Duane, Synchronized chaos in extended systems and meteorological teleconnections, *Phys. Rev. E,* **56,** 6475–6493 (1997)
8. G.S. Duane, J.J. Tribbia, synchronized chaos in geophysical fluid dynamics, *Phys. Rev. Lett.,* **86,** 4298–4301 (2001)
9. G.S. Duane, J.J. Tribbia, Weak Atlantic-Pacific teleconnections as synchronized chaos, *J. Atmos. Sci.,* **61,** 2149–2168 (2004)
10. R. Vautard, B. Legras, M. Déqué, On the source of midlatitude low-frequency variability. Part I: A statistical approach to persistence, *J. Atmos. Sci.,* **45,** 2811–2844 (1988)
11. U. Parlitz, Estimating model parameters from time series by autosynchronization, *Phys. Rev. Lett.,* **76,** 1232–1235 (1996)
12. G.S. Duane, D. Yu, L. Kocarev, Identical synchronization, with translation invariance, implies parameter estimation, *Phys. Lett. A,* **371,** 416–420 (2007)
13. C. Koch, J. Marroquin, A.Yuille, Analog "Neuronal" networks in early vision, *Proc. Nat. Acad. Sci.,* **83,** 4263–4267 (1986)
14. J.L. Anderson, An ensemble adjustment Kalman Filter for data assimilation, *Mon. Wea. Rev.,* **129,** 2884–2903 (2001)

Towards Robust Nonlinear Multivariate Analysis by Neural Network Methods

William W. Hsieh[1] and Alex J. Cannon[2]

[1] Dept. of Earth and Ocean Sciences, University of British Columbia
6339 Stores road, Vancouver, BC, Canada V6T 1Z4, whsieh@eos.ubc.ca
[2] Meteorological Service of Canada, Environment Canada, 201-401 Burrard Street, Vancouver, BC, Canada V6C 3S5, alex.cannon@ec.gc.ca

Abstract. While neural network models have provided nonlinear generalizations of classical linear multivariate models (e.g. regression, principal component and canonical correlation analyses), their applications to the analysis and prediction of real environmental and climate data are not always successful as many of the datasets are very noisy and/or contain relatively few independent observations. We review recent efforts directed towards making the nonlinear models more robust – the development of (1) an information criterion to alleviate overfitting in nonlinear principal component analysis, and (2) a robust version of nonlinear canonical correlation analysis. We also discuss two common causes undermining nonlinear models relative to linear models: (1) Time-averaging of data (e.g. from daily data to seasonal data) linearizes the relation between predictor and predictand due to the central limit theorem. (2) When new predictor data lies outside the training range, the nonlinear model may extrapolate poorly, thereby decreasing its forecast skills.

1 Introduction

The classical tools for multivariate statistical analysis include linear regression (LR), classification, principal component analysis (PCA) and canonical correlation analysis (CCA). These popular methods suffer from the limitation of being linear. Since the late 1980s, neural network (NN) methods have become popular for performing nonlinear regression (NLR) and classification [1]. More recently, NN methods have been extended to perform nonlinear PCA (NLPCA) and nonlinear CCA (NLCCA) [2].

AU: Please provide keywords.

In PCA, a given dataset is approximated by a straight line, which minimizes the mean square error (MSE) – pictorially, in a scatterplot of the data, the straight line found by PCA points in the dominant direction of the dataset. In NLPCA, the straight line in PCA is replaced by a curve which minimizes the MSE. NLPCA can be performed by a variety of methods, e.g. the autoassociative neural network (NN) model [3, 4], and the kernel PCA model [5]. NLPCA belongs to the class of nonlinear dimensionality reduction techniques,

which also includes principal curves [6], locally linear embedding (LLE) [7] and isomap [8]. Self-organizing map (SOM) [9] can also be regarded as a discrete verion of NLPCA.

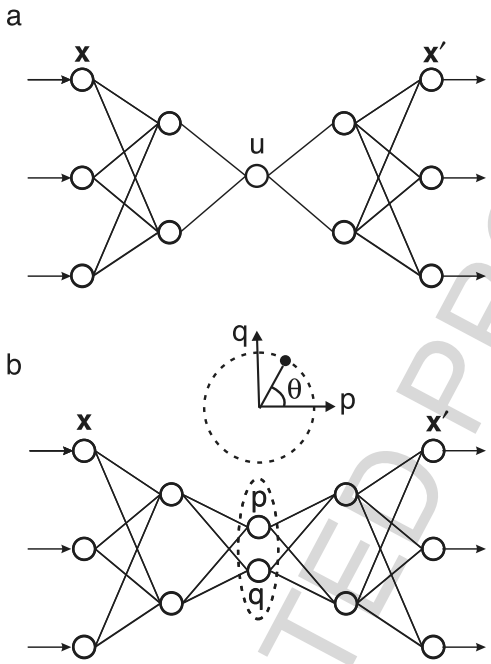


Fig. 1: (**a**) A schematic diagram of the autoassociative feed-forward multi-layer perceptron NN model for performing NLPCA. Between the input layer $\mathbf{x}$ on the left (the 0th layer) and the output layer $\mathbf{x}'$ on the far right (the 4th layer), there are 3 layers of "hidden" neurons (the 1st, 2nd and 3rd layers). Layer 2 is the "bottleneck" with a single neuron u giving the nonlinear principal component (NLPC). Layers 1 and 3 each have m hidden neurons. (**b**) The NN model used for extracting a *closed* curve NLPCA solution. At the bottleneck, there are now two neurons p and q constrained to lie on a unit circle in the p-q plane, giving effectively one free angular variable θ, the NLPC.

To perform NLPCA, the NN model (Fig. 1a) is a standard feed-forward (multi-layer perceptron) NN with 3 "hidden" layers of variables or "neurons" sandwiched between the input layer $\mathbf{x}$ on the left and the output layer $\mathbf{x}'$ on the right, where the middle hidden layer has only a single "bottleneck" neuron u. As an autoassociative model, the MSE between the output $\mathbf{x}'$ and the input $\mathbf{x}$ is minimized, and data compression is achieved by the bottleneck, yielding the nonlinear principal component (NLPC) u (see Appendix A for details). Model complexity can be increased by increasing m, the number of hidden neurons in layer 1 and in layer 3 of the NN (Fig. 1a). Common PCA

algorithms also extract higher PCA modes, with the directions of these modes being orthogonal to each other. In NLPCA, upon subtracting the NLPCA mode 1 solution from the data, the residual can be input into the NLPCA model again to extract the next mode, and so forth, although the orthogonality property of the PCA modes are lost in NLPCA.

Relative to the various other choices for nonlinear dimensionality reduction (kernel PCA, principal curves, LLE and isomap), NLPCA has some nice properties: For instance, its NN architecture provides a continuous (and differentiable) mapping function from $\mathbf{x}$ to the NLPC u, and also from u to $\mathbf{x}'$. In contrast, in kernel PCA, the inverse mapping from u to $\mathbf{x}'$ is a very difficult problem, which until recently lacked numerically stable algorithms [10]. When a new datum $\mathbf{x}_{\text{new}}$ becomes available, its NLPC and its projection onto the NLPCA mode are easily obtained from the NN mapping functions in NLPCA, whereas methods such as principal curves, LLE and isomap do not automatically provide mapping functions to handle the new datum.

The simple correlation between two variables x and y has been generalized in multivariate analysis by CCA, which finds the strongest correlated mode(s) between two sets of variables, $\mathbf{x}$ and $\mathbf{y}$. The first CCA mode extracts the linear oscillation in the $\mathbf{x}$-space which is most strongly correlated with a linear oscillation in the $\mathbf{y}$-space. NLCCA removes the restriction of only looking for linear oscillations in the two spaces. Various approaches based on NN and kernel methods have been proposed for NLCCA [11, 12, 13, 14, 15]. Reference [12] used three feedforward NN mappings to perform NLCCA (Fig. 2) (details in Appendix B).

When using nonlinear machine learning methods such as NN, the presence of noise in the data can lead to overfitting (i.e. fitting to the noise). Regularization (e.g. the addition of weight penalty or decay terms in the cost functions in NN models) has been commonly used to control overfitting by limiting model complexity (i.e. the effective number of model parameters) via the size of the weight penalty parameter(s) [1]. A larger weight penalty parameter P tends to give less nonlinear solutions than a smaller P. Typically, to find the appropriate P in nonlinear regression and classification, a number of models are trained with different P values. The models' MSE are validated on independent data not used in the model training stage, and the model with the lowest validated MSE is selected as the best. Alternatively, Bayesian methods have been developed to automatically estimate the size of the weight penalty parameter in nonlinear regression and classification problems [16, 17]. Since overfitting is much more serious in NLPCA than in NLR [18], a different approach is needed.

In this chapter, we review the recent efforts directed towards making the nonlinear multivariate methods more robust in dealing with noisy data – the use of an information criterion to alleviate overfitting in NLPCA in Sect. 2, and a robust version of nonlinear canonical correlation analysis in Sect. 3 – new developments since the review of [2]. We also discuss two common causes undermining nonlinear models relative to linear models:

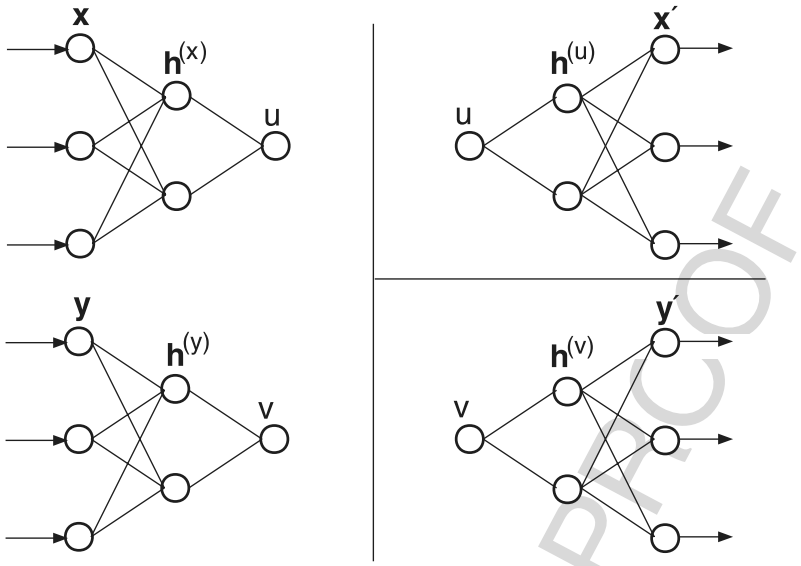


Fig. 2: The three feed-forward NNs used to perform NLCCA. The double-barreled NN on the *left* maps from the inputs $\mathbf{x}$ and $\mathbf{y}$ to the canonical variates u and v. The cost function J_1 forces the correlation between u and v to be maximized. On the *right* side, the two NNs map inversely from u and v to the original $\mathbf{x}$ and $\mathbf{y}$ spaces. The *top* NN maps from u to the output layer $\mathbf{x}'$, with the cost function J_2 basically minimizing the MSE of $\mathbf{x}'$ relative to $\mathbf{x}$. The third NN maps from v to the output layer $\mathbf{y}'$, with the cost function J_3 basically minimizing the MSE of $\mathbf{y}'$ relative to $\mathbf{y}$.

(a) Time-averaging of data (e.g. from daily data to seasonal data) linearizes the relation between predictor and predictand, due to the central limit theorem in statistics (Sect. 4). (b) When new predictor data lies outside the training range, the nonlinear model may extrapolate more poorly than the linear model, thereby lowering the forecast skill of the nonlinear model (Sect. 5).

2 Alleviating Overfitting in NLPCA

In the limit of infinite sample size, overfitting is not a problem when performing nonlinear regression on noisy data, since it can be shown that the output of a flexible enough NLR model approximates the conditional mean of the target data (Sect. 6.1.3 of [1]). While overfitting can also occur in NLPCA [4, 18, 19, 20], the situation is actually far worse than in NLR, because even in the limit of infinite sample size, overfitting is a problem when applying NLPCA to noisy data. As illustrated in Fig. 3, overfitting in NLPCA can arise from the geometry of the data distribution, instead of from the relative

scarsity of observations. Here for a Gaussian-distributed data cloud, a nonlinear model with enough flexibility will find the zigzag solution of Fig. 3b as having a smaller MSE than the linear solution in Fig. 3a. Since the distance between the point A and a, its projection on the NLPCA curve, is smaller in Fig. 3b than the corresponding distance in Fig. 3a, it is easy to see that the more zigzags there are in the curve, the smaller is the MSE. However, the two neighbouring points A and B, on opposite sides of an "ambiguity" line [6, 21], are projected far apart on the NLPCA curve in Fig. 3b. Thus simply searching for the solution which gives the smallest MSE is not a sufficient criterion for NLPCA to find the best solution in a highly noisy dataset.

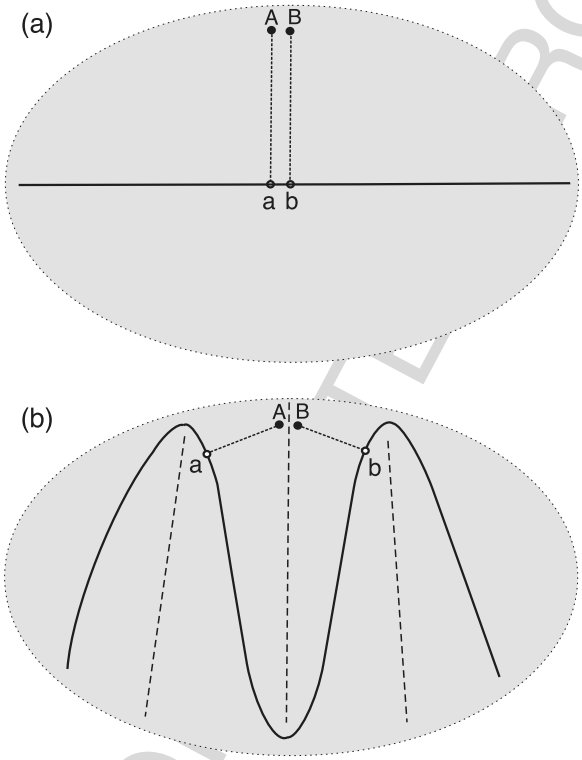


Fig. 3: Schematic diagram illustrating how overfitting can occur in NLPCA of noisy data (even in the limit of infinite sample size). (**a**) PCA solution for a Gaussian data cloud (*shaded in grey*), with two neighbouring points A and B shown projecting to the points a and b on the PCA straight line solution. (**b**) A zigzag NLPCA solution found by a flexible enough nonlinear model. *Dashed lines* illustrate "ambiguity" lines where neighbouring points (e.g. A and B) on opposite sides of these lines are projected to a and b, far apart on the NLPCA curve.

As the NLPCA model of [3] tends to extract zigzag solutions, [4] added weight penalty to the NLPCA model, which brought the overfitting under control. Unfortunately, there was no simple way to objectively estimate the appropriate P value needed to avoid overfitting (and underfitting), because with NLPCA, if the overfitting arises from the data geometry (as in Fig. 3b) and not from the relative scarsity of observations, using independent data to validate the MSE from the various models is not a viable method for selecting the appropriate P. Instead, [18] proposed a new "inconsistency" index for detecting the projection of neighbouring points to distant parts of the NLPCA curve:

For each data point $\mathbf{x}$ and its nearest neighbour $\tilde{\mathbf{x}}$, the NLPC for $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are u and $\tilde{u}$, respectively. With $C(u, \tilde{u})$ denoting the (Pearson) correlation between all the pairs $(u, \tilde{u})$, the inconsistency index I is defined by

$$I = 1 - C(u, \tilde{u})\,. \tag{1}$$

If for some nearest neighbour pairs, u and $\tilde{u}$ are assigned very different values, $C(u, \tilde{u})$ would have a lower value, leading to a larger I, indicating greater inconsistency in the NLPC mapping. With u and $\tilde{u}$ standardized to having zero mean and unit standard deviation, (1) is equivalent to

$$I = \frac{1}{2}\left\langle (u - \tilde{u})^2 \right\rangle, \tag{2}$$

where $\langle \cdots \rangle$ denotes averaging over all observations.

In statistics, various criteria, often in the context of linear models, have been developed to select the right amount of model complexity so neither overfitting nor underfitting occurs. These criteria are often called "information criteria" (IC), e.g. the Akaike IC [22], the Bayesian IC [23], etc. An IC is typically of the form

$$\text{IC} = \text{MSE} + \text{complexity term}, \tag{3}$$

where MSE is evaluated over the training data and the complexity term is larger when a model has more free parameters. The IC is evaluated over a number of models with different free parameters, and the model with the minimum IC is selected as the best. As the presence of the complexity term in the IC penalizes models which use excessive number of free parameters to attain low MSE, choosing the model with the minimum IC would rule out complex models with overfitted solutions.

Due to the presence of multiple minima in the cost function, we randomly divide the data into a training data set and a validation set (containing 85% and 15% of the original data, respectively, in the following examples), and for every given value of P and m, we train the model a number of times from random initial weights, and discard model runs where the MSE evaluated over the validation data is larger than the MSE over the training data. To choose among the model runs which have passed the validation test, a new holistic

IC to deal with the type of overfitting arising from the broad data geometry (Fig. 3b) is introduced as

$$H = \text{MSE} + \text{inconsistency term} \tag{4}$$

$$= \text{MSE} - C(u, \tilde{u}) \times \text{MSE} \;=\; \text{MSE} \times I\,, \tag{5}$$

where MSE and C are evaluated over all (training and validation) data, inconsistency is penalized, and the model run with the smallest H value is selected as the best. The general tendency as more model parameters are used is for MSE to decrease but eventually I increases sharply, thereby producing a minimum in H. (I itself may also have a minimum, but that minimum tends to choose a model with too few parameters, thus underfitting the data.) There is some randomness in the computed H value, since local minima in the cost function introduce randomness in the MSE. Furthermore, [21] showed that the NLPC u is not uniquely defined, since $v = g(u)$ for any invertible function g would give the same MSE and NLPCA approximation. Hence u is also a source of randomness for I and H. As we have restrained u by adding normalization conditions in the cost function (A.2), the randomness introduced by u in I and H does not appear to affect their effectiveness in practice.

Note that as the inconsistency term only prevents overfitting arising from the broad data geometry, validation data were still needed to prevent "local" overfitting from excessive number of model parameters, since H, unlike (3), does not contain a complexity term.

A test problem was set up in [18]: For a random number t uniformly distributed in the interval $(-1, 1)$, the signal $\mathbf{x}^{(s)}$ was generated by using a quadratic relation

$$x_1^{(s)} = t, \qquad x_2^{(s)} = \frac{1}{2}\, t^2. \tag{6}$$

Isotropic Gaussian noise (with variance being one half the average variance of $x_1^{(s)}$ and $x_2^{(s)}$) was then added to the signal $\mathbf{x}^{(s)}$ to give the noisy data $\mathbf{x}$ with 500 observations. NLPCA was performed on the data using the network in Fig. 1a with $m = 4$ (m being the number of hidden neurons in the first and in the third hidden layers of the NN) and with the weight penalty parameter P at various values (10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0). For each value of P, the model training was done 30 times starting from random initial weights, and model runs where the MSE evaluated over the validation data was larger than the MSE over the training data were deemed ineligible. In the traditional approach, among the eligible runs over the range of P values, the one with the lowest MSE over all (training and validation) data was selected as the best. Figure 4a shows this solution where the zigzag curve retrieved by NLPCA is very different from the theoretical parabolic signal (6), demonstrating the pitfall of selecting the lowest MSE run.

In contrast, in Fig. 4b, among the eligible runs over the range of P values, the one with the lowest information criterion H was selected. This solution,

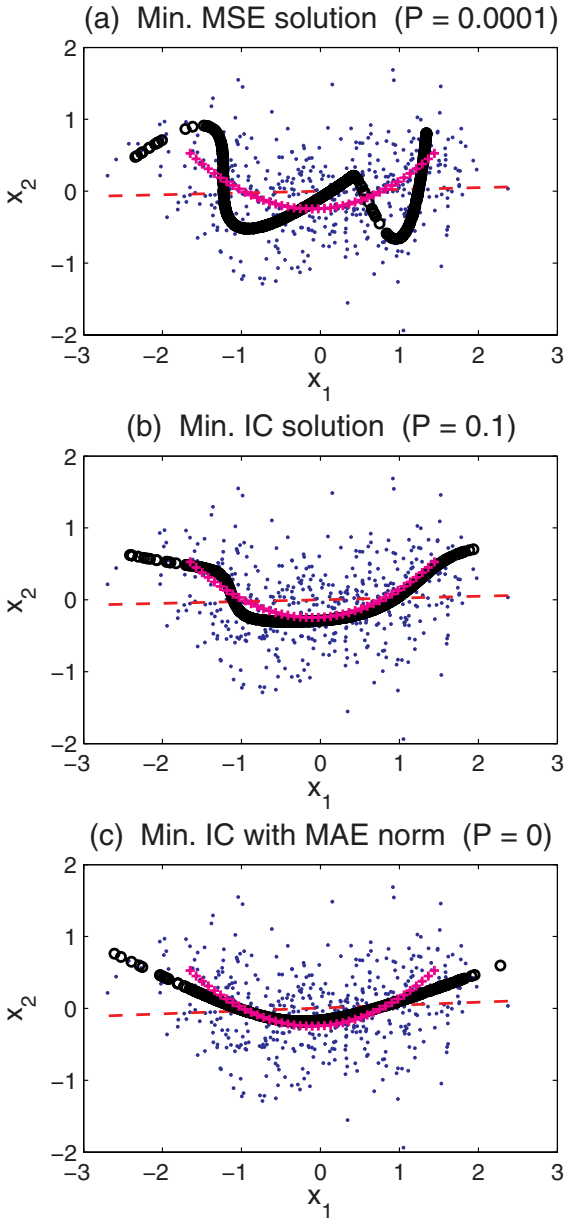


Fig. 4: The NLPCA solution (shown as densely overlapping *black circles*) for the synthetic dataset (*dots*), with the parabolic signal curve indicated by "+" and the linear PCA solution by the dashed line. The solution was selected from the multiple runs over a range of P values based on (**a**) minimum MSE, (**b**) minimum IC H, and (**c**) minimum IC together with the MAE norm.

which has a much larger weight penalty ($P = 0.1$) than that in Fig. 4a ($P = 10^{-4}$), shows less wiggly behaviour and better agreement with the theoretical parabolic signal.

Even less wiggly solutions can be obtained by changing the error norm used in the cost function from the mean square error to the mean absolute error (MAE), i.e. replacing $\langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle$ by $\langle \sum_j |x_j - x'_j| \rangle$ in Eq. (A.2). The MAE norm is known to be robust to outliers in the data (p. 210 of [1]). Fig. 4c is the solution selected based on minimum H with the MAE norm used. While wiggles are eliminated, the solution underestimates the curvature in the parabolic signal.

The H IC approach was also tested on a real climate dataset [18], namely the tropical Pacific sea surface temperature (SST), where the interannual variability is dominated by the El Niño-Southern Oscillation (ENSO) phenomenon [24]. The monthly SST anomalies (1948–2005) were obtained by removing the climatological seasonal cycle (i.e. subtracting from each monthly SST value the climatological mean value for that month). (NLPCA can be performed even if the climatological seasonal cycle is not removed, as was done in [4].) The 7 leading principal components (PC) containing 86.5% of the variance were retained, and served as the inputs for the NLPCA model.

NLPCA was performed over a range of m and P values ($m = 2, \ldots, 6$, and $P = 10$, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0). For each combination of m and P, 100 runs starting from random inital weights were made. Among all the runs made over the whole range of m and P values, the one with the lowest H was selected as the best (with $m = 5$, $P = 0$). The NLPCA solution is a curve in the 7-dimensional PC space.

We next compare the best solutions found for different hidden neuron number m. Since the overall best solution based on minimum H was for $m = 5$, we also showed the best solution found for $m = 2$ and $m = 6$ in the PC1-PC2 plane (Fig. 5). The (normalized) MSE for the 3 solutions in Fig. 5 are 0.898 ($m = 2$), 0.857 ($m = 5$) and 0.826 ($m = 6$), where for easy comparison with the linear mode, the values for the NLPCA solution have been divided by that from the PCA mode 1. For the (normalized) inconsistency index I, the values are 0.896 ($m = 2$), 0.879 ($m = 5$) and 0.946 ($m = 6$), while for the (normalized) H IC, the corresponding values are 0.804, 0.753 and 0.782 respectively. Hence the $m = 6$ solution has a lower MSE than the $m = 5$ solution, but the increased inconsistency from its wiggly curve (Fig. 5c) led to a larger I and a larger H. Compared to the $m = 2$ solution, the $m = 5$ solution has both lower MSE and lower I.

The tropical Pacific SST example illustrates that with a complicated oscillation like the El Niño-La Niña phenomenon, using a linear method such as PCA results in the nonlinear mode being scattered into several linear orthogonal modes (in fact, all 3 leading PCA modes are related to this phenomenon) [4]. This brings to mind the famous parable of the three blind men and their disparate descriptions of an elephant – hence the importance of the NLPCA as a unifier of the separate linear modes. In the study of climate variability,

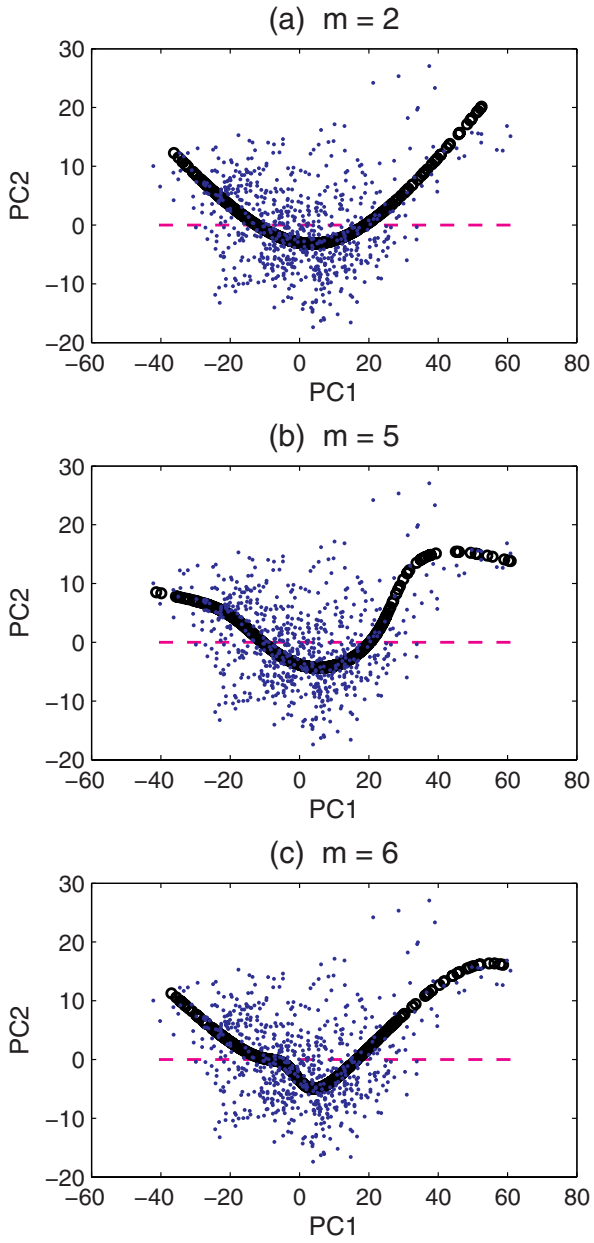


Fig. 5: The best NLPCA mode 1 solution (for the SST anomaly data) selected for (**a**) $m = 2$ (**b**) $m = 5$ and (**c**) $m = 6$. The solution is shown only in the PC1-PC2 plane, with the linear PCA mode 1 solution indicated by the dashed line. The warm El Niño episodes are represented by *dots* in the upper right corner and the cool La Niña episodes, in the upper left corner.

the wide use of PCA methods has created the not entirely accurate view that our climate is dominated by a number of spatially fixed oscillatory patterns, which is in fact due to the limitation of the linear method. Applying NLPCA to the tropical Pacific SSTA, one finds no spatially fixed oscillatory patterns, but an oscillation evolving in space as well as in time [2].

While the NLPCA is capable of finding a continuous open curve solution, there are many phenomena involving waves or quasi-periodic fluctuations, which call for a continuous closed curve solution. Reference [25] introduced an NLPCA with a circular node at the network bottleneck [henceforth referred to as the NLPCA(cir)], so that NLPCA(cir) is capable of approximating the data by a closed continuous curve. Fig. 1b shows the NLPCA(cir) network, which is identical to the NLPCA of Fig. 1a except at the bottleneck, where there are now two neurons p and q constrained to lie on a unit circle in the p-q plane, so there is effectively only one free angular variable θ, the NLPC (see Appendix A). This network has also been used to perform nonlinear singular spectrum analysis [26].

Although NLPCA(cir) is designed for extracting closed curve solutions, it is also capable of extracting an open curve solution. The reason is that if the input data mapped onto the p-q plane covers only a segment of the unit circle instead of the whole circle, then the inverse mapping from the p-q space to the output space will yield a solution resembling an open curve. Hence, NLPCA(cir) may extract either a closed curve or an open curve approximation to a dataset. The IC H not only alleviates overfitting in open curve solution, but also chooses between open and closed curve solutions. The inconsistency index and the IC are now obtained from

$$I = 1 - \frac{1}{2}\left[C(p,\tilde{p}) + C(q,\tilde{q})\right], \quad \text{and} \quad H = \text{MSE} \times I\,, \tag{7}$$

where p and q are from the bottleneck (Fig. 1b), and $\tilde{p}$ and $\tilde{q}$ are the corresponding nearest neighbour values.

For a test problem, consider a Gaussian data cloud (with 500 observations) in 2-dimensional space, where the standard deviation along the x_1 axis was double that along the x_2 axis. The data set was analyzed by the NLPCA(cir) model with $m = 2, \ldots, 5$ and $P = 10$, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0. From all the runs, the solution selected based on the minimum MSE has $m = 5$ (and $P = 10^{-5}$) (Fig. 6a), while that selected based on minimum H has $m = 3$ (and $P = 10^{-5}$) (Fig. 6b). The minimum MSE solution has (normalized) MSE = 0.370, $I = 9.50$ and $H = 3.52$, whereas the minimum H solution has the corresponding values of 0.994, 0.839 and 0.833, respectively. Thus the IC correctly selected a nonlinear solution (Fig. 6b) which is similar to the linear solution. (Due to finite sample size, the curve solution does not exactly match the straight line, which would require infinite sample size). The IC also rejected the closed curve solution of Fig. 6a, in favour of the open curve solution of Fig. 6b, despite its much larger MSE.

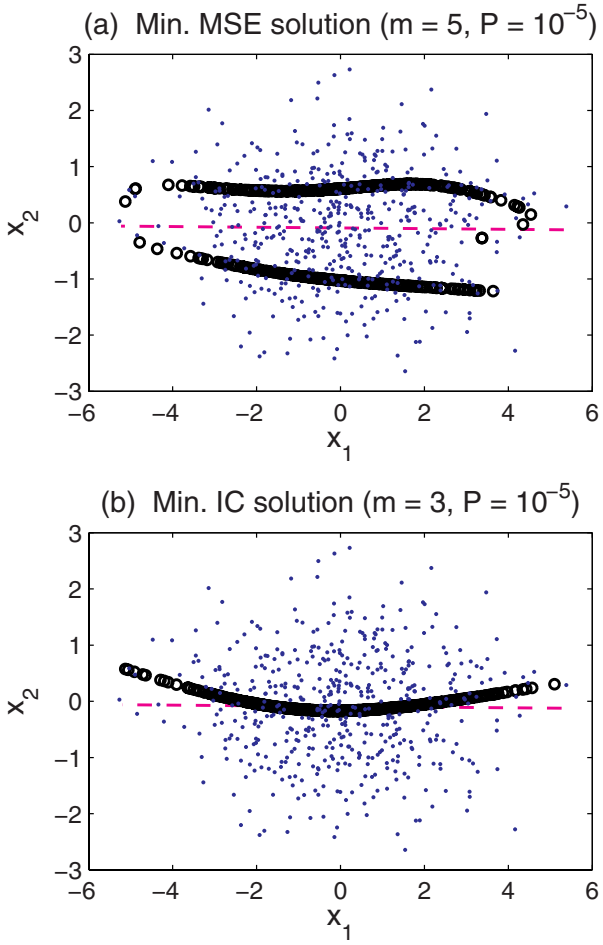


Fig. 6: The NLPCA(cir) mode 1 for a Gaussian dataset, with the solution selected based on (**a**) minimum MSE and (**b**) minimum IC. The PCA mode 1 solution is shown as a *dashed line*.

The IC also performed well when there is a strongly nonlinear signal in the noisy data, as demonstrated in Fig. 9 of [18]. For real data, the method was applied successfully to the quasi-biennial oscillation (QBO) in the equatorial stratospheric wind [18]. In summary, the application of the IC to NLPCA and NLPCA(cir) has been successful in model selection, i.e. choosing the best model which neither overfits nor underfits the data.

3 Robust NLCCA

NLCCA, the nonlinear analog of CCA outlined in Appendix B, is able to describe coupled nonlinear variability between two multivariate datasets. It has been used to study the nonlinear relation in the tropical Pacific between the sea level pressure (SLP) field and the SST field [27], and between the wind stress and the SST [28], as well as the nonlinear relation between the tropical Pacific SST and the extratropical atmospheric variability [29].

Due to its complicated architecture, NLCCA is prone to overfitting, particularly when applied to the short, noisy datasets common in climate studies. We explore the use of robust cost functions as a means of improving the performance of NLCCA. The basic model architecture is kept intact. Instead, the cost functions used to set the model parameters are replaced with more robust versions. A cost function based on the biweight midcorrelation [30] replaces one based on the Pearson correlation, and cost functions based on MAE can be used to replace ones based on MSE.

The Pearson correlation is not a robust measure of association between two variables, as its estimates can be affected by the presence of a single outlier [30]. For short, noisy datasets the cost function J_1 [Eq. (B.3) in Appendix B] using the Pearson correlation ($\mathrm{cor}(u, v)$) may lead to overfitting in NLCCA. For instance, when applying NLCCA to detect the relation between the tropical Pacific SLP and SST, [31] found that a spurious correlation of 1.00 was obtained by NLCCA. In this case, both the SLP and SST data contained the very strong El Niño signal during 1997–1998, the strongest El Niño in the data record from 1948–2003. The double-barreled NN on the left hand side of Fig. 2 then used strongly nonlinear mapping functions to produce canonical variates u and v with extremely large magnitude during 1997–1998, leading to the spuriously high value of $\mathrm{cor}(u, v) = 1.00$. In contrast, the correlation obtained after excluding the (u, v) values during 1997–1998 was only 0.28.

Robust correlation coefficients, including the Spearman rank correlation and the biweight midcorrelation, are reviewed by [30]. After testing both the Spearman correlation and the biweight midcorrelation, [31] proposed replacing the non-robust Pearson correlation by the robust biweight midcorrelation "bicor" as defined by Eq. (B.12), i.e. the cost function J_1 for the NN on the left hand side of Fig. 2 has $\mathrm{cor}(u, v)$ in (B.3) replaced by $\mathrm{bicor}(u, v)$. For the two NNs on the right hand side of Fig. 2, one further has the option of replacing the L_2 error norm with the robust L_1 norm in (B.6) and (B.7), i.e. replacing the MSE by the MAE in these cost functions.

Reference [31] used the synthetic test problem of [12] to compare the performance of the robust and non-robust versions of NLCCA. The synthetic data contains two correlated modes plus noise. The first correlated mode ($\mathbf{x}$ and $\mathbf{y}$) is given by

$$x_1 = t - 0.3t^2, \quad x_2 = t + 0.3t^2, \quad x_3 = t^2, \tag{8}$$

$$y_1 = t^3, \quad y_2 = -t + 0.3t^3, \quad y_3 = t + 0.3t^2, \tag{9}$$

where t is a uniformly distributed random number in $[-1, 1]$. The second correlated mode ($\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$) is given by

$$\tilde{x}_1 = -s - 0.3s^2, \quad \tilde{x}_2 = s - 0.3s^3, \quad \tilde{x}_3 = -s^4, \tag{10}$$

$$\tilde{y}_1 = \mathrm{sech}(4s), \quad \tilde{y}_2 = s + 0.3s^3, \quad \tilde{y}_3 = s - 0.3s^2, \tag{11}$$

where s is a uniformly distributed random number in $[-1, 1]$. The shapes of the correlated modes are given in [12] and [31].

To test the performance of the NLCCA models, 50 training and test datasets, each with 500 observations, were randomly generated from Eqs. (8–11). The signal in each dataset was produced by adding the second mode to the first mode, with the variance of the second equal to one third that of the first. Normally distributed random noise with standard deviation equal to 50% of the signal standard deviation was added to the data. The variables were then standardized to zero mean and unit standard deviation.

NLCCA models with different combinations of the non-robust (cor and MSE) and robust (bicor and MAE) cost functions were developed on the training datasets and applied to the test datasets. All NNs had three neurons in their hidden-layers and were trained without weight penalty terms. To avoid local minima in the cost functions, each network in Fig. 2 was trained 30 times from different random initial weights and biases. The network with the lowest value of its associated cost function was then selected for use and applied to the test data.

Root MSE (RMSE) values between the first synthetic mode and the first mode extracted by NLCCA models with different combinations of non-robust and robust cost functions are shown in Fig. 7 for the 50 test datasets. On average, all models performed approximately the same, although, for the leading NLCCA mode of the $\mathbf{x}$ dataset, NLCCA with bicor/MSE cost functions yielded the lowest median RMSE (0.44), followed by NLCCA with bicor/MAE (0.45) and NLCCA with cor/MSE (0.45). NLCCA with cor/MAE performed worst with a median RMSE of 0.47. Median RMSE values and relative rankings of the models were the same for the leading NLCCA mode of the $\mathbf{y}$ dataset.

Of the four models, NLCCA with the robust cost functions (bicor/MAE) was the most stable. No trial yielded an RMSE in excess of the series standard deviation of one, with the maximum value under 0.6 for the $\mathbf{x}$ mode. The other models had at least one trial with an RMSE value greater than one, which is indicative of severe overfitting. Maximum values for the $\mathbf{x}$ mode ranged from 1.8 for NLCCA with bicor/MSE, to 47.4 for NLCCA with cor/MSE, and 49.6 for cor/MAE. NLCCA with bicor/MAE performed similarly for the $\mathbf{y}$ mode, although two trials with RMSE greater than 20 were found for NLCCA with bicor/MSE cost functions.

Overall, results for the synthetic dataset suggest that replacing the cor/MSE cost functions in NLCCA with bicor/MAE cost functions leads to a more sta-

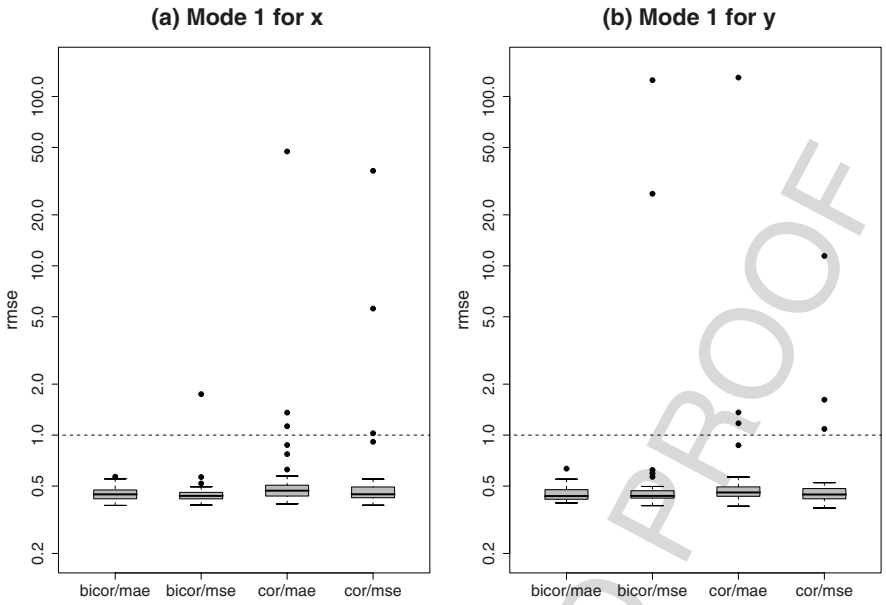


Fig. 7: Boxplots showing the distribution of RMSE between the first synthetic mode and the first mode extracted by NLCCA models for (**a**) **x** and (**b**) **y** with different combinations of non-robust and robust cost functions over 50 trials. Boxes extend from the 25th to 75th percentiles, with the line indicating the median. Whiskers represent the most extreme data within ± 1.5 times the interquartile range (i.e., the box height); values outside this range are plotted as dots. The dashed line indicates a RMSE equal to one. The ordinate is log-scaled to accomodate the large range in RMSE.

ble model that was less susceptible to overfitting and poor test performance. All models were run without weight penalty terms in this comparison. In practice, the non-robust models will need weight penalty terms to reduce overfitting, as is done in the next test.

Reference [31] applied NLCCA to tropical Pacific monthly SLP and SST data from 1948 to 2003. The climatological seasonal cycle was removed, data were detrended by removing the long-term linear trend, and a 3-month running mean filter was applied. After PCA, the first 6 SST PCs (accounting for 73% of the total SST variance) and the 6 SLP PCs (accounting for 80% of the variance) were retained for further analysis.

Three variants of the NLCCA model were applied to the SLP and SST datasets. The first, representing the standard NLCCA model, incorporated both non-robust cost functions (cor/MSE). The second and third used the bicor cost function to train the double-barreled network and either the MAE or MSE cost function to train the inverse mapping networks.

To assess the usefulness of the three variants of NLCCA for seasonal forecasting, models were validated on the basis of their forecast performance. PC scores from the SLP dataset were used to predict PC scores from the SST dataset at lead times of 0, 3, 6, 9, and 12-months. (Lead times are defined as the number of months from the predictor observation to the predictand observation, e.g., a forecast with a 3-month lead time from January would be for April.) Taking $\mathbf{x}$ to be historical values of the SLP PC scores and $\mathbf{y}$ to be historical values of the SST PC scores, forecasts for a new case $\mathbf{y}'$ were made as follows. First, the double-barreled network was trained with $\mathbf{x}$ and $\mathbf{y}$ as inputs and the resulting values of u and v were used to train the inverse mapping networks. Given a new SLP data point $\mathbf{x}$, a new value of the canonical variate u was obtained from the double-barreled network. Regression equations [e.g. Eq. (B.9)] were then used to predict a new value of v', which was entered into the appropriate inverse mapping network to give $\mathbf{y}'$. For the second and higher NLCCA modes, the same procedure was followed using residuals from the previous mode as inputs.

Following [27], NNs were trained both with and without weight penalty terms using two hidden neurons. To avoid overfitting in models trained with weight penalty, values of the coefficients P_1, P_2, and P_3 in Eqs. (B.3), (B.6) and (B.7) were determined via 10-fold cross-validation on the training dataset. The training record was split into 10 contiguous segments. Models were trained on 9 of the 10 segments using weight penalties from the set $\{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 0\}$. Forecasts on the remaining segment were then recorded for each weight penalty coefficient. While fixing the weight penalties, these steps were repeated 9 times, each time making forecasts on a different segment. Finally, forecasts for all 10 segments were combined and validated against observations. Weight penalties that minimized the aggregated cross-validation error were recorded, NNs were retrained using these penalties on all 10 segments combined. Ten models were trained in this manner to assess sensitivity to initial weights and biases.

A second round of cross-validation was used to estimate out-of-sample forecast performance of the models. The historical record was split into 5 segments (each approximately 11 years in length). Models were trained on 4 of the 5 segments using the cross-validation procedure outlined above. Forecasts on the remaining segment were then recorded. These steps were repeated 4 times, each time making forecasts on a different segment. Finally, forecasts for all 5 segments were combined and compared with observations.

Results from NLCCA models with one extracted mode are shown in Fig. 8. Cross-validated Pearson correlation skill is averaged over the entire tropical Pacific domain following reconstruction of the SST field from the predicted SST PCs. Results with weight penalty are only given for the NLCCA model with cor/MSE cost functions as the addition of penalty terms to models with the bicor cost function did not generally lead to significant changes in skill.

Without weight penalty, the NLCCA model with cor/MSE cost functions performed poorly, exhibiting mean skills worse than CCA at all lead times. Even with concurrent predictor/predictand fields, the mean correlation skill was lower than 0.2. NLCCA with bicor/MSE cost functions and bicor/MAE cost functions performed much better, with mean correlation skills exceeding 0.5 at the 0-month lead time. Over the 10 trials, minimum skills from models incorporating the bicor cost function were higher than maximum skills from the corresponding cor/MSE models without weight penalty.

For NLCCA with cor/MSE cost functions, minimum correlations were lower than zero (i.e., no cross-validation skill) for 6, 9, and 12-month lead times. All NLCCA models with bicor/MSE and bicor/MAE cost functions, even those at a 12-month lead time, showed positive skill. In general, NLCCA models with bicor exhibited the least variability in skill between repeated trials. In no case was the range between minimum and maximum skill greater than 0.2. For NLCCA with cor/MSE cost functions, the range in skill exceeded 0.2 at all lead times, indicating a very unstable model.

Little difference in skill was evident between bicor/MSE and bicor/MAE models, which suggests that the switch from cor to bicor in the double-barreled network cost function was responsible for most of the increase in skill relative to the standard NLCCA model.

Results discussed to this point have been for NLCCA models without weight penalty. Addition of weight penalty to the standard NLCCA model resulted in improvements in the mean correlation skill, although performance still lagged behind NLCCA with the bicor cost function at 9 and 12-month lead times. At 0, 3, and 6-month lead times, maximum skill over the 10 trials did, however, exceed the mean level of skill of the bicor-based models, which suggests that an appropriate amount of weight penalty can result in a good performing model. However, the wide range in performance over the 10 trials (e.g., at 0 and 6-month lead times) reflects the instability of the training and cross-validation steps needed to choose the weight penalty coefficients. In practice, it may be difficult to consistently reach the performance level of the robust model by relying solely on weight penalty to control overfitting of the standard NLCCA model.

Returning to the NLCCA models with bicor/MSE and bicor/MAE cost functions, little difference in skill between the models is apparent from Fig. 8. At short lead times (0 and 3-months), when the signal is strongest, the bicor/MSE model performed slightly better than the bicor/MAE model, whereas at the longest lead time (12-months), when the signal is weakest, the bicor/MAE model performed best (and with less variability among runs).

NLCCA models with the bicor/MSE and bicor/MAE cost functions tended to perform slightly better than CCA. For the bicor/MAE model, the small improvement in performance was significant (i.e., minimum skill over the 10 trials exceeded CCA skill) at 0, 3, 6, and 12-month lead times, while the same was true of the bicor/MSE model at 0 and 3-month lead times.

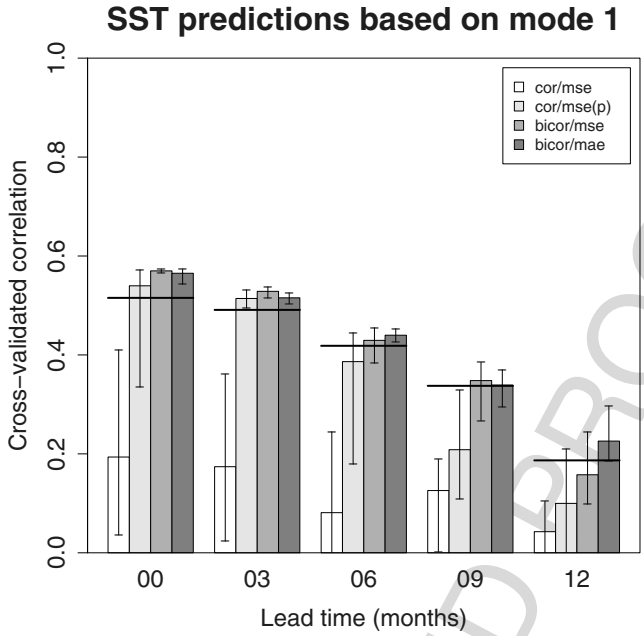


Fig. 8: Cross-validated correlation skill for NLCCA models trained with cor/MSE, bicor/MSE, and bicor/MAE cost functions. Weight penalty was applied to the model denoted cor/MSE(p). *Bars* show the mean correlation over the spatial domain, averaged over the 10 trials. *Vertical lines* extend from the minimum to the maximum spatial mean correlation from the 10 trials. *Horizontal lines* show correlation skill from the CCA model for comparison. The ordinate is limited to showing positive cross-validated skill.

4 Effects of Time-Averaging

In this section, we discuss one of two main factors undermining the advantage of nonlinear models over linear models. Time-averaging is widely used to reduce noise in the data; however, it also linearizes the relations in the dataset. In a study of the nonlinear relation between the precipitation rate (the predictand) and 10 other atmospheric variables (the predictors) in the NCEP/NCAR reanalysis data [32, 33] examined the daily, weekly and monthly averaged data by nonlinear multiple regression using NN over 3 regions (British Columbia, Canada, Middle East and northeastern China), and discovered that the strongly nonlinear relations found in the daily data became dramatically reduced by time-averaging to the almost linear relations found in the monthly data.

To explain this phenomenon, [33] invoked the well-known central limit theorem from statistics. For simplicity, consider the relation between two variables x and y. If $y = f(x)$ is a nonlinear function, then even if x is a normally distributed random variable, y will in general not have a normal distribution. Now consider the effects of time-averaging on the (x, y) data. The bivariate central limit theorem [34] says that if $(x_1, y_1), \ldots, (x_n, y_n)$ are independent and identically distributed random vectors with finite second moments, then (X, Y), obtained from averaging $(x_1, y_1), \ldots, (x_n, y_n)$, will, as $n \to \infty$, approach a bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, where μ_1 and μ_2 are the mean of X and Y, respectively, σ_1^2 and σ_2^2 are the corresponding variance, and ρ the correlation between X and Y.

From the bivariate normal distribution, the conditional probability distribution of Y (given X) is also a normal distribution [34], with mean

$$\mathrm{E}[Y|X] = \mu_2 + (X - \mu_1)\rho\sigma_2/\sigma_1. \tag{12}$$

This linear relation in X explains why time-averaging tends to linearize the relationship between the two variables. With more variables, the bivariate normal distribution readily generalizes to the multivariate normal distribution.

To visualize this effect, consider the synthetic dataset

$$y = x + x^2 + \epsilon, \tag{13}$$

where x is a Gaussian random variable with unit standard deviation and ϵ is Gaussian noise with a standard deviation of 0.5, so each day's value is independent of that of the next day. Averaging this "daily" data over 7 consecutive days and over 30 days reveals a dramatic weakening of the nonlinear relation (Fig. 9), and the shifting of the y density distribution towards Gaussian with the time-averaging. With real data, there is autocorrelation in the time series, so the monthly data will be effectively averaging over far fewer than 30 independent observations as done in this synthetic dataset. Seasonal data in the extratropics, however, will probably involve averaging about 30 independent observations.

If the data has strong autocorrelation, so that the integral time scale from the autocorrelation function is not small compared to the time-averaging window, then there are actually few independent observations used during the time-averaging, and the central limit theorem does not apply. For instance, the eastern equatorial Pacific sea surface temperatures have an integral time scale of about a year, hence nonlinear relations can be detected from monthly or seasonal data, as found by NLPCA and NLCCA. In contrast, the midlatitude weather variables have integral time scales of about 3–5 days, so monthly averaged data would have effectively averaged over about 6–10 independent observations, and seasonal data over 20–30 independent observations, so the influence of the central limit theorem cannot be ignored.

While time-averaging tends to reduce the nonlinear signal, it also smooths out the noise. Depending on the type of noise (and perhaps on the type of

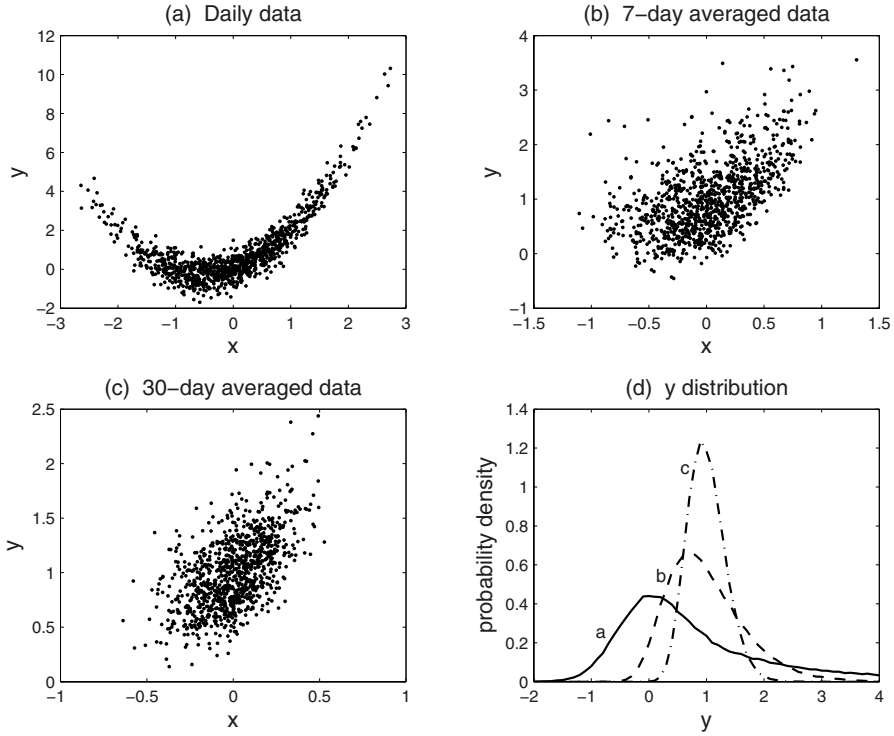


Fig. 9: Effects of time-averaging on a nonlinear relation. (**a**) Synthetic "daily" data from a quadratic relation between x and y, and the data time-averaged over (**b**) 7 observations and (**c**) 30 observations. The probability density distribution of y is shown in (**d**) for cases (**a**), (**b**) and (**c**).

nonlinear signal), it is possible that time-averaging may nevertheless enhance the detection of a nonlinear signal above the noise for some datasets. In short, researchers should be aware that time-averaging could have a major impact on our modelling or detection of nonlinear empirical relations.

5 Extrapolation

We next discuss another factor which undermines the advantage of nonlinear models over linear models, namely extrapolation. For NLR problems, NN models (with proper weight penalty so there is neither overfitting nor underfitting) perform nonlinear interpolation well. However when presented with new data where the predictor lies beyond the range of (predictor) values used in model training, the NN model is then *extrapolating* instead of interpolating. We will illustrate the extrapolation behaviour with a simple test problem.

Let the signal be

$$y = x + \frac{1}{5}x^2. \tag{14}$$

We choose 300 observations, with x having unit standard deviation and Gaussian distribution, and y given by (14) plus Gaussian noise (with the noise standard deviation the same as the signal standard deviation). With 6 hidden neurons, the Bayesian NN model from the MATLAB Neural Network Toolbox was used to solve this NLR problem. In Fig. 10a, upon comparing with the true signal (dashed curve), it is clear that the NN model interpolated better than the LR model (solid line), but for large x values, NN extrapolated worse than LR. Fig. 10b shows the same data fitted by a fourth order polynomial, where for strongly negative x values, the polynomial extrapolated worse than LR. Hence nonlinear models which interpolate better than LR provide no guarantee that they extrapolate better than LR. In fact, Wu et al. [35] found that for seasonal forecasting of the North American surface air temperature, the NN model extrapolated worse than LR.

How the nonlinear model extrapolates is dependent on the type of nonlinear model used. With a polynomial fit, as $|x| \to \infty$, $|y| \to \infty$. However, for NN models (with 1 hidden layer $\mathbf{h}$), where the kth hidden neuron

$$h_k = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b}_k), \quad y = \widetilde{\mathbf{w}} \cdot \mathbf{h} + \widetilde{b}\,, \tag{15}$$

once the model has been trained, then as $\|\mathbf{x}\| \to \infty$, the tanh function remains bounded within ± 1, hence y remains bounded – in sharp contrast to the unbounded behaviour with polynomial extrapolation (Fig. 10).

6 Summary and Conclusion

The nonlinear generalization of classical multivariate statistical methods by machine learning methods such as NN is exciting. However, when applied to environmental sciences, the datasets may be very noisy and/or contain relatively few independent observations, and the nonlinear methods may fail. Thus it is essential that more robust nonlinear methods be developed.

With noisy data, not having plentiful observations could cause a flexible nonlinear model to overfit. In the limit of infinite sample size, overfitting cannot occur in nonlinear regression, but can still occur in NLPCA due to the geometric shape of the data distribution. A new inconsistency index I for detecting the projection of neighbouring points to distant parts of the NLPCA curve has been introduced, and incorporated into a holistic IC H to choose the model with the appropriate weight penalty parameter and the appropriate number of hidden neurons [18]. Tests with synthetic data and real climate data indicated that this IC is effective in model selection, and in deciding between open curve and closed curve solutions.

To make NLCCA more robust, non-robust cost functions in the model are replaced by robust cost functions – the Pearson correlation is replaced by

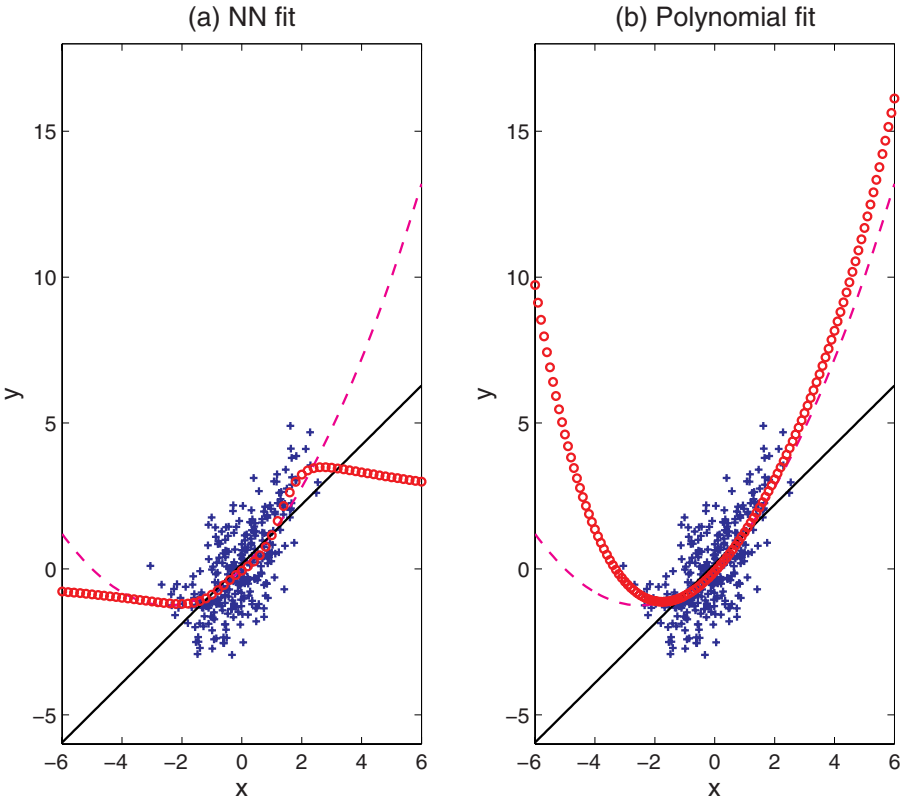


Fig. 10: (**a**) Nonlinear regression fit by Bayesian NN. The data are indicated by crosses and the Bayesian NN solution by *circles*. *Dashed curve* indicates the theoretical signal and *solid line* the LR solution. (**b**) Fit to the same dataset but by a fourth order polynomial.

the biweight midcorrelation, while the MSE in the inverse mapping network can be replaced by the MAE [31]. Tests showed that replacing the Pearson correlation by the biweight midcorrelation greatly improved the stability of the NLCCA model. In contrast, the choice between the MSE or MAE cost function appears to be more problem dependent, and should be considered as part of the model selection process. The MATLAB codes for NLPCA and NLCCA are downloadable from `http://www.ocgy.ubc.ca/projects/clim.pred/download.html`.

Two common causes undermining nonlinear models relative to linear models have also been highlighted. (1) Time-averaging of data (e.g. from daily data to seasonal data) linearizes the relation between predictor and predictand due to the central limit theorem [33]. (2) When new predictor data lies outside the training range, the nonlinear model may extrapolate poorly, thereby un-

dermining the forecast capability of the nonlinear model [35]. How to improve the nonlinear model forecast at extrapolation points is currently being investigated.

Acknowledgements. W. Hsieh has been supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada, and a project grant from the Canadian Foundation for Climate and Atmospheric Sciences.

Appendix A: NLPCA

With the input variables forming the 0th layer of the network in Fig. 1a, a neuron $v_j^{(i)}$ at the ith layer ($i = 1, 2, 3, 4$) receives its value from the neurons $\mathbf{v}^{(i-1)}$ in the preceding layer, i.e.

$$v_j^{(i)} = f^{(i)}(\mathbf{w}_j^{(i)} \cdot \mathbf{v}^{(i-1)} + b_j^{(i)}), \tag{A.1}$$

where $\mathbf{w}_j^{(i)}$ is a vector of weight parameters and $b_j^{(i)}$ a bias or offset parameter, and the transfer or activation functions $f^{(1)}$ and $f^{(3)}$ are the hyperbolic tangent functions, while $f^{(2)}$ and $f^{(4)}$ are simply the identity functions. Effectively, a nonlinear function $u = F(\mathbf{x})$ maps from the higher dimension input space to the lower dimension bottleneck space, followed by an inverse transform $\mathbf{x}' = \mathbf{G}(u)$ mapping from the bottleneck space back to the original space, as represented by the outputs. To make the outputs as close to the inputs as possible, the cost function J, basically the MSE, is minimized. More precisely, [2] used

$$J = \langle \|\mathbf{x} - \mathbf{x}'\|^2 \rangle + \langle u \rangle^2 + (\langle u^2 \rangle - 1)^2 + P \sum_j \|\mathbf{w}_j^{(1)}\|^2 , \tag{A.2}$$

where on the right hand side, the first term is the MSE (with $\langle \cdots \rangle$ denoting an observation or time mean), the second and third terms are for restraining u towards $\langle u \rangle = 0$ and $\langle u^2 \rangle = 1$, and the final term is a weight penalty or regularization term, with P the weight penalty parameter. [4] found that penalizing just the first layer of weights is sufficient to limit the nonlinear modelling capability of the model. By minimizing J, the values of the weight and bias parameters are solved. The nonlinear optimization was carried out by the quasi-Newton algorithm fminunc.m in the MATLAB Optimization Toolbox. A number of optimization runs was made with random initial values of the weight and bias parameters, and only runs where the MSE evaluated over the validation data was not larger than the MSE over the training data were deemed eligible, with the best solution selected as the one with the smallest value of H, calculated from (5).

To obtain closed curve solutions, we use NLPCA with a circular bottleneck node (Fig. 1b). At the bottleneck, we first calculate the pre-states p_o and q_o by

$$p_o = \mathbf{w}_1^{(2)} \cdot \mathbf{v}^{(1)} + b_1^{(2)}\,, \quad \text{and} \quad q_o = \mathbf{w}_2^{(2)} \cdot \mathbf{v}^{(1)} + b_2^{(2)}\,, \tag{A.3}$$

then with

$$r = (p_o^2 + q_o^2)^{1/2}\,, \tag{A.4}$$

the circular node is defined by

$$p = p_o/r\,, \quad \text{and} \quad q = q_o/r\,, \tag{A.5}$$

which satisfies the unit circle equation $p^2 + q^2 = 1$. Thus, although there are two variables p and q at the bottleneck, there is only one angular degree of freedom (θ) from the circle constraint. For more details, see the review by [2]. The model run having the smallest H, as computed from (7), is selected as the best solution.

Appendix B: NLCCA

Consider a dataset $\{x_i(t)\}$ with i variables and another dataset $\{y_j(t)\}$ with j variables, where each dataset has $t = 1, ..., N$ observations. The variables $\{x_i(t)\}$ can be grouped to form the vector $\mathbf{x}(t)$ and the variables $\{y_j(t)\}$ can be grouped to form the vector $\mathbf{y}(t)$. CCA looks for the linear combinations

$$u(t) = \mathbf{a} \cdot \mathbf{x}(t), \quad v(t) = \mathbf{b} \cdot \mathbf{y}(t) \tag{B.1}$$

such that the Pearson correlation between the canonical variates u and v, i.e., cor(u, v), is maximized.

In NLCCA, the nonlinear analog of linear CCA, the linear mappings in Eq. (B.1) are replaced with nonlinear mappings performed by NN (Fig. 2). The double-barreled network on the left-hand side nonlinearly maps $\mathbf{x}$ to u and $\mathbf{y}$ to v by

$$\begin{aligned} h_k^{(x)} &= \tanh[(\mathbf{W}^{(x)}\mathbf{x} + \mathbf{b}^{(x)})_k], \quad u = \widetilde{\mathbf{w}}^{(x)} \cdot \mathbf{h}^{(x)} + \widetilde{b}^{(x)} \\ h_l^{(y)} &= \tanh[(\mathbf{W}^{(y)}\mathbf{y} + \mathbf{b}^{(y)})_l], \quad v = \widetilde{\mathbf{w}}^{(y)} \cdot \mathbf{h}^{(y)} + \widetilde{b}^{(y)} \end{aligned} \tag{B.2}$$

where $h_k^{(x)}$ and $h_l^{(y)}$ are the hidden-layer neurons; $\mathbf{W}^{(x)}$ and $\mathbf{W}^{(y)}$ are the hidden-layer weight matrices; $\mathbf{b}^{(x)}$ and $\mathbf{b}^{(y)}$ are the hidden-layer bias vectors; $\widetilde{\mathbf{w}}^{(x)}$ and $\widetilde{\mathbf{w}}^{(y)}$ are the output-layer weight vectors; $\widetilde{b}^{(x)}$ and $\widetilde{b}^{(y)}$ are the output-layer biases. The number of hidden-layer neurons controls the overall complexity of the network; the hidden-layer must contain more than one neuron ($2 \le k \le K$ and $2 \le l \le L$) to obtain a nonlinear solution [27].

Weight and bias parameters in the double-barreled network are obtained by minimizing the cost function

$$J_1 = -\mathrm{cor}(u,\, v) + \langle u \rangle^2 + \langle v \rangle^2 + \left(\langle u^2 \rangle^{\frac{1}{2}} - 1 \right)^2 + \left(\langle v^2 \rangle^{\frac{1}{2}} - 1 \right)^2 + P_1 \left[\sum_{ki} \left(W_{ki}^{(x)} \right)^2 + \sum_{lj} \left(W_{lj}^{(y)} \right)^2 \right]. \quad \text{(B.3)}$$

The first term maximizes the correlation between the canonical variates u and v; the second, third, fourth, and fifth terms are normalization constraints that force u and v to have zero mean and unit variance; the sixth term is a weight penalty whose relative magnitude is controlled by the parameter P_1. Larger values of P_1 lead to smaller weights (i.e., fewer effective model parameters), which results in a more linear model. If $\tanh(\cdot)$ is replaced by the identity function, then Eq. (B.2) reduces to Eq. (B.1) and the network performs linear CCA.

Once the canonical variates u and v have been found, the inverse mappings to $\mathbf{x}'$ and $\mathbf{y}'$ are given by the two NNs on the right-hand side of Fig. 2:

$$h_k^{(u)} = \tanh[(\mathbf{w}^{(u)} u + \mathbf{b}^{(u)})_k], \quad \mathbf{x}' = \widetilde{\mathbf{W}}^{(u)} \mathbf{h}^{(u)} + \widetilde{\mathbf{b}}^{(u)}, \quad \text{(B.4)}$$

$$h_l^{(v)} = \tanh[(\mathbf{w}^{(v)} v + \mathbf{b}^{(v)})_l], \quad \mathbf{y}' = \widetilde{\mathbf{W}}^{(v)} \mathbf{h}^{(v)} + \widetilde{\mathbf{b}}^{(v)}. \quad \text{(B.5)}$$

Weight and bias parameters in these two networks are found by minimizing the cost functions

$$J_2 = \left\langle \|\mathbf{x}' - \mathbf{x}\|^2 \right\rangle + P_2 \sum_k \left(w_k^{(u)} \right)^2, \quad \text{(B.6)}$$

$$J_3 = \left\langle \|\mathbf{y}' - \mathbf{y}\|^2 \right\rangle + P_3 \sum_l \left(w_l^{(v)} \right)^2, \quad \text{(B.7)}$$

respectively, where $\|\cdot\|^2$ is the square of the L_2-norm, with the L_p-norm given by

$$L_p(\mathbf{e}) = (\|\mathbf{e}\|^p)^{1/p} = \left(\sum_i |e_i|^p \right)^{1/p}. \quad \text{(B.8)}$$

J_2 and J_3 thus give the MSE between the model predictions and the observed $\mathbf{x}$ and $\mathbf{y}$ variables subject to weight penalty terms whose magnitudes are controlled by the parameters P_2 and P_3. Once the first mode has been extracted from the data, the next leading mode can be extracted from the model residuals, and so on for higher modes.

For seasonal climate prediction tasks, where the goal is to predict values of a multivariate predictand dataset from a multivariate predictor dataset, e.g., $\mathbf{y}' = \mathbf{f}(\mathbf{x})$, values of the canonical variate v' must be predicted from values of the canonical variate u. For canonical variates normalized to unit variance and zero mean, the linear least-squares regression solution is given by [36].

$$v' = u\, \mathrm{cor}(u,\, v) \quad \text{(B.9)}$$

For robust NLCCA, one needs to calculate the biweight midcorrelation: First rescale x and y as

$$p = \frac{x - M_x}{9\,\mathrm{MAD}_x}\,, \quad q = \frac{y - M_y}{9\,\mathrm{MAD}_y}\,, \tag{B.10}$$

where M_x and M_y are the median values of x and y respectively and MAD_x and MAD_y are the median values of $|x - M_x|$ and $|y - M_y|$ respectively. Next, the sample biweight midcovariance is given by

$$\mathrm{bicov}(x,\, y) = \frac{N \sum_t a(t)b(t)c(t)^2 d(t)^2 (x(t) - M_x)(y(t) - M_y)}{[\sum_t a(t)c(t)(1 - 5p(t)^2)]\,[\sum_t b(t)d(t)(1 - 5q(t)^2)]}\,, \tag{B.11}$$

where $a(t) = 1$ if $-1 \le p(t) \le 1$, otherwise $a(t) = 0$; $b(t) = 1$ if $-1 \le q(t) \le 1$, otherwise $b(t) = 0$; $c(t) = (1 - p(t)^2)$; and $d(t) = (1 - q(t)^2)$. The biweight midcorrelation is then given by

$$\mathrm{bicor}(x,\, y) = \frac{\mathrm{bicov}(x,\, y)}{\sqrt{\mathrm{bicov}(x,\, x)\,\mathrm{bicov}(y,\, y)}}\,. \tag{B.12}$$

The biweight midcorrelation, like the Pearson correlation, ranges from -1 to $+1$.

For robust NLCCA, $\mathrm{bicor}(u, v)$ replaces $\mathrm{cor}(u, v)$ in (B.3). One can also replace the L_2 norm with the robust L_1 norm in (B.6) and (B.7), i.e. replace the MSE by the MAE in these cost functions.

References

1. Bishop, C.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
2. Hsieh, W.: Nonlinear multivariate and time series analysis by neural network methods. Reviews of Geophysics **42** (2004) RG1003, doi:10.1029/2002RG000112
3. Kramer, M.: Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal **37** (1991) 233–243
4. Hsieh, W.: Nonlinear principal component analysis by neural networks. Tellus **53A** (2001) 599–615
5. Schölkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10** (1998) 1299–1319
6. Hastie, T., Stuetzle, W.: Principal curves. Journal of the American Statistical Association **84** (1989) 502–516
7. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290** (2000) 2323–2326
8. Tenenbaum, J., de Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–2323
9. Kohonen, T.: Self-organzing formation of topologically correct feature maps. Biological Cybernetics **43** (1982) 59–69

10. Kwok, J.T.Y., Tsang, I.W.H.: The pre-image problem in kernel methods. IEEE Transactions on Neural Networks **15** (2004) 1517–1525
11. Lai, P., Fyfe, C.: A neural implementation of canonical correlation analysis. Neural Networks **12** (1999) 1391–1397
12. Hsieh, W.: Nonlinear canonical correlation analysis by neural networks. Neural Networks **13** (2000) 1095–1105
13. Lai, P., Fyfe, F.: Kernel and non-linear canonical correlation analysis. International Journal of Neural Systems **10** (2000) 365–377
14. Suykens, J., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.: Least Squares Support Vector machines. World Scientific, New Jersey (2002)
15. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
16. MacKay, D.: Bayesian interpolation. Neural Computation **4** (1992) 415–447
17. Foresee, F., Hagan, M.: Gauss-Newton approximation to Bayesian regularization. In: Proceedings of the 1997 International Joint Conference on Neural Networks (1997)
18. Hsieh, W.: Nonlinear principal component analysis of noisy data. Neural Networks **20** (2007) 434–443
19. Christiansen, B.: The shortcomings of nonlinear principal component analysis in identifying circulation regimes. Journal of Climate **18(22)** (2005) 4814–4823
20. Christiansen, B.: Reply to Monahan and Fyfe's comment on "The shortcomings of nonlinear principal component analysis in identifying circulation regimes". Journal of Climate (2007) (in press)
21. Malthouse, E.: Limitations of nonlinear PCA as performed with generic neural networks. IEEE Transactions on Neural Networks **9** (1998) 165–173
22. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **AC-19** (1974) 716–723
23. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6** (1978) 461–464
24. Monahan, A.: Nonlinear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure. Journal of Climate **14** (2001) 219–233
25. Kirby, M., Miranda, R.: Circular nodes in neural networks. Neural Computation **8** (1996) 390–402
26. Hsieh, W., Wu, A.: Nonlinear multichannel singular spectrum analysis of the tropical Pacific climate variability using a neural network approach. Journal of Geophysical Research **107** (2002) DOI: 10.1029/2001JC000957
27. Hsieh, W.: Nonlinear canonical correlation analysis of the tropical Pacific climate variability using a neural network approach. Journal of Climate **14** (2001) 2528–2539
28. Wu, A., Hsieh, W.: Nonlinear canonical correlation analysis of the tropical Pacific wind stress and sea surface temperature. Climate Dynamics **19** (2002) 713–722. DOI:10.1007/s00382–002–0262–8
29. Wu, A., Hsieh, W., Zwiers, F.: Nonlinear modes of North American winter climate variability detected from a general circulation model. Journal of Climate **16** (2003) 2325–2339
30. Wilcox, R.: Robust Estimation and Hypothesis Testing. Elsevier, Amsterdam (2004)

AU: Please provide location for Ref. [17]

AU: Please update Ref. [20].

31. Cannon, A., Hsieh, W.: Robust nonlinear canonical correlation analysis: Application to seasonal climate forecasting. (Nonlinear Processes in Geophysics (submitted))
32. Kalnay, E., et al.: The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society **77** (1996) 437–471
33. Yuval, Hsieh, W.: The impact of time-averaging on the detectability of nonlinear empirical relations. Quarterly Journal of the Royal Meterological Society **128** (2002) 1609–1622
34. Bickel, P., Doksum, K.: Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day, Oakland (1977)
35. Wu, A., Hsieh, W., Cannon, A., Shabbar, A.: Improving neural network predictions of North American seasonal climate by outlier correction. (Nonlinear Processes in Geophysics (submitted))
36. von Storch, H., Zwiers, F.: Statistical Analysis in Climate Research. Cambridge University Press, Cambridge (1999)

AU: Please update Refs. [31] and [35]

Complexity of Spatio-Temporal Correlations in Japanese Air Temperature Records

Reik Donner[1,3], Takahide Sakamoto[2,3], and Noboru Tanizuka[3]

[1] Institute for Transport and Economics, Dresden University of Technology, Andreas-Schubert-Str. 23, 01062 Dresden, Germany, donner@vwi.tu-dresden.de
[2] Electrical Engineering Course, Osaka Municipal Ikuno Technical High School, 2-3-66 Ikuno-higashi, Ikuno-ku, Osaka-shi, 544-0025 Japan, taka_sakamoto@mem.iee.or.jp
[3] Graduate School of Science, Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai-shi, 599-8531 Japan, tanizuka@mi.s.osakafu-u.ac.jp

Abstract. The variability of meteorological observables is known to crucially depend on the geographical conditions and the considered spatial as well as temporal scales. In this contribution, we explicitly take the spatial dimension into account. Recent studies on this aspect have considered individual investigations of spatially distributed records from different stations, which form network structures with interesting statistical properties. However, the results of such studies are strongly influenced by the preprocessing of the time series, in the case of temperature records particularly by the applied deseasonalisation strategy. As a complementary approach, we investigate whether the interdependences between pairs of meteorological records can be used to extract additional information about the regularity of temporal variations of the regional climate and its potential change with time. As an alternative to the consideration of univariate estimates of fractal dimensions, the concept of multivariate dimension estimates is introduced. Different quantitative measures for the complexity of linear correlations are introduced and thoroughly compared. After studying the results for stationary model systems, our approach is used to characterise the variability of temperature records from 13 Japanese meteorological stations. The complexity of the complete record varies on an annual period with a larger complexity during the summer season, which is possibly related to the action of the East Asian monsoonal circulation.

Keywords: Temperature records, Spatio-temporal correlations, Multivariate dimension estimates, Variability, Japan

1 Introduction

The climate of the Earth is a high-dimensional complex system which is subjected to different global and local forcings and nonlinear internal feedback mechanisms that act on very different temporal as well as spatial scales.

Therefore, its behaviour is chaotic [1, 2, 3, 4, 5, 6] and thus characterised by a strong sensitivity with respect to relatively small changes of certain environmental parameters. Such changes are known to be able to lead to sudden transitions in the dynamics of the entire system, with the possible breakdown of the North Atlantic thermo-haline circulation as the probably best studied example [7, 8]. Time series recording the variability of climatological observables are therefore often characterised by a very strong and irregular variability and rather high levels of observational as well as "dynamical" noise. This holds in particular for the case of meteorological data obtained from either direct measurements since the start of the instrumental period or reconstructions of earlier time intervals. Moreover, the variability of the corresponding observables in both measurements and climate models often shows properties like non-Gaussian probability distribution functions or long-term persistence.

Atmospheric patterns are characterised by scales in both time and space on which some meteorological quantities like temperature or air pressure vary only weakly. If one analyses the temporal evolution of such variables at different locations influenced by the same pattern, it is therefore likely that the corresponding time series are more or less strongly correlated, with a maximum correlation at a time lag corresponding to the spatial distance between the sites and the typical velocity with which the pattern moves in space. Due to the dynamic evolution of the observed structures during their spatial motion, the strength of correlations between records decays with increasing distance between the considered locations. This statement holds in general for very different spatial as well as temporal scales:

- On a global scale, the interrelationships between sea-level pressure records obtained from reanalysis data have been utilised to derive a network-like structure [9]. Similar features are likely to be found in simulations of climate models as well. However, the behaviour of such models is known to differ from reanalysis data not only in terms of absolute variabilities and correlations, but also with respect to their non-linear features like the local predictability [10].
- On continental scales (i.e., several hundreds to thousands of kilometers), simple linear cross-correlation functions may (depending on the particular geographic situation) not necessarily be an optimal measure for describing the interrelationships and exactly detecting the delay corresponding to the maximum dependence between meteorological time series. As an alternative, one may consider different other measures which are sensitive to nonlinear correlations. Recent results on noisy electrophysiological data demonstrated that a maximisation of the linear spectral coherence may also be well-suited for an appropriate detection of delays between different signals [11]. For temperature and precipitation records, Rybski et al. [12] have suggested that the concept of phase synchronisation [13] may also be applied for detecting these time lags. Whereas this suggestion is underlined by inspections on model systems

comparing the phase synchronisation and correlation function approaches [14], it is likely that the concept of synchronisation analysis may lead to erroneous results if there is no well-defined high-frequency oscillatory component in the time-series [15, 16, 17].

In this chapter, we analyse long-term daily maximum, minimum, and mean air temperatures from different meteorological stations distributed all over Japan, which cover the last 30 years. Complementarily to other studies focussing exclusively on the temporal characteristics of such records, we use the entire multivariate data set to study the temporally varying complexity of the spatial correlations. For this purpose, the novel concept of multivariate dimension estimates is introduced and thoroughly applied to our data. In addition, we study the mutual spatio-temporal interdependences between the individual records in order to identify the dynamic skeleton of the Japanese temperature network. According to these aims, this manuscript is organised as follows: In Sect. 2, we briefly summarise arguments for an appropriate preprocessing of the data, which in particular reflects the problem of deseasonalisation. The approach of multivariate dimension estimates is introduced in Sect. 3. Finally, in Sect. 4, we apply different methods for uni- as well as multivariate assessments of the spatio-temporal correlations of air temperatures in Japan. Possible implications of our results for the understanding of nonlinear interactions between different components of the climate system are discussed.

2 Preprocessing and Data Analysis

In the case of temperature records, there is the conceptual problem that the relevant dynamics occur on at least two very different time scales. Besides the daily fluctuations on which we would like to focus in this contribution, there is the annual cycle that dominates the variation amplitude on longer time scales [18]. In order to analyse short-term correlation features of time series of air temperatures, it is therefore necessary to apply a sophisticated preprocessing of the data which removes the annual cycle component as good as possible and leaves the short-term variability unchanged.

In order to separate the dynamics on different time scales, a variety of approaches can be found in the literature [19]. For example, in order to remove long-term trends in the case of temperature records, it is convenient to firstly apply a long-term moving average filter (with a width of ≥ 1 year) to the time series which extracts such trend components. If the residual is subtracted from the original time series, the remaining signal (including the annual cycle component) remains almost invariant.

For the problem of deseasonalisation, i.e., the removal of periodic long-term components from a time series, there are several statistical methods which may be rougly distinguished into the following groups:

- Heuristic methods: In this case, the data are considered separately for the respective calendar dates. For each day of the year, mean values and eventually also higher-order moments are computed. By standardising the original data for every calender day according to these statistical quantities, one approaches a deseasonalised series. For example, the subtraction of the respective mean values for every calendar day corresponds to the so-called phase averaging method. It has to be noted that the heuristic methods do not yield a perfect deseasonalisation, as the amplitude of the annual component may change on both subannual and interannual time scales [20].
- Nonparametric smoothing methods: Like the heuristic methods, this type does not assume any functional form of the annual cycle. Two different subtypes may be distiguished: Methods where the exact cycle length is not taken into account include the traditional unweighted as well as weighted moving average filters, whose bandwidth has to be significantly smaller than the annual period. However, in this case, the temporal correlations may be essentially changed. As an alternative, one may also consider filters which explicitly refer to the length of the seasonal period [21, 22].
- Spectral methods: In contrast to the heuristic and non-parametric smoothing approaches, spectral methods implicitly assume a particular shape of the seasonal cycle. In the most convenient case of a high-low-pass filter based on a Fourier transform, a harmonic function with a period of $t_0 = 365.25$ days is fitted to the time series by linear regression (i.e., by minimising the quadratic residual between model and observations) and then removed from the record. However, spectral methods assume a particular shape of the periodic function, which is not necessarily present in natural signals. As an alternative, one may use a wavelet decomposition of the record and reconstruct a signal by taking only the components into account which vary on time scales that are significantly shorter than the annual cycle.
- Empirical mode decomposition (EMD): The concept of empirical mode decomposition has been suggested as a novel tool for time-scale separation in nonstationary systems [23]. By this heuristic algorithmic approach, the time series is successively decomposed into components which vary on significantly different time scales. According to this, the corresponding approach might be helpful to filter annual components as well as interannual long-term trends from meteorological data [24, 25, 26, 27]. However, in the case of EMD, it is not a priori clear that the extracted long-term components do not contain residual information from shorter time scales and vice versa, such that additional testing is required before applying this method as a standard tool for deseasonalisation in climatological time series [28].

In this contribution, we will not perform a detailed evaluation of the different mentioned approaches. However, in order to briefly illustrate the

importance of initial deseasonalisation for the outcome of both, linear and nonlinear methods of time-series analysis [29], Fig. 1 shows the linear auto-correlation function

$$C_X(\tau) = \frac{\frac{1}{M-1-\tau}\sum_{i=1}^{M-\tau}(X_{i+\tau}-\bar{X})(X_i-\bar{X})}{\frac{1}{M-1}\sum_{i=1}^{M}(X_i-\bar{X})^2} \quad \text{with} \quad \bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i \quad (1)$$

and the non-linear mutual information

$$I_X(\tau) = \sum_{s,s'=(1)}^{(S)} \frac{P(X_{i+\tau} \in s, X_i \in s')}{P(X_{i+\tau} \in s)\ P(X_i \in s')} \quad (2)$$

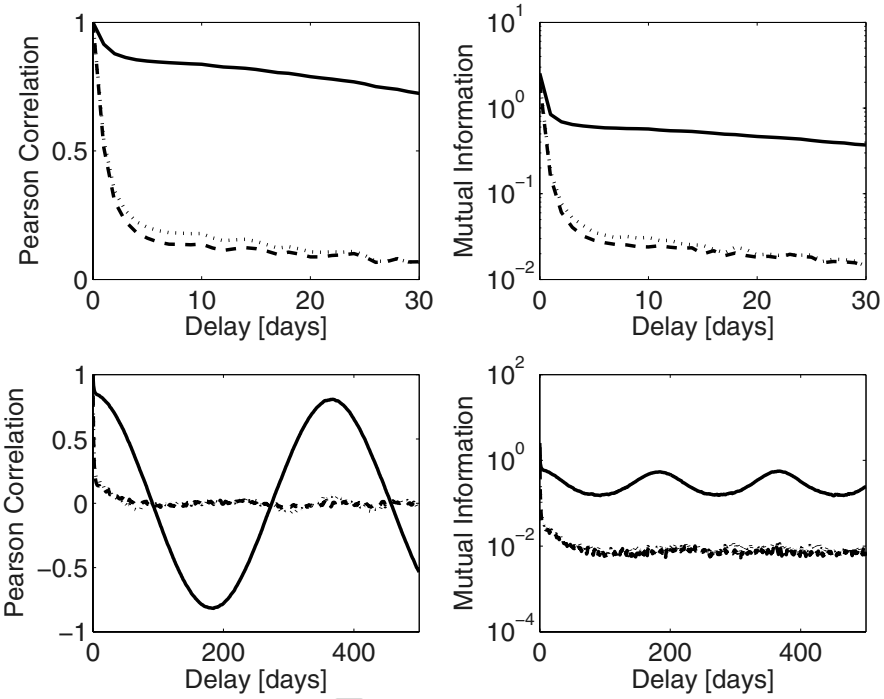


Fig. 1: Standard Pearson (auto-) correlation function (*left*) and mutual information (*right*) of the daily maximum air temperature record from Tokyo, Japan, between January 1975 and December 2005. The *upper* and *lower panels* show the same functions on different time-scales. Different line styles correspond to the original data (*solid*, after subtraction of the yearly mean obtained with a running one-year moving-average filter), the residual time series after applying the phase averaging method (*dashed*), and the time series resulting from subtraction of a sinusoidal signal with a period of $T = 365.25$ days whose phase and amplitude have been estimated by a linear least-squares approach (*dotted*).

(where $s \in \mathcal{S} = \{(1), \ldots, (S)\}$ is a partition of the range of the considered observable X - in our case, the temperature – into S mutually disjoint classes) for a 31-years time series of daily maximum air temperatures recorded in Tokyo, Japan. For comparison, the results are displayed for the original time series and the deseasonalised ones after application of the phase-averaging method and after the subtraction of a sinusoidal signal with a period of $T = 365.25$ days fitted to the data. It can be easily seen that in the original time series, the annual component clearly dominates the results, whereas the deseasonalised data are characterised by a successive decay of both linear and nonlinear correlations as the considered temporal distance between two observations increases. Note that there are still small residuals of the annual cycle in the record in the case of a narrow-banded spectral filter, which can be seen by a small, but significant increase of the linear correlations at delays of about 365 days.

3 Multivariate Dimension Analysis

Let us now consider the question of how to quantify the complexity of interrelationships within or between time series. Within the framework of self-similar processes, fractal dimensions [30] are often used to quantify the complexity of univariate time series, with recent generalisations to multivariate data [31, 32, 33, 34]. However, for real-world systems, the underlying self-similarity assumption is often violated, such that this approach may not be capable to give a sophisticated characterisation of the dynamics. In the context of univariate climatological time series, Broomhead and King [35] and Fraedrich [2] independently suggested to use the number of statistically relevant components in the suitably embedded time series as a measure for dimensionality (and, hence, the complexity of temporal interrelationships) in terms of the so-called singular spectrum analysis [36, 37, 38]. Apart from the corresponding problems of defining such an appropriate embedding [39, 40, 41], the underlying approach may be directly generalised to the case of multivariate time series and spatio-temporal or ensemble correlations.

In the following, we will discuss a potential approach to assess the number of relevant components in multivariate time series and a possible measure for the strength of ensemble correlations or, more general, spatio-temporal interdependences.

3.1 Statistical Decomposition of Multivariate Data Sets

The characterisation of ensemble correlations by a single statistical parameter requires an appropriate statistical decomposition of the corresponding multivariate time series. In principle, this decomposition can be performed by a variety of different approaches, including purely linear methods like Karhunen-Loéve decomposition (KLD) (which is often referred to as principal component analysis (PCA) or empirical orthogonal function (EOF) method) [42, 43],

multi-dimensional scaling (MDS) [44], or, referring to a separate consideration of patterns in the frequency domain, multi-channel singular spectrum analysis (MSSA) [45], a combination of the "standard" singular spectrum analysis (SSA) [35] with PCA. All these methods have the common concept that some matrix (which is suitably constructed from the observational data) is subjected to a singular value decomposition (SVD), i.e., is decomposed into its eigenvalues and the corresponding eigenvectors. In the case of KLD, one makes use of the correlation (or scatter) matrix of the observed data set. For MDS, a transformed matrix of the squared linear inter-point distances is used, whereas MSSA is based on a Toeplitz-type lag-covariance matrix obtained from every univariate component time series.

Whereas the SVD step of all these methods may be easily and computationally efficiently performed, there are also different nonlinear generalisations. One possible way to obtain such generalisations is replacing the Euclidean metric by one defined by the local neighborhood, e.g., in terms of isometric feature mapping (ISOMAP) [46] or locally linear embedding (LLE) [47]. An alternative is realising the decomposition in terms of neural networks, including methods like nonlinear principal component analysis (NLPCA) [48] or independent component analysis (ICA) [49]. However, these nonlinear variants require a much larger amount of data for computation, while the linear methods can be applied to rather short time series as well. In addition, the methods based on neural networks do not necessarily lead to well-defined component variances. It has therefore to be noted that the approach described in the following is not directly applicable in such cases.

In the following, let us consider the Karhunen-Loéve decomposition as an example for which the derived components have a rather intuitive interpretation. As its principal idea has been introduced about 100 years ago (see, e.g., [43] for some historical remarks), KLD is today frequently applied as a standard method for compressing spatio-temporal data by finding the largest linear subspace that contains substantial statistical variations of the data. In the case of observations with N simultaneously measured variables and M points in time, the $M \times N$-dimensional data matrix A (rescaled from the original observations to have zero mean and unit variance in any component time series) is used to define the $N \times N$-dimensional symmetric and positive semidefinite scatter matrix $S = A^T A$. This matrix can be completely described by its non-negative eigenvalues σ_i^2 $(i = 1, \ldots, N)$ and their corresponding eigenvectors (which are in geosciences usually referred to as the empirical orthogonal functions (EOF)). Without loss of generality, the eigenvalues σ_i^2 of S may already be given in decreasing order $\sigma_1^2 \geq \cdots \geq \sigma_N^2 \geq 0$. For our following considerations, we will use the component variances λ_i, explained variances e_i, and remaining (or residual) variances r_i, which are defined as

$$\lambda_i = \frac{\sigma_i^2}{\sum_{j=1}^{N} \sigma_j^2} \qquad \left(\sum_{i=1}^{N} \lambda_i = 1 \right) \tag{3}$$

$$e_i = \sum_{j=1}^{i} \lambda_j = \frac{\sum_{j=1}^{i} \sigma_j^2}{\sum_{j=1}^{N} \sigma_j^2} \qquad (e_0 = 0,\ e_N = 1) \tag{4}$$

$$r_i = 1 - e_i = \frac{\sum_{j=i+1}^{N} \sigma_j^2}{\sum_{j=1}^{N} \sigma_j^2} \qquad (r_0 = 1,\ r_N = 0)\,. \tag{5}$$

In order to examine the dynamic features of the data, it is common to additionally study the time-dependence of the amplitudes corresponding to the respective EOF (if thus considered dynamically, KLD is usually referred to as principal component analysis (PCA)). However, this approach still exclusively reflects linear properties.

3.2 Karhunen-Loève Decomposition (KLD) Dimension

The idea of using Karhunen-Loève decomposition for estimating the number of degrees of freedom in spatially extended systems is already presented in [50]. Since in the case of weakly turbulent systems, the same quantity may be represented with methods based on fractal dimensions [51] or Lyapunov exponents [52, 53], this number is conveniently referred to as "the" dimension of the considered system. Following this line of argumentation, one may extend the application of Karhunen-Loève decomposition beyond the purely linear point of view described above.

To determine the number of degrees of freedom in spatially extended systems, Zoldi et al. [54] introduced the concept of KLD dimension for a quantitative characterisation of spatio-temporal chaos [55, 56, 57]. The KLD dimension may be defined as the number of eigenvalues required to capture some specified fraction $0 \leq f \leq 1$ of the total variance $\sum_{i=1}^{N} \lambda_i = 1$ of the data, i.e.,

$$D_{KLD}(f) = \min \{ i : e_i \geq f \}\,, \tag{6}$$

with the limiting cases $D_{KLD}(0) = 0$ and $D_{KLD}(1) = N$. The value of D_{KLD} may serve as an upper bound for the true dimensionality of a system, as the decomposition into orthogonal components in terms of Karhunen-Loéve decomposition may yield particularly redundant components which might be reduced if a nonlinear method is applied.

It should be noted that the above definition (6) is modified with respect to the original one introduced by Zoldi and co-workers who considered $D_{KLD}(f)$ being the maximum number of eigenmodes describing less than a fraction of f of the total variance. This modification is motivated by the fact that for applications in data analysis, for a given f the minimum number of modes that explains a given amount of total variance is usually the quantity of interest. Moreover, this redefinition leads to a more "natural" behaviour of the KLD dimension at the limiting cases $f = 0$ and $f = 1$ as described above.

In the case of simulations of spatio-temporally chaotic systems (i.e., systems which exhibit self-similarity over a large range of spatial scales), Zoldi et al. observed (for any f) a linear scaling of D_{KLD} with the system size N. Whereas the KLD dimension is otherwise restricted to integer values, this finding suggested to study a normalised version, the KLD dimension density $\delta_{KLD} = D_{KLD}/N$ [56], whose values are bounded to the unit interval. Note that in the case of systems with a typical scale, the corresponding behaviour of D_{KLD} may show a scaling which is restricted to a certain range. Under certain conditions, there may even exist more than one distinct scaling region, which corresponds to the behaviour of estimates of fractal dimensions in systems like the chaotic Roessler oscillator [58]. As an example from ecology, Wilson and Keeling [59] studied the spatial dynamics of a predator-prey-resource model and found a different scaling behaviour on small and large scales. In a similar way, such distinct scalings may be identified in images of ecological systems [60]. The transition between the different scaling regions may then be identified as the characteristic spatial (or temporal) scale of the system.

The KLD dimension has mainly been used to characterise the dynamics of spatially extended model systems in the extensive chaotic state [54], spiral-defect chaos [55], and reaction-diffusion systems [56]. Recently, Varela et al. [57] applied D_{KLD} for an investigation of spatiotemporal data from electrochemical oscillator experiments (with $M \geq 6000$ and $N = 50$). It has been demonstrated that this measure is well suited for quantifying differences between regular and turbulent states.

To adapt the concept of KLD dimension for the analysis of possibly instationary multivariate time series, one may additionally consider the temporal variability of the observations for a temporally localised characterisation of the dynamics. While the consideration of S for the complete data set loses any temporal information about the variations in the complexity of interrelationships between the different components (which may be significant especially if $M \gg N$), a separate computation of the KLD dimension for sliding windows in time [56] allows a resolution of the varying complexity down to the scale of N points in time or even below.

3.3 Linear Variance Decay (LVD) Dimension

Whereas the KLD dimension density can be widely applied to characterise large data sets from spatio-temporally chaotic systems, its direct use for the characterisation of observational records is problematic in the case of small data sets (i.e., small N) or time windows (small M) due to different reasons:

- δ_{KLD} has a possible range of only $N + 1$ different, equally spaces values. Thus, the number of possible values becomes very small for the considered data. As a consequence, small changes of the structure of interrelationships between the component time series are not detected by this measure,

whereas it changes discontinuously (with a step size of $1/N$) when these modifications of the data increase over a certain threshold. Thus, if N is rather small, only rather strong changes within the data are detected by a dramatic change of δ_{KLD}.

- There is no natural choice of the cutoff parameter f which has to be specified separately for each application. Thus, it is useful not to consider δ_{KLD} as an *absolute,* but rather as an *relative* dimension density. However, for applications where only a qualitative detection and description of changes of the complexity of interrelationships within multivariate data is requested, this subtile difference is no major problem.
- Due to the small amount of observational data in time, certain finite-size effects have to be expected which may cause any quantitative interpretation of δ_{KLD} to fail.

The above arguments call for the definition of more general estimates for *relative* dimension densities, which can already be applied to short multivariate time series. As one possible approach, one may consider the scaling of δ_{KLD} with the cutoff parameter f by fitting a suitable parametric function to the respective curve. In this case, one should note that for a given value of $\delta_{KLD}(f) = i/N$ $(i = 0, \ldots, N)$, $1 - f$ plays the role of the remaining variances r_i for $i = 1, \ldots, N$ with i/N being the relative number of components considered.

For the component variances λ_i, the scaling behaviour has been investigated in some detail for random matrices [61, 62] as well as real-world geoscientific data [63] in terms of the logarithmic eigenvalue (LEV) curves or scree graphs (for an overview, see [43]). Commonly, these graphical methods are used as a simple possibility for graphically checking whether the component variances decay sufficiently smooth, which is an important prerequisite for a meaningful interpretation of KLD-based dimension estimates. Furthermore, for a certain class of multivariate random processes, Preisendorfer has derived analytical results on the scaling of the eigenvalues λ_i [43]. However, in the case of general multivariate data sets, only the leading eigenvalues of the covariance matrix are typically considered as being dynamically relevant in terms of statistically significant orthonormal basis vectors of a certain linear subspace. The remaining eigenvalues λ_i are assumed to represent stochastic variations and, hence, to follow a distinct scaling law which depends on the length of the time series. Although this assumption yields a fundamental restriction to the analysis, it is usually not checked in applications, for example, by comparing the distribution of all eigenvalues to that expected for multivariate Gaussian white noise.

In contrast to the component variances λ_i, there are no studies analysing the scaling of the remaining variances r_i in some detail. However, as we will show later for some examples, a rough inspection of the corresponding values for both random matrices as well as observational data shows that the decay corresponding to the major components (i.e., the consideration of the compo-

nents with the highest variances) is often reasonably well approximated by an exponential decay law. As a consequence, one can make the following ansatz:

$$r_i = 10^{-\frac{i}{N}/\delta} \quad \text{for} \quad i \leq i_{max} < N. \tag{7}$$

In this expression, the choice of the decadal logarithm, i.e., an explained fraction of 90% of the total variance of the data as a reference value, is motivated by the fact that a corresponding threshold yields a reasonable number for the effective degree of freedom in spatially extended systems, cf. [50] (93%), [54] (between 81% and 95%), or [64] (90%). The values of $\delta \cdot N$ can therefore be considered as an estimate of the degrees of freedom. As it quantifies the decay of remaining variances of the linear principal components of a multivariate data set, the scaling coefficient δ is called the *linear variance decay* (LVD) dimension density of the considered multivariate data set [16, 65, 66].

The value of δ may be roughly estimated by an ordinary linear least square approach to Eq. (7) [16]. However, if N is rather small, there are only few points to interpolate the respective model function. Moreover, there are again only $N-1$ possible choices of the threshold i_{max} for fitting this function (as $r_0 = 1$ and $r_N = 0$ by definition, an exponential decay law must be subjected to a certain cutoff at $i_{max} < N$). To overcome this difficulty and define the model function with respect to a continuously distributed cutoff parameter f (which is important when δ should be considered dynamically), one can make use of the relationship between r_i and $1 - f$, which is illustrated in panels (a) and (b) of Fig. 2: reversing the axes in (b) and multiplying δ_{KLD} by N, one approaches a continuously defined equivalent of the logarithmic representation of the remaining variances in panel (a) (where the illustrated function is defined only for integer values of i). A scaling law of the KLD dimension density corresponding to that of the remaining variances then looks as follows:

$$\delta_{KLD}(\phi) = -\delta(f) \ \log_{10}(1-\phi) \quad \text{for} \quad \phi \in [0, f]. \tag{8}$$

Ordinary linear least squares estimate. As $\delta_{KLD}(f)$ is well-defined for $f \in [0, 1]$, the defining expression of the LVD dimension density expression allows to calculate δ as a function of the maximum considered value of f for any $f \in (0, 1)$. As a particularly suited approach, one may apply a continuous least-square approach minimizing the functional

$$F_\alpha(f) = \int_{\log_{10}(1-f)}^{0} (\delta_{KLD}(x) + \alpha x)^2 \ 10^x \ dx \tag{9}$$

with respect to α (here, the transformation $x = \log_{10}(1-\phi)$ has been used). One easily convinces oneself that $F_\alpha(f)$ has (for any value of f) a unique global minimum at

$$\hat{\delta}^{uw}(f) = \arg\min_\alpha F_\alpha(f) = -\frac{\int_{\log_{10}(1-f)}^{0} \delta_{KLD}(x) \ x \ 10^x \ dx}{\int_{\log_{10}(1-f)}^{0} x^2 \ 10^x \ dx} \tag{10}$$

which is easily computed by separately evaluating the integrals over all ranges of x where $\delta_{KLD}(x)$ has a constant value. The above expression for $\hat{\delta}^{uw}(f)$ has already been considered as a rough estimate of the exponential decay scale $\delta(f)$ in [16, 65, 66, 67]. However, in the following, we will show that the properties of this estimator call for a further improvement.

Normalisation. As it follows from its definition, there is a special behaviour of the estimated LVD dimension density $\hat{\delta}^{uw}$ as the maximum considered variance fraction f goes to 0 or 1, respectively: $\hat{\delta}^{uw} \to +\infty$ for $f \to 0$ as $\log_{10}(1-f) \to 0$, and $\hat{\delta}^{uw} \to 0$ for $f \to 1$ because $\log_{10}(1-f) \to -\infty$ (see Fig. 2c). From the first observation, it follows that the unweighted estimate defined above is not appropriately normalised to values within $[0, 1]$. In order to correct this fact for the continuous estimate, one may consider the minimally possible value $\hat{\delta}^{uw}_{min}$ of the estimator $\hat{\delta}^{uw}$ (which would occur if $\delta_{KLD}(f) = 1/N$) by setting

$$\hat{\delta}^{norm}(f) = \frac{1}{N} \frac{\hat{\delta}^{uw}(f)}{\hat{\delta}^{uw}_{min}(f)} \tag{11}$$

with

$$\begin{aligned} \hat{\delta}^{uw}_{min}(f) &= -\frac{1}{N} \frac{\int_{\log_{10}(1-f)}^{0} x\ 10^x\ dx}{\int_{\log_{10}(1-f)}^{0} x^2\ 10^x\ dx} \\ &= \frac{1}{N} \frac{(1-f)(\ln 10)^2 \log_{10}(1-f) - f \ln 10}{(1-f)(\ln 10)^2 (\log_{10}(1-f))^2 - 2(1-f) \ln 10 \log_{10}(1-f) + 2f}. \end{aligned} \tag{12}$$

Weighted linear least squares estimate. According to the logarithmic transformation from a non-linear to a linear least-squares problem, the estimated values of $\hat{\delta}^{uw}$ are not only non-normalised, but also show a systematic bias compared to the *true* LVD dimension density. In order to approach more reliable estimates, one has to either directly use a nonlinear least-squares approach or to explicitly correct the estimator by introducing a proper weight function. Whereas in the case of an unweighted estimate, all possible values of the explained variance $\leq f$ contribute equally, the underlying exponential model implicitly requires a much higher sensitivity with respect to larger values of f (i.e., low values of the remaining variance). A proper choice of the weight function which takes this idea into account is $w(f) = (1-f)^{-1}$. Introducing this factor into the integrand in Eq. (9), the exponential factors in the original unweighted estimate $\hat{\delta}^{uw}$ are eliminated as $w(x) = 10^{-x}$ after the substitution described above, i.e.,

$$\hat{\delta}(f) = \arg\min_{\alpha} F_\alpha(f) = -\frac{\int_{\log_{10}(1-f)}^{0} \delta_{KLD}(x)\ x\ dx}{\int_{\log_{10}(1-f)}^{0} x^2\ dx}. \tag{13}$$

Note that if expanding the exponential terms in the original unweighted functional $F_\alpha(f)$ into a Taylor series, the weighted estimate corresponds to the zeroth-order term in this expansion.

Normalised weighted linear least squares estimate. Combining our normalisation with the weighted linear least squares approach, we finally end up with the following estimate, which is referred to as the *relative LVD dimension density* δ^{rel}_{LVD}:

$$\hat{\delta}^{rel}_{LVD}(f) = \frac{\int_{\log_{10}(1-f)}^{0} \delta_{KLD}(x)\ x\ dx}{\int_{\log_{10}(1-f)}^{0} x\ dx}. \tag{14}$$

Using the notation of remaining variances r_i and $i_{max} := D_{KLD}(f) - 1$, one may derive the following equivalent expression:

$$\hat{\delta}^{rel}_{LVD}(f) = -\frac{\sum_{i=1}^{i_{max}} \frac{i}{N} \left[x^2\right]_{\log_{10} r_i}^{\log_{10} r_{i-1}} + \delta_{KLD}(f) \left[x^2\right]_{\log_{10}(1-f)}^{\log_{10} r_{i_{max}}}}{(\log_{10}(1-f))^2} \tag{15}$$

Note that unlike the original unweighted estimate $\hat{\delta}^{uw}$, the value of $\hat{\delta}^{rel}_{LVD}$ does not depend on the specific choice of a particular base for the logarithm. However, although its definition incorporates an appropriate weighting and normalisation, some conceptual problems remain in interpreting $\hat{\delta}^{rel}_{LVD}$. In particular, the estimated values do still depend on the maximally explained variance fraction f, which motivates the term *relative* dimension density. However, unlike for the KLD dimension density, this dependence is continuous. Despite this potential point of criticism, the relative LVD dimension density $\hat{\delta}^{rel}_{LVD}$ can be considered as a meaningful measure for the strength of linear interdependences between the components of arbitrary multivariate time series.

3.4 Generalisations of the LVD Dimension Density

The formalism described above is rather general and might be adapted to study the eigenvalues of any symmetric matrix of interaction coefficients. In the following, we will give some examples for possible modifications and fields of application.

If the values in the different component time series deviate strongly from a Gaussian distribution, the consideration of the linear Bravais-Pearson correlation coefficients as above may lead to biased results. To avoid the corresponding problems, one may replace these coefficients by a nonparametric (rank-order) correlation coefficient like Spearman's Rho [68]. Alternatively, measures of concordance (e.g., Kendall's Tau [69]) might be considered. Both measures have the advantage that their values do not depend on the exact values of the observables and are invariant against any strongly monotonous transformation of the data. However, in the limit of infinitely long time series, rank-order correlation and Pearson correlation converge to each other. Hence, the consideration of any of these correlation coefficients should yield

qualitatively consistent results. Using a matrix of pairwise rank-order correlation coefficients instead of the standard ones in our formalism then leads to the *non-parametric linear variance decay dimension density* δ_{NLVD}. This measure has recently been used to study spatio-temporal interrelationships of river runoffs in a common catchment [67], which are a typical example of hydro-meteorological time series with strongly non-Gaussian distributions.

The LVD dimension density can also be generalised using measures that are not exclusively sensitive with respect to linear correlations, but can also detect non-linear statistical dependences. As an example, one may consider *information variance decay dimension densities* $\delta_{IVD}^{(q)}$ for which the linear correlation coefficients between the component time series are replaced by the respective (generalised) mutual information of order q [70].

To evaluate the degree of (phase) synchronisation between more than two interacting oscillatory subsystems, one may consider the eigenvalues of matrices of pairwise (phase) synchronisation indices like the mean resultant length [71, 72]. In contrast to other measures of multivariate (phase) synchronisation, this approach does not explicitly assume a spatially homogeneous synchronisation process, but gives additional information about the potential heterogeneity.

Finally, it is also possible to define optimally lagged variance decay dimension densities $\delta_*^{(lag)}$ in which the previously used equal-time interaction measures (correlation coefficient, mutual information, etc.) are replaced by the maximum values of the corresponding measures as a function of the time shift τ between each pair of component time series.

It has to be noted that the above list of possible generalisations is far from being complete and gives rise to a variety of different fields of application.

3.5 Example 1: Multivariate Gaussian Random Processes

As a first illustrative example, let us consider the case of multivariate Gaussian white noise, for which the individual components are pairwise independent from each other. For such a record, Preisendorfer [43] has already given explicit expressions for the distribution of the eigenvalues σ_i^2 of the covariance matrix. In particular, as the number of available data becomes large (i.e., $M \to \infty$), the components are recognised as being pairwise independent, such that the orthogonal components identified by the Karhunen-Loéve decomposition can be identified with the original components of the record. As these components have been assumed to have unit variance, it follows that

$$\lambda_i \to \frac{1}{N} \qquad \Rightarrow \qquad r_i \to 1 - \frac{i}{N}. \tag{16}$$

In contrast to the exponential decay model assumed by the LVD dimension density approach, in the case of independent Gaussian random variables, the residual variances thus decay *linearly*. In Fig. 2, it is shown that this leads to

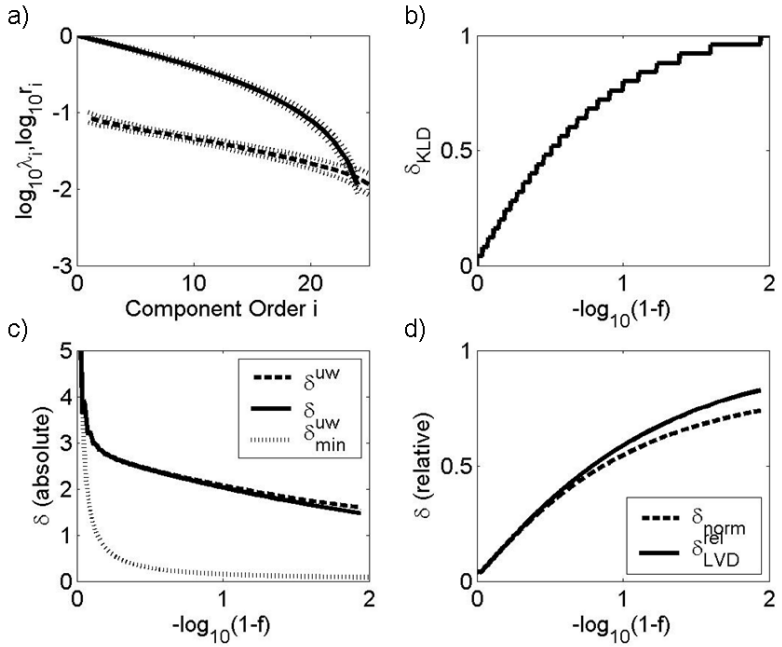


Fig. 2: Results of the linear statistical decomposition and multivariate dimension analysis for $N = 25$ pairwise independent Gaussian white noise components with a length of $M = 100$ data points. (**a**) Logarithmic representation of the component variances λ_i (*dashed*) and remaining variances r_i (*solid*), including the $\pm 3\sigma$ significance levels estimated from a set of 100 realisations. (**b**) KLD dimension density δ_{KLD} in dependence of the explained variance fraction f (the scaling of the axis has been chosen for a better visualibity of the different values and corresponds to a linear scaling in $-x$ after the substitution described in Sect. 3.3). (**c,d**) Different estimates of the "absolute" and relative LVD dimension density as described in the text. Note that the values differ from those given in [16] due to the choice of the decadal instead of the natural logarithm.

a systematic deviation already for $M = 100$, which however mainly influences the residual variances at large component orders. For the leading orders, in the case of finite time series the exponential model is still a reasonable approximation. However, for $M = 100$ this is not true anymore for the range of $r_i \lesssim 10\%$, where the corresponding approximation breaks down. Consequently, in this range, the different estimates of the LVD dimension density are not very reliable. In general, we therefore recommend to always discuss the goodness-of-fit of the exponential model in addition to the estimated decay scale. In particular, in order to identify signatures of stochasticity, such goodness-of-fit statistics may be used as corresponding parameters.

The finite-size corrections to the eigenvalue distributions might be used to theoretically derive correction terms to the scaling of the residual variances for finite M as well. However, this is beyond the scope of the presented work. With respect to real-world applications from the geosciences, an application to more complicated stochastic processes such as multivariate auto-regressive processes might also be of considerable interest [73]. We plan to study the corresponding questions in our future research.

3.6 Example 2: The Politi-Witt Model Revisited

The example of stochastic component time series with Gaussian distributions discussed above is rather generic. In contrast to this, many observational data from geoscientific systems are likely to have some deterministic, but eventually highly dimensional chaotic components. To demonstrate the power of our multivariate dimension analysis approach, we will reconsider a model system which approximates the behaviour of spatio-temporal chaos with a prescribed dimension density $d \in [0, 1]$. The corresponding model has been originally introduced by [74] and was already studied by [16, 65, 66] to test for the applicability of multivariate dimension estimates.

Let $\{F_1, \ldots, F_n\}$ be the basis of a sufficiently high-dimensional Fourier space whose elements may be expressed as

$$F_{kj} = \begin{cases} 1/\sqrt{n}, & \text{if } k = 1, \\ \sqrt{2/n} \cos\left(\frac{2\pi}{n} \left[\frac{k}{2}\right] j\right), & \text{if } k > 1 \text{ and odd}, \\ \sqrt{2/n} \sin\left(\frac{2\pi}{n} \left[\frac{k}{2}\right] j\right), & \text{if } k \text{ even}, \end{cases} \tag{17}$$

where $[\cdot]$ denotes the integer part. $j = 1, \ldots, N \leq n$ gives the "spatial" position on a regular one-dimensional lattice, which is used to construct a multivariate data set as follows:

$$x_{ij} = \sum_{k=1}^{dn} \xi_{ik} F_{kj}. \tag{18}$$

Here, ξ_{ik} (where $i = 1, \ldots, M$ corresponds to the position in time) is a set of random numbers taken from an appropriate distribution. If $|\xi_{ik}| < 1$, the set of values x_{ij} is contained in a dn-dimensional hypercube and forms a $M \times N$-dimensional data matrix. If M is sufficiently large, the eigenvalues λ_i of the associated covariance matrix (which has a Toeplitz structure) show an abrupt decay at the component index dn, corresponding to the dimension of the underlying hypercube [16, 65]. However, even under these conditions, the decay of the remaining variances can be approximated by our exponental model with a reasonable accuracy.

Following [16, 65, 66], let the ξ_{ik} be taken from a uniform distribution on $[-3^{1/3}, 3^{1/3}]$. This setting corresponds to the system originally studied by [74] both analytically and numerically. A detailed investigation of the dependence

of the unweighted estimate δ^{uw} on both the length of the time series M and the true dimension density d can be found in [16, 66]. In particular, in the aforementioned references, it has been shown that in the case of an appropriately chosen value of f, the KLD dimension density δ_{KLD} may give a slightly better quantitative estimate of the true dimension d than δ_{LVD}, while both characteristics converge to a asymptotically constant value if the length of the component time series becomes sufficiently large. However, whereas in the long-term limit, both types of dimension estimates give reasonable values, the LVD dimension density is clearly superior for detecting small changes within the system. The latter observation is of a particular importance for a possible short-term characterisation of geophysical time series in the case of instationary conditions.

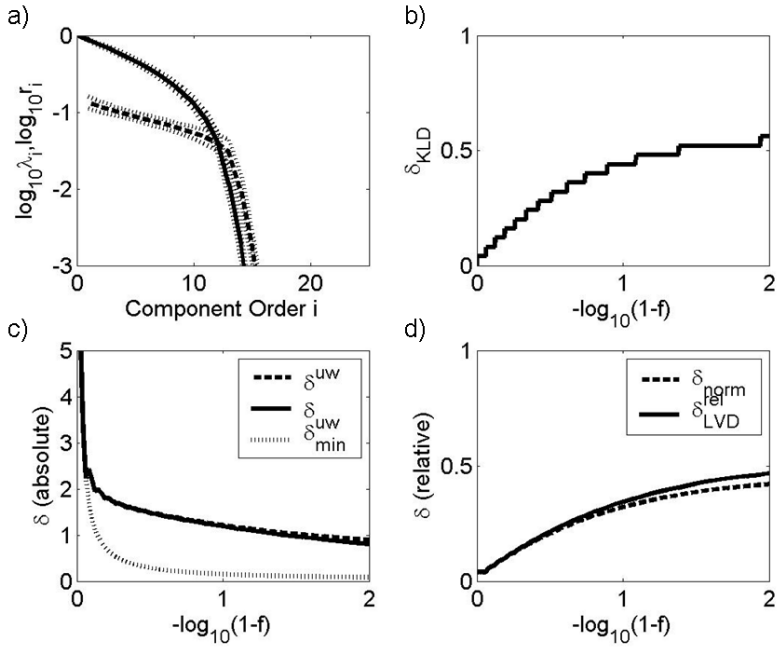


Fig. 3: As Fig. 2, for the Politi-Witt model with a true dimension density of $d = 0.5$.

Comparing the behaviour of the different estimates with that in the case of multivariate random processes (see Figs. 2 and 3), one may observe that the qualitative behaviour is rather similar for both types of systems. However, for similar values of the variance threshold f, for the Politi-Witt model all dimension estimates are significantly lower compared to the the case of multivariate noise. This lower dimensionality is obviously related to a lower

dynamic complexity, which is reflected by a lower number of statistically significant components. Note that both KLD dimension density and relative LVD dimension density do not change very much between $f = 90\%$ and 99% in the case of the Politi-Witt model. In particular, in Fig. 3, in can be seen that their values in the considered range are rather close to the "theoretical" dimension density of $d = 0.5$. In contrast to this, in the case of a completely stochastic system, all estimates do significantly increase within this range of f and show only a very slow convergence towards the values near 1 as f becomes large.

4 Correlations of Japanese Temperature Records

In the following, we will study the spatio-temporal correlations between air temperature records across the Japanese islands. A sketch of the Japanese archipel is shown in Fig. 4, including the approximate locations of the 13 meteorological stations which we will use in our analysis. The data contain daily minimum, maximum, and mean air temperatures for the time interval between 1975 and 2005 provided by the Japanese Meteorological Agency. It has

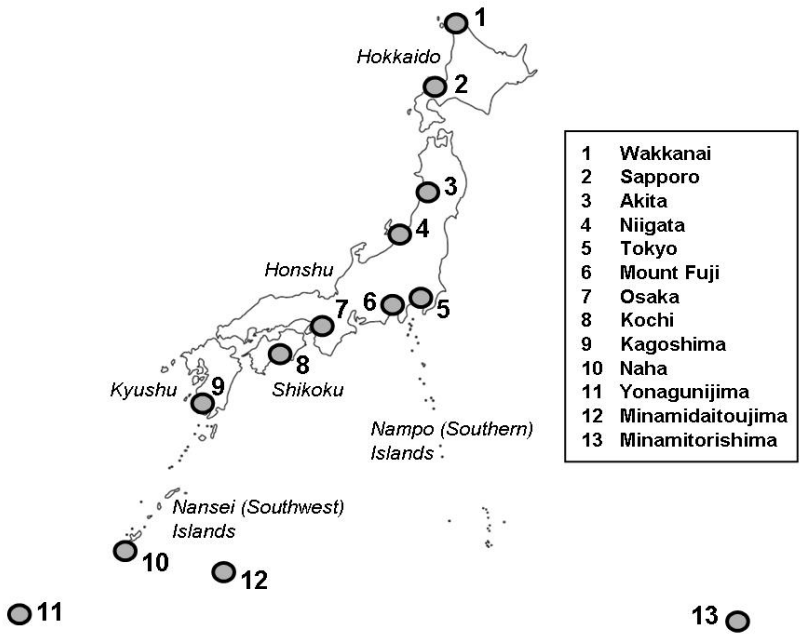


Fig. 4: Map on the Japanese archipel, including the approximate locations of the 13 meteorological stations used in this study.

to be noted that the covered area is rather large and includes the small southern islands with a tropical climate, whereas the northernmost main island Hokkaido is characterised by mild conditions during the summer, but rather cold temperatures during the winter season. The regional climate system is strongly influenced by the East Asian monsoonal circulation, which roughly means a very high humidity in summer and significantly dryer conditions in winter.

According to our results from Sect. 2, a sophisticated preprocessing of all time series is necessary before any further analysis. As our corresponding analysis revealed that the linear statistical features do not depend significantly on the particular deseasonalisation approach, in the following, we will consider all time series to be subjected to the phase averaging method after a removal of long-term trends extracted by a one-year running moving average filter. Moreover, we will standardise all data afterwards in the usual way to have zero means and unit variances.

As a first step of our analysis, let us consider the linear (Bravais-Pearson) equal-time cross-correlation coefficients

$$C_{XY} = \sum_{i=1}^{M} X_i Y_i \tag{19}$$

between all pairs (X, Y) of records in our data set. In addition, we define the *correlation distances* $d_{XY} = 1 - C_{XY}$ between all stations, which are then used for a one-dimensional agglomerative cluster analysis with the single-linkage method. Our corresponding results are shown in Fig. 5. It turns out that there are three major groups of stations, whose mutual correlations are most pronounced in the case of daily maximum and mean temperatures. These three groups can easily be attributed to specific geographical regions: Hokkaido and Northern Honshu (stations 1–4), Western Honshu/Kyushu/Shikoku (7–9), and the Southwestern Islands (10–12). In addition, there are the special cases of Tokyo, Mount Fuji, and Minamitorishima (also known as Marcus' Island) which show more specific temperature variations that can be explained by their special geographical features (metropolis region, mountain, small isolated island). In general, the observed correlations in the mean temperature records are significantly stronger than those between the respective maximum or minimum temperatures. With respect to the reported geographical clusters, the daily minimum values reveal a less pronounced structure, as in particular the records from the smaller southwestern islands are much less correlated than the corresponding mean or maximum temperatures.

As already stated in the introduction, the spatial extention of the studied area leads to the fact that simple climatological patterns influence different locations at different times. Consequently, the study of equal-time correlations as above does not necessarily yield an optimum representation of the mutual correlation pattern of our records. As an alternative, we consider the maximum values of the cross-correlation functions

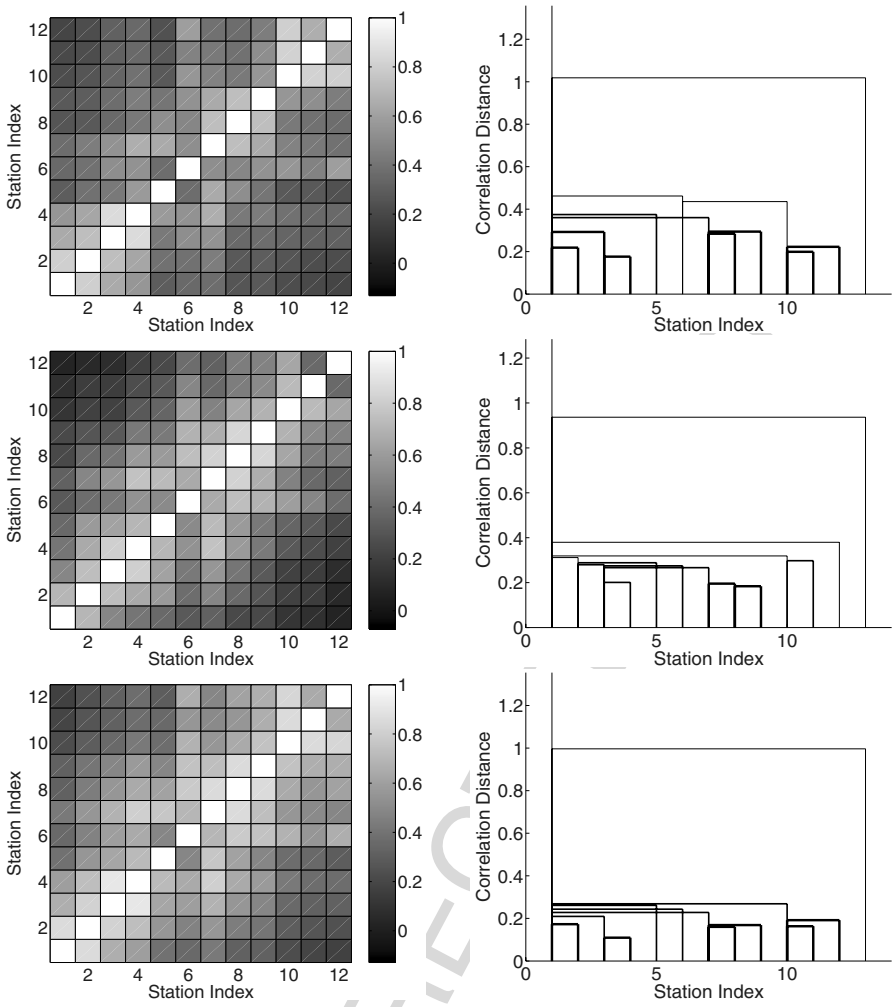


Fig. 5: Map of the mutual equal-time correlation coefficients (*left panels*) and the resulting dendrograms of a one-dimensional agglomerative cluster analysis based on the correlation distance (*right panels*) for the daily maximum, minimum, and mean temperatures (from *top* to *bottom*) recorded at 13 Japanese stations (see Fig. 4) between January 1975 and December 2005. In the dendrograms, the identified clusters are visualised by *thicker lines*, which correspond to larger mutual correlation coefficients. The station 13 (Minamitorishima) is only very weakly correlated to the other locations and therefore not shown in the correlation maps.

$$C_{XY}(\tau) = \sum_{i=1}^{M} X_{i+\tau} Y_i \tag{20}$$

that may occur for $\tau \neq 0$. For this purpose, we use the complete information contained in these functions and search for the maximum values by approximating the available values (for integer numbers of the respective delay τ in a basic unit of one day) by applying cubic spline interpolation. Indeed, the resulting correlation patterns are qualitatively consistent with those of the equal-time correlations, but include additional information in terms of the "optimal" delay between each pair of stations (see Fig. 6).

In order to give a condensed view on the correlation pattern of the Japanese climatology, we suggest to combine the information about the maximum correlation and the corresponding temporal delay into one graphical representation. For this purpose, we propose a visualisation in terms of a "correlation network" which is shown in Fig. 7. In this representation, links between two "nodes" (here: meteorological stations) located in a two-dimensional plane are shown in terms of three-dimensional arrows whose colors and heights give information about the minimum correlation distances d_{XY} (obtained from the maxima of the cross-correlation functions) and the corresponding optimal delays τ, respectively. This network representation is inspired from the three-dimensional visualisation of airborne traffic networks as well as the "climate networks" recently studied by Tsonis and co-workers [9, 75, 76]. In our case, a detailed inspection shows that there is a strong coincidence between high correlations and low delays, which is a characteristic feature of records from stations with a relatively small spatial distance, i.e., the geographical clusters already identified above.

Our presented approach can be easily generalised to other measures of interdependences, including non-parametric rank-order correlation functions based on Spearman's Rho or Kendall's Tau, nonlinear mutual information, or generalised correlation functions obtained from recurrence plots. For a review of these approaches with applications to hydrological records (river runoffs from different gauges in a common catchment), we refer to Ref. [67]. In this work, we will further focus exclusively on the linear correlation properties.

Whereas up to this point, we have restricted our interest to the matrix of mutual correlations between the temperature records from different stations, in the following we will go one step further. In particular, we will consider the entire multivariate data set as a whole and investigate its statistical properties by means of the multivariate dimension estimates introduced in Sect. 3. In order to distinguish this approach from the consideration of *mutual* correlations, we will refer to it to as *ensemble correlations* [67].

Figure 8 shows the eigenvalues and remaining variances of the covariance matrices for daily maximum and minimum temperatures. It can be seen that for both observables, the decay of the residual variance is reasonably well approximated by an exponential function (even better than in case of the two model systems discussed in the previous section). In addition, the resulting

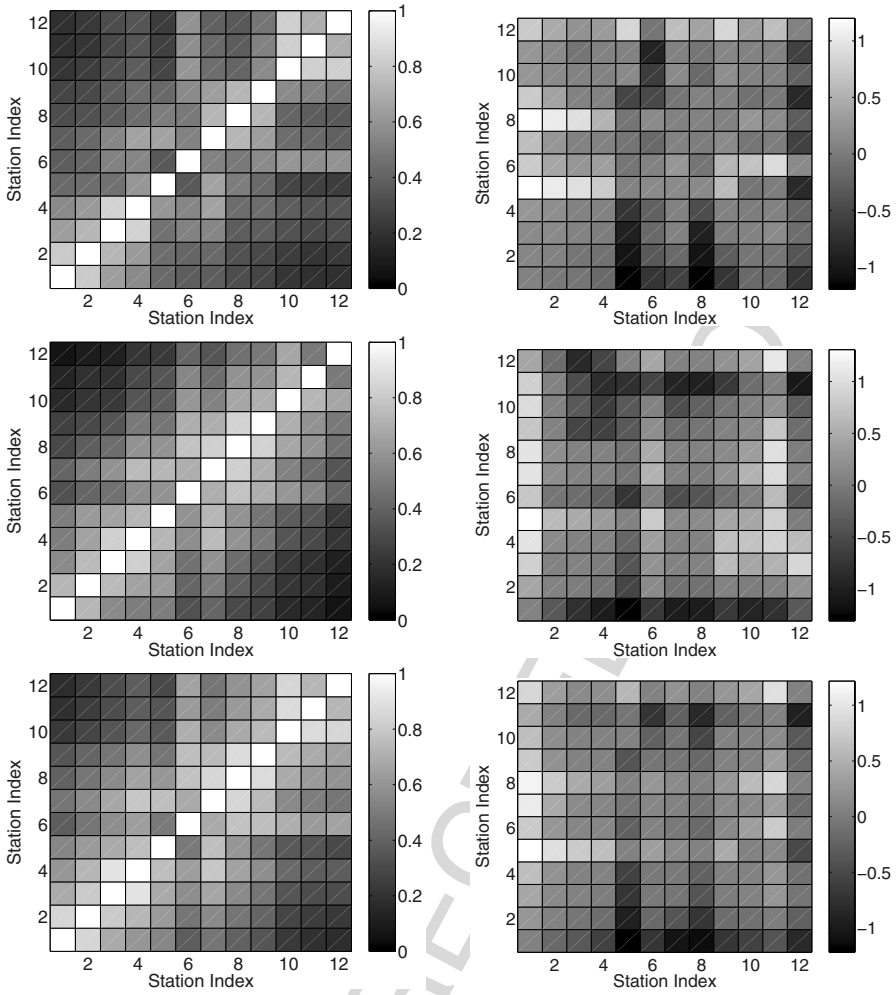


Fig. 6: Map of the maximum values of the cross-correlation functions (*left panels*) and the corresponding optimal mutual delays (*in days*) of the considered records (*right panels*) for the daily maximum, minimum, and mean temperatures (from *top* to *bottom*) recorded at 12 Japanese stations (without Minamitorishima) between January 1975 and December 2005.

KLD and relative LVD dimension densities are shown as a function of the explained variance fraction f. Considering the resulting values for typical choices of f of about 0.9 to 0.95, the estimated dimension densities are between the values of the two model systems. This indicates that although the mutual correlations between the individual records are rather strong, there are residual

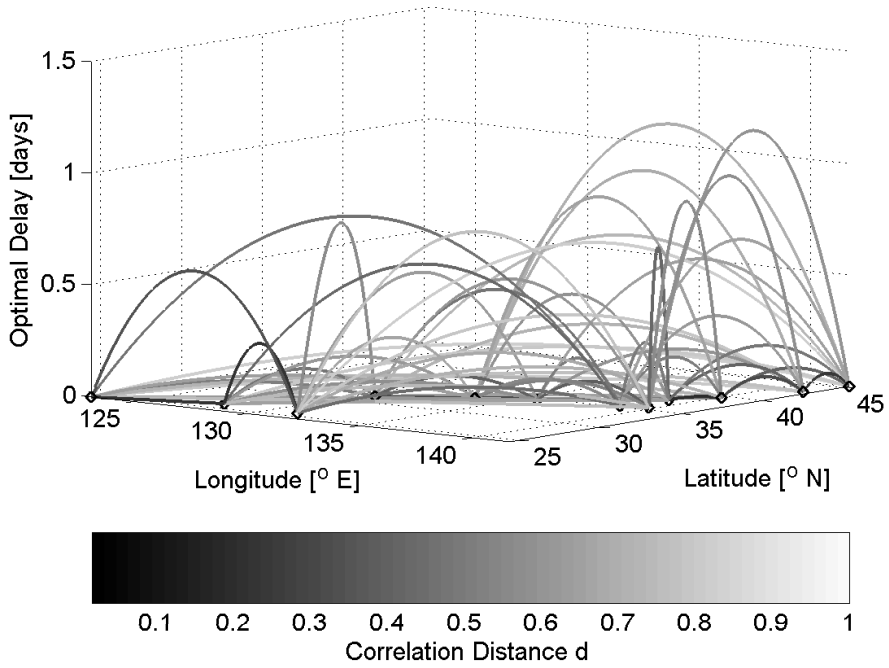


Fig. 7: Network representation obtained from the Japanese daily maximum temperature records from 12 meteorological stations (without Minamitorishima). The color of the arrows corresponds to the minimum correlation distance (or, alternatively, the maximum of the cross-correlation function), whereas the height represents the mutual delay between two stations. Note that in this representation, there is no information about the direction of this delay.

fluctuations that resemble stochastic components rather than signatures of complex deterministic behaviour.

In Fig. 9, we have shown the resulting values of δ^{rel}_{LVD} calculated for sliding windows of 14 or 28 days width, respectively. One may clearly observe that the estimated dimensionality of the data varies with an annual period, although the annual cycle components have been filtered out by previous deseasonalisation. This observation suggests that the spatial correlations between the different locations are different during different seasons, relating to the large-scale atmospheric circulation patterns influencing Eastern Asia. In particular, the summer conditions are characterised by a significantly larger dimensionality, i.e., a larger number of dynamically significant patterns, which possibly relates to a common influence of the monsoon. However, one has to mention that our analysis does not yet provide enough evidence for a corresponding

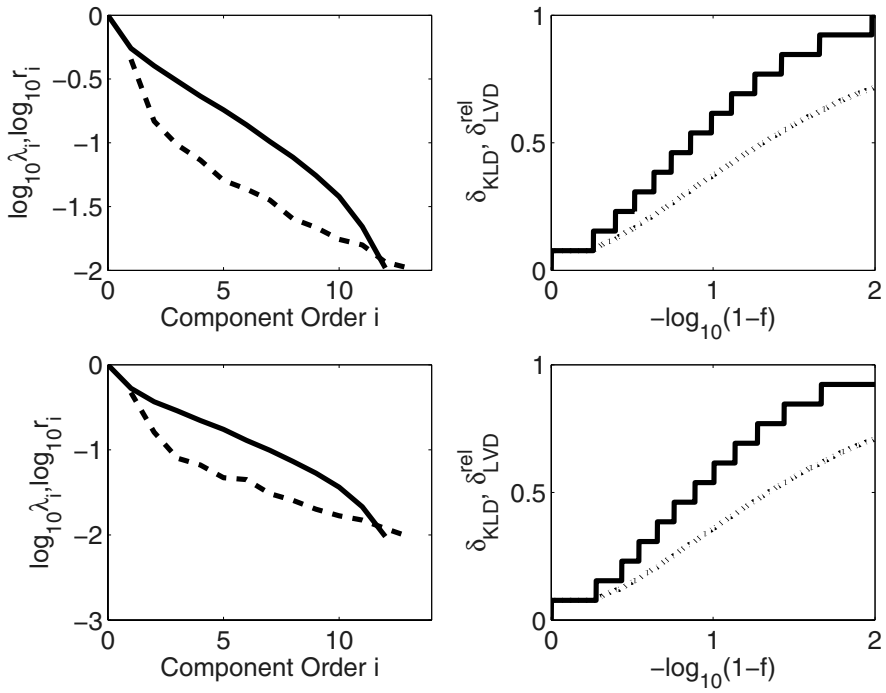


Fig. 8: *Left panels*: Normalised eigenvalues λ_i (*dashed*) and residual variances r_i (*solid*) of the covariance matrices of the 31-years records of daily maximum (*top*) and minimum (*bottom*) temperatures from 13 Japanese meteorological stations. *Right panels*: The corresponding values of the KLD (*solid*) and relative LVD (*dotted*) dimension density in dependence on the explained variance fraction f.

conclusion. In order to present a more detailed explanation, a deeper analysis of the corresponding spatial patterns is necessary in terms of the spatial EOFs obtained from time series from a much larger set of meteorological stations.

In order to study the temporal evolution of the annual component of the variations in the dimensionality of our data, a continuous wavelet analysis has been performed. The results shown in Fig. 9 reveal that the strength of the "locking" to the annual cycle of insolation indeed significantly varies over the entire length of the record, with a maximum coherence in 1987/88 and a minimum coherence around winter 1993/94. We have tried to correlate this locking to different possible influences like the El Niño Southern Oscillation or the solar activity cycle. However, no substantial indication has been found so far that any of these phenomena causes the quantitative variations of the observed locking to the annual period in a direct or indirect way.

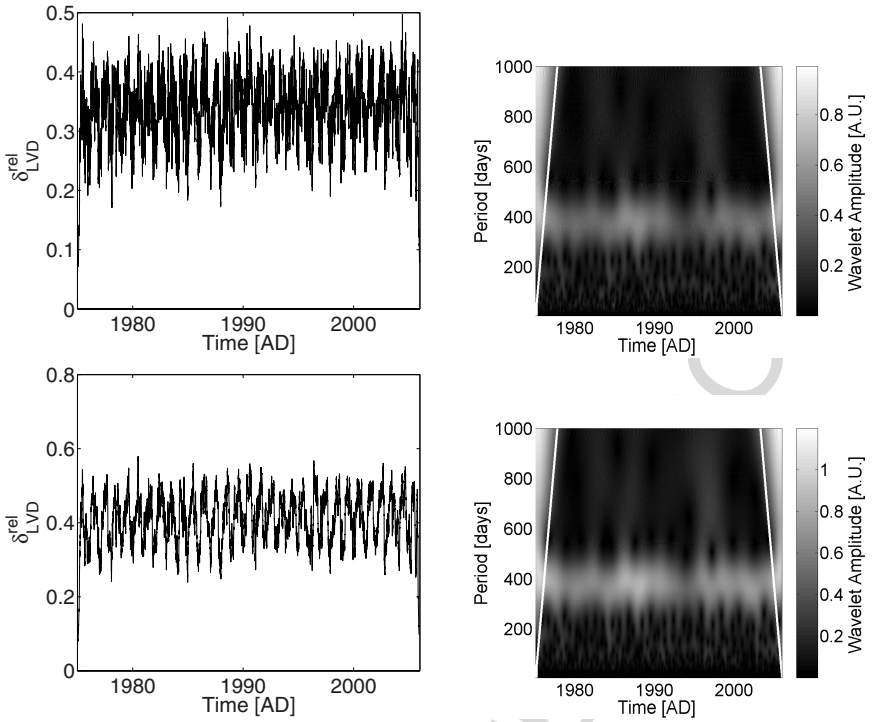


Fig. 9: *Left panels*: Relative LVD dimension density δ^{rel}_{LVD} (for $f = 0.95$) obtained from maximum air tempreature time series of 12 Japanese stations (without Minamitorishima) calculated for running windows of 14 (*upper panels*) and 28 days (*lower panels*) width, respectively, for the time interval between January 1975 and December 2005. *Right panels*: Corresponding wavelet spectrograms, estimated with a complex Morlet wavelet. The *white lines* indicate the respective cones of influence.

5 Summary and Outlook

In this work, we have presented some conceptual ideas for the investigation of correlations between spatially distributed climatological time series. As an example, we have studied a set of daily temperature records from different Japanese meteorological stations. On the one hand, the mutual correlations between all pairs of stations have been considered. As a novel point of view, we have introduced the ideas of climatological correlation networks and correlation cluster analysis, which allow to derive qualitative as well as quantitative statements about the significance of such mutual interdependences. On the other hand, we have theoretically developed a new class of multivariate dimension estimates, which can be used to characterise the complexity of interrelationships between the component time series in multivariate data sets.

In particular, our new estimates are well suited for a dynamic characterisation of temporally varying ensemble correlations. With respect to our considered example, our approach allows a unified assessment of the nonlinear dynamics of a regional climate system.

One has to mention that our presented analyses have been based on linear tools (correlation analysis, principal component analysis), although these concepts have been extended using ideas from nonlinear dynamics. This procedure is apparently in conflict with the inherently nonlinear nature of the climate system. Hence, the statistical uncertainty of correlations triggered by non-normality, sampling errors, and other features of the data may cause problems in quantitatively interpreting our results. In future studies, it is therefore necessary to explicitly take potential errors in the estimated correlations into account, in particular, by statistically quantifying them in terms of confidence intervals.

As a potential disadvantage of the concept of multivariate dimension estimates, one has to mention that this integrated view on the complexity of mutual correlations loses information about the detailed spatial variability patterns. In order to compensate this, we would like to mention only two possible approaches: (i) a combination with an analysis of the empirical orthogonal functions (EOF) resulting from principal component analysis or an alternative statistical decomposition, and (ii) the generalisation of the presented concept to univariate dimension estimates. For the latter purpose, it is possible to use properly embedded univariate time series whose eigenvalues are then separately quantified with our approach (i.e., the KLD step in our analysis is substituted by a singular system analysis). A corresponding approach may be used to quantify the spatial variations of the complexity of meteorological records, which is usually done in terms of fractal theory by calculating similarity or correlation dimensions. However, the use of SSA-based estimates is by far less demanding with respect to the required amount of data.

In order to further validate our results, additional information from a larger set of stations with a larger time coverage is necessary. If such data become available, it will be of particular interest to extend our approach of climatological correlation networks to a detailed investigation of the statistical properties of such networks. Similar studies have been recently performed by Tsonis and co-workers [9, 75, 76] based on reanalysis data and variations of distinct climatological oscillation indices. However, as several recent studies suggest that the nonlinear properties of direct observations, reanalysis data, and climate models may differ from each other, it might be of considerable interest to compare not only the "traditional" linear and non-linear measures, but also the characteristics of the resulting networks.

Acknowledgements. This work has been financially supported by the Japanese Society for the Promotion of Science (JSPS) (project no. PE 06066) and the German Research Foundation. We are grateful to the Japanese

Meteorological Agency for providing their observational time series publically on the web. The calculations of the mutual information functions have been performed using the software package TISEAN. Discussions with A. Galka and T. Ozaki are gratefully acknowledged.

References

1. C. Nicolis, G. Nicolis, Is there a climatic attractor? *Nature,* 311, 529–532 (1984)
2. K. Fraedrich, Estimating the Dimensions of Weather and Climate Attractors. *J. Atmos. Sci.,* 43, 419–432 (1986)
3. C. Essex, T. Lookman, M.A.H. Nerenberg, The climate attractor over short timescales. *Nature,* 326, 64–66 (1987)
4. A.A. Tsonis, J.B. Elsner, Chaos, strange attractors, and weather. *Bull. Amer. Meteor. Soc.,* 70, 16–23 (1989)
5. X. Zeng, R.A. Pielke, R. Eykholt, Estimating the fractal dimension and the predictability of the atmosphere. *J. Atmos. Sci.,* 49, 649–659 (1992)
6. R. Kleeman, Statistical predictability in the atmosphere and other dynamical systems. *Physica D,* 230, 65–71 (2007)
7. W.S. Broecker, Thermohaline circulation, the Achilles heel of our climate system: will man-made CO_2 upset the current balance? *Science,* 278, 1582–1588 (1997)
8. P.U. Clark, N.G. Pisias, T.F. Stocker, A.J. Weaver, The role of the thermohaline circulation in abrupt climate change. *Nature,* 415, 863–869 (2002)
9. A.A. Tsonis, P.J. Roebber, The architecture of the climate network. *Physica A,* 333, 497–504 (2004)
10. W. von Bloh, M.C. Romano, M. Thiel, Long-term predictability of mean daily temperature data. *Nonlin. Proc. Geophys.,* 12, 471–479 (2005)
11. R.B. Govindan, J. Raethjen, F. Kopper, J.C. Claussen, G. Deuschl, Estimation of time delay by coherence analysis. *Physica A,* 350, 277–295 (2005)
12. D. Rybski, S. Havlin, A. Bunde, Phase synchronization in temperature and precipitation records. *Physica A,* 320, 601–610 (2003)
13. A. Pikovsky, M.G. Rosenblum, J. Kurths, *Synchronization – A Universal Concept in Nonlinear Sciences,* Cambridge University Press, Cambridge (2003)
14. L. Cimponeriu, M. Rosenblum, A. Pikovsky, Estimation of delay in coupling from time series. *Phys. Rev. E,* 70, 046213 (2004)
15. R. Donner, Interdependences between daily European temperature records: Correlation or phase synchronization? In: P. Marquié (ed.), *Nonlinear Dynamics of Electronic Systems (NDES 2006),* Université de Bourgogne, Dijon, 26–29 (2006)
16. R. Donner, *Advanced Methods for Analysing and Modelling Multivariate Palaeoclimatic Time Series,* PhD thesis, University of Potsdam (2007)
17. R. Donner, M. Thiel, Scale-resolved phase coherence analysis of hemispheric sunspot activity: a new look onto the north-south asymmetry. *Astron. Astrophys.,* 475, L33–L36 (2007)
18. V. Lucarini, T. Nanni, A. Speranza, Statistics of the seasonal cycle of the 1951–2000 surface temperature records in Italy. *Il Nuovo Cimento C,* 27, 285–298 (2004)
19. K. Edel, K.A. Schäffer, W. Stier, *Analyse Saisonaler Zeitreihen,* Physica, Heidelberg (1997)

20. S. Pezzulli, D.B. Stephenson, A. Hannachi, The variability of seasonality. *J. Clim.,* 18, 71–88 (2005)
21. W.M. Persons, Indices of business conditions. *Rev. Econom. Stat.,* 1, 5–107 (1919)
22. A. Wald, *Berechnung und Ausschaltung von Saisonschwankungen,* Springer, Vienna (1936)
23. N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London A,* 454, 903–995 (1998)
24. D.G. Duffy, The application of Hilbert-Huang transforms to meteorological datasets. *J. Atmos. Ocean. Technol.,* 21, 599–611 (2004)
25. H. El-Askary, S. Sarkar, L. Chiu, M. Kafatos, T. El-Ghazawi, Rain gauge derived precipitation variability over Virginia and its relation with the El Nino southern oscillation. *Adv. Space Res.,* 33, 338–342 (2004)
26. K. Coughlin, K.K. Tung, Empirical mode decomposition of climate variability. In: N. Huang, S. Shen (eds.), *Hilbert-Huang Transform and Its Applications,* World Scientific, Singapore, 149–166 (2005)
27. M.K.I. Molla, M.S. Rahman, A. Sumi, P. Banik, Empirical mode decomposition analysis of climate changes with special reference to rainfall data. *Discr. Dyn. Nature Soc.,* 2006, 45348 (2006)
28. I.M. Jánosi, R. Müller, Empirical mode decomposition and correlation properties of long daily ozone records. *Phys. Rev. E,* 71, 056126 (2005)

AU: Please update Ref. [29].

29. V.N. Livina, Y. Ashkenazy, A. Bunde, S. Havlin, Seasonality effects on nonlinear properties of hydrometeorological records. In: J. Kropp, H.J. Schellnhuber (eds.), *In Extremis: Extremes, Trends and Correlations in Hydrology and Climate,* Springer, Berlin, 2008, (submitted)
30. J.C. Sprott, *Chaos and Time Series Analysis,* Oxford University Press, Oxford (2003)
31. M. Bauer, H. Heng, W. Martienssen, Characterization of spatiotemporal chaos from time series. *Phys. Rev. Lett.,* 71, 521–524 (1993)
32. E. Olbrich, R. Hegger, H. Kantz, Analysing local observations of weakly coupled maps. *Phys. Lett. A,* 244, 538–544 (1998)
33. C. Raab, J. Kurths, Estimation of large-scale dimension densities. *Phys. Rev. E,* 64, 016216 (2001)
34. C. Raab, N. Wessel, A. Schirdewan, J. Kurths, Large-scale dimension densities for heart rate variability analysis. *Comput. Cardiol.,* 32, 985–988 (2005)
35. D.S. Broomhead, G.P. King, Extracting qualitative dynamics from experimental data. *Physica D,* 20, 217–236 (1986)
36. R. Vautard, M. Ghil, Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D,* 35, 395–424 (1989)
37. R. Vautard, P. Yiou, M. Ghil, Singular-spectrum analysis: a toolkit for short, noisy chaotic signals. *Physica D,* 58, 95–126 (1992)
38. J.B. Elsner, A.A. Tsonis, *Singular Spectrum Analysis: A New Tool in Time Series Analysis,* Springer, New York (1996)
39. A.I. Mees, P.E. Rapp, L.S. Jennings, Singular-value decomposition and embedding dimension. *Phys. Rev. A,* 36, 340–346 (1987)
40. A.M. Albano, J. Muench, C. Schwartz, A.I. Mees, P.E. Rapp, Singular-value decomposition and the Grassberger-Procaccia algorithm. *Phys. Rev. A,* 38, 3017–3026 (1988)

41. M. Paluš, I. Dvořák, Singular-value decomposition in attractor reconstruction: pitfalls and precautions. *Physica D,* 221–234 (1992)
42. J.T. Joliffe, *Principal Component Analysis,* Springer, New York (1986)
43. R.W. Preisendorfer, *Principal Component Analysis in Meteorology and Oceanography,* Elsevier, Amsterdam (1988)
44. T.F. Cox, M.A.A. Cox, *Multidimensional Scaling,* 2nd ed., Chapman and Hall, London (2000)
45. G. Plaut, R. Vautard, Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere. *J. Atmos. Sci.,* 51, 210–236 (1994)
46. J. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction. *Science,* 290, 2319–2323 (2000)
47. S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding. *Science,* 290, 2323–2326 (2000)
48. M.A. Kramer, Nonlinear principal component analysis using autoassociative neural networks. *Am. Inst. Chem. Engin. J.,* 37, 233–243 (1991)
49. A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis,* Wiley, New York (2001)
50. S. Ciliberto, B. Nicolaenko, Estimating the number of degrees of freedom in spatially extended systems. *Europhys. Lett.,* 14, 303–308 (1991)
51. Y. Pomeau, Measurement of the information density in turbulence. *Compt. Rend. Acad. Sci. Ser. II,* 300, 239–241 (1985)
52. K. Kaneko, Spatiotemporal chaos in one-dimensional and two-dimensional coupled map lattices. *Physica D,* 37, 60–82 (1989)
53. G. Mayer-Kress, K. Kaneko, Spatiotemporal chaos and noise. *J. Stat. Phys.,* 54, 1489–1508 (1989)
54. S.M. Zoldi, H.M. Greenside, Karhunen-Loève decomposition of extensive chaos. *Phys. Rev. Lett.,* 78, 1687–1690 (1997)
55. S.M. Zoldi, J. Liu, K.M.S. Bajaj, H.S. Greenside, G. Ahlers, Extensive scaling and nonuniformity of the Karhunen-Loève decomposition for the spiral-defect chaos state. *Phys. Rev. E,* 58, 6903–6906 (1998)
56. M. Meixner, S.M. Zoldi, S. Bose, E. Schöll, Karhunen-Loève local characterization of spatiotemporal chaos in a reaction-diffusion system. *Phys. Rev. E,* 61, 1382–1385 (2000)
57. H. Varela, C. Beta, A. Bonnefort, K. Krischer, Transitions to electrochemical turbulence. *Phys. Rev. Lett.,* 94, 174104 (2005)
58. M. Thiel, *Recurrences: Exploiting Naturally Occurring Analogues,* PhD thesis, University of Potsdam (2004)
59. H.B. Wilson, M.J. Keeling, Spatial scales and low-dimensional deterministic dynamics. In: U. Dieckmann, R. Law, J.A.J. Metz (eds.), *The Geometry of Ecological Interactions: Simplifying Spatial Complexity,* Cambridge University Press, Cambridge, 209–226 (2000)
60. R. Proulx, P. Cote, L. Parrott, in prep.
61. S.A. Farmer, An investigation into the results of principal component analysis of data derived from random numbers. *The Statistician,* 20, 63–72 (1971)
62. J.R. Probert-Jones, Orthogonal pattern (Eigenvector) analysis of random and partly random fields. *Conf. Prob. Stat. Atmos. Sci.,* 3, 187–192 (1973)
63. J.M. Craddock, C.R. Flood, Eigenvectors for representing the 500 mb geopotential surface over the Northern Hemisphere. *Quart. J. Roy. Meteor. Soc.,* 95, 576–593 (1969)

AU: Please update Refs. [60], [65] [67] and [72]

64. P.V. Baily, B.H. KenKnight, J.M. Rogers, E.E. Johnson, R.E. Ideker, W.M. Smith, Spatial organization, predictability, and determinism in ventricular fibrillation. *Chaos,* 8, 103–115 (1998)
65. R. Donner, A. Witt, Temporary dimensions of multivariate data from paleoclimate records – A novel measure for dynamic characterization of long-term climate change. *Int. J. Bifurcation Chaos,* in press
66. R. Donner, A. Witt, Characterisation of long-term climate change by dimension estimates of multivariate palaeoclimatic proxy data. *Nonlin. Proc. Geophys.,* 13, 485–497 (2006)
67. R. Donner, Spatial correlations of hydro-meteorological records in a river catchment. In: J. Kropp, H.J. Schellnhuber (eds.), *In Extremis: Extremes, Trends and Correlations in Hydrology and Climate,* Springer, Berlin, 2008, (submitted)
68. C. Spearman, The proof and measurement of association between two things. *Amer. J. Psychol.,* 15, 72–101 (1904)
69. M.G. Kendall, A new measure of rank correlation. *Biometrika,* 30, 81–93 (1938)
70. B. Pompe, Measuring statistical dependences in a time series. *J. Stat. Phys.,* 73, 587–610 (1993)
71. R. Donner, A. Hofleitner, J. Höfener, S. Lämmer, D. Helbing, Dynamic stabilization and control of material flows in networks and its relationship to phase synchronization. *Proc. PhysCon 2007,* 1188 (2007).
72. R. Donner, Multivariate analysis of spatially heterogeneous phase synchronisation in complex systems: application to self-organised control of material flows in networks. *Eur. Phys. J. B,* (submitted)
73. S.M. Barbosa, M.E. Silva, M.J. Fernandes, Multivariate autoregressive modelling of sea level time series from TOPEX/Poseidon satellite altimetry. *Nonlin. Proc. Geophys.,* 13, 177–184 (2006)
74. A. Politi, A. Witt, Fractal dimension of space-time chaos. *Phys. Rev. Lett.,* 82, 3034–3037 (1999)
75. A.A. Tsonis, K.L. Swanson, P.J. Roebber, What do networks have to do with climate? *Bull. Amer. Meteorol. Soc.,* 87, 585–595 (2006)
76. A.A. Tsonis, K. Swanson, S. Kravtsov, A new dynamical mechanism for major climate shifts. *Geophys. Res. Lett.,* 34, L13705 (2007)

Part II

Applications in Oceanography and Seismology

Time Series Analysis of Sea-Level Records: Characterising Long-Term Variability

Susana M. Barbosa, Maria Eduarda Silva, and Maria Joana Fernandes

Universidade do Porto, Faculdade de Ciências, susana.barbosa@fc.up.pt

Abstract. The characterisation and quantification of long-term sea-level variability is of considerable interest in a climate change context. Long time series from coastal tide gauges are particularly appropriate for this purpose. Long-term variability in tide gauge records is usually expressed through the linear slope resulting from the fit of a linear model to the time series, thus assuming that the generating process is deterministic with a short memory component. However, this assumption needs to be tested, since trend features can also be due to non-deterministic processes such as random walk or long range dependent processes, or even be driven by a combination of deterministic and stochastic processes. Specific methodology is therefore required to distinguish between a deterministic trend and stochastically-driven trend-like features in a time series. In this chapter, long-term sea-level variability is characterised through the application of (i) parametric statistical tests for stationarity, (ii) wavelet analysis for assessing scaling features, and (iii) generalised least squares for estimating deterministic trends. The results presented here for long tide gauge records in the North Atlantic show, despite some local coherency, profound differences in terms of the low frequency structure of these sea-level time series. These differences suggest that the long-term variations are reflecting mainly local/regional phenomena.

AU: Please provide Keywords.

1 Introduction

Sea-level is a fundamental geophysical parameter that is relevant for many geosciences sub-disciplines including geodesy, oceanography, marine biogeosciences and climatology. Sea-level, or the height of the sea surface above a reference level, is measured by tide gauges at coastal sites and through radar altimeters on-board satellite platforms.

Tide gauges are the only historical source of precise sea-level measurements, some dating back to the 19th century. Tide gauge data have been traditionally used for navigational purposes, for the prediction of tides at a given location, and in the definition of levelling systems [1]. The increasing interest in environmental issues and climate change in the second half of the 20th century lead to a renewed interest and new applications of tide gauge

measurements, including the study of currents, storm surges and sea-level extremes, and the estimation of sea-level change (e.g. [2, 3, 4]).

Satellite altimetry [5] has the enormous advantage of being able to measure sea-level at a global scale, yielding a uniform space-time dataset of sea surface heights. However, the conversion of the radar measurements into a estimate of the height of the sea surface involves a large number of steps and therefore a considerable number of potential error sources, including errors in the geophysical corrections applied to the satellite measurements, errors in recovering the orbit of the satellite, and instrumental drifts affecting the stability of the altimeter [6, 7]. Furthermore, high quality and continuous satellite altimetry measurements are only available since 1993, hindering the analysis of long-term variability from satellite time series.

Sea-level is considered a key indicator of climate change and an important observational constraint for global climate models [8]. From a climate change perspective, the quantification of long-term sea-level variability is of paramount importance. According to the 4th assessment report of the IPCC [9], there is high confidence that the rate of sea-level rise has increased between the mid-19th and mid-20th centuries. For the 1993–2003 period, the rate of sea-level rise derived from satellite altimetry is significantly higher than the average rate, but not unprecedented, as concluded from inspection of the tide gauge record [10, 11]. Thus, it is unknown whether the higher rate from 1993 to 2003 is due to decadal variability or an increase in the longer-term trend.

In nearly all geosciences problems, the interaction between mathematical/statistical methodology and application-specific knowledge is vital for scientific advancement. However, a fruitful interplay between geosciences (physical) and time series analysis (statistical) perspectives is seldom easy to achieve, as emphasised in the still pertinent insight of Sir Gilbert Walker in 1927 "*There is, today, always a risk that specialists in two subjects, using languages full of words that are unintelligible without study, will grow up not only, without knowledge of each others work, but also will ignore the problems which require mutual assistance*" [12]. Although it is not straightforward, narrowing the gap between the physical and statistical perspectives is essential for the characterisation of long-term sea-level variability. From a geoscientific point of view, the main question is whether sea-level is rising or falling at a specific site, or in a given area, or globally. This question is often translated into the goal of determining long-term variability or even more often the "trend" in sea-level. However, from a time series analysis point of view, although a trend seems to be a feature that can be easily recognised in a time series plot, it lacks a precise, rigorous definition. A caricatured definition of trend is given by [13]: "*a trend is a trend is a trend ...*". In much of time series literature, trend is conceived as that part of a series which changes relatively slowly over time, and loosely defined as "long-term change in the mean level"; what is meant by long-term involves a subjective assessment, and different authors use the term trend in different ways (e.g. [14, 15]).

An alternative to the subjective notion of trend in time series analysis is to consider the somewhat opposite concept of stationarity. Stationarity is not only a mathematically well defined property of a time series, but actually a very fundamental one, since most methods of time series analysis are based on the theory of stationary stochastic processes. The geoscientific question of determining the trend in sea-level time series can therefore be translated into a first statistical question of whether the sea-level time series are stationary (no trend).

For a time series that cannot be considered stationary, the next obvious statistical question concerns the type of nonstationarity. Characterising a time series as nonstationary does not translate directly into having a deterministic (often linear) trend, although this notion is still quite common in geosciences, and in particular in sea-level research. In fact, many different processes, including deterministic, random walk, and long range dependent processes, can engender trend or trend-like features in a time series. Specific methodology is therefore required to distinguish between a deterministic trend and stochastically-driven trend-like features in a time series.

In most geosciences problems, the need to quantify long-term variability prompts the computation of linear trends trough ordinary linear regression. However, from a time series analysis point of view, when considering the estimation of a deterministic trend the time series character of the data needs to be taken into account in the regression framework. Autocorrelation is an ubiquitous feature in most geosciences time series, and needs therefore to be appropriately included in the estimation procedure.

In this chapter, the characterisation of long-term variability in sea-level is addressed from a time series analysis perspective. The methodology is described in Sect. 2. Time series of sea-level heights are first tested for stationarity through parametric statistical tests (Sect. 2.1). The scaling properties of the series are then examined in the wavelet domain (Sect. 2.2) in order to assess persistent or long-memory features. The estimation of linear trends in a time series context is considered in Sect. 2.3. Results on the characteristics of North Atlantic long-term sea-level variability are presented in Sect. 3 and discussed in Sect. 4.

2 Characterisation of Long-Term Variability

Long-term variability in sea-level records is often expressed through the linear slope resulting from the fit of a (deterministic) linear trend model to the sea-level time series (e.g. [16, 17]). Then, it is assumed that the process generating the sea level time series is deterministic with a short memory stochastic component. This assumption needs to be tested since trend features can also be due to non-deterministic processes such as random walk or long-range dependent processes, or even be driven by a combination of deterministic and

stochastic processes. The application of parametric statistical tests for discriminating between stationarity (no trend), a deterministic linear trend and a stochastic trend in the form of a random walk is addressed in Sect. 2.1. However, the discrimination between a deterministic trend and a stochastic alternative exhibiting significant low frequency variability such as long range dependence can be particularly challenging in the time domain and is more easily handled in the wavelet domain. The wavelet spectrum is blind to deterministic trends and is particularly useful for assessing the scaling features of a time series (e.g. [18, 19]), complementing the parametric statistical tests. Such wavelet-based approach is considered in Sect. 2.2. Finally, if the parametric tests and the wavelet analysis indicate that the assumption of a deterministic trend is plausible, the estimation procedure needs to take into account the time series nature of the data (Sect. 2.3).

2.1 Stationarity

The concept of stationarity plays a key role in time series analysis and is a basic assumption of most time series models. Nevertheless, most geosciences time series are non-stationary. One of the most common approaches to assess non-stationarity in the form of monotonic trends is the rank-based Mann-Kendall non-parametric test for a random process null hypothesis against a monotonic alternative (e.g. [20]). However, this test is not robust to autocorrelation, and serial correlation induces the identification of spurious trends. Parametric statistical tests of stationarity taking serial correlation into account have been developed mainly in econometrics, in order to discriminate between wide sense stationarity (no trend), deterministic trends plus stationary stochastic noise, and non-stationarity in the form of a unit root (including random walk).

Standard parametric tests for stationarity such as the Dickey-Fuller (DF) test [21], the augmented Dickey-Fuller (ADF) test [22] or the Phillips-Perron (PP) test [23] have been designed to test the null hypothesis of a random walk against a stationary alternative. The PP test has the advantage over the classical DF and ADF tests of handling serial correlation and heteroscedasticity directly in the test statistic. The Phillips-Perron (PP) test is based on the model

$$X_t = \eta + \beta t + \pi X_{t-1} + \psi_t \tag{1}$$

with the unit root null hypothesis expressed by $H_0 : \pi = 1$; the stationary process ψ_t is not assumed to be white noise and serial correlation and heteroscedasticity in the ψ_t term are handled directly in the test statistic.

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [24] complements the previous tests by testing a stationary null hypothesis in the form of a constant level or a deterministic trend. The test assumes that the time series can be decomposed into the sum of a deterministic trend, a random walk (r_t) and a stationary stochastic noise (ν_t):

$$X_t = \beta t + r_t + \nu_t \qquad r_t = r_{t-1} + \varepsilon_t \qquad \nu_t \sim \mathcal{N}(0, \sigma_\nu^2) \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad (2)$$

allowing to test a level-stationary hypothesis

$$H_0 : \sigma_\varepsilon^2 = 0 \qquad (\beta = 0) \qquad H_1 : \sigma_\varepsilon^2 \neq 0 \quad (3)$$

and a trend-stationary hypothesis

$$H_0 : \sigma_\varepsilon^2 = 0 \qquad (\beta \neq 0) \qquad H_1 : \sigma_\varepsilon^2 \neq 0. \quad (4)$$

Statistical tests assuming a random walk null hypothesis and tests assuming a stationary null hypothesis are complementary and therefore their joint application is recommended. In this work, the PP test and the KPSS test are jointly applied for testing stationarity of a sea-level time series. If both tests reject the null hypothesis then alternative parametrisations such as long range dependence should be considered. If both tests fail to reject the null hypothesis, then the time series (or the tests) are not sufficiently informative for discriminating the kind of stationary behaviour. Rejection of the unit root hypothesis in the PP test and no rejection of KPSS's test null hypothesis points to a deterministic trend, while no rejection of the unit root null hypothesis in the PP test and rejection of KPSS's null hypothesis indicates a unit root process.

2.2 Scaling

Long range dependence (or long-memory) was first noted in hydrology from the study of the water levels of the Nile river as the tendency for a flood year to be followed by another flood year [25, 26]. Long range dependence is one of the most important manifestations of scale invariance. A process exhibits scale invariance if its spectral density function S is a power law for frequencies approaching zero:

$$lim_{f \to 0} S(f) = C f^\alpha \quad (5)$$

where $C > 0$ and α are constants. The value of the scaling exponent α defines not only long memory but also other kinds of scaling behaviour (Table 1). Thus, the estimation of the scaling exponent of a time series provides an alternative and complementary way of characterising its low-frequency structure.

The discrete wavelet transform is a natural tool for scaling processes, since the wavelet spectrum (corresponding to the variance of the wavelet coefficients as a function of scale) provides a summary of the spectral density function, reproducing in the wavelet domain the power laws underlying the scaling processes [18, 19]. Furthermore, the discrete wavelet transform is insensitive to deterministic features and acts as a decorrelating transform, converting long range dependence in the time domain into short range statistical dependence in the wavelet domain, rendering its application to the analysis of long range

Table 1: Values of the scaling exponent α for different stochastic processes (adapted from [18], p. 286).

α	Process
$\alpha = 0$	white noise
$\alpha \geq 0$	short range stationary
$-1 < \alpha < 0$	long memory
$\alpha = -2$	random walk
$\alpha = -1$	1/f or flicker noise

dependence particularly appealing [18]. In this study, the scaling exponent α is estimated from the slope of the wavelet spectrum as described in [27]. The estimation of the scaling exponent from the wavelet spectrum rather than from the direct Fourier spectrum is particularly advantageous in the case of nonstationarity (e.g.[28, 29]). Furthermore, estimates based on the wavelet spectrum are more robust to the presence of trends and periodicties [30].

2.3 Trend Estimation

Long-term sea-level variability is usually quantified through a deterministic linear model, that can be written in matrix form as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(0, \Sigma) \tag{6}$$

where $\mathbf{y}$ is a length-n vector of sea level observations, X is a $n \times 2$ matrix ($X = [\mathbf{1}\ \mathbf{t}]$ and $\mathbf{t}$ denotes time), $\boldsymbol{\beta}$ is the vector of parameters and $\boldsymbol{\varepsilon}$ is a length-n vector of errors with symmetric and positive definite covariance matrix Σ.

This linear model is commonly fitted to a sea-level time series by ordinary least squares (OLS). The regression model underlying ordinary least squares (denoted OLS model) is the linear model (6) with uncorrelated errors, i.e diagonal covariance matrix $\Sigma = \sigma^2 I$. Then the trend estimator $\widehat{\boldsymbol{\beta}}_{OLS}$ is given by

$$\widehat{\boldsymbol{\beta}}_{OLS} = (X^T X)^{-1} X^T \mathbf{y} \tag{7}$$

with variance

$$V[\widehat{\boldsymbol{\beta}}_{OLS}] = \sigma^2 (X^T X)^{-1}. \tag{8}$$

In the OLS model the observations are assumed to be independent, but this assumption is not valid, in general, for a time series. For a non-diagonal covariance matrix, $\Sigma \neq \sigma^2 I$, the estimator of $V[\widehat{\boldsymbol{\beta}}_{OLS}]$ (8) is biased and inconsistent, affecting statistical significance and the estimated standard errors.

The effect of serial correlation on regression is a well known problem. One of the earliest approaches for handling serial correlation in time series regression was through transformations to the ordinary least squares estimator,

such as the Cochrane-Orcutt method [31]. Another approach to deal with serial correlation is to consider an effective sample size: the number of degrees of freedom is reduced by considering instead of the original series length an effective sample size computed from e.g. lag-1 auto-correlation. This approach was used in the estimation of the linear trend in Key West tide gauge record by [16] and to account for serial correlation in altimetry time series (e.g. [32]). Still within the ordinary least squares framework, corrections to the mean and variance of the estimates can be derived under an assumed autocorrelation structure in order to correct for serial correlation [6, 33].

Generalised least squares (GLS) is a more general approach for the estimation of a linear trend. The regression model underlying generalised least squares (denoted GLS model) is the linear model (6) with correlated errors, i.e. non-diagonal covariance matrix $\Sigma \neq \sigma^2 I$. The trend estimator $\widehat{\boldsymbol{\beta}}_{GLS}$ is then given by

$$\widehat{\boldsymbol{\beta}}_{GLS} = (X^T \widehat{\Sigma}^{-1} X)^{-1} X^T \widehat{\Sigma}^{-1} \mathbf{y} \tag{9}$$

with variance

$$V[\widehat{\boldsymbol{\beta}}_{GLS}] = (X^T \widehat{\Sigma}^{-1} X)^{-1}. \tag{10}$$

Since estimation of the covariance matrix Σ requires $n(n+1)/2$ parameters, $\widehat{\Sigma}$ cannot be obtained from a sample of size n and restrictive parametrisations must be assumed, through specification of a stationary process for the error correlation structure. In this work, a set of four different stationary processes is considered in the specification of Σ: a first order autoregressive process, a second order autoregressive process, a first order moving average process and a first order autoregressive/moving average process. Although the restriction to these low order models is limiting, simulation studies show that differences in estimation efficiency between correct and misspecified correlation structures are small, suggesting that there may not much to be gained in trying very high order parametrisations for the errors correlation structure [34]. For each time series the model for the error correlation structure is selected from this set of models using the Akaike Information Criterion (AIC). Numerical maximisation of log-likelihood allows the simultaneous estimation of both regression coefficients and parameters of the error covariance process.

3 Long-Term Variability of North Atlantic Sea-Level

3.1 Data

Sea-level time series from sixteen tide gauge stations in the North Atlantic with long (> 50 years) and continuous records (gaps < 1 year and missing values $< 2.5\%$) are analysed (Fig. 1, Table 2). Although longer time series are available for some of the records (Brest, Halifax) shorter periods have been selected for analysis in order to avoid large gaps in the time series and maintain

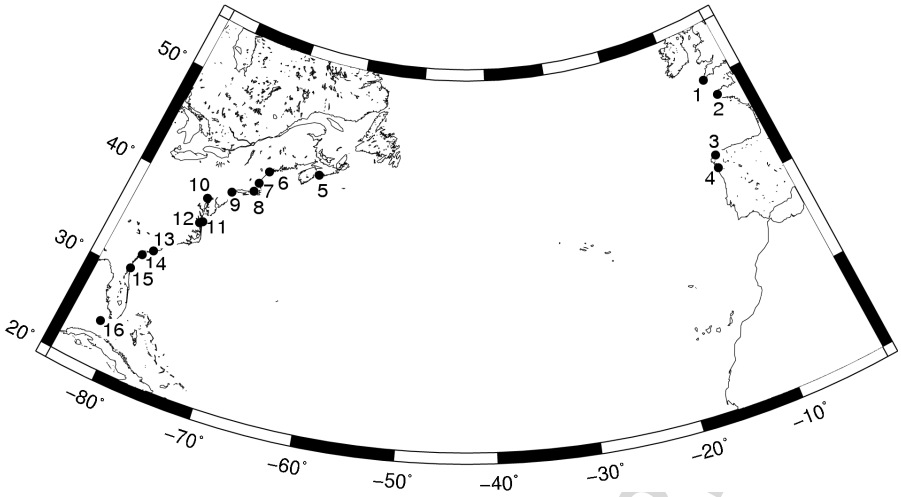


Fig. 1: Map of tide gauge locations: 1-Newlyn; 2-Brest; 3-Coruña; 4-Vigo; 5-Halifax; 6-Portland; 7-Boston; 8-Newport; 9-New York; 10-Baltimore; 11-Kiptopeke; 12-Hampton; 13-Charleston; 14-Fort Pulaski; 15-Mayport; 16-Key West.

coherency with criteria for missing observations. Monthly time series are obtained from the Permanent Service for Mean Sea Level (PSMSL) database [35] of Revised Local Reference (RLR) tide gauge measurements. Seasonality is handled by subtracting from each observation the average value for the corresponding month. Missing observations are filled-in by linear interpolation.

Table 2: Analysed tide gauge records.

	Longitude (E)	Latitude (N)	Period	No. observations	% missing
Newlyn	−05.55	50.10	1916–2003	1056	0.19
Brest	−04.50	48.38	1953–2000	576	0.3
Coruña	−08.40	43.37	1944–2001	696	1.6
Vigo	−08.73	42.23	1944–2001	696	0.86
Halifax	−63.58	44.67	1920–2002	996	1.30
Portland	−70.25	43.67	1912–2003	1104	0.27
Boston	−71.05	42.35	1921–2003	996	0.80
Newport	−71.33	41.50	1931–2003	876	1.26
New York	−74.02	40.70	1927–2003	924	1.08
Baltimore	−76.58	39.37	1903–2003	1212	0.16
Kiptopeke	−75.98	37.17	1952–2003	624	0.80
Hampton	−76.33	36.95	1928–2003	912	0
Charleston	−79.93	32.78	1922–2003	984	0
Fort Pulaski	−80.90	32.03	1935–2003	828	1.21
Mayport	−81.43	30.40	1929–2000	864	0.35
Key West	−81.80	24.55	1913–2003	1092	0.73

Seasonally-adjusted sea-level records exhibit for most stations a seemingly increasing trend (Fig. 2). Stationarity is tested in the time domain in Sect. 3.2 and scaling features of the series are analysed in the wavelet domain in Sect. 3.3. Linear trends are estimated in Sect. 3.4.

3.2 Stationarity Tests

Stationarity of sea level time series is examined through PP and KPSS statistical tests. The results for the two tests, in terms of the corresponding p-values, are given in Table 3.

For all records the PP test rejects the unit root null hypothesis (p-value < 0.05) indicating that a stochastic trend from an underlying random walk process can be discarded. For Newlyn, Coruña, Vigo, Kiptopeke, Hampton and Fort Pulaski the KPSS null hypothesis is not rejected indicating a deterministic trend, but long range dependence, often present in sea-level records [36, 37], cannot be excluded. For the remaining records, the KPSS test leads to the rejection of trend stationarity, indicating that these time series are not well represented either by a stationary, random walk or trend-stationary process and therefore that an alternative parametrisation (such as long range dependence) needs to be considered.

3.3 Scaling Exponent

Sea-level is known to exhibit scale-invariance over a wide range of frequencies [37]. In order to examine the scaling properties of the sea-level series, a

Table 3: The p-values from KPSS (H_0: deterministic trend) and PP (H_0: random walk) statistical tests.

	KPSS test	PP test
Newlyn	> 0.1	< 0.01
Brest	< 0.01	< 0.01
Coruña	> 0.1	< 0.01
Vigo	0.067	< 0.01
Halifax	< 0.01	< 0.01
Portland	< 0.01	< 0.01
Boston	< 0.01	< 0.01
Newport	0.023	< 0.01
New York	< 0.01	< 0.01
Baltimore	< 0.01	< 0.01
Kiptopeke	> 0.1	< 0.01
Hampton	> 0.1	< 0.01
Charleston	< 0.01	< 0.01
Fort Pulaski	> 0.1	< 0.01
Mayport	0.012	< 0.01
Key West	0.05	< 0.01

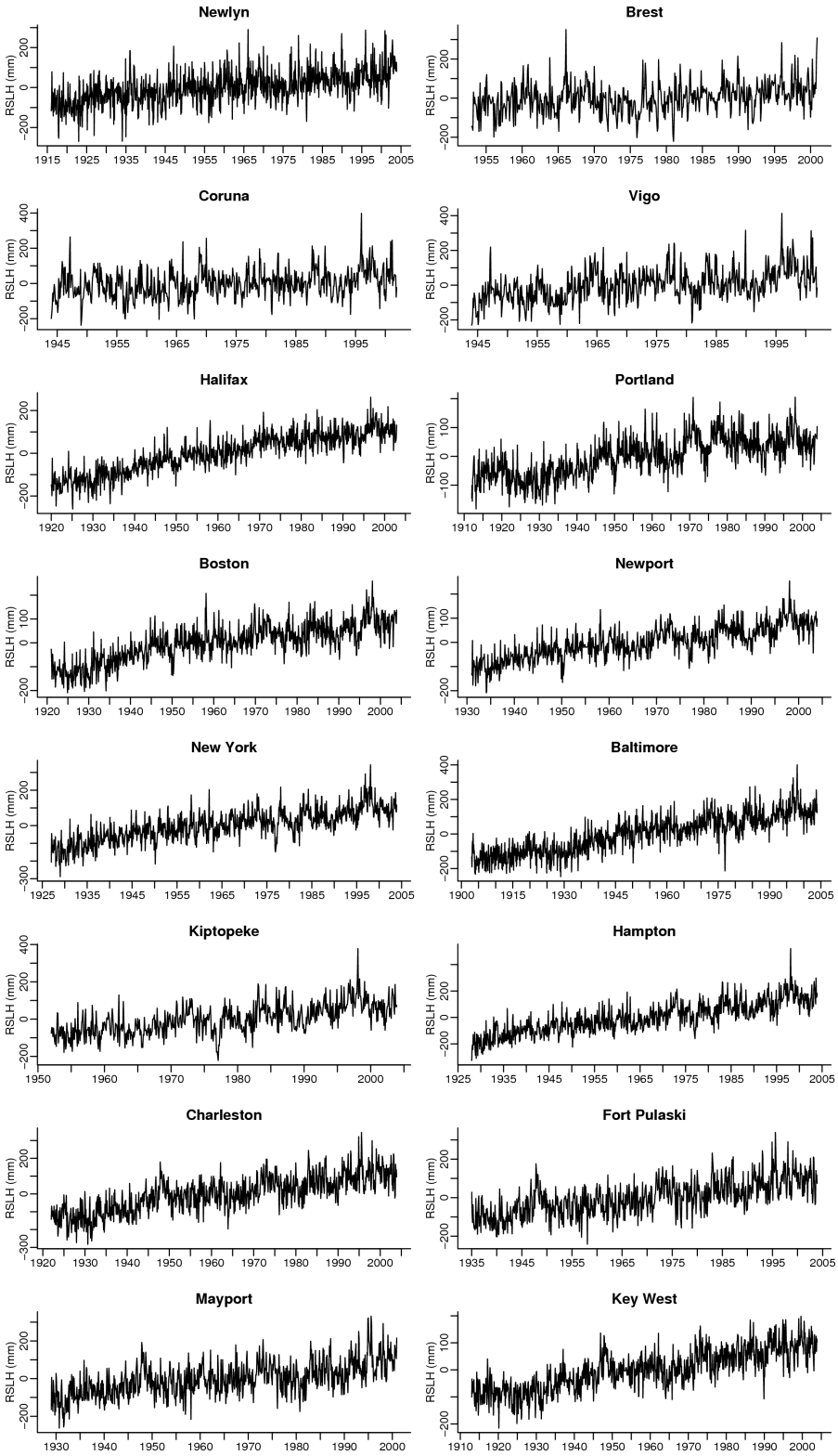


Fig. 2: Monthly (seasonally-adjusted) time series of relative sea level heights (RSLH).

wavelet analysis based on the maximal overlap version of the discrete wavelet transform [38] is carried out using a Daubechies wavelet filter of length $L = 4$ [39]. Brick-wall boundary conditions are applied for unbiased estimates [40]. The wavelet spectrum is constructed by representing on a log10-log10 graph the wavelet variance estimated from the resulting wavelet coefficients [41] versus scale, along with the corresponding 95% confidence intervals (Fig. 3). An alternative wavelet-based estimation of the scaling exponent would consist in considering the continuous rather than the discrete wavelet transform in order to have a larger number of scales over which to estimate the scaling exponent (e.g. [42]).

For most records the wavelet spectrum exhibits a linear behaviour within some scale range. The slope of the wavelet spectrum is estimated by a weighted least squares estimator that takes into account the large sample properties

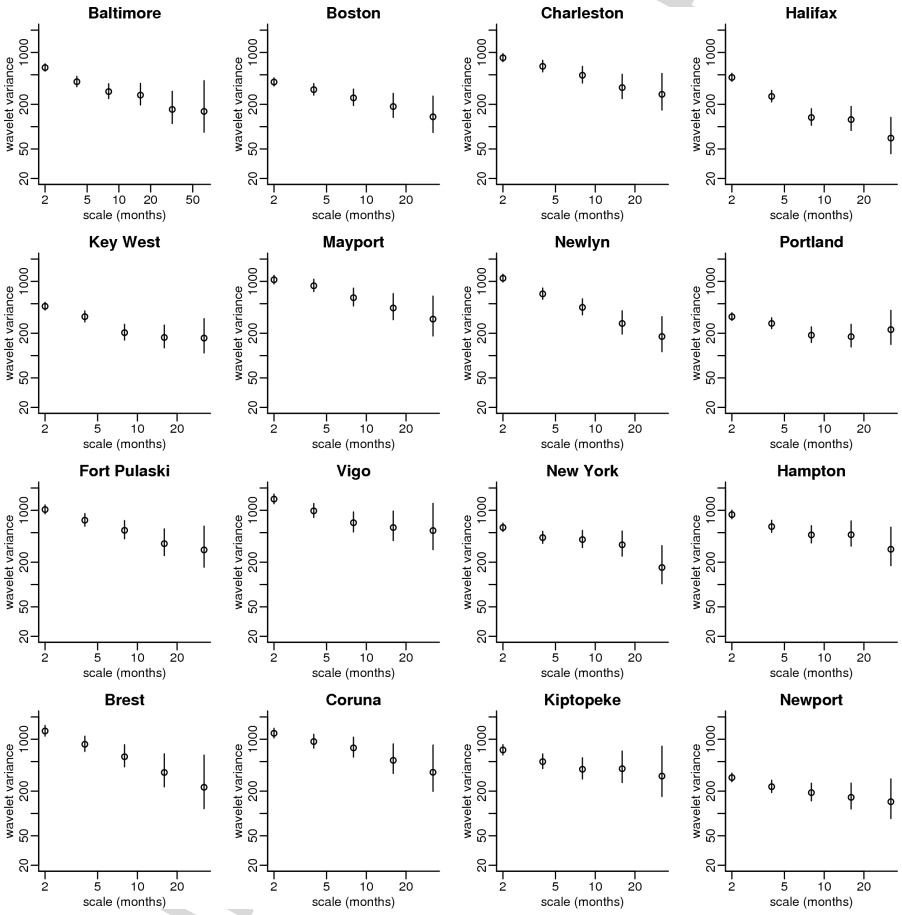


Fig. 3: Wavelet spectrum for each tide gauge record.

Table 4: Estimates of the scaling exponent α and associated standard error (s.e.).

	$\widehat{\alpha}$	s.e.	Bootstrap s.e
Newlyn	−0.34	0.062	0.014
Brest	−0.41	0.088	0.023
Coruña	−0.62	0.079	0.051
Vigo	−0.57	0.079	0.086
Halifax	−0.28	0.064	0.089
Portland	−0.72	0.061	0.160
Boston	−0.64	0.064	0.020
Newport	−0.70	0.069	0.035
New York	−0.68	0.067	0.130
Baltimore	−0.55	0.053	0.041
Kiptopeke	−0.67	0.084	0.087
Hampton	−0.63	0.067	0.055
Charleston	−0.59	0.064	0.031
Fort Pulaski	−0.53	0.071	0.025
Mayport	−0.60	0.070	0.030
Key West	−0.54	0.061	0.120

of the wavelet variance estimator ([18], pp. 374–378), yielding the scaling exponent (α) and corresponding standard error (Table 4). For comparison, standard errors derived by bootstrap resampling (e.g. [43, 44]) considering 50 bootstrap replicates are also included in Table 4. For some stations (Key West, Portland and New York) the bootstrap standard errors are fairly large, suggesting that the power-law assumption is questionable in these cases. For all records the values obtained for the scaling exponent α (Table 4) are consistent with long memory behaviour (within the $]-1, 0[$ range) but with distinct degrees of stochastic persistence from weak $(]-0.4, -0.2[)$ to moderate $(]-0.6, -0.4[)$ and strong persistence $(]-0.8, -0.6[)$.

3.4 Linear Trends

For the records for which a deterministic linear trend is plausible, the rate of sea-level change is estimated by generalised least squares. The estimates obtained from ordinary least squares regression are also shown for comparison (Table 5). The main difference in using generalised rather than ordinary least squares lies in the magnitude of the estimated standard errors: as a result of positive serial correlation, ordinary least squares errors are biased downward. This influences the statistical significance of the estimated trends, although the trend values themselves are not affected.

Here, linear trends have been derived assuming a deterministic trend (as suggested by the statistical tests) and a short-range dependent process for the stochastic component. This assumption of serially-correlated (but not long-range correlated errors) is however, only valid for Newlyn and Halifax records,

Table 5: Linear trends (mm/yr) estimated by GLS. The error correlation structure is represented for Key West by a AR(1) process and for the remaining series by a ARMA(1,1) process.

	$\widehat{\beta}_{GLS}$	s.e$_{GLS}$	$\widehat{\beta}_{OLS}$	s.e$_{OLS}$
Newlyn	1.73	0.13	1.72	0.083
Coruña	1.40	0.33	1.37	0.17
Vigo	2.57	0.36	2.56	0.18
Halifax	3.25	0.18	3.29	0.059
Kiptopeke	3.34	0.40	3.36	0.16
Hampton	4.37	0.24	4.33	0.10
Fort Pulaski	2.98	0.20	3.00	0.12
Key West	2.23	0.086	2.23	0.052

since the remaining stations exhibit long range dependence (although in the form of moderate stochastic persistence). This means that a deterministic linear model can be estimated for Newlyn and Halifax (e.g. as in [17] and [45]) while for the remaining records the linear model could be adapted in order to include a long range-dependent stochastic component [46]. An alternative approach for estimating linear slopes along with realistic uncertainties would be the joint estimation of the linear slope and of the scaling exponent by maximum likelihood.

Table 5 indicates that the sea-level slope at Coruña is considerably lower than the one obtained for Vigo, although the two sites are very close. According to [47] the discrepancy is explained by a jump in the reference level of the Coruña record. On the western boundary, the estimated sea level trends are higher for the stations in Chesapeake Bay (Kiptopeke, Hampton). The large linear trends obtained at Chesapeake Bay may result from land subsidence. Local subsidence is caused by groundwater extraction [48] while regional subsidence of the entire Mid-Atlantic coast results from post-glacial adjustment [49]. Furthermore, Chesapeake Bay has been identified as a tectonically active area [50]. At Hampton subsidence is possibly enhanced by compaction of the filling of a large buried impact crater [51].

4 Discussion

Long-term sea-level variability has been characterised through the application of parametric statistical tests for stationarity, wavelet analysis for assessing scaling features, and generalised least squares for estimating deterministic trends.

Parametric tests of stationarity are based on asymptotic properties which are not necessarily met in practice, and therefore require a large sample size to be efficient. Therefore, results from statistical tests alone must be viewed

with caution, particularly for short records (Brest, Coruña, Vigo). Moreover, statistical tests are designed in a way that the null hypothesis is rejected only if there is strong evidence against it. No rejection of a deterministic trend does not indicate the existence of a trend, but only that such feature cannot be ruled out. Although the application of parametric statistical tests can give some insight on the stationary features of a time series, in practice the distinction between a nearly non-stationary stochastic process, such as a long range dependent process or a near unit root process, and a non-stationary deterministic process is very difficult, since both type of processes yield similar features: an empirical autocorrelation function dying out slowly and a spectrum with large spectral content at zero frequency. Therefore, these tests are known to have low power, particularly against near unit root and fractionally differenced alternatives [52, 53, 54]. The discrete wavelet transform on the other hand is blind to polynomial trends and is particularly useful for assessing the scaling features of a time series, including long range dependence.

The results obtained from the stationarity tests and the wavelet analysis show that the analysed tide gauge records exhibit distinct low-frequency characteristics. The stationarity tests indicate a deterministic trend for Newlyn, Coruña, Vigo, Kiptopeke, Hampton, Fort Pulaski and Key West (although only marginally for Key West). Except for Newlyn, all the other records also exhibit stochastic variability in the form of long range dependence.

For Newlyn, the stochastic dependence is consistent with only a very weak long-memory process. Thus the trend component for Newlyn can be represented by a (deterministic) linear trend plus a stochastic stationary noise. In the case of Brest, the value of α is similar to the value from the nearby station of Newlyn, although the results from the stationarity tests are quite different for the two records: for Brest, the trend stationarity null hypothesis is rejected while for Newlyn a deterministic trend cannot be ruled out. This is a consequence of the different length of the two time series and of the sensitivity of the stationarity tests to time series length; the wavelet approach is fairly insensitive to sample size, indicating a similar stochastic component for the two records. For Portland, Boston, Newport and New York the stochastic variability is characterised by strong long range dependence. While a deterministic feature cannot be entirely ruled out, the persistent behaviour from the stochastic component alone is able to explain the trend in these records with no need for a deterministic generating process. Thus a long memory model rather than a deterministic linear model should be considered in the description of sea level variability from these records. Since for these sites the changes in the relative height of the sea surface are stochastically persistent in time, eventual disturbances in the coastal system can impact long-term sea-level variability, influencing sea-level variations even after the actual disturbance has ceased.

The results presented here for long tide gauge records in the North Atlantic show, despite some local coherency, profound differences in terms of the low frequency structure of these sea-level time series. These differences suggest

that the trend structure reflects mainly local/regional phenomena. Therefore, each record must be analysed individually, and results from several tide gauge records (for example to obtain a regional estimate of sea-level change in the North Atlantic) should only be jointly considered if the corresponding records exhibit the same type of low-frequency properties.

The characterisation of long-term sea-level variability is pertinent for the understanding, estimation and forecasting of sea-level change. For realistic estimates of the long-term rate of sea-level and for forecasting future variations, both deterministic and stochastic contributions need to be taken into account. This is a challenging task, but this study shows how the combination of different methodologies, in time and wavelet domain, can be used to extract additional information in terms of low-frequency characteristics from a sea-level time series.

References

1. D. Pugh: *Tides, Surges and Mean Sea-Level* (John Wiley and Sons, Chichester UK 1996)
2. D. Pugh: *Changing Sea Levels – Effects of Tides, Weather and Climate* (Cambridge University Press, Cambridge New York 2004)
3. P.L. Woodworth, D.T. Pugh, M.P. Meredith et al., Sea level changes at Port Stanley, Falkland Islands, *J. Geophys. Res.* **110**, C06013 (2005)
4. G. Woppelmann, S. Zerbini, M. Marcos, Tide gauges and Geodesy: a secular synergy illustrated by three present-day case studies, *C. R. Geosci.* **338**, 980–991 (2006)
5. L. Fu, A. Cazenave: *Satellite Altimetry and Earth Sciences* (Academic press, San Diego 2001)
6. B.D. Beckley, F.G. Lemoine, S. B. Luthcke et al., A reassessment of global and regional mean sea level trends from TOPEX and Jason-1 altimetry based on revised reference frame and orbits, *Geophys. Res. Lett.* **34**, L14608 (2007)
7. M.J. Fernandes, S.M. Barbosa, C. Lazaro, Impact of altimeter data processing on sea level studies, *Sensors* **6**, 131–163 (2006)
8. A. Cazenave, R.S. Nerem, Present-day sea level change: observations and causes, *Rev. Geophys.* **42**, RG3001 (2004)
9. N.L. Bindoff, J. Willebrand, V. Artale et al., 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* ed by S. Solomon, D. Qin, M. Manning, et al. (Cambridge University Press, Cambridge, United Kingdom and New York, USA 2007)
10. S.J. Holgate, P.L. Woodworth, Evidence for enhanced coastal sea level rise during the 1990s, *Geophys. Res. Lett.* **31**, L07305, 2004
11. S.J. Holgate, On the decadal rates of sea level change during the twentieth century, *Geophys. Res. Lett.* **34**, L0160 (2007)
12. R. Katz, Sir Gilbert Walker and a connection between El Nino and statistics, *Statistical Science* **17**, 97–112 (2002)

13. A. Cairncross, *Essays in Economic Management*, (Allen & Unwin, London 1971)
14. C. Chatfield, *The Analysis of Time Series: An Introduction*, 6th edn, (Chapman & Hall/CRC, Boca Raton 2003)
15. A.C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, (Cambridge University Press, Cambridge New York 1991)
16. G. Maul, D. Martin, Sea level rise at Key West, Florida, 1846–1992: America's longest intrument record?, *Geophys. Res. Lett.* **20**, 1955–1958 (1993)
17. P.L. Woodworth, M.N. Tsimplis, R.A. Flather et al., A review of the trends observed in British Isles mean sea level data measured by tide gauges, *Geophys. J. Int.* **136**, 651–670 (1999)
18. D.B. Percival, A.T. Walden, *Wavelet Methods for Time Series Analysis* (Cambridge University Press, Cambridge New York 2000)
19. P. Abry, P. Flandrin, M.S. Taqqu et al., Self-similarity and long range dependence through the wavelet lens. In: *Theory and applications of long range dependence*, ed by P. Doukhan, G. Oppenheim, M.S. Taqqu (Birkhäuser, Boston 2003) pp. 527–556
20. D. Burn, A. Mohamed, H. Elnur, Detection of hydrologic trends and variability, *J. Hydrol.* **255**, 107–122 (2002)
21. D.A. Dickey, W.A. Fuller, Distribution of the estimators for autoregressive time series with a unit root, *J. Am. Stat. Assoc.* **74**, 427–431 (1979)
22. S.E. Said, D.A. Dickey, Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika* **71**, 599–607 (1984)
23. P.C.B. Phillips, P. Perron, Testing for a unit root in time series regression, *Biometrika* **75**, 335–346 (1988)
24. D. Kwiatkowski, P. Phillips, P. Schmidt et al., Testing the null hypothesis of stationarity against the alternative of a unit root, *J. Econometrics*, 159–178 (1992)
25. H.E. Hurst, Long term storage capacity of reservoirs, *Trans. Amer. Soc. Civil Eng.* **116**, 770–799 (1951)
26. B.B. Mandelbrot, J.R. Wallis, Fractional Brownian motion, fractional noises and applications, *Water Resour. Res.* **4**, 909–918 (1968)
27. S.M. Barbosa, M.J. Fernandes, M.E. Silva, Long range dependence in North Atlantic sea level, *Physica A* **371**, 725–731 (2006)
28. B. Whitcher, M.J. Jensen, Wavelet estimation of a local long memory parameter, *Explor. Geophys.* **31**, 94–103 (2000)
29. S. Stoev, M.S. Taqqu, C. Park, G. Michailidis, J.S. Marron, LASS: a tool for the local analysis of self-similarity, *Comput. Stat. Dat. An.* **50**, 2447–2471 (2006)
30. A. Almasri, B. Holmquist, S. Hussain, Impact of the periodicity and trend on the FD parameter estimation, *J. Stat. Comput. Sim.* **77**, 79–87 (2007)
31. D. Cochrane, G. Orcutt, Application of least squares regression to relationships containing autocorrelated error terms, *J. Am. Stat. Assoc.* **44**, 32–61 (1949)
32. R.S. Nerem, Measuring very low frequency sea level variations using satellite altimeter data, *Global and Planet. Change* **20**, 157–171 (1999)
33. J. Lee, R. Lund, Revisiting simple linear regression with autocorrelated errors, *Biometrika* **91**, 240–245 (2004)
34. S. Koreisha, Y. Fang, Generalized least squares with misspecified serial correlation structures, *J. R. Statist. Soc. B* **63**, 515–531 (2001)
35. P.L. Woodworth, R. Player, The permanent service for mean sea level: an update to the 21st century, *J. Coastal Res.* **19**, 287–295 (2003)

36. A.T. Hsui, K.A. Rust, G.D. Klein, A fractal analysis of quaternary, cenozoic-mesozoic, and Late Pennsylvanian sea level changes, *J. Geophys. Res.* **98**, B12, 21963–21968 (1993)
37. C.G.A. Harrisson, Power spectrum of sea level change over fifteen decades of frequency, *Geochem. Geophy. Geosy.* **3**, 1047 (2002)
38. D.B. Percival, H. Mojfeld, Analysis of subtidal coastal sea level fluctuations using wavelets, *J. Am. Stat. Assoc.* **92**, 868–880 (1997)
39. I. Daubechies, Orthonormal bases of compactly supported wavelets, *Commun. Pur. Appl. Math.* **41**, 909–996 (1988)
40. R.W. Lindsay, D.B. Percival, D.A. Rothrock, The discrete wavelet transform and the scale analysis of the surface properties of sea ice, *IEEE T. Geosci. Remote* **34**, 771–787 (1996)
41. D.B. Percival, On the estimation of the wavelet variance, *Biometrika* **82**, 619–631 (1995)
42. T.-H. Li, H.-S. Oh, Wavelet spectrum and its characterization property for random processes, *IEEE T. Inform. Theory* **48**, 2922–2937 (2002)
43. B. Efron, R. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall/CRC, Boca Raton 1993)
44. A.C. Davison, D.V. Hinkley, *Bootstrap Methods and Their Application* (Cambridge University Press, Cambridge 1997)
45. K. Hilmi, T. Murty, M.E. Sabh et al., Long-term and short-term variations of sea level in eastern Canada: a review, *Mar. Geod.* **25**, 61–78 (2002)
46. P.M. Robinson, F.J. Hidalgo, Time series regression with long range dependence, *Ann. Stat.* **25**, 77–104 (1997)
47. M. Marcos, D. Gomis, S. Monserrat et al., Consistency of long sea level time series in the northern coast of Spain, *J. Geophys. Res.* **110**, C03008 (2005)
48. V. Gornitz, L. Seeber, Vertical crustal movements along the east coast, North America, from historic and Late Holocene sea level data, *Tectonophysics* **178**, 127–150 (1990)
49. W.R. Peltier, Global sea level rise and glacial isostatic adjustement: an analysis of data from the east coast of North America, *Geophys. Res. Lett.* **23**, 717–720 (1996)
50. B. Douglas, Global sea level change: determination and interpretation, *Rev. Geophys.* **33**, 1425–1432 (1995)
51. C. Koeberl, C.W. Poag, W.U. Reimold et al., Impact origin of the Chesapeake Bay structure and the source of the North American tektites, *Science* **271**, 1263–1266 (1996)
52. F. Diebold, G. Rudebusch, On the power of Dickey-Fuller tests against fractional alternatives, *Econ. Lett.* **35**, 155–160 (1991)
53. D. Lee, P. Schmidt, On the power of the KPSS test of stationarity agains fractionally-integrated alternatives, *J. Econometrics* **73**, 285–302 (1996)
54. S.J. Leybourne, P. Newbold, The behaviour of Dickey-Fuller and Phillips-Perron tests under the alternative hypothesis, *Economet. J.* **2**, 92–106 (1999)

Empirical Global Ocean Tide and Mean Sea Level Modeling Using Satellite Altimetry Data Case Study: A New Empirical Global Ocean Tide and Mean Sea Level Model Based on Jason-1 Satellite Altimetry Observations

Alireza A. Ardalan[1] and Hassan Hashemi[1]

Department of Surveying and Geomatics Engineering, Center of Excellence in Surveying Engineering and Disaster Prevention, Faculty of Engineering, University of Tehran, P. O. Box: 11155-4563, Tehran-Iran, Tel: 0098-21-82084383, Fax: 0098-21-88337642, ardalan@ut.ac.ir and hashemih@ut.ac.ir

Abstract. In this contribution an empirical approach to global ocean tide and Mean Sea Level (MSL) modeling based on satellite altimetry observations is presented with all details. Considering the fact that the satellite altimetry technique can provide sea level observations at the global scale, spherical harmonics defined for the whole range of spherical coordinates ($0 \leq \lambda \leq 2\pi$, and $-\pi/2 \leq \phi \leq +\pi/2$) could be among the possible choices for global ocean tide modeling. However, when applied for modeling of global ocean tide, spherical harmonics lose their orthogonality due to the following reasons: (1) Observation of sea surface is made over discrete points, and not as a continuous function, which is needed for having the orthogonality property of spherical harmonics in functional space. (2) The range of application of spherical harmonics for global ocean tide modeling is limited to the sea areas covered by satellite altimetry observations and not the whole globe, which is also required for the fluffiness of the orthogonality of spherical harmonics. In this contribution we show how a set of orthonormal base functions at the sea areas covered by the satellite altimetry observations can be derived from spherical harmonics in order to solve the lack of orthogonality. Using the derived orthonormal base functions, a global MSL model, and empirical global ocean tide models for six major semidiurnal and diurnal tidal constituents, namely, S2, M2, N2, K1, P1, and O1 as well as three long term tidal components, i.e., Mf, Mm, and Ssa, are developed based on six years of Jason-1 satellite altimetry sea level data as a numerical case study.

Keywords: Empirical tidal modeling, Harmonic analysis, Spherical harmonics, Orthonormal base functions, Gram-Schmidt, Tidal constituents, Mean Sea Level (MSL), Satellite altimetry, TOPEX/Poseidon, Jason-1.

1 Introduction

Accurate knowledge of the ocean tide is essential in various geodetic and geophysical applications, especially for removal of tidal effects from terrestrial, airborne, and satellite gravimetry observations. Moreover, computation of the Mean Sea Level (MSL) as the zero frequency tidal constituent allows, together with global knowledge on Earth's gravitational field, e.g. in terms of a geopotential model, to arrive at various products such as the geoid's gravity potential value W_0, Sea Surface Topography (SST) and the marine geoid from sea level observations (for example [1, 2, 3, 4, 5]). Besides, engineering activities at sea, such as building harbours, offshore oil platforms, placing under water pipelines, and many other applications require accurate knowledge about sea level and currents, which cannot be derived without a thorough knowledge of the tide. Information about the tide, currents and seawater circulation are also of great importance for navigation. For a review on sea level observations and the history of tidal analysis we refer to NOAA web-page at http://co-ops.nos.noaa.gov/predhist.html. Due to the importance of tidal studies, establishment of tide gauge stations at harbours has a long time history. For instance historical tide gauge stations are in operation since early 19th century at San Francisco (USA), Cascais (Portugal), Brest (France), Newlyn (UK), Hoek Van Holland (Netherlands), Stockholm (Sweden), and Aberdeen (Scotland).

It is important to note that coastal tide gauge stations along continental coastlines and islands can only provide local information about the sea tide at the coastal areas and close vicinity of the tide gauge stations. For that reason, even application of submerged tide gauges, which have been in use at offshore sea areas since mid 1960s, could not provide enough sea level information towards global ocean tide modeling. Before satellite altimetry, the lack of globally distributed sea level observations was the main reason for the application of differential equations for global ocean tide modeling and for the development of hydrodynamic and hybrid ocean tide models, purely based on tide gauge observations and also bottom pressure recorders at deep ocean sites, such as Sch80, which is derived by Schwiderski 1980 [6], FES94.1 which is computed by Le Provost et al. 1994 [7], and FES98 which is developed by Lefèvre et al. 2000 [8]. The advent of satellite altimetry technique and application of the early satellite altimetry sea level observations provided the possibility of ocean tide studies at the global scale. As pioneer global ocean tidal studies based on satellite altimetry sea level observations, we may refer to Le Provost 1983 [9] via SEASAT satellite altimetry and Ray and Cartwright 1990 [10] based on GEOSAT satellite altimetry mission.

Today, altimetry satellites have made it possible to access uniformly distributed global information on temporal sea level variations, and as such satellite altimetry information is widely used for global ocean tide and MSL modeling. Modern altimetry satellites are equipped with range altimeter measuring instruments, which can measure the distance between satellite's

altimeter antenna and sea surface with an accuracy up to few centimetres. For example, the altimeter of TOPEX/Poseidon satellite is reported to be accurate as ±2.5 cm [11]. Thanks to geodetic positioning systems such as SLR (Satellite Laser Ranging), GPS (Global Positioning System), and DORIS (Doppler Orbit determination and Radiopositioning Integrated on Satellite) and also availability of accurate global gravitational field models such as Earth Gravity Model 1996 (EGM96) [12], the geodetic position of the altimetry satellites can be determined to a very high degree of certainty. For example, TOPEX/Poseidon altimetry satellite can be positioned in its orbit with an accuracy level of up to ±3 cm [11].

As a result of availability of versatile satellite altimetry information and also tide gauge information, tremendous efforts have been made towards global ocean tide modeling and therefore referring to only outstanding contributions may result in the following lengthy list: Andersen 1995 [13], Andersen 1995 [14], Cartwright and Ray 1990 [10], Cartwright and Ray 1991 [15], Cartwright et al. 1991 [16], Desai and Wahr 1995 [17], Eanes 1994 [18], Eanes 2002 [19], Eanes and Bettadpur 1996 [20], Egbert et al. 1994 [21], Egbert 1997 [22], Egbert et al. 1999 [23], Egbert and Erofeeva 2002 [24], Egbert and Ray 2000 [25], Egbert and Ray 2001 [26], Egbert and Ray 2003 [27], Kagan and Kivman 1993 [28], Kantha 1995 [29], Knudsen 1994 [30], Krohn 1984 [31], Le Provost et al. 1994 [7], Le Provost et al. 1998 [32], Le Provost 2002 [33], Lefèvre et al. 2000 [8], Lefèvre et al. 2002 [34], Letellier 2004 [35], Letellier et al. 2004 [36], Ma et al. 1994 [37], Matsumoto et al. 1995 [38], Matsumoto et al. 2000 [39], Mazzega et al. 1994 [40], Ray et al. 1994 [41], Ray 1999 [42], Sanchez and Pavlis 1995 [43], Schrama and Ray 1994 [44], Schwiderski 1980 [6], Schwiderski 1980 [45], Schwiderski 1980 [46], Tierney et al. 2000 [47], and Wang and Rapp 1994 [48]. To be able to provide a systematic review over the mentioned contributions, we have arranged Table 1 and Table 2. Short technical specification of the abovementioned efforts towards global ocean tide modeling is provided in Table 1, while Table 2 is dedicated to a brief description of the computational procedure leading to global ocean tide models. There are also quite a few number of authors who have tried to assess the validity of the available global ocean tidal models by making use of various geophysical information and geodetic observations such as pelagic tide gauge data, and GPS observations as well as modern sources of gravity measurements. As a sample of those activities we may refer to Andersen et al. 1995 [49], Baker and Bos 2003 [50], Bos et al. 2002 [51], King and Padman 2005 [52], King et al. 2005 [53], Llubes and Mazzega 1997 [54], Shum et al. 1997 [55], and Urschl et al. 2005 [56]. To be able to provide a brief review over various approaches towards ocean tide modeling and also various types of ocean tide solutions, Table 3 and Table 4 have been arranged. Table 3 provides a list of various approaches to ocean tide modeling along with a brief description, while Table 4 presents a list of various types of ocean tide models.

Table 1: Summary list of recent global ocean tide models.

Model	Ref.	Data Sources		Num. of	Spatial	Spatial	Analysis	Dynamic
		SA[1]	TG[2]	Waves	Coverage	Resolution	Approach	Assumption
Sch80	[6]	no	yes	11	global	$1^\circ \times 1^\circ$	hydrodynamical interpolation	linear
CR91	[10]	GEOSAT	no	60	$-69^\circ \le \phi \le +69^\circ$	$1^\circ \times 1.5^\circ$	response/orthotides functions	none
CSR2.0	[18]	T/P	no	60	global	$1^\circ \times 1^\circ$	response/orthotides functions	none
TPXO.2	[21]	T/P	no	8	$-80^\circ \le \phi \le +70^\circ$	$0.58^\circ \times 0.70^\circ$	inverse theory/hydrodynamics	linear
Knudsen	[30]	T/P	no	4	$-65^\circ \le \phi \le +65^\circ$	$1^\circ \times 1.5^\circ$	harmonic/spherical harmonics	none
FES94.1	[7]	no	yes	8	global	$0.5^\circ \times 0.5^\circ$	hydrodynamics/finite elements	nonlinear
Mazzega	[40]	T/P	yes	8	$-66^\circ \le \phi \le +66^\circ$	$0.5^\circ \times 0.5^\circ$	inversion	none
RSC94	[41]	T/P	no	60	$-65^\circ \le \phi \le +65^\circ$	$1^\circ \times 1^\circ$	response/orthotides functions	none
SR95.0	[44]	T/P	no	5	$-65^\circ \le \phi \le +65^\circ$	$1^\circ \times 1^\circ$	harmonic	none
Rapp	[48]	T/P	no	4	$-65^\circ \le \phi \le +65^\circ$	$1^\circ \times 1.5^\circ$	harmonic/spherical harmonic	none
Andersen	[13]	ERS1,T/P	no	60	$-82^\circ \le \phi \le +82^\circ$	$0.75^\circ \times 0.75^\circ$	response/orthotides functions	none
DW95.0	[17]	T/P	no	60	$-66^\circ \le \phi \le +66^\circ$	$1^\circ \times 1^\circ$	response/orthotides functions	none
Kantha.1	[29]	T/P	yes	8	$-80^\circ \le \phi \le +66^\circ$	$0.2^\circ \times 0.2^\circ$	hydrodynamic	nonlinear
ORI96	[38]	T/P	no	8	global	$0.5^\circ \times 0.5^\circ$	response	linear
GSFC94A	[43]	T/P	no	8	$-77^\circ \le \phi \le +69^\circ$	$2^\circ \times 2^\circ$	finite difference/proudman functions	none
CSR3.0	[20]	T/P	no	8	$-78^\circ \le \phi \le +90^\circ$	$0.5^\circ \times 0.5^\circ$	response/orthotides functions	none
TPXO.3	[22]	T/P	no	8	$-80^\circ \le \phi \le +70^\circ$	$0.58^\circ \times 0.70^\circ$	inverse theory/hydrodynamics	linear
FES95.2.1	[32]	T/P	no	8	$-86^\circ \le \phi \le +90^\circ$	$0.5^\circ \times 0.5^\circ$	hydrodynamics/finite elements	nonlinear
GOT99.2b	[42]	T/P	no	8	global	$0.5^\circ \times 0.5^\circ$	hydrodynamics/finite elements	nonlinear
GOT00.2	[42]	ERS1/2,T/P	no	8	$-86^\circ \le \phi \le +90^\circ$	$0.5^\circ \times 0.5^\circ$	hydrodynamics	nonlinear
FES98	[8]	no	yes	8	global	$0.25^\circ \times 0.25^\circ$	hydrodynamics/finite elements	nonlinear
NAO.99b	[39]	T/P	no	16	$-83^\circ \le \phi \le +90^\circ$	$0.5^\circ \times 0.5^\circ$	response/orthotides functions	linear
CSR4.0	[19]	T/P	no	8	global	$0.5^\circ \times 0.5^\circ$	response/orthotides functions	none
TPXO.5	[22]	T/P	no	10	$-86^\circ \le \phi \le +90^\circ$	$0.5^\circ \times 0.5^\circ$	inverse theory/hydrodynamics	linear
TPXO.6.2	[24]	T/P	yes	10	$-86^\circ \le \phi \le +90^\circ$	$0.25^\circ \times 0.25^\circ$	inverse theory/hydrodynamics	linear
TPXO.7.0	[24]	T/P,Jason-1	yes	10	$-86^\circ \le \phi \le +90^\circ$	$0.25^\circ \times 0.25^\circ$	inverse theory/hydrodynamics	linear
FES99	[34]	T/P	yes	8	$-86^\circ \le \phi \le +90^\circ$	$0.25^\circ \times 0.25^\circ$	hydrodynamics/finite elements	nonlinear
FES2004	[35]	T/P,ERS	yes	14	$-86^\circ \le \phi \le +90^\circ$	$0.125^\circ \times 0.125^\circ$	hydrodynamics/finite elements	nonlinear

[1] Satellite Altimetry

[2] Tide Gauge

Table 2: Brief description of the computation procedures used for the global ocean tide models mentioned in Table 1.

Model	Ref.	Description
Sch80	[6]	Computed by Schwiderski (1980) as the first global hydrodynamic ocean tide model purely based on tide gauge sea level observations, it has become a standard reference for the comparison of ocean tide models. The Schwiderski global ocean tide model is available on a regular grid of latitude-longitude $1^\circ \times 1^\circ$.
CR91	[10]	Computed by Cartwright and Ray (1990) as a global ocean tide solution for the diurnal and semidiurnal tidal components with latitude-longitude $1^\circ \times 1.5^\circ$ resolution, it has been derived based on the first year of the GEOSAT satellite altimetry mission using orthotide expansion [57].
CSR2.0	[18]	Computed by Eanes (1994) as the global ocean tide model of the Centre for Space Research (CSR), University of Texas at Austin, it is based on two years of TOPEX/Poseidon satellite altimetry sea level observations, and applying orthotide functions in response analysis approach. This model provides diurnal and semidiurnal ocean tidal constituents for the whole globe at a $1^\circ \times 1^\circ$ grid intervals, and outside the coverage area of TOPEX/Poseidon it is extended to $+66^\circ \leq \phi \leq +72^\circ$ and $-72^\circ \leq \phi \leq -66^\circ$ using the CR91 model [10] and to $\phi \geq +72^\circ$ and $\phi \leq -72^\circ$ via the Sch80 model [6].
TPXO.2	[21]	Computed by Egbert et al. (1994) as the global ocean tide model of the Oregon State University (OSU) based on a global inverse solution that best fits hydrodynamical solutions and sea level observations. Sea surface measurements are provided by a homogeneous selection of the first 40 cycles of TOPEX/Poseidon satellite altimetry crossover data. This model includes eight major tidal constituents, namely, M2, S2, N2, K2, K1, O1, P1, and Q1 and is given over a grid of latitude-longitude $0.58^\circ \times 0.70^\circ$ within the area bounded by $-80^\circ \leq \phi \leq +70^\circ$.
Knudsen	[30]	Computed by Knudsen (1994) as a global ocean tide model based on harmonic analysis approach and surface spherical harmonics expansions up to degree and order $n_{\max} = 18$ as base functions for 34 cycles of TOPEX/Poseidon satellite altimetry sea level observations.

Continued on the next page...

Table 2 – Continued

Model	Ref.	Description
FES94.1	[7]	Computed by Le Provost et al. (1994) as a pure hydrodynamic global ocean tide model, tuned to fit to the globally distributed tide gauges data. This model is the earliest version of the Finite Element Solution (FES) global ocean tide models and has been derived upon a finite element grid with very fine resolution near the coast. The design of the model is based on a non-linear formulation of the shallow water equations. The model is given over a grid of $0.5^\circ \times 0.5^\circ$, and contains eight major tidal constituents, namely, M2, S2, N2, K2, 2N2, K1, O1, and Q1. The model covers the global ocean areas, except for some minor marginal seas such as the Bay of Fundy.
Mazzega	[40]	Computed by Mazzega et al. (1994) based on one year of TOPEX/Poseidon satellite altimetry data, including 40 cycles as well as coastal and deep sea tide gauge data to derive global estimates for eight major tidal constituents, namely, M2, S2, N2, K2, K1, O1, Q1, and P1 over a grid of $0.5^\circ \times 0.5^\circ$ within TOPEX/Poseidon coverage area.
RSC94	[41]	Computed by Ray et al. (1994) as a global ocean tide model based on 65 cycles of TOPEX/Poseidon satellite altimetry data using orthotide functions in response analysis approach. This model which is a product of NASA Goddard Space Flight Center (GSFC) provides diurnal and semidiurnal tidal components over $1^\circ \times 1^\circ$ grid within TOPEX/Poseidon coverage area.
SR95.0	[44]	Computed by Schrama and Ray (1994) as a global ocean tide model developed by NASA Goddard Space Flight Center (GSFC), using approximately 12 months of TOPEX/Poseidon satellite altimetry data based on harmonic analysis approach. The model includes major short-period tidal constituents, namely, M2, S2, N2, K1, and O1 within a grid of $1^\circ \times 1^\circ$ over the coverage area of TOPEX/Poseidon. SR95.0 is a deep global ocean tide model for the sea areas over 250 m depth.
Rapp	[48]	Computed by Wang and Rapp (1994) as the global ocean tide model of the Ohio State University (OSU) using 50 cycles of TOPEX/Poseidon satellite altimetry data, and harmonic analysis approach. This model is available over a 1° (in latitude directions) by 1.5° (in longitude direction) grid within the coverage area of TOPEX/Poseidon.

Continued on the next page...

Table 2 – Continued

Model	Ref.	Description
Andersen	[13]	Computed by Andersen (1995) as a global ocean tide model representing major diurnal and semidiurnal tidal constituents, within a grid of latitude-longitude $0.75^\circ \times 0.75^\circ$ spatial resolution bounded inside the interval $-82^\circ \leq \phi \leq +82^\circ$. The data used for computation of this solution are from the first 1.5 years of ERS-1 and TOPEX/Poseidon altimetry satellites.
DW95.0	[17]	Computed by Desai and Wahr (1995) at the University of Colorado as the global ocean tide model for diurnal and semidiurnal tidal constituents, using 1.7 years of TOPEX/Poseidon satellite altimetry sea level observations and applying orthotide response method [57]. The model is given on a grid of latitude-longitude $1^\circ \times 1^\circ$ spatial resolution within the geographical area, bounded by $-66^\circ \leq \phi \leq +66^\circ$.
Kantha.1	[29]	Computed by Kantha (1995) as a high-resolution global ocean tidal model, not including the Arctic region, computed for semidiurnal constituents, namely, M2, S2, N2, and K2 and diurnal tidal components, namely, K1, O1, P1, and Q1, using two years of TOPEX/Poseidon satellite altimetry sea surface measurements within sea areas deeper than 1000 m, and coastal tide gauge sea level observations. The model is given on a grid of $0.2^\circ \times 0.2^\circ$ within the geographical area $-77^\circ \leq \phi \leq +69^\circ$.
ORI96	[38]	Computed by Matsumoto et al. (1995) as a global ocean tide model for eight major tidal constituents, namely, M2, S2, N2, K2, K1, O1, P1, and Q1, with $0.5^\circ \times 0.5^\circ$ spatial resolution using TOPEX/Poseidon sea surface height data of cycles 009 to 094 by applying response analysis to tidal constituents at crossover points and hydrodynamical interpolation.
GSFC94A	[43]	Computed by Sanchez and Pavlis (1995) as the global ocean tide model of Goddard Space Flight Center (GSFC) for the main diurnal and semidiurnal tidal constituents, namely, M2, S2, N2, K2, K1, O1, P1, and Q1, using approximately 15 months of TOPEX/Poseidon satellite altimetry sea level data. The GSFC94A model is given over a grid of $2^\circ \times 2^\circ$ for the global ocean areas within $-76.75^\circ \leq \phi \leq +69.25^\circ$ and deeper than 250 m.

Continued on the next page...

Table 2 – Continued

Model	Ref.	Description
CSR3.0	[20]	Computed by Eanes and Bettadpur (1996) as the global ocean tide model of the University of Texas for diurnal and semidiurnal tidal constituents using response analysis technique applied to 2.4 years of TOPEX/Poseidon satellite altimetry observations, within cycles 001–089. The model is available over a grid of latitude-longitude $0.5^\circ \times 0.5^\circ$ spatial resolution and preserves the fine details of the FES94.1 model, whilst being more accurate for long wavelength signals. The model within the area $\phi \geq +66^\circ$ and $\phi \leq -66^\circ$ is exactly the same as FES94.1 ocean tide model [7].
TPXO.3	[22]	Computed by Egbert (1997) as the global ocean tide model which in least squares sense best fits the hydrodynamic tidal solution to cross-over data from the first 116 cycles of TOPEX/Poseidon satellite altimetry mission. This model provides eight primary tidal constituents, namely, K1, O1, P1, Q1, M2, S2, N2, and K2, within a grid of latitude-longitude $0.58^\circ \times 0.70^\circ$ spatial resolution over the sea areas bounded by $-79.71^\circ \leq \phi \leq +69.71^\circ$.
FES95.2.1	[32]	Computed by Le Provost et al. (1998) as an upgraded version of the FES94.1 [7] tidal solution, via assimilation of altimeter-derived sea level data and hydrodynamic model. The model is given over latitude-longitude $0.5^\circ \times 0.5^\circ$ grid for the eight major tidal constituents, namely, K1, O1, Q1, M2, S2, N2, K2, and 2N2.
GOT99.2b	[42]	Computed by Ray (1999) as the global ocean tide model which derives its long wavelength signals from FES94.1 hydrodynamic model [7] and uses TOPEX/Poseidon data to adjust the hydrodynamic solution. The model is given over $0.5^\circ \times 0.5^\circ$ grid. GOT99.2b model within the area $\phi \geq +66^\circ$ and $\phi \leq -66^\circ$ is purely the hydrodynamic solution and for the sea areas bounded by $-66^\circ \leq \phi \leq +66^\circ$ is based on both the satellite altimetry data and the hydrodynamic solution.
GOT00.2	[42]	Computed by Ray (1999) as the updated version of GOT99.2 [42] that assimilates TOPEX/Poseidon, ERS-1, and ERS-2 satellite altimetry data. This model uses 286 cycles of TOPEX/Poseidon and 81 cycles of ERS-1 and ERS-2 satellite altimetry data to adjust a priori FES94.1 hydrodynamic model [7]. The model is given over a grid of $0.5^\circ \times 0.5^\circ$ globally.

Continued on the next page...

Table 2 – Continued

Model	Ref.	Description
FES98	[8]	Computed by Lefèvre et al. (2000) as a version of the Final Element Solution (FES) hydrodynamic ocean tide models, which is based on assimilated tidal models at approximately 700 coastal, island and deep ocean tide gauges sea level observations, with the hydrodynamic solution. This model provides eight major tidal constituents, namely M2, S2, N2, K2, 2N2 , K1, O1 and Q1, over a grid of latitude-longitude $0.25^\circ \times 0.25^\circ$ spatial resolution.
NAO.99b	[39]	Computed by Matsumoto et al. (2000) as a global ocean tide model for 16 major tidal constituents over a grid of latitude-longitude $0.5^\circ \times 0.5^\circ$ spatial resolution based on assimilation of five years of TOPEX/Poseidon satellite altimetry sea level measurements with hydrodynamic model Sch80 [6].
CSR4.0	[19]	Computed by Eanes (2002) as the updated version of CSR3.0 global ocean tide model [20], via application of a longer time span of TOPEX/Poseidon satellite altimetry sea level data. CSR4.0 model provides diurnal and semidiurnal major tidal constituents on a $0.5^\circ \times 0.5^\circ$ grid and has been developed using 6.5 years of TOPEX/Poseidon satellite altimetry sea level observations within cycles 001 to 239.
TPXO.5	[22]	Computed by Egbert (1997) from Oregon State University (OSU) is a global ocean tide model derived by application of inverse tidal theory to tide gauge data as well as TOPEX/Poseidon satellite altimetry observations to make optimum balance between sea level observations and the linearized hydrodynamics theory. This model is available over a $0.5^\circ \times 0.5^\circ$ grid. The methods used for the computations of this model are described in detail by Egbert et al. (1994) [21] and by Egbert and Erofeeva (2002) [24].
TPXO.6.2	[24]	Computed by Egbert and Erofeeva (2002) is an updated version of TPXO.5 model, provided over a $0.25^\circ \times 0.25^\circ$ grid.
TPXO.7.0	[24]	Computed by Egbert and Erofeeva (2002) it is an updated version of TPXO.6.2 model developed over a $0.25^\circ \times 0.25^\circ$ grid. TPXO.7.0 model is the current version, which best-fits in least-squares sense to the Laplace Tidal Equations (LTE) and along track averaged data from TOPEX/Poseidon and Jason-1 satellite altimetry sea level observations.

Continued on the next page...

Table 2 – Continued

Model	Ref.	Description
FES99	[34]	Computed by Lefèvre et al. (2002) as the hydrodynamic global ocean tide model of the kind Finite Element Solution (FES), following prior solutions of the kind, i.e., FES95.2.1 [32], and FES98 [8]. The model is the result of assimilation of the hydrodynamic tidal solution derived from barotropic equations and 700 tide gauges and 687 cycles of TOPEX/Poseidon satellite altimetry observations. The model includes 8 major tidal constituents, namely, M2, S2, N2, K2, 2N2, K1, O1, and Q1.
FES2004	[35]	Computed by Letellier (2004) as a global $0.125^\circ \times 0.125^\circ$ ocean tide model, of the kind of Finite Element Solution (FES), for 14 tidal constituents, namely, M2, S2, K2, N2, 2N2, O1, P1, K1, Q1, Mf, Mtm, Mm, Msqm and M4.

Here our focus will be on the harmonic analysis approach using orthonormal base functions over the sea areas for the computation of global ocean tide models. Application of orthonormal base functions to global sea surface studies has previously been proposed and applied by for example Mainville 1987 [65], Hwang 1991 [66], Hwang 1993 [67], Hwang 1995 [68], Rapp et al. 1995 [69], Rapp et al. 1996 [70], and Rapp 1999 [71]. We differ in our approach from previous contributions to global ocean tide modeling via harmonic analysis in the type of model applied as basis, namely orthonormal base functions over the sea areas. Here we will apply the idea of using orthonormal base functions to represent MSL and sine and cosine coefficients of the nine main tidal constituents, namely, S2, M2, N2, K1, P1, O1, Mf, Mm, and Ssa at the global scale based on six years of Jason-1 satellite altimetry observations for cycles 001–200.

In the following section, the underlying mathematical theory of our approach is presented. Section 3 entitled “Case study” is devoted to technical details and our numerical results, while the Sect. 4 entitled “Assessments” covers the numerical tests for checking the validity and the accuracy of the derived geophysical models. Final conclusions and remarks are given in the last section.

2 Mathematical Setup and Modeling Scheme

Let us start the explanation of our approach by the mathematical setup for modeling global sea level variation through harmonic analysis using Fourier sine and cosine expansion as follows:

Table 3: Summary list of various approaches for ocean tide modeling.

Tidal Modeling Scheme	**Description**
Harmonic Analysis	Invented by Darwin in 1883 [58] as an efficient tool for the study and modeling of the ocean tide from sea level observations. The harmonic analysis approach implements the Fourier sine and cosine base functions and computes the projection of the time series of sea level variations at a point onto the base functions, by using the orthogonality of the Fourier base functions or least squares computations. More details can be found for example in Cartwright and Ray 1990 [10], Cartwright and Ray 1991 [15], Cherniawsky et al. 2001 [59], Cherniawsky et al. 2004 [60], Knudsen 1994 [30], Ponchaut et al. 2001 [61], and Schrama and Ray 1994 [44].
Response Analysis	This method solves for the response of the ocean surface to the tidal forcing instead of computing the tidal constituents. See for example Desai and Wahr 1995 [17], Eanes and Bettadpur 1996 [20], Ray et al. 1994 [41], Smith et al. 1997 [62], Smith et al. 1999 [63], and Matsumoto et al. 1995 [38] for details on the response analysis approach towards ocean tide modeling. Ma et al. 1994 [37], Matsumoto et al. 1995 [38], Smith 1997 [64], Smith et al. 1997 [62], and Smith et al. 1999 [63] have prove that "response analysis" and "harmonic analysis" methods are compatible towards ocean tide modeling when applied to satellite altimetry sea level observations.
Dynamic Approach	Begins with Isaac Newton in 1687 and his hydrostatic equilibrium theory for the synthesis of tide phenomenon based on its driving forces. Half a century later Pierre-Simon Laplace in 1775 established a system of partial differential equations referred to Laplace Tidal Equations (LTE) to describe flow of the water mass due to the tidal forces. Laplace Tidal Equations (LTE) are based on the bathymetry and the shape of the ocean boundaries and are still used for the hydrodynamic modeling of the ocean tide. For the details see for example Andersen 1995 [14], Egbert et al. 1994 [21], Egbert 1997 [22], Egbert and Erofeeva 2002 [24], Kantha 1995 [29], Le Provost et al. 1994 [7], Le Provost et al. 1998 [32], Le Provost 2002 [33], Lefèvre et al. 2000 [8], Lefèvre et al. 2002 [34], Letellier 2004 [35], Matsumoto et al. 1995 [38], Matsumoto et al. 2000 [39], Ray 1999 [42], Schwiderski 1980 [6], Schwiderski 1980 [45], and Schwiderski 1980 [46]

Table 4: List of various types of ocean tide models.

Ocean Tide Solution	Description
Empirical models	Based on sea level observations and not the driving forces of the tide phenomenon.
Hydrodynamic models	Based on gravitational forces driving the tide phenomenon and interaction of sea bottom topography, the shape of the ocean boundaries, and friction between the sea bottom and tidal currents in a system of partial differential equations.
Assimilation models	Dynamical models assimilating tide gauge and satellite altimetry seal level observations. In other words, the general dynamics of the sea, due to the tide, is combined with the sea surface measurements.

$$ssh(\lambda,\phi;t) = U_0(\lambda,\phi) + \sum_{k=1}^{N} A_k(\lambda,\phi) f_k \cos(\omega_k t + \psi_k(\lambda,\phi) + u_k). \quad (1)$$

In Eq. (1), $ssh(\lambda,\phi;t)$ is the Sea Surface Height (SSH) with respect to a reference ellipsoid, (λ,ϕ) are geodetic longitude and latitude, and t is the time. The integer value N represents the total number of tidal constituents considered in the mathematical model. $\omega_k = 2\pi/T_k$ is the angular velocity of the tidal constituent k and T_k denotes the time period of the tidal constituent k. Thanks to astronomical tidal studies, the frequencies and the periods of the tidal constituents are accurately known. $A_k(\lambda,\phi)$ and $\psi_k(\lambda,\phi)$ in Eq. (1) are respectively the amplitude and phase lag of the tidal constituent k, which are both functions of the geodetic coordinates (λ,ϕ) and are considered as unknown parameters of the Eq. (1). Factors f_k and u_k express the nodal modulations due to the lunar tides of the tidal constituent k. Factors f_k and u_k are for 18.6-year regression of the lunar nodal point and both are depending on the position of the lunar node. These factors can be computed for the major lunar tidal constituents, namely M2, N2, K1, O1, Mf, and Mm by using the formula derived by Doodson 1928 [72] (See Appendix A). The second part of the Eq. (1) is also called Sea Level Anomaly in the oceanographic literature. For the purpose of solving the unknown amplitude $A_k(\lambda,\phi)$ and phase lag $\psi_k(\lambda,\phi)$ using the least squares approximation, and to avoid the singularities of the amplitude at the amphidromic points, it is more common to write the Eq. (1) in the following form using cosine and sine functions:

$$\begin{aligned} &ssh(\lambda,\phi;t) = U_0(\lambda,\phi) + \\ &\sum_{k=1}^{N} \{U_k(\lambda,\phi) f_k \cos(\omega_k t + u_k) + V_k(\lambda,\phi) f_k \sin(\omega_k t + u_k)\}. \end{aligned} \quad (2)$$

Equation (2) can be considered as the basic observation equation in harmonic analysis approach. Another advantage of applying the above equation as a mathematical model, as compared to Eq. (1) is the linearity of the equation with respect to unknown parameters, which avoids iteration of adjustment computations. Our criterion for the number of considered tidal constituents, was to produce a balance between: (i) the gained accuracy of the tidal model, which should be compatible with the accuracy of satellite altimetry observations, and (ii) the computation labour. Theoretically, by including more tidal constituents, a more accurate tidal model would be obtained. However, in practice one should bear in mind that the final accuracy of the model cannot exceed the accuracy of the input data. In Eq. (2) functions $U_k(\lambda, \phi)$ and $V_k(\lambda, \phi)$ are coefficients of sine and cosine functions of the Fourier expansion respectively, and $U_0(\lambda, \phi)$ is the Mean Sea Level (MSL) as the constant part of the Fourier expansion, i.e., zero frequency oceanic wave, which are all functions of geodetic longitude λ, and latitude ϕ. Having derived these functions, the amplitude $A_k(\lambda, \phi)$ and phase $\psi_k(\lambda, \phi)$ functions of the tidal constituents can be computed as follows:

$$A_k(\lambda, \phi) = \sqrt{U_k(\lambda, \phi)^2 + V_k(\lambda, \phi)^2} \tag{3}$$

$$\psi_k(\lambda, \phi) = \tan^{-1} \frac{V_k(\phi, \lambda)}{U_k(\lambda, \phi)}. \tag{4}$$

In order to develop mathematical models for the unknown functions, namely, $U_k(\lambda, \phi)$, $V_k(\lambda, \phi)$, and $U_0(\lambda, \phi)$, some base functions have to be selected. Owing to the coverage of the oceans over the Earth, an ideal set of base functions for the modeling should be those which are defined for the whole range of spherical angles, i.e., $0 \leq \lambda \leq 2\pi$, and $-\pi/2 \leq \phi \leq +\pi/2$. Possible candidates are, for example, the surface spherical harmonics, or the spheroidal surface harmonics developed by Thong and Grafarend 1989 [73]. In this study, we first select surface spherical harmonics as a basis, and the unknown functions $U_k(\lambda, \phi)$, $V_k(\lambda, \phi)$, and $U_0(\lambda, \phi)$ are mathematically formulated as the expansion of these functions up to the degree and order $n_{\max}$, shown in Eqs. (5), and (6).

$$\begin{aligned} U_k(\lambda, \phi) &= \sum_{n=0}^{n_{\max}} \sum_{m=0}^{n} a_{nm}^k \bar{C}_{nm}(\lambda, \phi) + b_{nm}^k \bar{S}_{nm}(\lambda, \phi) \\ V_k(\lambda, \phi) &= \sum_{n=0}^{n_{\max}} \sum_{m=0}^{n} c_{nm}^k \bar{C}_{nm}(\lambda, \phi) + d_{nm}^k \bar{S}_{nm}(\lambda, \phi) \\ &\forall\, k = 1, 2, \ldots, N \end{aligned} \tag{5}$$

$$U_0(\lambda, \phi) = \sum_{n=0}^{n_{\max}} \sum_{m=0}^{n} a_{nm}^0 \bar{C}_{nm}(\lambda, \phi) + b_{nm}^0 \bar{S}_{nm}(\lambda, \phi) \tag{6}$$

where a_{nm}^0, b_{nm}^0, a_{nm}^k, b_{nm}^k, c_{nm}^k, and d_{nm}^k are the unknown coefficients to be determined, and $\bar{C}_{nm}(\lambda, \phi)$, and $\bar{S}_{nm}(\lambda, \phi)$ are normalised surface spherical harmonic functions from degree n and order m, defined by Eq. (7) [74]

$$\begin{Bmatrix} \bar{C}_{nm}(\lambda,\phi) \\ \bar{S}_{nm}(\lambda,\phi) \end{Bmatrix} = \bar{P}_{nm}(\sin\phi) \times \begin{Bmatrix} \cos m\lambda \\ \sin m\lambda \end{Bmatrix} \tag{7}$$

subject to $\forall n = 0, 1, ..., n_{\max}$ and $m = 0, 1, ..., n$. In Eq. (7), $\bar{P}_{nm}$ are fully normalised associated Legendre functions of the first kind from degree n and order m, defined in Appendix B. If we assume that the surface spherical harmonics are expanded up to the degree and order $n_{\max}$ and we have N tidal constituents included in the mathematical model, then the total number of the unknown parameters that have to be computed within Eq. (2) is equal to $(2N+1) \times (n_{\max}+1)^2$.

Here, it should be mentioned that surface spherical harmonics are orthogonal base functions when the spatial domain of their application is the whole sphere, i.e., spherical angles $0 \leq \lambda \leq 2\pi$, and $-\pi/2 \leq \phi \leq +\pi/2$. However, no satellite altimetry mission gives such a coverage and in any case the Earth is not fully covered by water. Besides, satellite altimetry sea level observations are made over discrete points and not continuously, therefore the continuity condition, needed for the orthogonality property of spherical harmonics would be violated. Therefore, if the surface spherical harmonics are used for global ocean tidal modeling as the basis, then they lose their orthogonality. Hence, owing to the non-orthogonality of surface spherical harmonic functions over the oceans, spectral analysis of oceanic signals using such a representation may lead to misleading results and implications [67]. Therefore, it can be concluded, surface spherical harmonics are not optimal for the data defined only over the oceans. To resolve this problem, a different set of base functions has to be used for the representation of oceanic tides. The ideal base functions will be a set of orthonormal base functions over the oceans, more precisely the study area. Therefore, as the first step in global ocean tide modeling, one has to follow a mathematical procedure which could lead to functions defined over the study area and that are orthogonal. Application of orthonormal base functions to oceanic studies at the global scale, e.g. for Sea Surface Topography (SST) modeling, has previously been proposed by Mainville 1987 [65], Hwang 1991 [66], Hwang 1993 [67], Hwang 1995 [68], Rapp et al. 1995 [69], Rapp et al. 1996 [70], and Rapp 1999 [71]. Here we summarise our reasons for application of normalised surface spherical harmonics within the Eqs. (5), and (6) as follows: (i) These functions are well studied and are widely used for representation of the oceanic signals [30]. (ii) These functions can be orthogonalised within orthonormalising procedures. For domains of irregular geometry such as the oceans, the Gram-Schmidt orthonormalising process can be successfully applied [67]. Next, such orthonormal base functions can be considered to derive generalised Fourier sine and cosine functions within Eqs. (5), and (6). In this study, this process is going to be used and related problems will be addressed. The principle of the Gram-Schmidt orthonormalising process is well documented, for example by Davis 1975 [75] and Kreyszig 1978 [76]. A brief summary of the fundamentals of the Gram-Schmidt orthonormalising process is presented in Appendix C.

Let us start by applying Gram-Schmidt orthonormalising process to spherical harmonics, when applied to global ocean tide modeling, in order to obtain a set of orthonormal base functions over the study area, i.e., the sea area covered by satellite altimetry measurements. Spherical harmonics are orthogonal functions and as such are independent functions within space covered by whole the sphere. Therefore, they can be considered as input functions of the Gram-Schmidt orthonormalising process. Let us define a sequence $\{X_i\}$ for $i = 0, 1, 2, \ldots$, consists of all possible normalised surface spherical harmonic functions $\bar{C}_{nm}(\lambda, \phi)$, and $\bar{S}_{nm}(\lambda, \phi)$ up to maximum degree and order $n_{\max}$ as follows:

$$\begin{aligned}\{X_i\} = \{&\bar{C}_{00}, \bar{C}_{10}, \bar{C}_{11}, \bar{S}_{11}, \bar{C}_{20}, \bar{C}_{21}, \bar{S}_{21}, \bar{C}_{22}, \bar{S}_{22},\\ &\bar{C}_{30}, \bar{C}_{31}, \bar{S}_{31}, \bar{C}_{32}, \bar{S}_{32}, \bar{C}_{33}, \bar{S}_{33},\\ &\ldots, \bar{C}_{n_{\max} n_{\max}}, \bar{S}_{n_{\max} n_{\max}}\}.\end{aligned} \tag{8}$$

Considering the normalised surface spherical harmonic functions $\bar{C}_{nm}(\lambda, \phi)$, and $\bar{S}_{nm}(\lambda, \phi)$, the Gram matrix $\mathbf{G}$ can be defined by Eq. (9) as follows:

$$\mathbf{G} = \begin{bmatrix} \langle \bar{C}_{00}|\bar{C}_{00}\rangle_\zeta & \langle \bar{C}_{00}|\bar{C}_{10}\rangle_\zeta & \langle \bar{C}_{00}|\bar{C}_{11}\rangle_\zeta & \langle \bar{C}_{00}|\bar{S}_{11}\rangle_\zeta & \cdots \\ \langle \bar{C}_{10}|\bar{C}_{00}\rangle_\zeta & \langle \bar{C}_{10}|\bar{C}_{10}\rangle_\zeta & \langle \bar{C}_{10}|\bar{C}_{11}\rangle_\zeta & \langle \bar{C}_{10}|\bar{S}_{11}\rangle_\zeta & \cdots \\ \langle \bar{C}_{11}|\bar{C}_{00}\rangle_\zeta & \langle \bar{C}_{11}|\bar{C}_{10}\rangle_\zeta & \langle \bar{C}_{11}|\bar{C}_{11}\rangle_\zeta & \langle \bar{C}_{11}|\bar{S}_{11}\rangle_\zeta & \cdots \\ \langle \bar{S}_{11}|\bar{C}_{00}\rangle_\zeta & \langle \bar{S}_{11}|\bar{C}_{10}\rangle_\zeta & \langle \bar{S}_{11}|\bar{C}_{11}\rangle_\zeta & \langle \bar{S}_{11}|\bar{S}_{11}\rangle_\zeta & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \langle \bar{C}_{nn}|\bar{C}_{00}\rangle_\zeta & \langle \bar{C}_{nn}|\bar{C}_{10}\rangle_\zeta & \langle \bar{C}_{nn}|\bar{C}_{11}\rangle_\zeta & \langle \bar{C}_{nn}|\bar{S}_{11}\rangle_\zeta & \cdots \\ \langle \bar{S}_{nn}|\bar{C}_{00}\rangle_\zeta & \langle \bar{S}_{nn}|\bar{C}_{10}\rangle_\zeta & \langle \bar{S}_{nn}|\bar{C}_{11}\rangle_\zeta & \langle \bar{S}_{nn}|\bar{S}_{11}\rangle_\zeta & \cdots \end{bmatrix} \tag{9}$$

where $\langle \cdot|\cdot\rangle_\zeta$ symbolises the inner products of the normalised spherical harmonic functions over the domain ζ, i.e., the oceans covered by satellite altimetry. The inner products in Eq. (9) can be written as

$$\begin{aligned}&\left\{\begin{array}{l} \langle \bar{C}_{nm}(\lambda,\phi)\,|\bar{C}_{rs}(\lambda,\phi)\rangle_\zeta \\ \langle \bar{S}_{nm}(\lambda,\phi)\,|\bar{S}_{rs}(\lambda,\phi)\rangle_\zeta \\ \langle \bar{C}_{nm}(\lambda,\phi)\,|\bar{S}_{rs}(\lambda,\phi)\rangle_\zeta \\ \langle \bar{S}_{nm}(\lambda,\phi)\,|\bar{C}_{rs}(\lambda,\phi)\rangle_\zeta \end{array}\right\} = \\ &\tfrac{1}{a_\zeta}\int\int_\zeta \bar{P}_{nm}(\sin\phi)\,\bar{P}_{rs}(\sin\phi)\left\{\begin{array}{l} \cos m\lambda\cos s\lambda \\ \sin m\lambda\sin s\lambda \\ \cos m\lambda\sin s\lambda \\ \sin m\lambda\cos s\lambda \end{array}\right\} d\sigma.\end{aligned} \tag{10}$$

In Eq. (10), a_ζ represents the total sea area covered by satellite altimetry observations, and $d\sigma = \cos\phi d\lambda d\phi$ denotes a surface differential element. The total sea area a_ζ can be computed as the sum of the area of the finite elements covering the whole sea area of interest as follows:

$$a_\zeta = \int\int_\zeta d\sigma = \sum_{k=1}^{K}\sum_{l=1}^{L}\omega_{kl}\int_{\lambda_l}^{\lambda_{l+1}}\int_{\phi_k}^{\phi_{k+1}}\cos\phi d\lambda d\phi. \tag{11}$$

The Gram matrix defined in Eq. (9) in terms of normalised surface spherical harmonics can be written as follows:

$$\begin{Bmatrix} \langle \bar{C}_{nm}(\lambda,\phi)\,|\bar{C}_{rs}(\lambda,\phi)\rangle_\zeta \\ \langle \bar{S}_{nm}(\lambda,\phi)\,|\bar{S}_{rs}(\lambda,\phi)\rangle_\zeta \\ \langle \bar{C}_{nm}(\lambda,\phi)\,|\bar{S}_{rs}(\lambda,\phi)\rangle_\zeta \\ \langle \bar{S}_{nm}(\lambda,\phi)\,|\bar{C}_{rs}(\lambda,\phi)\rangle_\zeta \end{Bmatrix} = \frac{1}{a_\zeta}\sum_{k=1}^{K}\sum_{l=1}^{L}\omega_{kl}\xi^k_{nmrs}\int_{\lambda_l}^{\lambda_{l+1}}\begin{Bmatrix} \cos m\lambda\cos s\lambda \\ \sin m\lambda\sin s\lambda \\ \cos m\lambda\sin s\lambda \\ \sin m\lambda\cos s\lambda \end{Bmatrix} d\lambda. \tag{12}$$

Here, K and L in Eqs. (11) and (12) are the number of blocks in the latitudinal and longitudinal direction. ω_{kl}, which can be called the "sea function", attains 1 when the integration is over the sea area of interest and 0 in other cases, i.e.,

$$\omega_{kl} = \begin{cases} 1, & (\lambda,\phi)\in\zeta \\ 0, & (\lambda,\phi)\notin\zeta \end{cases}. \tag{13}$$

In Eq. (12) ξ^k_{nmrs} represents the integral of the product of two fully normalised associated Legendre functions within an element as follows:

$$\xi^k_{nmrs} = \int_{\phi_k}^{\phi_{k+1}} \bar{P}_{nm}(\sin\phi)\,\bar{P}_{rs}(\sin\phi)\cos\phi d\phi. \tag{14}$$

Using symbolic operation tool-boxes, provided for example by Mathematica or Matlab, integral of the products of cosine and sine functions in Eq. (10) can be analytically computed, even up to a very high degree and order. However, there is not such a possibility for the integration of the products of two fully normalised associated Legendre functions, i.e., ξ^k_{nmrs} in Eq. (12). Here ξ^k_{nmrs} can only be computed by recursive formulae, which are given by Mainville 1987 [65], and applied by Hwang 1991 [66], Hwang 1993 [67], and Hwang 1995 [68] to Sea Surface Topography (SST) modeling as follows:

$$\begin{aligned} &\xi^k_{nmrs} = \\ &\tfrac{a(n,m)}{n+r+1}\{\tfrac{n-r-2}{a(n-1,m)}I^k_{n-2,mrs} + \tfrac{2r+1}{a(r,s)}I^k_{n-1,m,r-1,s} - \\ &(1-x^2)\,\bar{P}_{n-1,m}(x)\bar{P}_{rs}(x)\big|_{x_{k+1}}^{x_k}\} \end{aligned} \tag{15}$$

when $n \neq m$ and $r \neq s$,

$$\xi^k_{nmrs} = \frac{a(r,s)}{n+r+1}\{\frac{r-n-2}{a(r-1,s)} I^k_{nn,r-2,s} - (1-x^2)\bar{P}_{nn}(x)\bar{P}_{r-1,s}(x)\big|^{x_k}_{x_{k+1}}\} \tag{16}$$

when $n = m$ and $r \neq s$,

$$\xi^k_{nmrs} = \frac{a(n,m)}{n+r+1}\{\frac{n-r-2}{a(n-1,m)} I^k_{rr,n-2,m} - (1-x^2)\bar{P}_{rr}(x)\bar{P}_{n-1,m}(x)\big|^{x_k}_{x_{k+1}}\} \tag{17}$$

when $n \neq m$ and $r = s$,

$$\xi^k_{nmrs} = \frac{1}{n+r+1}\{(n+r)b(n)b(n-1)I^k_{n-2,n-2,rr} + x\bar{P}_{nn}(x)\bar{P}_{rr}(x)\big|^{x_k}_{x_{k+1}}\} \tag{18}$$

when $n = m$, $r = s$, and $n \neq 0, 1$. In the above equations $x = \sin\phi$, $x_k = \sin\phi_k$, $x_{k+1} = \sin\phi_{k+1}$, and $a(\cdot,\cdot)$, and $b(\cdot)$ are defined as follows:

$$\begin{aligned} a(n,m) &= \sqrt{\tfrac{(2n+1)(2n-1)}{(n+m)(n-m)}}, \forall n \neq m \\ b(1) &= \sqrt{3} \\ b(n) &= \sqrt{\tfrac{2n+1}{2n}}, \forall n > 1. \end{aligned} \tag{19}$$

In order to compute the needed initial values of Eqs. (15) to (18), i.e., ξ^k_{nmrs} for $n = 0, 1$, $m = 0, \ldots, n$, $r = n, \ldots, n_{\max}$, and $s = m, \ldots, r$, we used symbolic programming of Matlab. Alternatively, there is another method to compute the integral product of two fully normalised associated Legendre functions, i.e., ξ^k_{nmrs}, named as "Product-Sum Formulae" applied by Hwang 1991 [66] which is not used in this study. After computations of the Gram matrix $\mathbf{G}$ within the above steps, everything will be ready to derive the elements of the matrix $\mathbf{C}$ of combination coefficients c_{ij} via a Cholesky decomposition of the Gram matrix $\mathbf{G}$ as shown in Appendix C by Eqs. (C.5) and (C.6). Now let us define sequence $\{\bar{X}_i\}$ for $i = 0, 1, 2, \ldots$, consists of orthonormal base functions $\bar{O}_{nm}(\lambda,\phi)$, and $\bar{R}_{nm}(\lambda,\phi)$ over study area as follows:

$$\begin{aligned} \{\bar{X}_i\} = \{&\bar{O}_{00}, \bar{O}_{10}, \bar{O}_{11}, \bar{R}_{11}, \bar{O}_{20}, \bar{O}_{21}, \bar{R}_{21}, \bar{O}_{22}, \bar{R}_{22}, \\ &\bar{O}_{30}, \bar{O}_{31}, \bar{R}_{31}, \bar{O}_{32}, \bar{R}_{32}, \bar{O}_{33}, \bar{R}_{33}, \\ &\ldots, \bar{O}_{n_{\max}n_{\max}}, \bar{R}_{n_{\max}n_{\max}}\}. \end{aligned} \tag{20}$$

Here the elements of the matrix $\mathbf{C}$, i.e., combination coefficients in Gram-Schmidt orthonormalising process, as defined in Appendix C by Eqs. (C.2), (C.3), and (C.4), produce the set of orthonormal base functions $\{\bar{X}_i\}$ using the set of surface spherical harmonic functions $\{X_i\}$ as follows:

$$\bar{X}_i(\lambda,\phi) = \sum_{j=0}^{i} c_{ij} X_j(\lambda,\phi). \tag{21}$$

As such we can derive the orthonormal base functions $\bar{O}_{nm}(\lambda,\phi)$, and $\bar{R}_{nm}(\lambda,\phi)$ using Eqs. (22) to (24), (See Mainville 1987 [65] and Hwang 1995 [68]).

$$\bar{O}_{nm}\left(\lambda,\phi\right) = c_{ii}\bar{C}_{nm}\left(\lambda,\phi\right) + \sum_{j=0}^{i-1} c_{ij} X_j\left(\lambda,\phi\right) \tag{22}$$

subject to $i = n^2$ for $m = 0$, and $i = n^2 + 2m - 1$ for $m \neq 0$,

$$\bar{R}_{n0}(\lambda,\phi) = 0 \tag{23}$$

subject to $m = 0$,

$$\bar{R}_{nm}\left(\lambda,\phi\right) = c_{ii}\bar{S}_{nm}\left(\lambda,\phi\right) + \sum_{j=0}^{i-1} c_{ij} X_j\left(\lambda,\phi\right) \tag{24}$$

with $i = n^2 + 2m$ for $m \neq 0$. The maximum degree and order of spherical harmonic expansion can be determined from the numerically linearly independent columns of the Gram matrix (See for example Hwang 1993 [67], and Rapp 1999 [71]). Now the orthonormalised base functions $\bar{O}_{nm}(\lambda,\phi)$, and $\bar{R}_{nm}(\lambda,\phi)$ can be used in Eqs. (5) and (6) in order to obtain the coefficients of the orthonormal base functions:

$$\begin{aligned} U_k(\lambda,\phi) &= \sum_{n=0}^{n_{\max}} \sum_{m=0}^{n} a_{nm}^k \bar{O}_{nm}(\lambda,\phi) + b_{nm}^k \bar{R}_{nm}(\lambda,\phi) \\ V_k(\lambda,\phi) &= \sum_{n=0}^{n_{\max}} \sum_{m=0}^{n} c_{nm}^k \bar{O}_{nm}(\lambda,\phi) + d_{nm}^k \bar{R}_{nm}(\lambda,\phi) \\ &\forall k = 1,2,\ldots,N \end{aligned} \tag{25}$$

$$U_0(\lambda,\phi) = \sum_{n=0}^{n_{\max}} \sum_{m=0}^{n} a_{nm}^0 \bar{O}_{nm}(\lambda,\phi) + b_{nm}^0 \bar{R}_{nm}(\lambda,\phi). \tag{26}$$

Finally, coefficients of the orthonormal base functions, namely, a_{nm}^0, b_{nm}^0, a_{nm}^k, b_{nm}^k, c_{nm}^k, and d_{nm}^k as well as the covariance matrix of the unknown parameters can be estimated using the satellite altimetry observations in Eq. (2) as follows:

$$\begin{aligned} \hat{\mathbf{x}} &= (\mathbf{A}^T \mathbf{C}_{\mathbf{l}}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C}_{\mathbf{l}}^{-1} \mathbf{l} \\ \mathbf{C}_{\hat{\mathbf{x}}} &= (\mathbf{A}^T \mathbf{C}_{\mathbf{l}}^{-1} \mathbf{A})^{-1} \end{aligned} \tag{27}$$

where $\mathbf{l}$ represents the vector of observations, $\mathbf{A}$ is the design matrix, $\mathbf{P}$ is the weight matrix, $\hat{\mathbf{x}}$ is the vector of estimated unknown coefficients, and $\mathbf{C}_{\hat{\mathbf{x}}}$ is the covariance matrix of the estimated coefficients. Selection of an appropriate sampling interval is one of the most important issues to be addressed when dealing with harmonic phenomena. More precisely, aliasing and over sampling are two key issues, which must be avoided in the harmonic analysis. Over sampling has no benefit other than making the numerical calculations too

long and may even lead to matrix operations to become below the capacity of common computers. However, aliasing is a much more serious problem. According to Nyquist sampling theorem, the sampling frequency must be at least twice the maximum frequency to be measured. When the frequency is higher than the Nyquist limit, i.e. half the sampling frequency, aliasing occurs. Thus, fulfilling the sampling theorem is an essential condition for a complete estimate of the coefficients of the base functions. In our case, to avoid this potential source of error, the sampling interval of the satellite altimetry data along the track must be equal to at least half of the spatial resolution of the maximum degree of the spherical harmonic expansion. Because the spatial resolution of the maximum degree of surface spherical harmonic functions is half of the spatial wavelength, it follows that the sampling interval must be selected at least to be a quarter of the spatial wavelength. Thus the sampling interval can be computed by the formula $(6400 \times \pi / n_{\max})\,/2$ (km).

3 Case Study

In this section we present technical details related to the computation of the global ocean tide amplitude and phase models for six semidiurnal and diurnal major tidal constituents, namely, S2, M2, N2, K1, P1, and O1 together with three long term tidal components, namely, Mf, Mm, and Ssa, and Mean Sea Level (MSL) based on orthonormal base functions over the study area and first six years of Jason-1 satellite altimetry data, including cycles 001–200. Jason-1 is jointly conducted by the Centre National d'Etudes Spatiales (CNES) and the National Aeronautics and Space Administration (NASA). Jason-1 is a follow-on mission to the highly successful TOPEX/Poseidon project and overflies the TOPEX/Poseidon ground tracks. Jason-1 was launched on December 7, 2001, and its first cycle began on January 15, 2002, coinciding with TOPEX/Poseidon cycle 344 [77]. Our computations are based on Geophysical Data Records (GDR) provided by the Jet Propulsion Laboratory (JPL) from http://podaac.jpl.nasa.gov. To the observed altimeter range, i.e., the measured distance between satellite and sea level, the following corrections need to be applied: (i) Wet tropospheric delay, (ii) Dry tropospheric delay, (iii) Ionospheric delay, (iv) Electromagnetic bias, (v) Inverse barometer pressure, (vi) Solid earth tide, and (vii) Pole tide. In order to apply these corrections we used the standard correction formulas provided by JPL for Jason-1 GDR data records [77]. Using these formulas the corrected range, i.e., corrected distance between the satellite and sea level can be computed. The corrected range was then combined with the measured geodetic ellipsoidal height of the satellite, i.e., altitude, derived from its precise positioning systems, to determine the sea surface height $ssh(\lambda, \phi; t)$ with respect to a reference ellipsoid in a point-wise manner as follows:

$$ssh(\lambda, \phi; t) = altitude(\lambda, \phi; t) - range(\lambda, \phi; t). \tag{28}$$

The reference ellipsoid used for Jason-1 is an ellipsoid of revolution with equatorial radius of $a = 6378136.3$ m and flattening $f = 1/298.257$, the same reference ellipsoid as used by TOPEX/Poseidon. Jason-1 sea level measurement is reported to have an accuracy of ± 4.2 cm [77]. We excluded the data over shallow water within a band of 5 km from the shoreline to avoid high frequency noises. We also did not use data flagged by JPL as less accurate data points, according to [77]. As mentioned before, the frequencies of the tidal constituents are taken as known values in our study because their precise values are known from astronomical studies. Table 5 shows the periods of the nine tidal constituents used here, namely, S2, M2, N2, K1, P1, O1, Mf, Mm, and Ssa. Computation of the elements of the Gram matrix $\mathbf{G}$ requires determination of the inner products of normalised surface spherical harmonic functions defined by Eq. (9) over the sea areas of interest, i.e., the sea area covered by the satellite altimetry observations. Computation of the integrals can be readily done in analytical form if the sea areas are known. For this purpose first we covered the world with a latitude-longitude $1^\circ \times 3^\circ$ grid, and next using the Jason-1 satellite altimetry data, those grid cells residingfor which at least one altimetry data is flagged as sea, are considered over the sea areas. The criterion for selecting the $1^\circ \times 3^\circ$ search grid cells to distinguish land from sea was the minimum cross-track spatial resolution of Jason-1 data over the equator, which is about 3°. It is also important to note that in this way the areas outside the coverage of the Jason-1 satellite will be also excluded from the integration domain as is needed. Using the determined $1^\circ \times 3^\circ$ grid cells over the sea areas the elements of the Gram matrix are computed for each cell analytically and then summed up in order to have the surface integrals computed for the whole sea area of interest. Besides, for the later applications, a finer $1^\circ \times 1^\circ$ grid is also generated from the developed $1^\circ \times 3^\circ$ grid. Figure 1 shows the derived $1^\circ \times 1^\circ$ grid over the sea areas which is also limited to the sea areas covered by Jason-1, i.e., sea areas within $-66^\circ \leq \phi \leq +66^\circ$ and

Table 5: Periods of the nine main tidal constituents used in the global ocean tide modeling, namely, S2, M2, N2, K1, P1, O1, Mf, Mm, and Ssa.

Tidal Constituent	**Symbol**	**Tidal Period (hour)**
Principal solar semidiurnal constituent	S2	12.000000
Principal lunar semidiurnal constituent constituent	M2	12.420601
Larger lunar elliptic semidiurnal constituent	N2	12.658348
Lunar diurnal constituent	K1	23.934470
Solar diurnal constituent	P1	24.065890
Lunar diurnal constituent	O1	25.819342
Lunisolar fortnightly constituent	Mf	327.85898
Lunar monthly constituent	Mm	661.30927
Solar semiannual constituent	Ssa	4382.9065

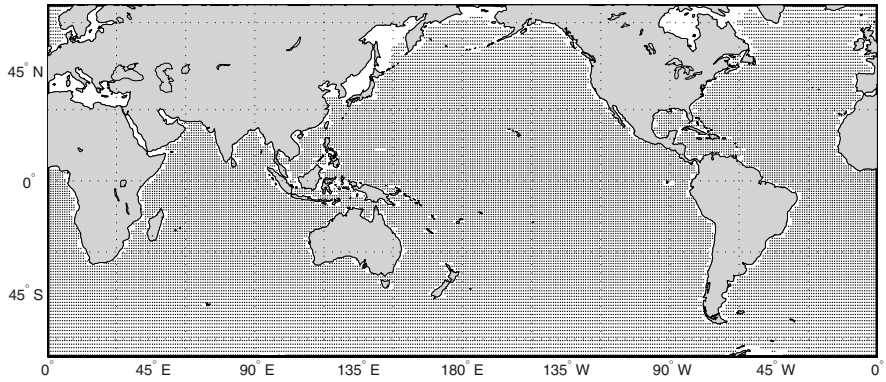


Fig. 1: latitude-longitude $1^\circ \times 1^\circ$ grid over the study area without inshore sea regions covered by Jason-1 satellite altimetry data within $-66^\circ \leq \phi \leq +66^\circ$ and $0^\circ \leq \lambda \leq +360^\circ$.

$0^\circ \leq \lambda \leq +360^\circ$. The maximum degree and order of surface spherical harmonic functions, which are numerically independent within the study area, was determined using numerically independent columns of the Gram matrix $\mathbf{G}$ based on checking the Gram matrix determinant $|\mathbf{G}|$, (See Appendix C for details). First, we computed Gram matrix $\mathbf{G}$ for normalised surface spherical harmonics up to maximum degree and order 25. Then we checked the Gram matrix determinant $|\mathbf{G}|$ for spherical harmonic functions at the different maximum degrees and orders and found that the Gram matrix determinant $|\mathbf{G}|$ becomes zero at degrees and orders over than $n = 20$. Therefore, normalised surface spherical harmonic functions up to degree and order $n_{\max} = 20$ are selected as the numerically independent functions in our study. Figure 2 shows the value of condition numbers of the Gram matrix $\mathbf{G}$ for different maximum degrees of normalised spherical harmonic expansion up to maximum degree and order $n_{\max} = 25$. Since the intersection of the lunar orbital plane with the earth's ecliptical plane, known as the nodal line, rotates once in every 18.6 years, this is an issue that must be considered in the ocean tide modeling. In the case of Jason-1, the variations of the factors f_k and u_k within a 10-day cycle are so small that the nodal modulations f_k and u_k can be computed for the average time of each 10-day cycle. Table 6 shows the values of these corrections for lunar tidal constituents, namely, M2, N2, K1, O1, Mf, and Mm which are estimated for cycle 198 on 2007.05.27, as an example. The sampling interval was selected according to the spatial wavelength of spherical harmonics expansion up to degree and order $n_{\max} = 20$. Using the formula $2 \times 6400 \times \pi / n_{\max}$, 2010.62 km was derived as the wavelength of the surface spherical harmonics of degree $n_{\max} = 20$. Therefore, we selected a sampling interval of 502.65 km, i.e., a quarter of the wavelength of the spherical harmonics for degree $n_{\max} = 20$. Considering the Jason-1 data spacing, which is

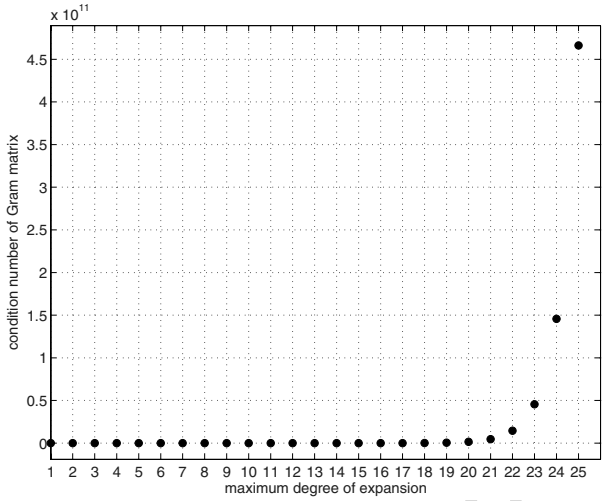


Fig. 2: The values of condition number of Gram matrix **G** for normalised surface spherical harmonic functions up to maximum degree and order $n_{\max} = 25$.

Table 6: Amplitude and phase nodal corrections for lunar tidal constituents, namely, M2, N2, K1, O1, Mf, and Mm within cycle 198 of Jason-1 on 2007.05.27.

Tidal Constituent	f_k	u_k (deg)
M2	0.9651	0.6636
N2	0.9651	0.6636
K1	1.1086	2.4034
O1	1.1757	−2.7131
Mf	1.4328	6.0899
Mm	0.8775	0.0000

every 5.8 km (with 1 Hz sampling rate), we selected a sample point every 45 data points. This sampling rate could result in 255.78 km along track data spacing and selection of 2018856 data points from the Jason-1 satellite altimetry observations within cycles 001 to 200, if all data were of good quality according to the Jason-1 data flags. Indeed we have used 37 flags defined on pages 25 and 26 of AVISO and PODAAC User Handbook [77] to find and use only those data among the selected 2018856 data points which are of good quality. Therefore, the observation points are selected with 255.78 km spacing along track according to the spatial wavelength of spherical harmonic expansion to degree and order $n_{\max} = 20$. Those data are used for the computation of the nine main tidal constituents, namely, S2, M2, N2, K1, P1, O1, Mf, Mm, and Ssa as well as MSL. Figure 3 shows the distribution of the observation points according to the specified sampling interval within cycle 190 including

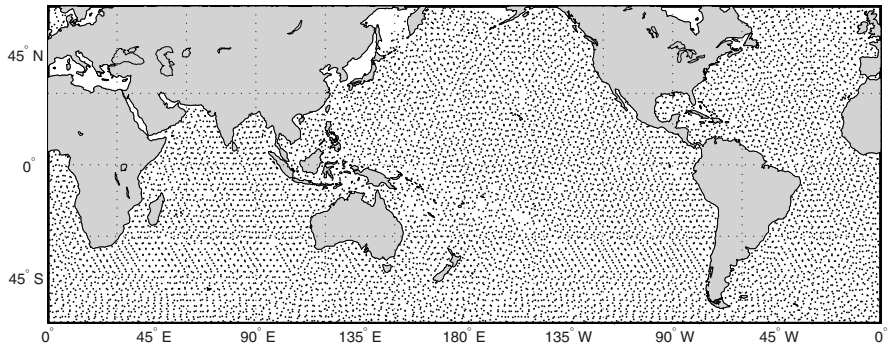


Fig. 3: Satellite altimetry observations within cycle 190 of Jason-1 including 11061 re-sampled data points with 255.78 km along-track spacing.

11061 sea level data records. Using the Jason-1 satellite altimetry observations within cycles 001–200, the orthonormal base functions over the study area, i.e., sea regions covered by Jason-1 satellite altimetry data records, bounded by $-66° \leq \phi \leq +66°$ and $0° \leq \lambda \leq +360°$, are computed and nine tidal constituents, namely, S2, M2, N2, K1, P1, O1, Mf, Mm, and Ssa as well as MSL were modeled. Figure 4 shows a plot of the computed model for MSL. The ellipsoidal heights shown in Fig. 4 are with respect to the WGS84 reference ellipsoid. Figures 5 to 10 show the computed co-range maps of the six major diurnal and semidiurnal tidal constituents S2, M2, N2, K1, P1, and O1.

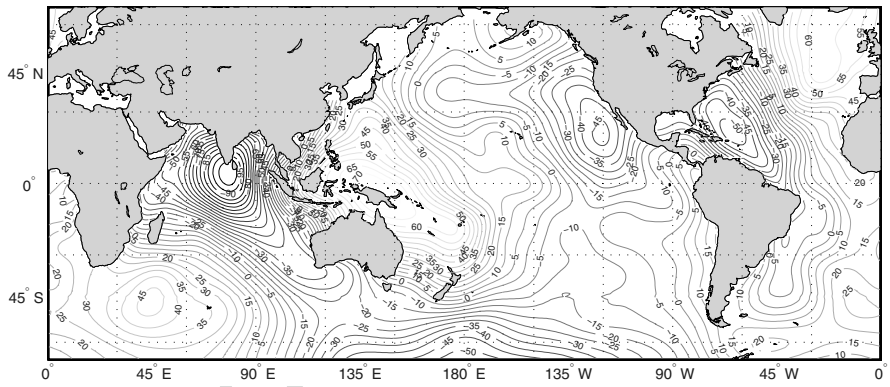


Fig. 4: Computed global MSL model with respect to WGS84 reference ellipsoid (contour intervals: 5 m).

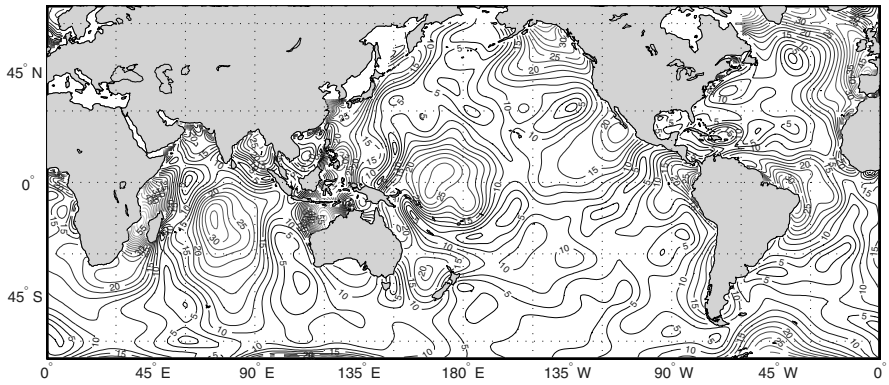


Fig. 5: Computed global co-range tidal model of S2 (contour intervals: 2.5 cm).

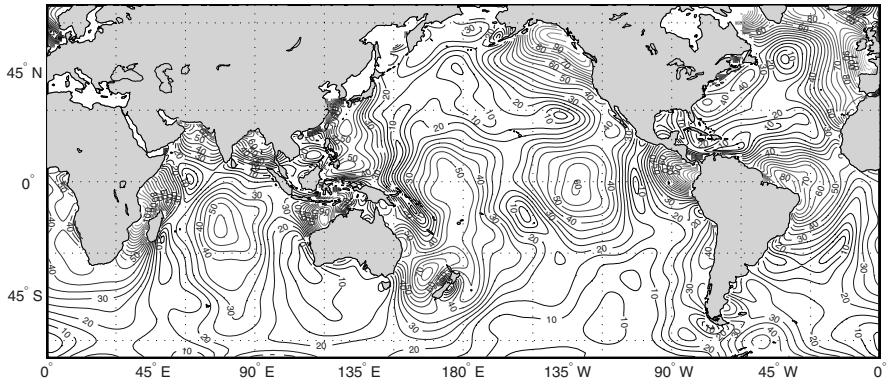


Fig. 6: Computed global co-range tidal model of M2 (contour intervals: 5 cm).

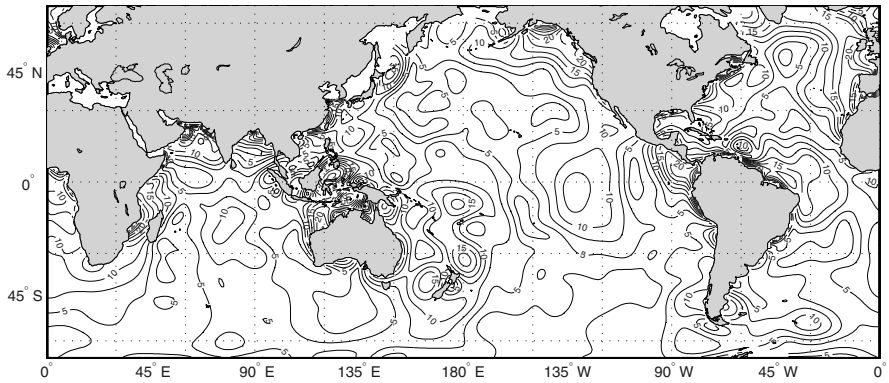


Fig. 7: Computed global co-range tidal model of N2 (contour intervals: 2.5 cm).

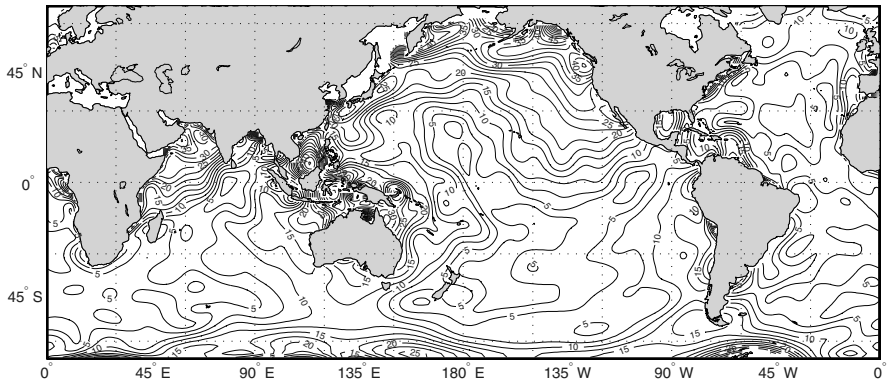


Fig. 8: Computed global co-range tidal model of K1 (contour intervals: 2.5 cm).

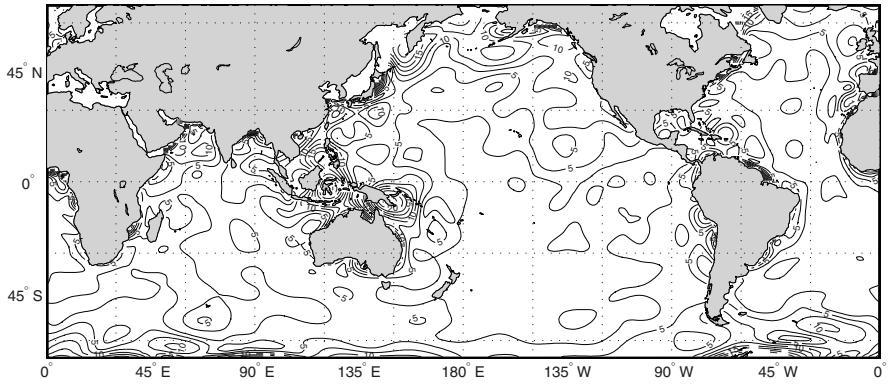


Fig. 9: Computed global co-range tidal model of P1 (contour intervals: 2.5 cm).

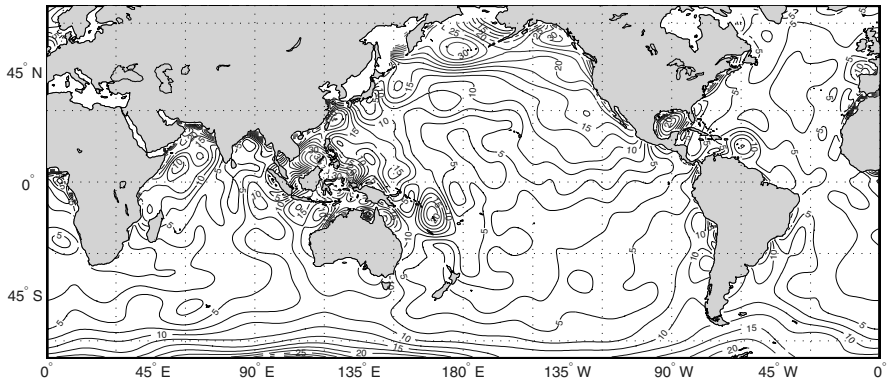


Fig. 10: Computed global co-range tidal model of O1 (contour intervals: 2.5 cm).

4 Assessments

In this section we are going to assess the Mean Sea Level (MSL), and six major diurnal and semidiurnal tidal constituents, namely, S2, M2, N2, K1, P1, and O1 as well as three long term tidal components, i.e., Mf, Mm, and Ssa, that are empirically modeled at the global scale, in terms of orthonormal base functions over the sea areas covered by Jason-1 altimetry satellite within the six years of its operation. To achieve this and also in order to verify the accuracy of the computed models, the results of six tests are going to be presented. In all tests we did not consider coastal areas, where one would expect to have large errors due to inaccurate satellite altimetry observations.

Test 1 is carried out by synthesising the sea surface observations within cycle 205 of Jason-1, that was not used in the modeling. Global distribution of those check points is shown in Fig. 11, which exactly corresponds to the ground tracks of Jason-1. The synthesised sea surface heights by the computed model within cycle 205 are presented in Fig. 12. The ellipsoidal heights shown in Fig. 12 are with respect to the WGS84 reference ellipsoid. Figure 13 shows the difference between the sea surface observations within cycle 205 and those synthesised by our computed MSL and tidal models using Eq. (2). Table 7 gives a statistical summary of the differences. The difference between the modeled sea surface heights and the observations is maximum 64.14 cm, with RMS 2.63 cm. The maximum difference shown in Table 7 may be caused by some short-term sea level variations. In fact, because we are comparing a tide model obtained via six years of satellite altimetry observations with the measurements within one cycle, such deviations can be considered quite justifiable. This test can be regarded as a combined verification of the amplitude, phase, and MSL models computed in this study.

As test 2 and test 3, we compare the MSL computed in this study with (i) OSUMSS95 (Yi 1995 [78], Rapp and Yi 1997 [79]), and (ii) GSFC00.1 MSS

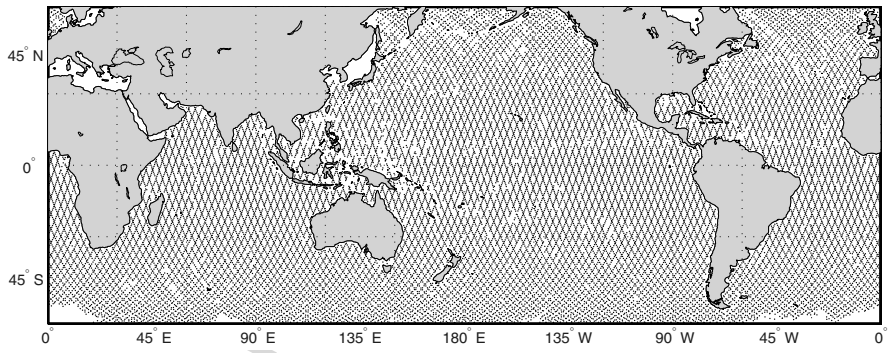


Fig. 11: Global distribution of the used check points for Test 1, selected within cycle 205 of Jason-1 satellite, including 30341 re-sampled data points with 85.26 km along-track spacing.

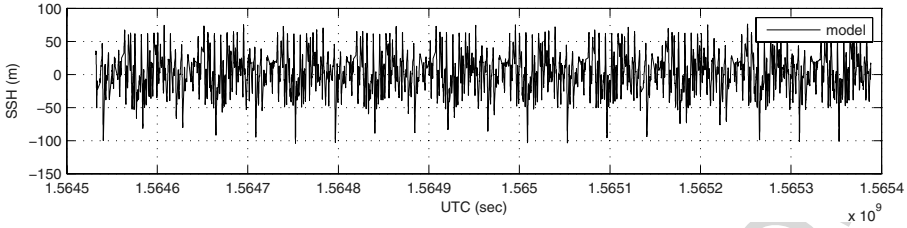


Fig. 12: Synthesised sea level heights by the computed MSL and tide models within cycle 205 of Jason-1 with respect to WGS84 reference ellipsoid.

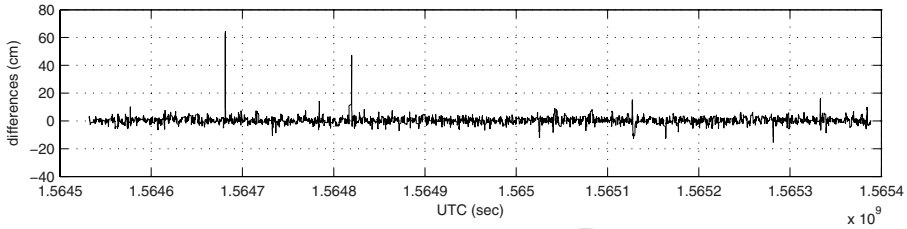


Fig. 13: Difference between the sea surface observations within cycle 205 and those synthesised by the computed MSL and tide models.

Table 7: Statistical summary of the difference between observed sea level heights within cycle 205 of Jason-1 and those synthesised by the computed MSL and tide models.

Total number of the check points	30341
Minimum of differences (cm)	−15.21
Maximum of differences (cm)	64.14
Mean of differences (cm)	0.35
Minimum of absolute differences (cm)	0.00
Maximum of absolute differences (cm)	64.14
Mean of absolute differences (cm)	1.53
RMS of differences (cm)	2.63

(Wang 2001 [80]) Mean Sea Surface (MSS). The OSUMSS95 MSS is based on satellite altimetry sea level data, provided by one year of TOPEX/Poseidon, one year of ERS-1, one year of GEOSAT, and the first cycle of ERS-1 satellite altimetry mission. The values are given on a $3.75' \times 3.75'$ grid within the geographical latitude $-80° \leq \phi \leq +82°$ globally. The OSUMSS95 MSL solution is the standard MSL model for TOPEX/Poseidon and is provided by JPL along with other TOPEX/Poseidon data. The details on the development of the OSUMSS95 MSL solution are given in Yi 1995 [78] where its comparisons with other solutions and its evaluations can be found. The GSFC00.1 MSS

derived by Wang 2001 [80] is computed based on satellite altimetry data from a variety of missions, including six years of TOPEX/Poseidon data, several years of ERS-1/2 data and GEOSAT data. The GSFC00.1 MSS is computed on a $2' \times 2'$ grid over the sea areas bounded by the geographical latitude $-80° \leq \phi \leq +80°$. The check points used in these two tests are the points of the $1° \times 1°$ grid over the study area shown in Fig. 1. The results of the two above comparisons are shown in Figs. 14 and 15 and Tables 8 and 9. The RMS values of the difference between the computed MSL and those of OSUMSS95

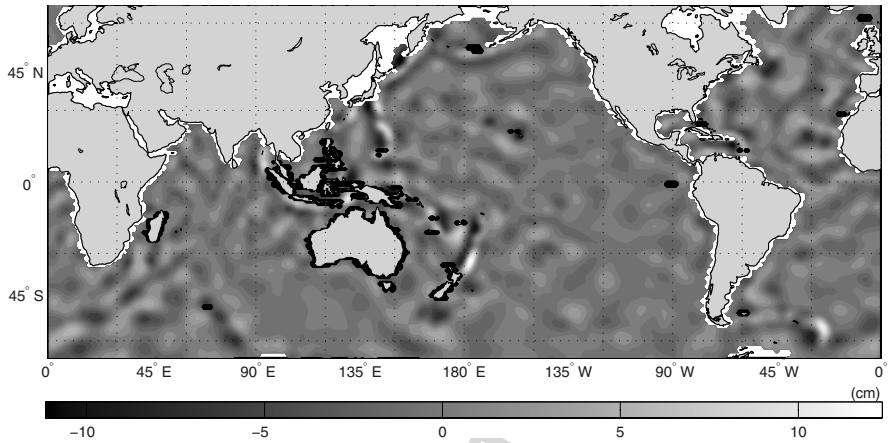


Fig. 14: The difference between the computed global MSL model in this study and OSUMSS95 MSL solution.

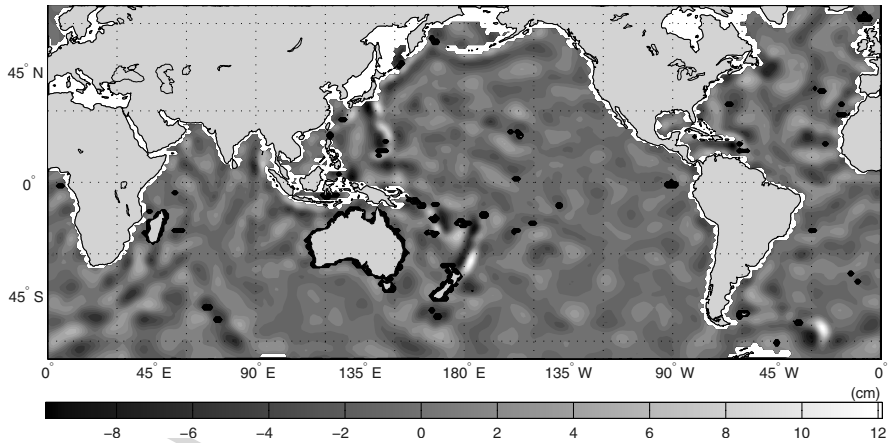


Fig. 15: The difference between the computed global MSL model in this study and GSFC00.1 MSL solution.

Table 8: Statistical summary of the difference between the computed global MSL model in this study and OSUMSS95 MSL solution.

Total number of the check points	31521
Minimum of differences (cm)	−11.15
Maximum of differences (cm)	13.53
Mean of differences (cm)	0.28
Minimum of absolute differences (cm)	0.00
Maximum of absolute differences (cm)	13.53
Mean of absolute differences (cm)	5.46
RMS of differences (cm)	7.33

Table 9: Statistical summary of the difference between the computed global MSL model in this study and GSFC00.1 MSL solution.

Total number of the check points	31521
Minimum of differences (cm)	−9.86
Maximum of differences (cm)	13.20
Mean of differences (cm)	0.26
Minimum of absolute differences (cm)	0.00
Maximum of absolute differences (cm)	13.20
Mean of absolute differences (cm)	5.31
RMS of differences (cm)	7.13

and GSFC00.1 MSL solutions are 7.33 cm, and 7.13 cm respectively, which are in good agreement with the RMS value obtained in test 1.

As test 4 we are going to compare the amplitude of the computed models in this study for the six major tidal constituents, namely, S2, M2, N2, K1, P1, and O1 with those computed by Goddard Space Flight Center (GSFC) via harmonic analysis of the tide gauge observations at 104 submerged (pelagic) tide gauge stations with the global distribution shown in Fig. 16. The data for this test was kindly supplied to the authors by Prof. Ray from GSFC. Among these submerged tide gauge stations, 98 stations are given with all six constituents mentioned above, and the remaining six tide gauges are provided with four tidal constituents, namely, S2, M2, K1, and O1. The results of this comparison are shown in Figs. 17 to 22. The statistical summary of this test is presented in Table 10. All the RMS values of the difference between amplitudes of the tidal constituents are at the sub-centimetre level, which supports the already derived RMS values of test 1, test2, and test 3. It is important to note that since we did not have any information about the phase of the tidal constituents in the tide gauge stations it has not been possible to present comparison of the phase lags.

As test 5 we compare the computed amplitude models for the six dominant tidal constituents (S2, M2, N2, K1, P1, and O1) in this study with those computed by TPXO.6.2 global ocean tide model [24]. As mentioned before, TPXO.6.2 global ocean tide model has been computed using inverse theory

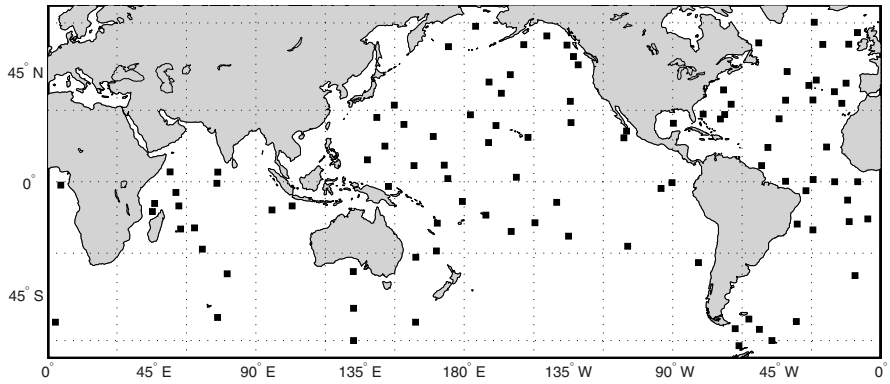


Fig. 16: 104 submerged tide gauge stations where the amplitude of the six major tidal constituents S2, M2, N2, K1, P1, and O1 of our model were compared with those computed by GSFC.

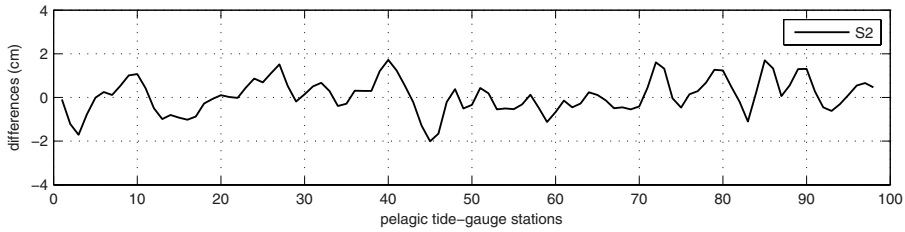


Fig. 17: The differences between the amplitude of the S2 constituent computed in this study and those computed by GSFC at the pelagic tide gauge stations.

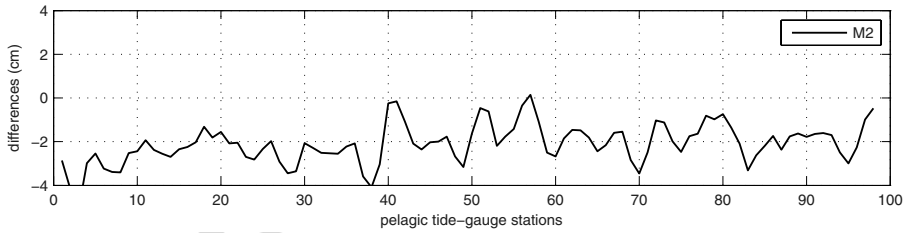


Fig. 18: The differences between the amplitude of the M2 constituent computed in this study and those computed by GSFC at the pelagic tide gauge stations.

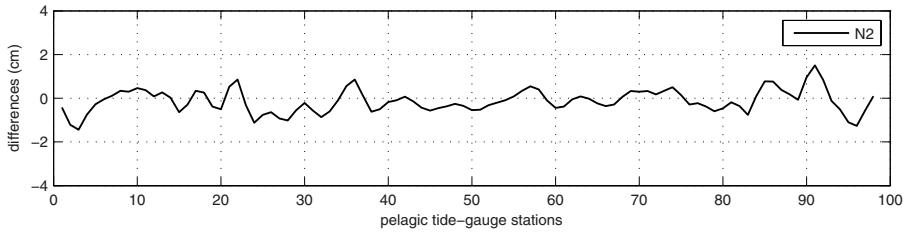


Fig. 19: The differences between the amplitude of the N2 constituent computed in this study and those computed by GSFC at the pelagic tide gauge stations.

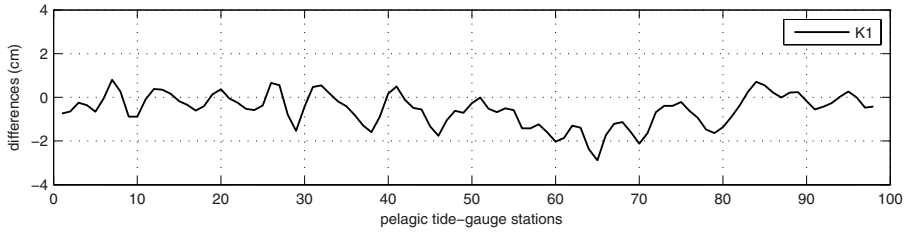


Fig. 20: The differences between the amplitude of the K1 constituent computed in this study and those computed by GSFC at the pelagic tide gauge stations.

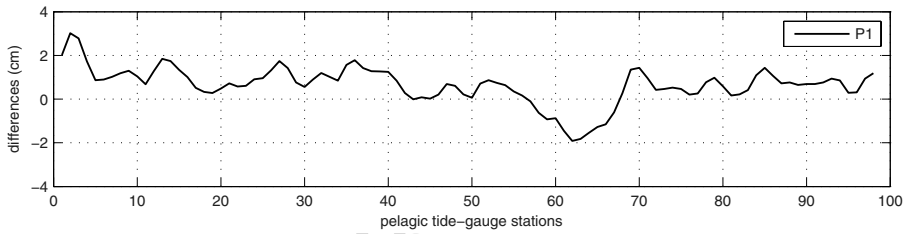


Fig. 21: The differences between the amplitude of the P1 constituent computed in this study and those computed by GSFC at the pelagic tide gauge stations.

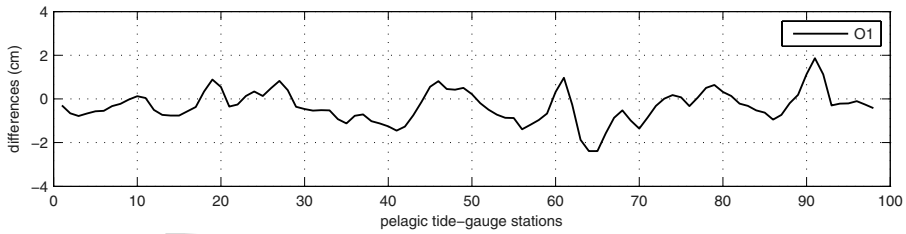


Fig. 22: The differences between the amplitude of the O1 constituent computed in this study and those computed by GSFC at the pelagic tide gauge stations.

Table 10: Statistical summary of the differences between the amplitude of the tidal constituents S2, M2, N2, K1, P1, and O1 computed in this study and those computed by GSFC via harmonic analysis of the tide gauge observations at the pelagic tide gauge stations.

Tidal Constituent	Minimum Difference (cm)	Maximum Difference (cm)	Mean Difference (cm)	Minimum Absolute Difference (cm)	Maximum Absolute Difference (cm)	Mean Absolute Difference (cm)	RMS (cm)
S2	−2.00	1.71	+0.06	0.00	2.00	0.61	0.77
M2	−5.33	0.69	−2.11	0.10	5.33	2.13	1.10
N2	−1.44	1.50	−0.13	0.02	1.50	0.43	0.52
K1	−2.88	0.80	−0.56	0.00	2.88	0.72	0.75
P1	−1.91	3.01	+0.64	0.01	3.01	0.89	0.84
O1	−2.38	1.86	−0.33	0.00	2.38	0.63	0.72

via tide gauge and TOPEX/Poseidon data by finding an optimum balance between sea level observations and hydrodynamics theory and has been presented on a grid of $0.5° \times 0.5°$. The maps and the statistical summary of the difference between the amplitudes of the tidal constituents S2, M2, N2, K1, P1, and O1 computed in this study and those derived by TPXO6.2 model, at 31521 check points over the $1° \times 1°$ grid (shown in Fig. 1), within the geographical area bounded by the latitude $-66° \leq \phi \leq +66°$ are given in Figs. 23 to 28 and in Table 11. The RMS values of the difference between the two models are less than three centimetres, which is in agreement with the results of previous tests. It should be noted that phase is not assessed by this test.

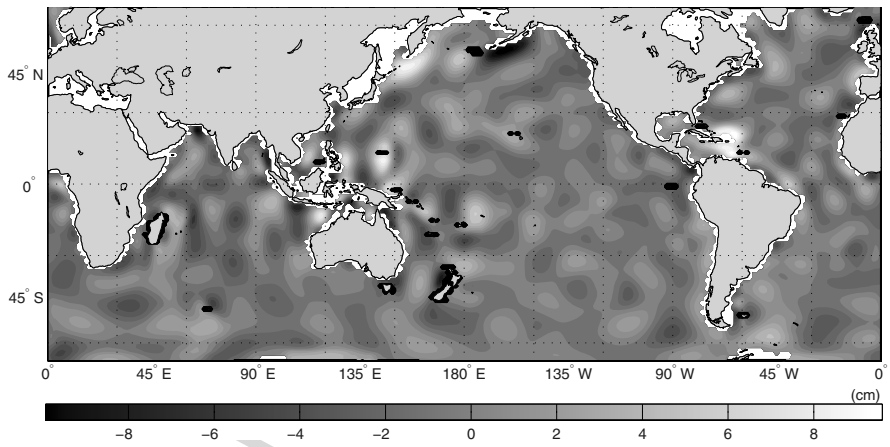


Fig. 23: The difference of computed S2 amplitude model from that derived by use of TPXO.6.2 tide solution.

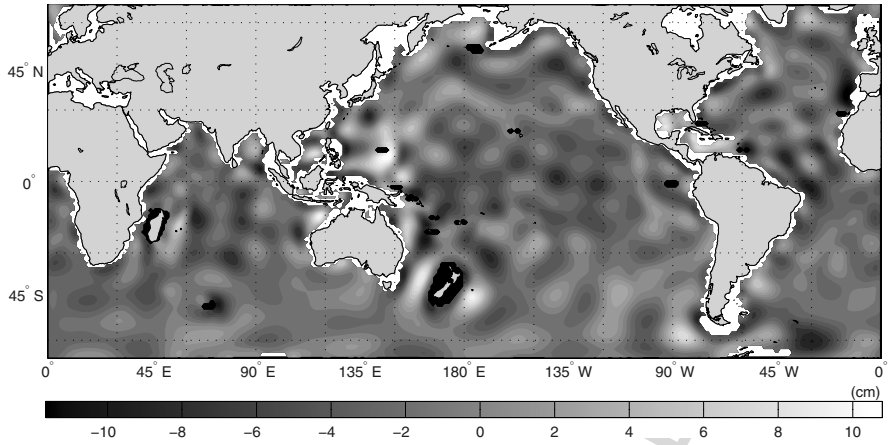


Fig. 24: The difference of computed M2 amplitude model from that derived by use of TPXO.6.2 tide solution.

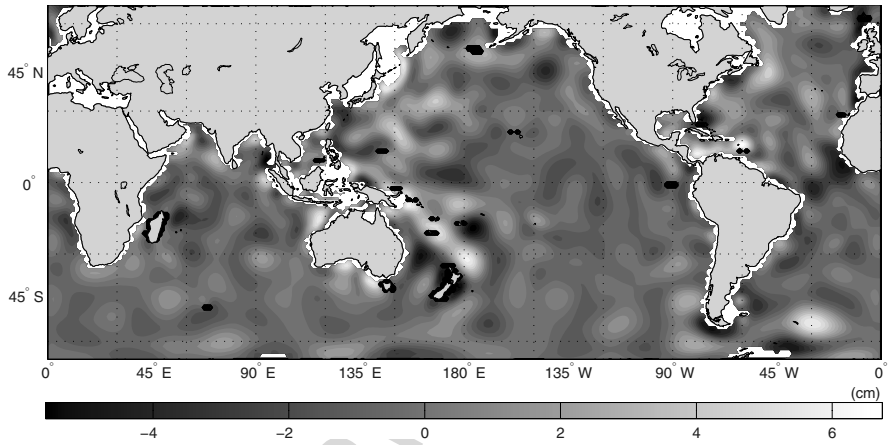


Fig. 25: The difference of computed N2 amplitude model from that derived by use of TPXO.6.2 tide solution.

So far we have compared our MSL and tidal constituent models with that of other solutions and it has been found out that the overall difference of our solutions with the already existing ones is at the order of centimetre in terms of RMS. However, those tests cannot say how or if our models have improved the already existing knowledge about MSL and tidal models. This can only be achieved if, for example, we consider the tidal models derived from tidal observations at the tide gauge stations as a bench mark to test the satellite

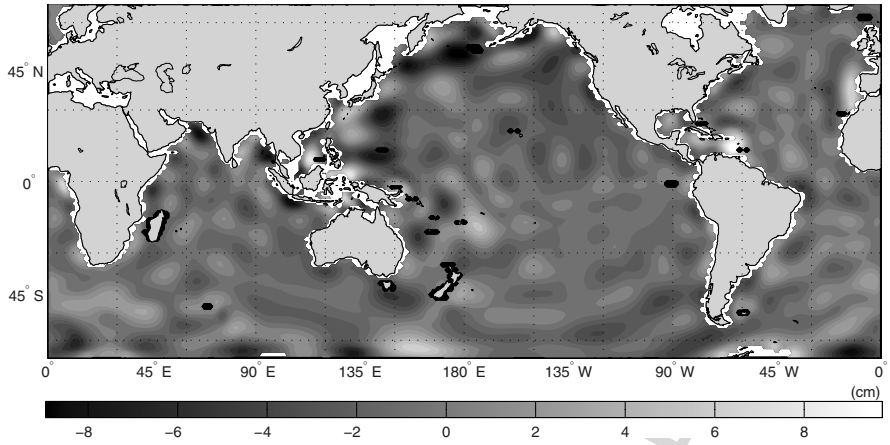


Fig. 26: The difference of computed K1 amplitude model from that derived by use of TPXO.6.2 tide solution.

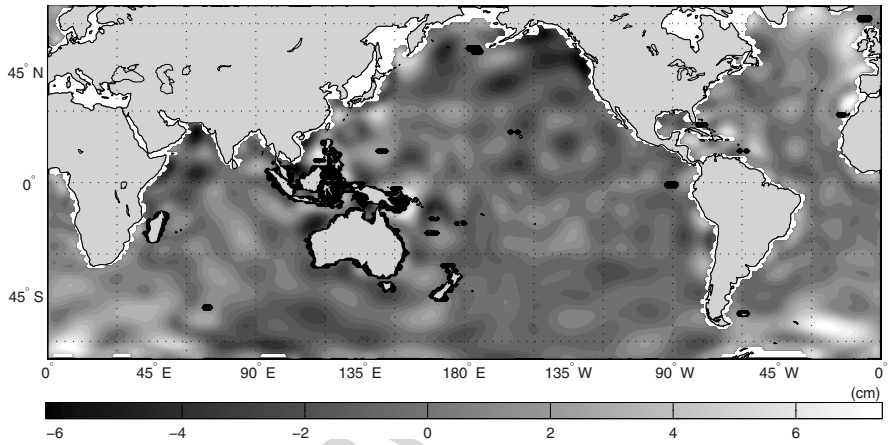


Fig. 27: The difference of computed P1 amplitude model from that derived by use of TPXO.6.2 tide solution.

altimetry derived models. To achieve this we have included test 6. Indeed within this final test we are going to compare the capability of our model with TPXO.6.2 in the synthesis of the amplitude of the six aforementioned major tidal constituents at the 104 submerged tide gauge stations. For this purpose first we compare the computed six tidal constituents by TPXO.6.2 with that of tide gauge stations. The results of this comparison are presented in Figs. 29 to 34 and Table 12. Comparing the RMS of the fit of our model to tidal

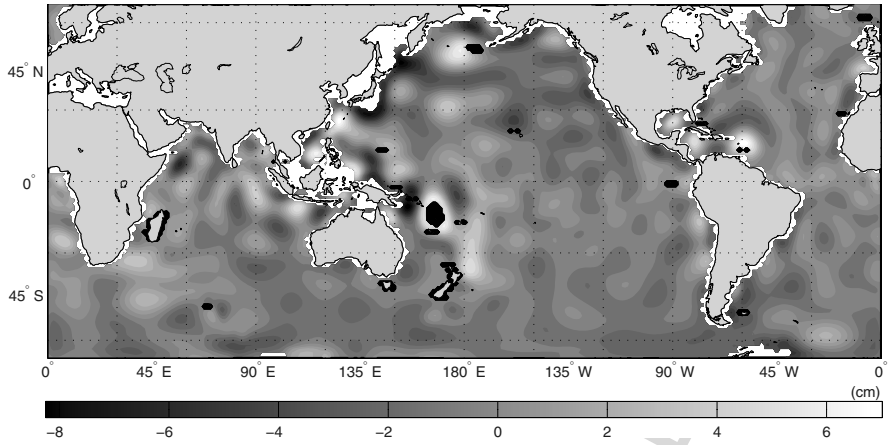


Fig. 28: The difference of computed O1 amplitude model from that derived by use of TPXO.6.2 tide solution.

Table 11: Statistical summary of the difference between the amplitude of the tidal constituents S2, M2, N2, K1, P1, and O1 computed in this study and those computed by TPXO6.2 at the check points of the $1^\circ \times 1^\circ$ grid.

Tidal Constituent	Minimum Difference (cm)	Maximum Difference (cm)	Mean Difference (cm)	Minimum Absolute Difference (cm)	Maximum Absolute Difference (cm)	Mean Absolute Difference (cm)	RMS (cm)
S2	−10.04	10.72	−0.05	0.00	10.72	1.34	1.86
M2	−11.75	11.96	−1.28	0.00	11.96	2.27	2.58
N2	−05.66	07.42	+0.04	0.00	07.42	1.00	1.40
K1	−09.01	10.70	−0.49	0.00	10.70	1.40	1.87
P1	−06.28	08.10	+0.44	0.00	08.10	1.21	1.65
O1	−08.31	07.81	−0.27	0.00	08.31	1.15	1.54

constituents derived from tide gauge stations to that of TPXO.6.2, see Table 10 and Table 12, the TPXO.6.2 is showing a slightly better fit. Naturally, since TPXO.6.2 global ocean tide model has been computed using inverse theory via tide gauge and TOPEX/Poseidon data by finding an optimum balance between sea level observations and hydrodynamics theory, it must reproduce the tide gauge observations better. However, as the results of Table 10 and Table 12 show, our model in spite of not having used any tide gauge data and assimilation of hydrodynamics theory is still following very well the accuracy of the TPXO.6.2 model. Of course, this test has been made over the submerged tide gauge stations and cannot be generalised to coastal areas.

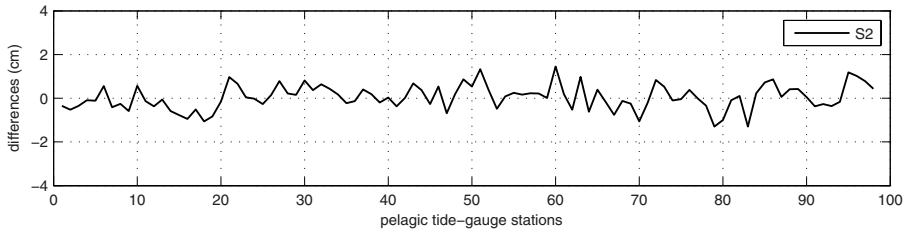


Fig. 29: The differences between the amplitude of the S2 tidal constituent computed by TPXO.6.2 solution and those computed by GSFC at the 104 pelagic tide gauge stations.

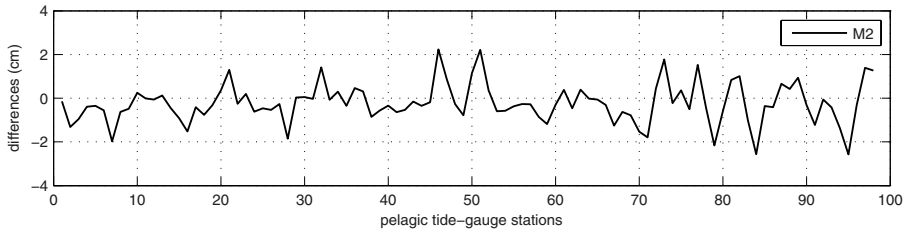


Fig. 30: The differences between the amplitude of the M2 tidal constituent computed by TPXO.6.2 solution and those computed by GSFC at the 104 pelagic tide gauge stations.

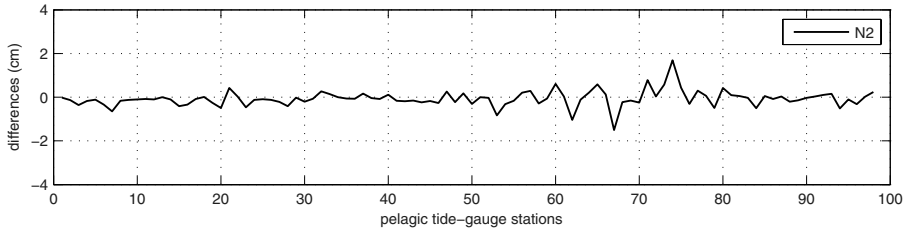


Fig. 31: The differences between the amplitude of the N2 tidal constituent computed by TPXO.6.2 solution and those computed by GSFC at the 104 pelagic tide gauge stations.

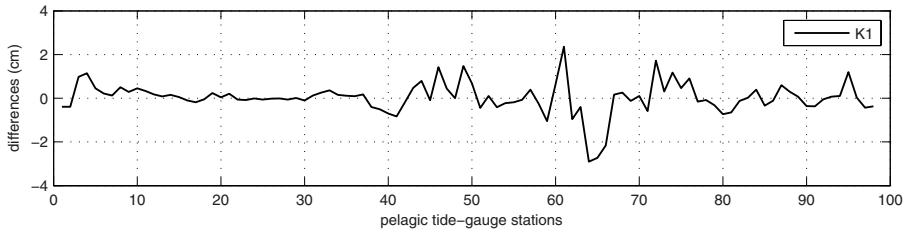


Fig. 32: The differences between the amplitude of the K1 tidal constituent computed by TPXO.6.2 solution and those computed by GSFC at the 104 pelagic tide gauge stations.

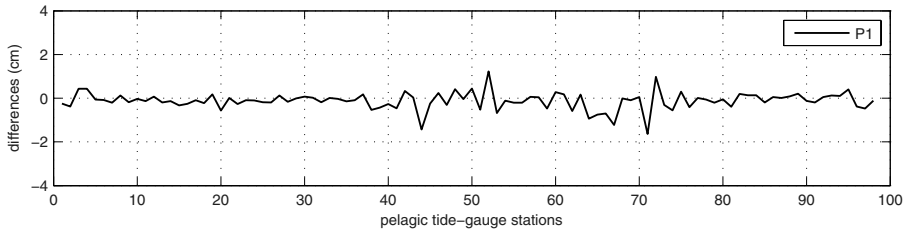


Fig. 33: The differences between the amplitude of the P1 tidal constituent computed by TPXO.6.2 solution and those computed by GSFC at the 104 pelagic tide gauge stations.

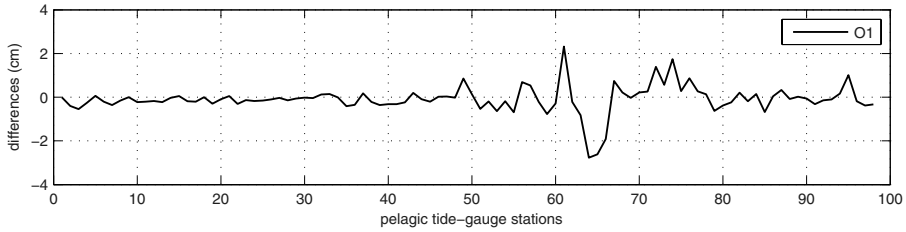


Fig. 34: The differences between the amplitude of the O1 tidal constituent computed by TPXO.6.2 solution and those computed by GSFC at the 104 pelagic tide gauge stations.

Table 12: Statistical summary of the differences between the amplitude of the tidal constituents S2, M2, N2, K1, P1, and O1 computed by TPXO.6.2 and those computed by GSFC via harmonic analysis of the tide gauge observations at the pelagic tide gauge stations.

Tidal Constituent	Minimum Difference (cm)	Maximum Difference (cm)	Mean Difference (cm)	Minimum Absolute Difference (cm)	Maximum Absolute Difference (cm)	Mean Absolute Difference (cm)	RMS (cm)
S2	−2.23	2.36	+0.04	0.01	2.36	0.68	0.87
M2	−2.57	2.23	−0.23	0.01	2.57	0.70	0.90
N2	−1.50	1.68	−0.06	0.00	1.68	0.24	0.36
K1	−2.90	2.35	+0.02	0.00	2.90	0.45	0.72
P1	−1.63	1.22	−0.12	0.01	1.63	0.28	0.39
O1	−2.76	2.31	−0.09	0.00	2.76	0.37	0.63

5 Discussions and Conclusions

In this study we presented an alternative approach to global tidal modeling using satellite altimetry observations, with all theoretical details. The usual approach to tidal analysis using satellite altimetry data is to generate time series of sea level observations from repeated measurements along the satellite tracks. In the usual approach the generated time series are next subject to spectral analysis for the computation of the tidal constituents and MSL. We refer to this approach as the point-wise tidal modeling scheme using satellite altimetry observations. Such ocean tide models are providing high resolution tidal information along satellite tracks. In contrast our contribution is providing spatial tidal information, or a 4-D tidal model, i.e., a model of time and position in terms of orthonormalised spherical harmonics for the amplitude and phase of the tidal constituents. As compared to the point-wise approach the resolution of our model may seem to be lower, however it should be realised that with the point-wise approach the high-resolution tidal information is only available along the satellite tracks where the time series of the sea level observations are constructed. Our approach gives MSL and ocean tidal models with uniform resolution, which is indeed the maximum global resolution that can be derived from, e.g. Jason-1 satellite altimetry observations considering its cross-track spacing. Naturally, if tide gauge observations and/or combination of other satellite altimetry data were used, the resolution of the solution could be further increased. It should be noted that here our intention has been to present a tool for global ocean tidal modeling using satellite altimetry data, which of course can be applied to any satellite altimetry data and/or tide gauge observations. In addition, it is also natural that if our method would be combined with hydrodynamically derived tidal models it could also cover the area outside satellite altimetry data coverage, which has not been our intention in this study.

The comparisons between our tidal model and other solutions based on different approaches showed that we have succeeded in deriving a new empirical global tidal model, consisting of a constant part, together with harmonic parts of up to nine major tidal constituents, i.e., S2, M2, N2, K1, P1, O1, Mf, Mm, and Ssa from six years of Jason-1 satellite altimetry data. According to the numerical tests, our tidal model has centimetre accuracy in estimating the sea surface height at the global scale. The constant part of the model, i.e. MSL, has proved to be in very good agreement with GSFC00.1 MSL solution [80] which is the standard JPL model for Jason-1. In addition, it is shown how Gram-Schmidt orthogonalisation procedure can be used to obtain a set of orthonormal base functions from surface spherical harmonics for global tidal modeling which can also be used for other oceanographic modeling applications.

Acknowledgements. The University of Tehran is gratefully acknowledged for the financial support of this study via grant number No. 8151007/1/02. The authors gratefully acknowledge Jet Propulsion Laboratory (JPL) for supplying the six years of Jason-1 satellite altimetry data used in this study at http://podaac.jpl.nasa.gov. We also would like to appreciate Prof. Richard Ray for providing us with pelagic tide gauge data, computed at Goddard Space Flight Center (GSFC). The fruitful comments of the anonymous reviewers of our paper are gratefully acknowledged. Besides, we would like to thank the editors for their comments and corrections which helped us to improve our contribution.

Appendix A: Nodal Modulations

Nodal corrections in the lunar tides, due to 18.6-year regression of the lunar nodal point, namely, factors f_k and u_k can be computed for the lunar major tidal constituents, namely M2, N2, K1, O1, Mf, and Mm using the formula provided by Doodson 1928 [72] given in Table A.1. Because these corrections are only applied for lunar tidal constituents, the factors f_k and u_k can be defined for solar tidal components, namely S2, P1, and Ssa as $f_k = 1$ and $u_k = 0$. In Table A.1, Ω represents the mean longitude of the Moon's ascending node, which can be computed by using Eq. (A.1) [72].

$$\Omega = 259.157 - 19.32818\,(y - 1900) - 0.05295\,(d + i) \qquad \text{(A.1)}$$

where y is the year, and d is the number of days elapsed since January 1st in the year y. Integer i in Eq. (A.1) can be computed as $i = [(y - 1901)/4]$, i.e., the integral part of $(y - 1901)/4$, which is the number of leap years between the year 1900 and the year y, excluding y as the leap day in the year, which is counted in d [72].

Table A.1: Nodal modulations of the lunar tides in terms of dominant lunar tidal constituents, namely, M2, N2, K1, O1, Mf, and Mm due to the 18.6-year regression of lunar nodal point [72].

Tidal Constituent	f_k	u_k (deg)
M2	$1.0004 - 0.0373\cos\Omega + 0.0002\cos 2\Omega$	$-2.14\sin\Omega$
N2	$1.0004 - 0.0373\cos\Omega + 0.0002\cos 2\Omega$	$-2.14\sin\Omega$
K1	$1.0060 + 0.1150\cos\Omega - 0.0088\cos 2\Omega + 0.0006\cos 3\Omega$	$-8.86\sin\Omega + 0.68\sin 2\Omega - 0.07\sin 3\Omega$
O1	$1.0089 + 0.1871\cos\Omega - 0.0147\cos 2\Omega + 0.0014\cos 3\Omega$	$+10.80\sin\Omega - 1.34\sin 2\Omega + 0.19\sin 3\Omega$
Mf	$1.0429 + 0.4135\cos\Omega - 0.0040\cos 2\Omega$	$-23.74\sin\Omega + 2.68\sin 2\Omega - 0.38\sin 3\Omega$
Mm	$1.0000 - 0.1300\cos\Omega + 0.0013\cos 2\Omega$	$+0.00\sin\Omega + 0.00\sin 2\Omega - 0.00\sin 3\Omega$

Appendix B: Fully Normalised Associated Legendre Functions of the First Kind

Here we define the fully normalised associated Legendre functions of the first kind $\bar{P}_{nm}(x)$ for an arbitrary argument $-1 \le x \le 1$ by means of the following relation [74]:

$$\bar{P}_{nm}(x) = \sqrt{(2-\delta_{m0})(2n+1)\frac{(n-m)!}{(n+m)!}}\frac{1}{2^n n!}(1-x^2)^{\frac{m}{2}}\frac{d^{n+m}}{dt^{n+m}}(x^2-1)^n \tag{B.1}$$

with

$$\delta_{m0} = \begin{cases} 1, & m = 0 \\ 0, & m \neq 0 \end{cases} \tag{B.2}$$

subject to $\forall n = 0, 1, ..., n_{\max}$ and $m = 0, 1, ..., n$. Normalised associated Legendre functions of the first kind $\bar{P}_{nm}(x)$ are usually computed by recursive formulas as shown by Eqs. (B.3) and (B.4) (See for example [81] and [82]).

$$\bar{P}_{nm}(x) = \sqrt{\frac{(2n+1)(2n-1)}{(n-m)(n+m)}}x\bar{P}_{n-1,m}(x) - \sqrt{\frac{(2n+1)(n+m-1)(n-m-1)}{(2n-3)(n+m)(n-m)}}\bar{P}_{n-2,m}(x) \tag{B.3}$$

for $\forall n \neq m$

$$\bar{P}_{mm}(x) = \sqrt{\frac{(2m+1)}{2m}}\sqrt{1-x^2}\bar{P}_{m-1,m-1}(x) \tag{B.4}$$

for $\forall n = m$ with starting values,

$$\bar{P}_{00}(x) = 1 \tag{B.5}$$

$$\bar{P}_{10}(x) = \sqrt{3}x \tag{B.6}$$

$$\bar{P}_{11}(x) = \sqrt{3(1-x^2)} \tag{B.7}$$

Appendix C: Gram-Schmidt Orthonormalisation

Considering a finite or a countably infinite set of linearly independent functions $\{f_i\} = \{f_1, f_2, ..., f_n\}$ defined over inner product space $\mathbb{L}^2_{\mathbb{D}}$, then a set of orthonormal base functions $\{g_i\} = \{g_1, g_2, ..., g_n\}$ over this inner product space can be obtained from the set of functions $\{f_i\}$ using Gram-Schmidt orthonormalising procedure as follows [76]:

$$\begin{array}{ll} g_1 = h_1 / \|h_1\|_{\mathbb{L}^2_{\mathbb{D}}} & h_1 = f_1 \\ g_2 = h_2 / \|h_2\|_{\mathbb{L}^2_{\mathbb{D}}} & h_2 = f_2 - \langle f_2 | g_1 \rangle_{\mathbb{L}^2_{\mathbb{D}}} g_1 \\ g_3 = h_3 / \|h_3\|_{\mathbb{L}^2_{\mathbb{D}}} & h_3 = f_3 - \langle f_3 | g_1 \rangle_{\mathbb{L}^2_{\mathbb{D}}} g_1 - \langle f_3 | g_2 \rangle_{\mathbb{L}^2_{\mathbb{D}}} g_2 \\ \vdots & \vdots \\ g_n = h_n / \|h_n\|_{\mathbb{L}^2_{\mathbb{D}}} & h_n = f_n - \sum_{i=1}^{n-1} \langle f_n | g_i \rangle_{\mathbb{L}^2_{\mathbb{D}}} g_i \end{array} \tag{C.1}$$

where $\langle \cdot | \cdot \rangle_{\mathbb{L}^2_{\mathbb{D}}}$ denotes the inner product of two functions defined as integral over the domain $\mathbb{D}$ and $\|\cdot\|_{\mathbb{L}^2_{\mathbb{D}}}$ expresses the L2-norm of a function defined over the domain $\mathbb{D}$. In general, one can write the above derivation as Eq. (C.2).

$$g_i = \sum_{j=1}^{i} c_{ij} f_j, \quad i = 1, 2, ..., n \tag{C.2}$$

where in the above expansion c_{ij} are known as the "Combination Coefficients" in the Gram-Schmidt orthonormalising procedure. One may write the Eq. (C.2) as follows:

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{C.3}$$

where $\mathbf{x} = (f_1, f_2, \ldots, f_n)^T$, $\mathbf{y} = (g_1, g_2, \ldots, g_n)^T$, and $\mathbf{C}$ is a lower triangular matrix containing the coefficients c_{ij} which can be defined as follows:

$$\mathbf{C} = \begin{bmatrix} c_{11} & 0 & 0 & \cdots & 0 \\ c_{21} & c_{22} & 0 & & 0 \\ c_{31} & c_{32} & c_{33} & & \\ & & & \ddots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix}. \tag{C.4}$$

Practically, an efficient method for computing combination coefficients c_{ij} is based on the Cholesky decomposition of the Gram matrix $\mathbf{G}(f_1, f_2, \ldots, f_n)$,

which for the set of linearly independent functions $\{f_i\} = \{f_1, f_2, ..., f_n\}$ can be defined as follows [83]:

$$\mathbf{G} = \begin{bmatrix} \langle f_1|f_1\rangle_{\mathbb{L}^2_{\mathbb{D}}} & \langle f_1|f_2\rangle_{\mathbb{L}^2_{\mathbb{D}}} & \cdots & \langle f_1|f_n\rangle_{\mathbb{L}^2_{\mathbb{D}}} \\ \langle f_2|f_1\rangle_{\mathbb{L}^2_{\mathbb{D}}} & \langle f_2|f_2\rangle_{\mathbb{L}^2_{\mathbb{D}}} & \vdots & \langle f_2|f_n\rangle_{\mathbb{L}^2_{\mathbb{D}}} \\ \langle f_n|f_1\rangle_{\mathbb{L}^2_{\mathbb{D}}} & \langle f_n|f_2\rangle_{\mathbb{L}^2_{\mathbb{D}}} & \cdots & \langle f_n|f_n\rangle_{\mathbb{L}^2_{\mathbb{D}}} \end{bmatrix}. \tag{C.5}$$

After deriving Gram matrix $\mathbf{G}$, combination coefficients c_{ij} can be readily computed by Eq. (C.6) [66]:

$$\mathbf{C} = (\mathbf{R}^{-1})^T \tag{C.6}$$

where $\mathbf{R}^T$ is the lower triangular matrix derived from Cholesky Decomposition of the Gram matrix $\mathbf{G}$. Equation (C.6) states that to find the combination coefficients c_{ij}, all that is required is to decompose $\mathbf{G}$ in the Cholesky sense and find the inverse of the lower triangular matrix $\mathbf{R}^T$. It remains to evaluate the inner products $\langle f_i|f_j\rangle_{\mathbb{L}^2_{\mathbb{D}}}$ to complete the construction of the orthonormal functions. The inner products of the base functions f_i and f_j shown in Eq. (C.5) can be defined within inner product space $\mathbb{L}^2_{\mathbb{D}}$ as follows:

$$\langle f_i|f_j\rangle_{\mathbb{L}^2_{\mathbb{D}}} = \frac{1}{a_{\mathbb{D}}} \int\int_{\mathbb{D}} f_i f_j d\sigma \tag{C.7}$$

where $d\sigma$ denotes surface differential element, and $a_{\mathbb{D}}$ represents the total sea areas of the study domain $\mathbb{D}$, which can be derived as follows:

$$a_{\mathbb{D}} = \int\int_{\mathbb{D}} d\sigma. \tag{C.8}$$

For computations of the above double integrals in Eqs. (C.7), and (C.8), geometric boundaries of the domain $\mathbb{D}$ have to be known. The above inner products can be computed over a finite element approach within the study domain as the summation of the inner products over the cells. Similar to vectors in linear algebra, the elements of a space of continuous functions have the problem of linear dependence [75]. Therefore one of the most important issues to be addressed when dealing with Gram-Schmidt orthonormalising procedure is to find the number of linearly independent functions $\{f_i\}$, defined over inner product space $\mathbb{L}^2_{\mathbb{D}}$, to construct the orthonormal base functions $\{g_i\}$. Selecting a number of functions $\{f_i\}$ higher than the number of linearly independent functions over the study domain $\mathbb{D}$, can lead to numerical instability in the computations and even singularity. One way to check the numerically dependence of the functions $\{f_i\}$ is by using the Gram matrix determinant $|\mathbf{G}|$. If the Gram matrix determinant $|\mathbf{G}|$ of the given set of functions $\{f_i\}$ becomes zero, then not all of the given functions are numerically linearly independent.

References

1. Ardalan, A.A., Grafarend, E.W., Kakkuri, J.: National height datum, the Gauss-Listing geoid level value w_0 and its time variation (Baltic sea level project: epochs 1990.8, 1993.8, 1997.4). J Geod **76** (2002) 1–28
2. Grafarend, E.W., Ardalan, A.A.: w_0: An estimate in the finnish height datum N60, epoch 1993.4, from twenty-five GPS points of the Baltic sea level project. J Geod **71** (1997) 673–679
3. Burša, M., Kouba, J., Radĕj, K., True, S.A., Vatrt, V., Vojt íšková, M.: Monitoring geoidal potential on the basis of TOPEX/Poseidon altimeter data and EGM96. Scientific Assembly of IAG, Rio de Janeiro (1997)
4. Burša, M., Radĕj, K., Šima, K., True, S.A., Vatrt, V.: Determination of the geopotential scale factor from TOPEX/Poseidon satellite altimetry. Stud Geoph et Geod **41** (1997) 203–216
5. Burša, M., Kouba, J., Muneendra, K., Müller, A., Radĕj, K., True, S.A., Vatrt, V., Vojtíšková, M.: Geoidal geopotential and world height system. Studia Geoph. et Geod **43** (2000) 327–337
6. Schwiderski, E.W.: Ocean tides, 1, global ocean tidal equations. Mar Geod **3** (1980) 161–217
7. Le Provost, C., Genco, M.L., Lyard, F., Vincent, P., Canceil, P.: Spectroscopy of the world ocean tides from a finite-element hydrodynamic model. J Geophys Res **99** (1994) 24777–24797
8. Lefèvre, F., Lyard, F.H., Le Provost, C.: FES98: A new global tide finite element solution independent of altimetry. Geophys Res Lett **27** (2000) 2717–2720
9. Le Provost, C.: An analysis of SEASAT altimeter measurements over a coastal area: The English channel. J Geophys Res **88** (1983) 1647–1654
10. Cartwright, D.E., Ray, R.D.: Oceanic tides from Geosat altimetry. J Geophys Res **95** (1990) 3069–3090
11. Benada, J.R.: PO.DAAC Merged GDR TOPEX-Poseidon Generation B user's handbook, version 2.0. Technical Report D-11007, Jet Propulsion Laboratory (JPL), 4800 Oak Grove Drive, Pasadena, California 91109 (1997)
12. Lemoine, F.G., Kenyon, S.C., Factor, J.K., Trimmer, R.G., Pavlis, N.K., Chinn, D.S., Cox, C.M., Klosko, S.M., Luthcke, S.B., Torrence, M.H., Wang, Y.M., Williamson, R.G., Pavlis, E.C., Rapp, R.H., Olson, T.R.: The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96. NASA Technical Paper NASA/TP-1998-206861, Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, MD (1998)
13. Andersen, O.B.: Global ocean tides from ERS-1 and TOPEX-POSEIDON altimetry. J Geophys Res **100** (1995) 25249–25260
14. Andersen, O.B.: New ocean tide models for loading computations. Bull Int Mare Terr **102** (1995) 9256–9264
15. Cartwright, D.E., Ray, R.D.: Energetics of global ocean tides from Geosat altimetry. J Geophys Res **96** (1991) 16897–16912
16. Cartwright, D.E., Ray, R.D., Sanchez, B.V.: Oceanic tide maps and spherical harmonic coefficients from Geosat altimetry. NASA Tech Memo 104544, Goddard Space Flight Center, National Aeronautics and Space Administration (NASA), Greenbelt, MD (1991)
17. Desai, S.D., Wahr, J.M.: Empirical ocean tide models estimated from TOPEX/Poseidon altimetry. J Geophys Res **100** (1995) 25205–25228

18. Eanes, R.J.: Diurnal and semidiurnal tides from TOPEX/Poseidon altimetry. Eos Trans AGU, Spring Meeting Suppl, Baltimore, MD **75** (1994)
19. Eanes, R.: The CSR4.0 global ocean tide model. ftp://www.csr.utexas.edu/pub/tide (2002)
20. Eanes, R.J., Bettadpur, S.V.: The CSR3.0 global ocean tide model: Diurnal and semi-diurnal ocean tides from TOPEX/Poseidon altimetry. Technical Report CRS-TM-96-05, Centre for Space Research, University of Texas, Austin, TX (1996)
21. Egbert, G.D., Bennett, A.F., Foreman, M.G.G.: TOPEX/Poseidon tides estimated using a global inverse model. J Geophys Res **99** (1994) 24821–24852
22. Egbert, G.D.: Tidal data inversion: Interpolation and inference. Prog Oceanogr **40** (1997) 81–108
23. Egbert, G.D., Bennett, A.F., Foreman, M.G.G.: TOPEX/POSEIDON tides estimated using a global inverse model. J Geophys Res **99** (1999) 24821–24852
24. Egbert, G.D., Erofeeva, L.: Efficient inverse modeling of barotropic ocean tides. J Atmos Ocean Tech **19** (2002) 183–204
25. Egbert, G.D., Ray, R.D.: Significant tidal dissipation in the deep ocean inferred from satellite altimeter data. Nature **405** (2000) 775–778
26. Egbert, G.D., Ray, R.D.: Estimates of M2 tidal energy dissipation from TOPEX/POSEIDON altimetry data. J Geophys Res **106** (2001) 22475–22502
27. Egbert, G.D., Ray, R.D.: Semi-diurnal and diurnal tidal dissipation from TOPEX-Poseidon altimetry. Geophys Res Lett **30** (2003) doi:10.1029/2003GL017676
28. Kagan, B.A., Kivman, G.A.: Modelling of global ocean tides with allowance for island effects. Ocean Dynam **45** (1993) 1–13
29. Kantha, L.H.: Barotropic tides in the global oceans from a nonlinear tidal model assimilating altimetric tides. 1. model description and results. J Geophys Res **100** (1995) 25283–25308
30. Knudsen, P.: Global low harmonic degree models of seasonal variability and residual ocean tides from TOPEX/Poseidon altimeter data. J Geophys Res **99** (1994) 24643–24655
31. Krohn, J.: A global ocean tide model with high resolution in shelf areas. Mar Geophys Res **7** (1984) 231–246
32. Le Provost, C., Lyard, F., Molines, J.M., Genco, M.L., Rabilloud, F.: A hydrodynamic ocean tide model improved by assimilating a satellite altimeter-derived data set. J Geophys Res **103** (1998) 5513–5529
33. Le Provost, C.: FES2002: A new version of the FES tidal solution series. Jason-1 Science Working Team Meeting, Biarritz, France (2002)
34. Lefèvre, F., Lyard, F.H., Le Provost, C.: FES99: A global tide finite element solution assimilating tide gauge and altimetric information. J Atmos Ocean Tech **19** (2002) 1345–1356
35. Letellier, T.: Etude des ondes de marée sur les plateux continentaux. PhD thesis, Université de Toulouse III, Ecole Doctorale des Sciences de l'Univers, de l'Environnement et de l'Espace (2004)
36. Letellier, T., Lyard, F., Lefèvre, F.: The new global tidal solution: FES2004. Ocean Surface Topography Science Team Meeting, Saint Petersburg, FL (2004)
37. Ma, X.C., Shum, C.K., Eanes, R.J., Tapley, B.D.: Determination of ocean tides from the first year of TOPEX/Poseidon altimeter measurements. J Geophys Res **99** (1994) 24809–24820

38. Matsumoto, K., Ooe, M., Sato, T., Segawa, J.: Ocean tide model obtained from TOPEX/Poseidon altimetry data. J Geophys Res **100** (1995) 25319–25330
39. Matsumoto, K., Takanezawa, T., Ooe, M.: Ocean tide models developed by assimilating TOPEX/Poseidon altimeter data into hydrodynamical model: A global model and a regional model around Japan. J Oceanogr **56** (2000) 567–581
40. Mazzega, P., Merge, M., Francis, O.: TOPEX/Poseidon tides: The OMP2 atlas. EOS Trans, AGU, Fall Meet suppl **75** (1994)
41. Ray, R.D., Sanchez, B.V., Cartwright, D.E.: Some extensions to the response method of tidal analysis applied to TOPEX/Poseidon altimetry. Eos Trans AGU, Spring Meet Suppl, Baltimore, MD **75** (1994)
42. Ray, R.D.: A global ocean tide model from TOPEX/Posidon altimetry: GOT99.2. NASA Tech Memo NASA/TM-1999-209478, Goddard Space Flight Center, Goddard Space Flight Center (1999)
43. Sanchez, B.V., Pavlis, N.K.: Estimation of main tidal constituents from TOPEX altimetry using a proudman function expansion. J Geophys Res **100** (1995) 25229–25248
44. Schrama, E.J.O., Ray, R.D.: A preliminary tidal analysis of TOPEX/Poseidon altimetry. J Geophys Res **99** (1994) 24799–24808
45. Schwiderski, E.W.: Ocean tides, 2, a hydronomical interpolations model. Mar Geod **3** (1980) 218–257
46. Schwiderski, E.W.: On charting global ocean tides. Rev Geophys Space Phys **18** (1980) 243–268
47. Tierney, C.C., Kantha, L.H., Born, G.H.: Shallow and deep water global ocean tides from altimetry and numerical modeling. J Geophys Res **105** (2000) 11259–11277
48. Wang, Y.M., Rapp, R.H.: Estimation of sea surface topography, ocean tides, and secular changes from Topex altimeter data. Technical Report 430, Dep Geod Sci Surv, Ohio State University, Columbus (1994)
49. Andersen, O.B., Woodworth, P.L., Flather, R.A.: Intercomparison of recent ocean tide models. J Geophys Res **100** (1995) 25261–25282
50. Baker, T.F., Bos, M.S.: Validating earth and ocean tide models using tidal gravity measurements. Geophys J Int **152** (2003) 468–485
51. Bos, M.S., Baker, T.F., Røthing, K., Plag, H.P.: Testing ocean tide models in the nordic seas with tidal gravity observations. Geophys J Int **150** (2002) 687–694
52. King, M.A., Padman, L.: Accuracy assessment of ocean tide models around antarctica. Geophys Res Lett **32** (2005) doi:10.1029/2005GL023901
53. King, M.A., Penna, N.T., Clarke, P.J., King, E.C.: Validation of ocean tide models around antarctica using onshore GPS and gravity data. Geophys Res Lett **110** (2005) doi:10.1029/2004JB003390
54. Llubes, M., Mazzega, P.: Testing recent global ocean tide models with loading gravimetric data. Prog Oceanogr **40** (1997) 369–383
55. Shum, C.K., Woodworth, P.L., Andersen, O.B., Egbert, G.D., Francis, O., King, C., Klosko, S.M., Le Provost, C., Li, X., Molines, J.M., Parke, M.E., Ray, R.D., Schlax, M.G., Stammer, D., Tierney, C.C., Vincent, P., Wunsch, C.I.: Accuracy assessment of recent ocean tide models. J Geophys Res **102** (1997) 25173–25194
56. Urschl, C., Dach, R., Hugentobler, U., Schaer, S., Beutler, G.: Validating ocean tide loading models using GPS. J Geod **78** (2005) 616–625
57. Groves, G.W., Reynolds, R.W.: An orthogonalized convolution method of tide prediction. J Geophys Res **80** (1975) 4131–4138

58. Darwin, G.H.: Report of a committee for the harmonic analysis of tidal observations. British Association Report (1883) 48–118
59. Cherniawsky, J.Y., Foreman, M.G.G., Crawford, W.R., Henry, R.F.: Ocean tides from TOPEX/Poseidon sea level data. J Atmos Ocean Tech **18** (2001) 649–664
60. Cherniawsky, J.Y., Foreman, M.G.G., Crawford, W.R., Beckley, B.D.: Altimeter observations of sea-level variability off the west coast of north america. Int J Remote Sens **25** (2004) 1303–1306
61. Ponchaut, F., Lyard, F., Prevost, C.L.: An analysis of the tidal signal in the WOCE sea level dataset. J Atmos Ocean Tech **18** (2001) 77–91
62. Smith, A.J.E., Ambrosius, B.A.C., Wakker, K.F., Woodworth, P.L., Vassie, J.M.: Comparison between the harmonic and response methods of tidal analysis using TOPEX-Poseidon altimetry. J Geod **71** (1997) 695–703
63. Smith, A.J.E., Ambrosius, B.A.C., Wakker, K.F., Woodworth, P.L., Vassie, J.M.: Ocean tides from harmonic and response analysis on TOPEX-Poseidon altimetry. Remote Sensing: Earth, Ocean and Atmosphere Advances in Space Research **22** (1999) 1541–1548
64. Smith, A.J.E.: Ocean tides from satellite altimetry. PhD thesis, Delft Institute for Earth-Oriented Space Research, Delft University of Technology, Delft, The Netherlands (1997)
65. Mainville, A.: The altimetry-gravimetry problem using orthonormal base functions. Technical Report 373, Dep Geod Sci Surv, Ohio State University, Columbus (1987)
66. Hwang, C.: Orthogonal functions over the oceans and applications to the determination of orbit error, geoid and sea surface topography from satellite altimetry. Technical Report 414, Dep Geod Sci Surv, Ohio State University, Columbus (1991)
67. Hwang, C.: Spectral analysis using orthonormal functions with a case study on the sea surface topography. Geophys J Int **115** (1993) 1148–1160
68. Hwang, C.: Orthonormal function approach for Geosat determination of sea surface topography. Mar Geod **18** (1995) 245–271
69. Rapp, R.H., Zhang, C., Yi, Y.: Analysis of dynamic ocean topography using TOPEX data and orthonormal functions. J Geophys Res **101** (1995) 22583–22598
70. Rapp, R.H., Zhang, C., Yi, Y.: Comparison of dynamic ocean topography using TOPEX data and orthonormal function. J Geophys Res **101** (1996) 22583–22598
71. Rapp, R.H.: Ocean domains and maximum degree of spherical harmonic and orthonormal expansions. Technical Report NASA/CR-1999-208628, Goddard Space Flight Center, National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, MD, Maryland 20771 (1999)
72. Doodson, A.T.: The analysis of tidal observations. Philosophical Transactions of the Royal Society of London, Series A, Containing papers of a Mathematical or Physical Character **227** (1928) 223–279
73. Thong, N.C., Grafarend, E.W.: A spheroidal model of the terrestrial gravitational field. Manuscr Geod **14** (1989) 285–304
74. Heiskanen, W.A., Moritz, H.: Physical Geodesy. W.H.Freeman, New York (1967)
75. Davis, P.J.: Interpolation and Approximation. Dover Publications (1975)

76. Kreyszig, E.: Introductory Functional Analysis with applications. John Wiley and Sons, New York, Chi Chester, Toronto (1978)
77. Picot, N., Case, K., Desai, S., Vincent, P.: AVISO and PO.DAAC user handbook, IGDR and GDR Jason products. Technical Report JPL D-21352, Jet propulsion Laboratory (2004)
78. Yi, Y.: Determination of gridded mean sea surface from TOPEX, ERS-1 and GEOSAT altimeter data. Technical Report 434, Dep Geod Sci Surv, Ohio State University, Columbus (1995)
79. Rapp, R.H., Yi, Y.: Role of ocean variability and dynamic ocean topography in the recovery of the mean sea surface and gravity anomalies from satellite altimeter data. J Geod **71** (1997) 617– 629
80. Wang, Y.M.: GSFC00 mean sea surface, gravity anomaly, and vertical gravity gradient from satellite altimeter data. J Geophys Res **106** (2001) 31167–31174
81. Rummel, R., Gelderen, M.V., Koop, R., Schrama, E.J.O., Sanso, F., Brovelli, M., Miggliaccio, F., Sacerdote, F.: Spherical harmonic analysis of satellite gradiometry. Technical Report 39, Netherlands Geodetic Commission, Delft University of Technology, Faculty of Geodetic Engineering (1993)
82. Tsoulis, D.: Spherical harmonic computations with topographic/isostatic coefficients. Technical Report IAPG/FESG No. 3, Institute of Astronomical and Physical Geodesy (IAPG), Technical University of Munich (1999)
83. Golub, G.H., Loan, C.F.V.: Matrix Computations. Third edn. John Hopkins University Press, Baltimore, MD (1996)

Fourier, Scattering, and Wavelet Transforms: Applications to Internal Gravity Waves with Comparisons to Linear Tidal Data

Jim A. Hawkins[1], Alex Warn-Varnas[2], and Ivan Christov[2,3]

[1] Planning Systems Inc., Slidell, LA 70458, USA, jhawkins@psislidell.com
[2] Naval Research Laboratory, Stennis Space Center, MS, 39529, USA
[3] Northwestern University, Evanston, IL 60208, USA

Abstract. Analysis of tides and internal waves from model studies in the South China Sea is done using three techniques. We summarize results from standard Fourier methods, continuous wavelet analysis and the direct scattering transform. Because the Fourier and wavelet analysis are inherently linear methods their utility in application to nonlinear dynamics is often questioned. Nevertheless, they have shown to be useful in delineating first order dynamics (for example finding fundamental modes). On the other hand the scattering transform, sometimes described as a 'nonlinear Fourier' technique, can in some cases succeed in elucidating nonlinear dynamics where linear methods have proven less successful. We apply these procedures to model results from Lamb's 2D non-hydrostatic model applied to the South China Sea and in some cases the multi-component tides used to force the Lamb model.

AU: Please provide Keywords.

1 Introduction

It is widely accepted that the first recorded internal wave was that described by J. Scott Russel. The correct mathematical framework for the phenomena came later with Korteweg and de Vries and their description of the KdV solutions to the one dimensional problem (see [1]) for a brief account of the early history of internal waves). Oceanic internal waves arise because of the naturally occurring stratification of the ocean's water column. As a result, internal waves arise throughout the earth's oceans. Well known examples include the internal waves observed in the Strait of Gibraltar and in the Sulu Sea. The University of Delaware maintains a website (http://atlas.cms.udel.edu/) containing an exhaustive catalogue of internal wave images gathered by satellite.

Internal gravity waves (IW) occur as a result of tidal flow over steep topography, for example, coastal shelves and deep water sills. As the tide flows over

the topography the thermocline is depressed resulting in the generation of a bore. The bore propagates and its leading edge steepens through nonlinear effects. Thereafter, the bore degenerates into solitary waves through frequency and amplitude dispersion [1, 2].

Dispersive effects become increasingly evident as the IWs propagate, causing the amplitudes and number of oscillations to vary over time. In this sense IWs are non-stationary, that is their spatial and temporal scales change as the IWs develop.

Dispersion of IWs is commonly summarized in amplitude, wavelength, and velocity relations. For example the amplitude of a solitary wave depression can be plotted against its width (or half-width) over a range of propagation distances [3]. The amplitudes and widths are often obtained by inspection. While useful (and widely used) there remains some subjectivity involved in determining the participant values.

Objective analyses exist to investigate non-stationary processes. They include statistical methods such as principle component analysis and time-frequency analysis including Fourier techniques, wavelets and multiscale analysis. Recent studies have employed these tools to study a variety of problems (see for example [4]).

This paper describes in some detail the application of three techniques to modeling results for IWs generated in the Strait of Luzon and propagating into the South China Sea. They are (1) the discrete Fourier transform (DFT), the direct scattering transform (DST), and (3) the wavelet transform (WT). Furthermore, analysis of tidal data used in driving the IW model is included for comparison.

A good deal of interest exists concerning the generation and propagation of internal gravity waves in the South China Sea. As part of the Asian Seas Acoustics Experiment (ASIAEX), field measurements (encompassing a variety of platforms) took place in 2001 in South China Sea to quantify acoustic volume interaction during presence of solitary waves. Analysis of the field data showed the presence of solitary waves with amplitudes up to 160 m, and phase speeds of .83 m/s to 1.6 m/s [5]. The recent 2005 and 2006 Windy Island Experiment [6] measured amplitudes of up to 250 m and phase speeds up to 3.4 m/s. Recent modeling studies predict the occurrence of solitary waves consistent with those observed. The internal waves (IW) appear to be generated by deep water sills in the Luzon Strait. The IWs travel across the South China Sea towards the coast of China, their structure evolving as they propagate (see Fig. 1).

In the following the model predictions are discussed in Sect. 2. The analysis met holds are briefly described in Sect. 3 in the following order, the DFT in Sect. 3.1, the DST in Sect. 3.2, and the WT in Sect. 3.3. Analysis of results are then discussed in Sect. 4. A concluding summary is contained in Sect. 5.

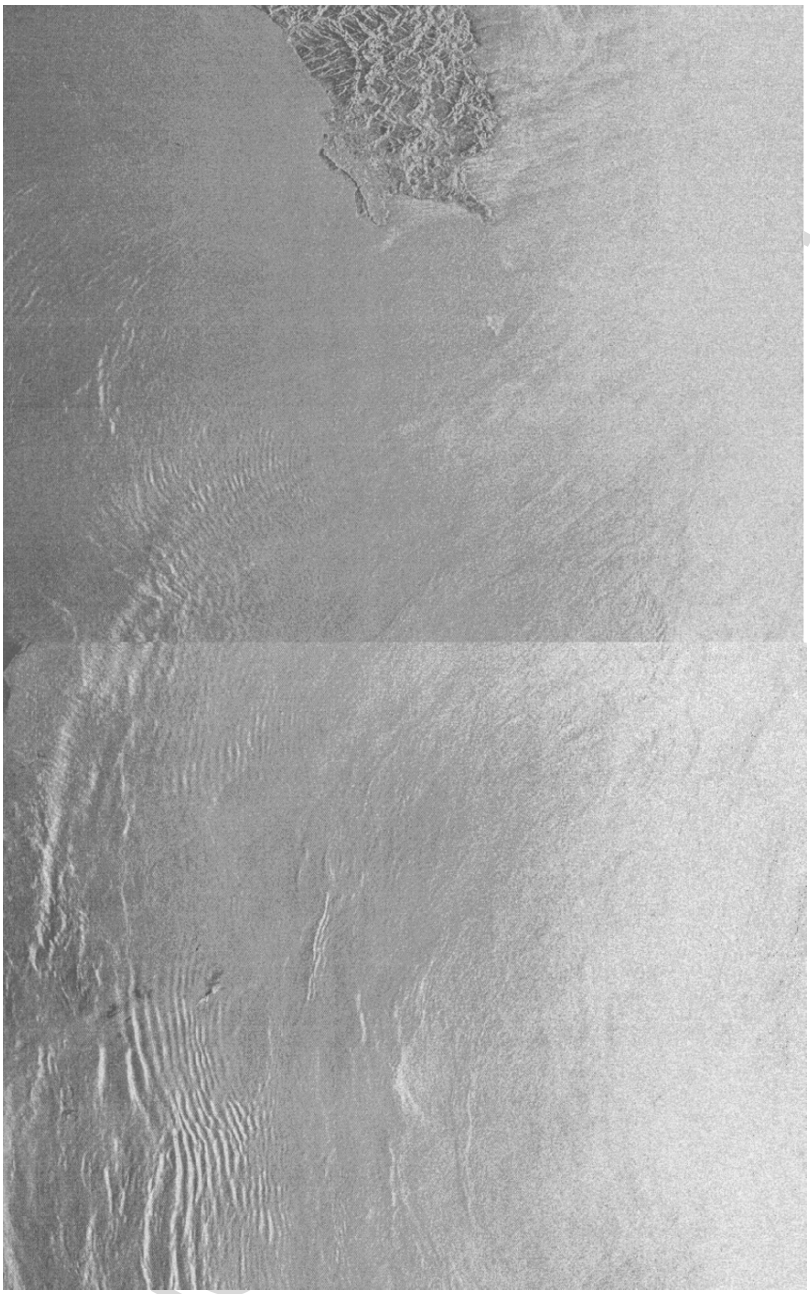


Fig. 1: Taiwan is located at the uppermost edge of the image. The Luzon Strait spans the region running south towards the Philippines (not shown) at the lower boundary. Internal waves can be seen in the lower left quadrant propagating westward (from University of Delaware, Center for Remote Sensing, url: atlas.cms.udel.edu).

2 Model Predictions

We have undertaken a model study of the area using the 2D ocean model developed by Kevin Lamb [2]. The model is initialized using analytic fits which approximate real density and bathymetry data. Internal waves are generated by tidal forcing from the Navy Coastal Prediction Model (NCOM) tidal model [7]. The results are discussed in the following paragraphs.

Figure 2 shows results from Lamb's 2D ocean model after 70 hours of simulation time. The density field is shown with several isopycnal lines spanning the domain of the upper 1000 m of ocean near the modeled sill of the Luzon Strait (the grey patch near the leftmost edge of the figure). Because the IWs described here begin as a tidal bore (a sharp depression of the pycnocline) and evolve into a group of solitary waves they can be identified throughout the domain as IW 'packets'. Three IW packets are easily noted located at ranges running east to west at −250 km, −550 km, and lastly near −700 km. The IWs are propagating toward a shelf located on the Chinese coast.

As the IWs propagate it is apparent that the nature of the IW packet is qualitatively changing over time. The IW at −250 km is tightly packed with numerous oscillations, at −550 km the oscillations have separated with large

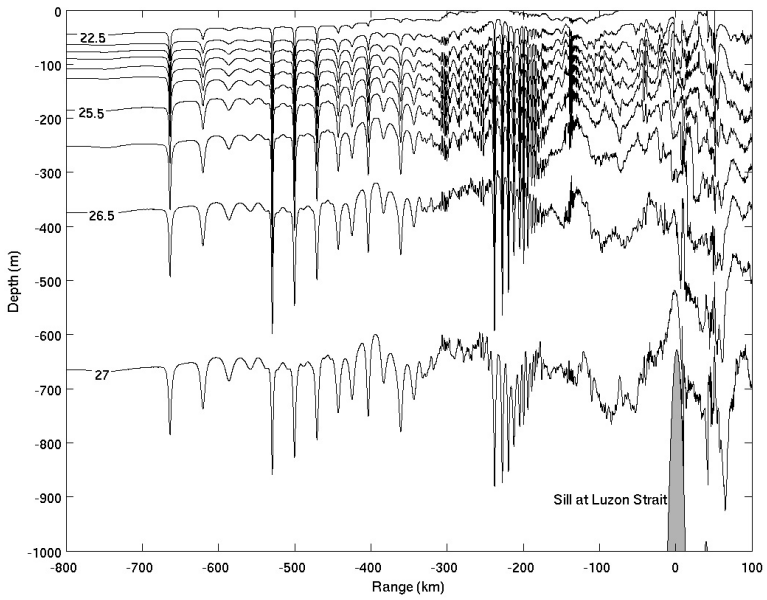


Fig. 2: Isopycnals (22.5–27, sigma-t units) are shown within the internal wave field. Results are from Lamb's 2D ocean model for the Luzon Strait.

amplitude oscillations at the leading edge of the IW packet, and still further at −700 km the separation continues, however the amplitude and number of oscillations has noticeably diminished.

The goal of the analysis in this paper is to compare the results from three techniques for quantifying the evolution of the internal waves. Each provides a different and complementary view of IW behavior.

3 Methods

In this section the techniques used for analysis are described. The DFT and WT are briefly summarized along with a more detailed development of the direct scattering transform. The descriptions here are provided as a point of reference for the discussion that follows. More detailed information can be found in the references.

3.1 Discrete Fourier Transform

An estimate of the energy or power at a particular Fourier frequency or wavelength characterizing a sequence is sought. First note that the squared value of a sequence integrated over time is a measure of energy. In this case we have the following expression for the energy, E, of a sequence $x(t)$ measured over time t with a period, L,

$$E = \int_0^L x(t)^2 dt. \tag{1}$$

The amount of power, W, in the sequence over the period is therefore given by the following equation,

$$P = \frac{1}{L}\int_0^L x(t)^2 dt. \tag{2}$$

The ideas above are implemented in a straight forward way by the *periodogram*. The discrete version of the periodogram, P_{xx}, can be written as follows [8],

$$P_{xx} = \frac{|X_L(f)|^2}{f_s L}, \tag{3}$$

where

$$X_L(f) = \sum_{n=0}^{L-1} x_L[n] \exp(-2\pi j f n/f_s), \tag{4}$$

is the discrete Fourier transform of the sequence $x(t)$ and f_s is the sampling frequency.

As will be seen in the following sections the DFT is more successfully applied to problems that are linear than to nonlinear problems like internal

waves. This leads one to speculate that perhaps a 'nonlinear' Fourier approach would prove more successful. The direct scattering transform is sometimes thought of as a nonlinear Fourier method in the sense that it uncovers constituent components in nonlinear problems.

3.2 Direct Scattering Transform

Next, we turn to the relationship between the (periodic, inverse) scattering transform and the (ordinary) Fourier transform, the interpretation of the former as a nonlinear generalization of the latter, and the algorithm for computing the DST spectrum of a data set. To this end, we begin by formulating the Korteweg–de Vries (KdV) equation, which describes the dynamics of weakly-nonlinear dispersive waves, for the internal-waves problem.

Under the assumption that the internal solitary waves are 'long' and that they are traveling in a 'shallow' layer (this will be made more precise below), the governing (KdV) equation of the pycnocline displacement, which we denote by $\eta(x,t)$, is

$$\eta_t + c_0\eta_x + \alpha\eta\eta_x + \beta\eta_{xxx} = 0, \quad 0 \le x \le L, \quad t \ge 0, \tag{5}$$

where $L(>0)$ is the spatial period (i.e., the length of the domain), and the subscripts denote partial differentiation with respect to an independent variable. In addition, $c_0(>0)$, $\alpha(<0)$, and $\beta(>0)$ are (constant) physical parameters (see, e.g., Apel [1] for their interpretation). The simplest way to evaluate them is to assume a two-layer (density) stratification [9]. Then we have

$$c_0^2 \simeq g\left(\frac{\varrho_2-\varrho_1}{\varrho_1}\right)\left(\frac{h_1h_2}{h_1+h_2}\right), \quad \alpha \simeq \frac{3c_0}{2}\left(\frac{h_1-h_2}{h_1h_2}\right), \quad \beta \simeq \frac{c_0h_1h_2}{6}, \tag{6}$$

where h_1 and $h_2(>h_1)$ are the distances from the *unperturbed* pycnocline to the free surface and to the ocean bottom, respectively, while ϱ_1 and $\varrho_2(>\varrho_1)$ are the fluid densities in the top and bottom layers, respectively. Furthermore, we are interested in the periodic initial-value problem. That is to say, given a data set $\eta(x,t=0)$ such that $\eta(x+L,0)=\eta(x,0)$ for $0\le x\le L$, we wish to determine its evolution $\eta(x,t)$ for $t>0$.

The strategy for solving the periodic KdV equation by the scattering transform can be split into two distinct steps: the *direct problem* and the *inverse problem*. The former, which is termed the *direct scattering transform* (DST), consists of solving the Schrödinger eigenvalue problem

$$\{-\partial_{xx} - \kappa\,\eta(x,0)\}\,\psi = \mathcal{E}\psi, \tag{7}$$

where $\kappa \equiv \alpha/(6\beta)$ is a nonlinearity-to-dispersion ratio, ψ is an eigenfunction, and $\mathcal{E}$ is a (real) spectral eigenvalue such that $\sqrt{\mathcal{E}}$ is a (complex) wavenumber. For periodic signals, as we have assumed, it is well-known that the eigenvalues fall into two distinct sets [10]: the *main spectrum*, which we write as the set $\{\mathcal{E}_n\}_{n=0}^{2N}$, and the *auxiliary spectrum*, which we write as the set $\{\mu_n^0\}_{n=0}^{N-1}$, where N is the number of degrees of freedom (i.e., nonlinear normal modes).

On the other hand, the inverse problem consists of constructing the *nonlinear Fourier series* from the spectrum $\{\mathcal{E}_n\} \cup \{\mu_n^0\}$ using Abelian hyperelliptic functions [10] or the Riemann Θ-function [11]. In former case, which is the so-called μ-representation of the scattering transform, the *exact* solution of (5), subject to periodic boundary conditions, takes the form

$$\eta(x,t) = \frac{1}{\kappa}\left\{2\sum_{n=0}^{N-1}\mu_n(x,t) - \sum_{n=0}^{2N}\mathcal{E}_n\right\}. \tag{8}$$

It is important to note that all *nonlinear* waves and their *nonlinear* interactions are accounted for in this *linear* superposition. Unfortunately, the computation of the nonlinear normal modes (i.e., the hyperelliptic functions $\mu_n(x,t)$, $0 \le n \le N-1$) is *highly* nontrivial; however, numerical approaches have been developed [12] and successfully used in practice [13, 14]. In addition, we note that the auxiliary spectrum, often referred to as the hyperelliptic function 'phases,' is such that $\mu_n^0 \equiv \mu_n(0,0)$ [10, 14].

Several special cases of (8) offer insight into why the latter is analogous to the (ordinary) Fourier series. In the small-amplitude limit, i.e., when $\max_{x,t}|\mu_n(x,t)| \ll 1$, we have $\mu_n(x,t) \sim \cos(x - \omega_n t + \phi_n)$, where ω_n is a frequency and ϕ_n a phase. Therefore, if we suppose that all the nonlinear normal modes fall in the small-amplitude limit, then (8) reduces to the ordinary Fourier series. This relationship is more than just an analogy, Osborne and Bergamasco [15] give a rigorous derivation of the (ordinary) Fourier transform from the scattering transform. Next, if there are no interactions, e.g., the spectrum consists of a single wave (i.e., $N = 1$), we have $\mu_0(x,t) = \mathrm{cn}^2(x - \omega_0 t + \phi_0|m_0)$, which is a Jacobian elliptic function with modulus m_0. In fact, it is the well-known *cnoidal wave* solution of the periodic KdV equation [1].

For the hyperelliptic representation of the nonlinear Fourier series, given by (8), the wavenumbers are *commensurable* with those of the ordinary Fourier series, i.e., $k_n = 2\pi(n+1)/L$ $(0 \le n \le N-1)$ [10, 13, 14]. However, this is not the only way to classify the nonlinear normal modes. One can use the 'elliptic modulus' (or, simply, modulus) m_n, termed the 'soliton index,' of each of the hyperelliptic functions, which can be computed from the discrete spectrum as

$$m_n = \frac{\mathcal{E}_{2n+2} - \mathcal{E}_{2n+1}}{\mathcal{E}_{2n+2} - \mathcal{E}_{2n}}, \quad 0 \le n \le N-1. \tag{9}$$

Then, each nonlinear normal modes falls into one of three distinct categories based on its soliton index:

1. $m_n \ge 0.99 \Rightarrow$ solitons, in particular, $\mathrm{cn}^2(x|m=1) = \mathrm{sech}^2(x)$;
2. $m_n \ge 0.5 \Rightarrow$ nonlinearly interacting cnoidal waves (e.g., moderate-amplitude Stokes waves);
3. $m_n \ll 1.0 \Rightarrow$ radiation, in particular, $\mathrm{cn}^2(x|m=0) = \cos^2(x)$.

Furthermore, it can be shown [13, 14] that the amplitudes of the hyperelliptic functions are given by

$$A_n = \begin{cases} \frac{2}{\kappa}(\mathcal{E}_{\text{ref}} - \mathcal{E}_{2n+1}), & \text{for solitons;} \\ \frac{1}{2\kappa}(\mathcal{E}_{2n+2} - \mathcal{E}_{2n+1}), & \text{otherwise (radiation);} \end{cases} \tag{10}$$

where $\mathcal{E}_{\text{ref}} = \mathcal{E}_{2n^*+2}$ is the *soliton reference level* with n^* being the largest n for which $m_n \geq 0.99$. Then, clearly, the number of solitons in the spectrum is $N_{\text{sol}} \equiv n^*$.

To summarize: the DST consists of computing the amplitudes and degrees of nonlinearity (moduli) of the nonlinear normal modes. Furthermore, if the KdV equation governs (at least to a good approximation) the evolution of the data set, then the DST spectrum characterizes the dynamics for *all* time. If that is not case, then the DST provides an instantaneous projection of the dynamics onto the solution space of the periodic KdV equation, giving us a nonlinear characterization of the data set at a particular *instant* of time.

In addition, the DST has been successfully employed in the Fourier-like decomposition of data from inherently nonlinear physical phenomena such as shallow-water ocean surface waves [13], laboratory-generated surface waves [14], and internal gravity waves in a stratified fluid [16]. Also, we note that the numerical implementation of the DST used in this paper is a modified version of Osborne's automatic algorithm [10], as described in [17].

Finally, we quantify the assumption of 'long, shallow-water' waves made above, so that the limits of the DST's applicability are clear. The latter assumption amounts to requiring that the largest wave amplitude (denoted by $\eta_{\max} \equiv \max_{x,t} |\eta(x,t)|$) is much smaller than the top layer's depth, i.e., $\eta_{\max}/h_1 \ll 1$, and that the characteristic width of the waves is much greater than the top layer's depth, i.e., $h_1/\Delta \ll 1$, where Δ can be taken to be, e.g., the largest half-width of the waves [1, 9]. Also, we may compute the (spatial) Ursell number of a data set, which is defined [14] as

$$\text{Ur} = \frac{3}{16\pi^2}\left(\frac{\eta_{\max}}{h_2}\right)\left(\frac{L}{h_2}\right)^2. \tag{11}$$

This gives an additional measure of the 'nonlinearity' of a wave train, with $\text{Ur} \simeq 1$ being the limit of linear theory.

3.3 Wavelet Transform

While characterizing the scale of internal waves is important, it is equally important to know how that scale changes over time. In this regard the wavelet transform proves particularly useful. Here we discuss the application of the continuous wavelet transform and leave aside other multiscale analysis which can be useful in analyzing IW [18]. The general development of the continuous wavelet transform is well described in the literature [19, 20].

The wavelet transform $W_g(s,x)$ of a spatial sequence $f(x)$ can be defined as follows,

$$W_g(s,x) = \int g_{sx'}(x) f(x') dx', \tag{12}$$

where the wavelets $g_{sx'}(x)$ are generated from the shifted and scaled versions of the *mother wavelet* $g(x)$,

$$g_{sx'}(x) = \frac{1}{\sqrt{s}} g\left(\frac{x - x'}{s}\right), \tag{13}$$

where s and x' are real values that scale and shift the wavelet, respectively. Note that the wavelet transform is in fact a convolution of the wavelet g with the sequence $f(x)$.

In the work described here the continuous wavelet transform is used with the mother wavelet chosen to be the Morlet wavelet. This provides two advantages. First, as noted above the WT is a convolution of the wavelet with the sequence to be analyzed. Thus, the WT can be implemented using the convolution property of the Fourier transform, that is, convolution in space becomes a product of transforms in Fourier space. This property is employed in the algorithm described by Torrence and Compo [21] which is used here. In this formulation the discrete wavelet transform is the inverse Fourier transform of the following product,

$$W_n(s) = \sum_{j=0}^{N-1} \hat{f}_j \hat{g}^*(sk_j) \exp(ik_j n \delta x), \tag{14}$$

where $\hat{f}$ and $\hat{g}$ are the Fourier transforms of the sequence and wavelet, respectively. The second advantage the Morlet wavelet affords is that there is an explicit relationship between the wavelet scale s of a sequence and the standard Fourier components. This allows a direct comparison between the familiar DFT Fourier components and those obtained via the wavelet transform. The power of a wavelet component W_n is given by the amplitude squared $|W_n(s)|^2$.

4 Analysis

The analytic methods just described can be applied to both linear and nonlinear problems. We discuss application to linear problems using tide data and to nonlinear problems using internal waves.

Results from analysis of the data here can be grouped into two broad regimes. Characteristic scales can be discerned over illustrative segments of data which are short compared to the complete data set. Other patterns can only be made out if relatively long sequences are examined. Hence, in the following the analysis is divided into short and long data sequences. First, we investigate short segments of data in Sect. 4.1 and then longer data segments in Sect. 4.2.

4.1 Analysis: Short Data Sequences

In this section we look at the frequency and spatial characteristics of short data segments using the DFT, DST, and WT.

4.1.1 DFT

For analysis, we use several datasets. First, we will look at DFT analysis of tidal velocities. This data is linear and is a good example of the strength of the DFT. We will then analyze a segment of the internal wave shown in Fig. 2 using the DFT, the DST, and the wavelet transform. Finally, we will analyze a longer segment of the IW field using the windowed DFT and the WT and compare the results.

First consider the tidal velocity over time (days) and its Fourier spectrum shown in Fig. 3. The upper panel shows tidal velocity taken from the NCOM tides model [7] sampled at roughly an hour (59 mins) over about 50 days. Qualitatively, we see many high frequency oscillations on the order of a day and a long (14 day) component modulating the entire time period. The lower panel is the power spectrum of the sequence. Note the large amplitudes near 0.04 and 0.08 (h^{-1}), these components correspond to 12.4 and 24 hr tidal components as expected. The long (14 day) modulation component is the so-called fortnight effect known to exist in this tide and can be seen very near the left edge of the plot.

The DFT results for the tidal velocity clearly show the tide's component parts and are a good example of the utility of the DFT. In this case, tidal velocity, the dynamics are very nearly linear and hence it is a good candidate for analysis with the DFT.

We now consider a segment of the internal wave field from Fig. 2. In the upper panel of Fig. 4 the segment shows the oscillating displacement of a single isopycnal (25.1 in sigma units) which is near a depth of –150 m when undisturbed and includes 8 distinguishable troughs, the largest of which at −530 km reaching nearly −400 m. The segment is restricted to include only the internal wave 'packet' spanning a range between −550 km and −350 km. Here and in the coming discussion we will repeatedly examine this internal wave segment by a number of techniques. The lower panel shows the Fourier spectrum for the IW. It can be seen by inspection that the separation of the troughs of the IW in the upper panel are on the order of 25 km. The DFT spectrum shows the tides's Fourier components unevenly spread over a range near 25 km. This moderate spectral resolution giving the power in a range of components rather than clear peaks presents a limitation in the application of the DFT to IWs. The Welch spectrum is overlayed on the periodogram for comparison. While smoother, it nevertheless suffers the same problem with resolution.

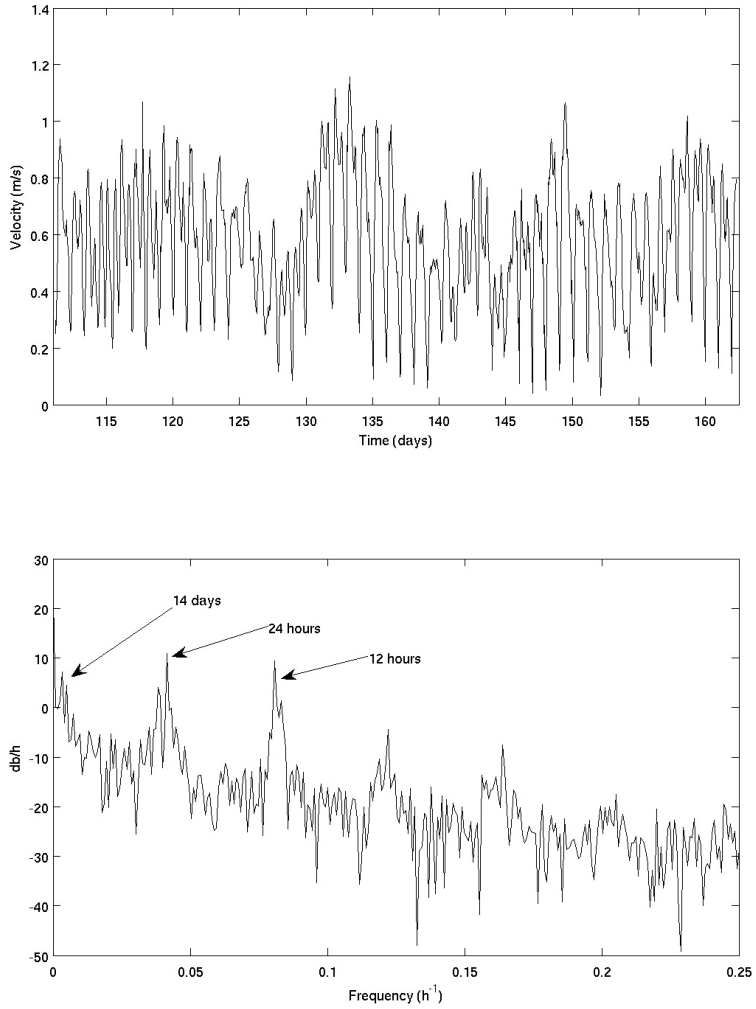


Fig. 3: *Upper panel* shows average tidal velocity in the Luzon Strait. The *lower panel* shows the Fourier components for the tide obtained from the DFT.

4.1.2 DST

Because of the physical basis of the DST, it only makes sense to apply it to nonlinear wave phenomena that are governed (at least to a good approximation) by the KdV equation. Therefore, we only consider the application of the

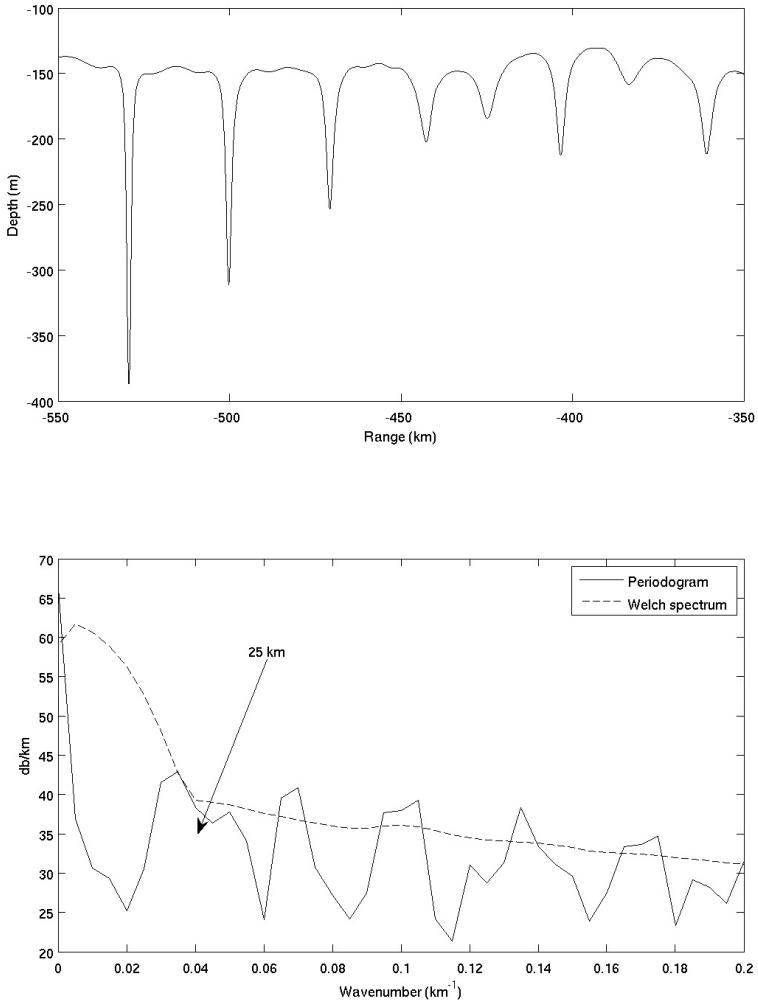


Fig. 4: *Upper panel* shows short segment of IW field taken from Fig. 1. *Lower panel* shows Fourier components calculated from the DFT. The periodogram (*solid curve*) and the Welch spectrum (*dashed curve*) are shown for comparison. Note that 25 km component 'disappears' as a result of smoothing the Welch spectrum.

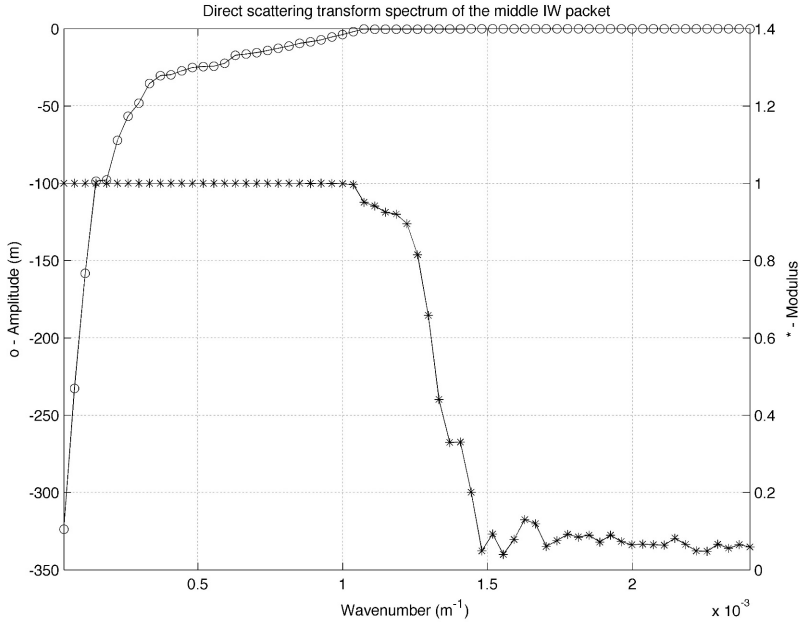


Fig. 5: Scattering transform spectrum of the middle wave packet (range −375 km to −545 km).

DST to the internal wave segment. The spectrum of the middle wave packet (recall Fig. 2) is shown in Fig. 5.

The DST finds 28 solitons in the data set traveling on a 'reference level' of −158.8 m; the Ursell number is 3.209. Physically, the reference level (shown as a black, dashed horizontal line in Fig. 5 and those that follow below) can be understood as the location of the undisturbed isopycnal in the absence of anything but non-interacting (well-separated) solitons. All this means is that the amplitudes of the soliton nonlinear oscillation modes are measured with respect to this reference level.

What is interesting about the DST spectrum is that it not only immediately captures the six solitary waves visible in the data set but also finds a number of 'hidden' modes that cannot be found by observation. Moreover, the DST spectrum reveals that the visible solitary waves fall into two distinct groups — the leftmost three waves and the one near −405 km form one group, while the ones near −445 and −425 km are part of another group. We can make this distinction because of the *trends* in the amplitude versus wavenumber plot of the spectrum given in Fig. 5. In other words, we see that the first four amplitudes' absolute values decrease essentially linearly with the wavenumber, and the slope of the line connecting them is about that of the line which connects the the crests of the leftmost three waves (and the one

near −405 km if it is 'moved' to be next to the latter ones). However, after the fourth mode in the spectrum, the trend of the nonlinear oscillations' amplitudes changes abruptly, which signifies a break in the pattern, and the rest of the modes cannot be grouped with the first four.

It may be surprising that there are 28 solitons in the spectrum of this wave packet, thus, one must keep in mind that these internal waves are highly nonlinear structures, while the KdV equation, which is the basis of the DST, governs the weakly-nonlinear limit. Therefore, we cannot say with certainty that there are *precisely* 28 solitons present in the data. However, we *can*, with a high degree of certainty, conclude that there are 'hidden' solitons and that solitons represent the energetic part of the spectrum (i.e., moderately nonlinear waves and radiation are hardly present, if at all).

4.1.3 WT

Here we will consider the wavelet transform of the IW segment previously discussed shown in the upper panel of Fig. 4 (the WT of tidal data will be considered in a later section). The results are shown in Fig. 6. Note that the x-axis duplicates that shown with the data sequence (between −550 and −350). In this sense the spectrum power is co-located near it's associated IW. The y-axis shows the Fourier wavelength associated with the wavelet scale. The

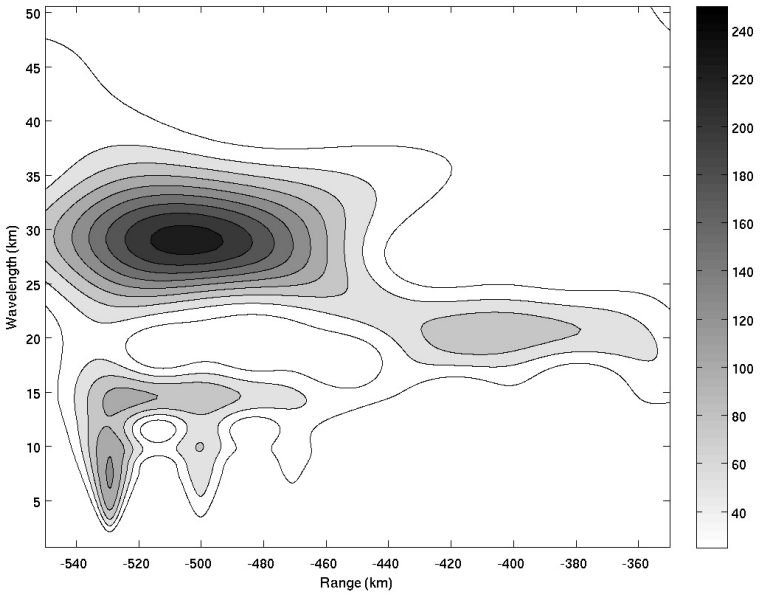


Fig. 6: Wavelet spectrum for internal wave segment. *Darker colors* represent greater wavelet power.

spectrum shows components with significant energy spread over the region between −460 km and −540 km peaking near −510 km at a wavelength of near 30 km. The location of these components indicates that the primary characteristic lengths associated with the first few depressions are about 30 km.

Note small increased areas of wavelet power near the ranges −530 km, −500 km, −470 km. These components are associated with individual troughs of the IW. This can be understood by recalling that the wavelet transform is a convolution of the wavelet with the waveform being analyzed. These small peaks come about when the probing wavelet becomes situated inside the IW troughs. As a consequence, the wavelet scale is on the order of the width of the troughs of IW. This feature is not observed with the DFT.

The great advantage that the continuous wavelet transform enjoys is the ability to isolate the characteristic scales of IW. While the resolution is not to the extent that we have seen in the Fourier analysis of the tidal data (Fig. 3), nevertheless, the WT is able to locate the characteristic scales of IWs. Moreover, the WT localizes these scales in space. This feature holds the possibility that the scale of the IW can be tracked over time. This phenomenon is investigated more closely in the following sections.

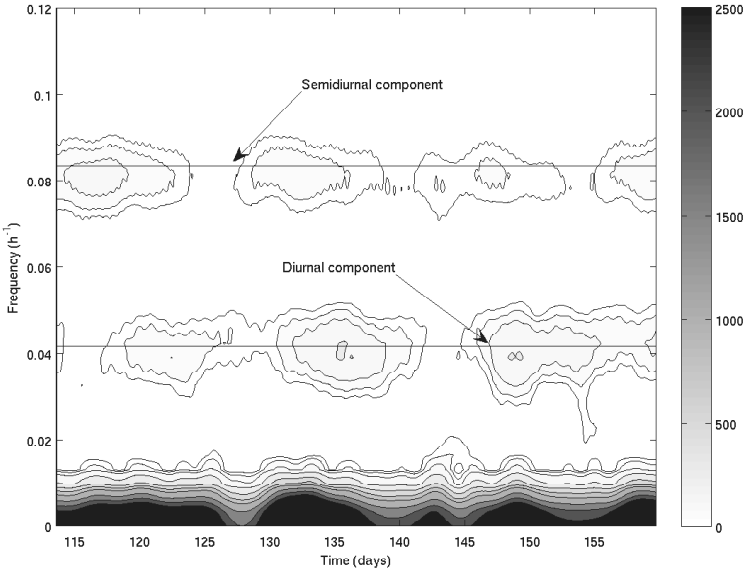
4.2 Analysis: Long Data Sequences

As previously noted, internal waves are nonstationary in the sense that their characteristic spatial and temporal scales evolve over time. At its inception the IW packet is a single depression (or bore) in the isopycnal, which, upon propagation developes into a series of solitary waves through nonlinear dispersion. These solitons grow in amplitude and separate, effectively lengthening the packet. The fully developed IW packet analyzed in the above sections is of this type. Further propagation leads to an IW packet whose constituent solitons has diminished in both number and amplitude. Thus, the three IW packets observed in Fig. 2 can be thought of as snapshots of a single IW packet over its lifetime. This pattern is repeated to varying degrees in most naturally observed IWs. The evolutionary aspect of IW dynamics is a good example of a nonstationary system. For this reason, it is instructive to investigate long data sequences to elucidate this behavior.

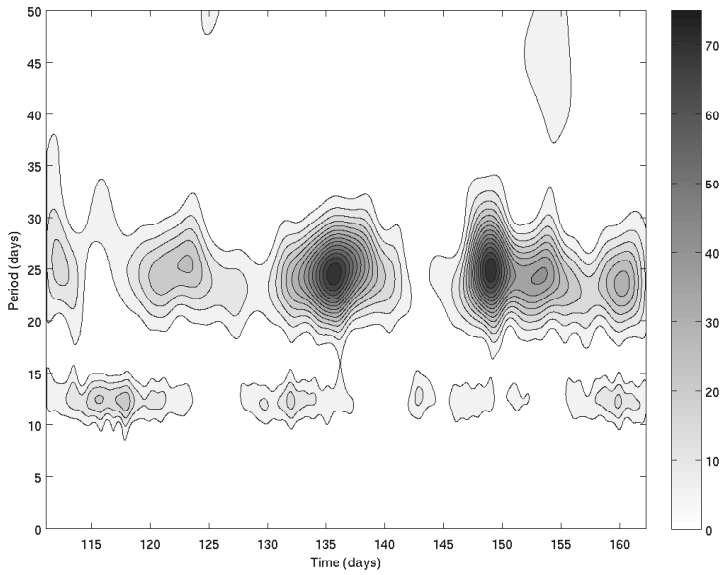
In the following section, the windowed discrete Fourier transform is used to investigate tidal data. Subsequently, the long data sequence of the IW field is analyzed with the WT and and the DST.

4.2.1 Windowed Discrete Fourier Transform

We noted earlier that the DFT does not discern variations in time in the sense that the components discovered via the DFT occurred throughout the tidal time sequence. Some resolution in time can be gained by repeatedly applying the DFT within a short window which is 'slid' along the waveform being analyzed. This is the idea behind the windowed Fourier transform (WFT). Fig. 7a shows the results of applying the WFT to tidal data obtained from



(a) Windowed Fourier transform of tidal data.



(b) Wavelet transform of tidal data.

Fig. 7: Tidal data analysis results: (**a**) windowed Fourier transform (**b**) wavelet transform. *Darker colors* indicate greater relative power.

NCOM and used to drive the Lamb model to generate internal waves. The x-axis is time and ranges over about 50 days and the y-axis is the usual Fourier frequency (h^{-1}). With the spectrogram we can see variation over time of the frequency components of the tide. There are at least two peaks in the spectrum, one at 0.04 and another at 0.08 (h^{-1}) associated with 12.4 and 24 hr tidal period respectively. In analysis of the tidal data the WFT yields good time-frequency information.

4.2.2 Wavelet Transform

The wavelet transform can be considered a refinement of the WFT. Recall that the mother wavelet is scaled and shifted along the waveform to be tested yielding the wavelet spectrum. In this sense the WFT represents a crude wavelet which is a square wave that can be scaled and shifted along the waveform, the resulting spectrum varying with both time and frequency. Noting this similarity it is not surprising that, we expect the wavelet transform to yield results similar to those of the spectrogram.

Figure 7b shows the wavelet transform for the tidal data previously analyzed. Again we can make out the diurnal and semi-diurnal components of the tide. Qualitatively, the results are almost identical to those found using the WFT (excepting that the WFT returns the reciprocal of the period).

Lastly we consider a series of three internal wave packets and analyze the result with wavelets. In Fig. 8 the upper panel shows the series of IWs and the lower panel the associated wavelet spectrum. The results show the generation and evolution of the internal wave packets as they propagate toward the leftmost edge of the domain.

The general features we saw previously (Fig. 6) are repeated for each of the packets (the middle packet being the one previously described). The peaks in the spectrum most closely associated with the leading edge of each of the IW packets, (−670 km, −500 km, −225 km) correspond to the characteristic wavelengths of the individual IW packets. The length increases from about 10 km for the first packet to about 30 km for the middle packet and roughly 35 km as the packet reaches the left boundary. The increase reflects the gradual increase in distance between troughs within each packet.

Referring to the peaks in the wavelet spectrum allows us to draw attention to the wavelet component with the maximum intensity. However the peaks are surrounded by areas of high (relative to the background) intensity reflecting the fact that the spectrum is spread across wavelengths and ranges. In Fig. 8 the concentration of spectral intensity 'spreads' with time so that we see the intensity of the spectrum for the IW packet at −225 km is well concentrated in range and wavelength, at −500 km the intensity measurably broadens and finally the intensity of the packet at −670 km is quite diffuse. The cause of this general dissipation could be from attenuation of the internal wave packet through either physical or numerical mechanisms.

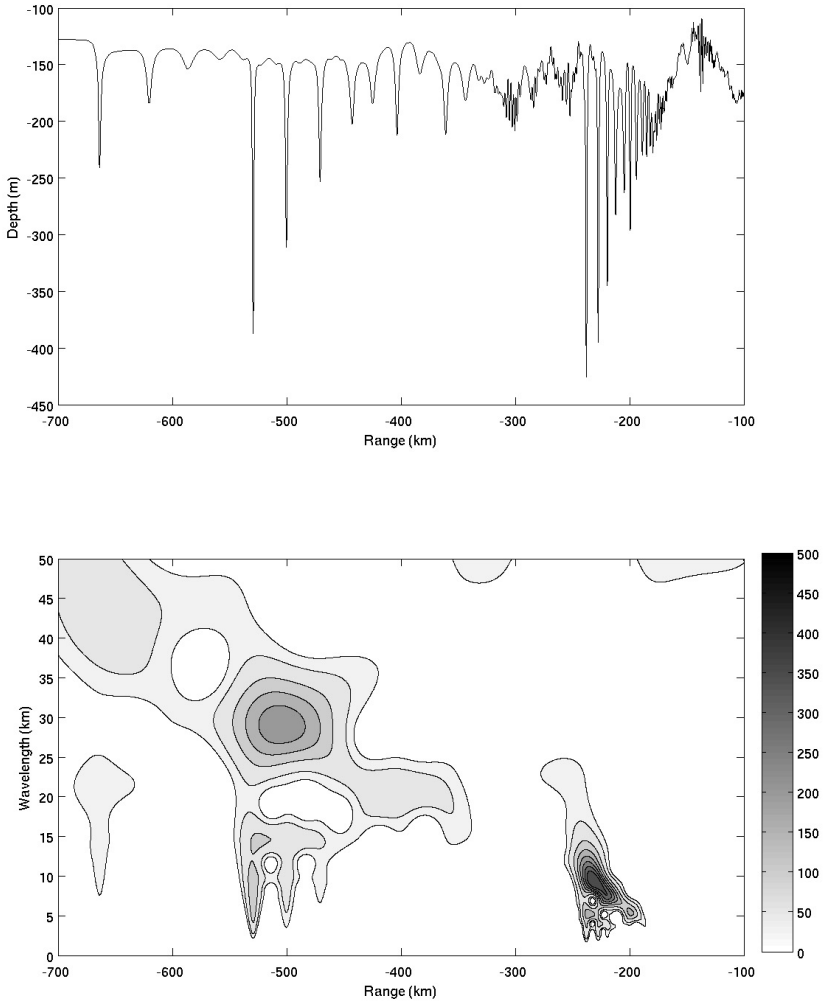


Fig. 8: Wavelet transform spectrum for long time internal wave sequence.

Finally note the peaks associated with individual troughs within each of the three packets where the wavelets 'fit' just inside the individual troughs. The characteristic wavelength for these troughs does not appear to change significantly over the propagation distance. This indicates the general shape of the troughs is somewhat constant throughout the domain.

4.2.3 Direct Scattering Transform

Because the DST identifies the KdV-based nonlinear normal modes of the data set and their evolution, it would only make sense to perform a DST analysis of the entire isopycnal if it were governed by the KdV equation. Clearly, that is not the case as the solitary waves can 'age'. Therefore, in this subsection, we perform a 'windowed' scattering transform analysis of the full data set. That is to say, we take three snapshots of the evolution of the internal solitary waves and compute the DST spectrum of each. This approach is similar to that of Zimmerman and Haarlemmer [16], who computed the DST spectrum of their data at different times in order to identify the nonlinear normal modes that are invariants of the motion (i.e., those that do not change in time).

To this end, in the top panel of Fig. 9, we show the DST analysis of the leftmost (farthest away from the sill) wave packet of the isopycnal under consideration. The middle wave packet, which was the subject of Sect. 3.2, is given in Fig. 5. And, the rightmost (closest to the sill) wave packet's DST spectrum is shown in the bottom panel of Fig. 9. For the leftmost packet, the DST finds 27 solitons traveling on a reference level of −151.3 m; the Ursell number of the data set is 2.008. On the other hand, for the rightmost wave

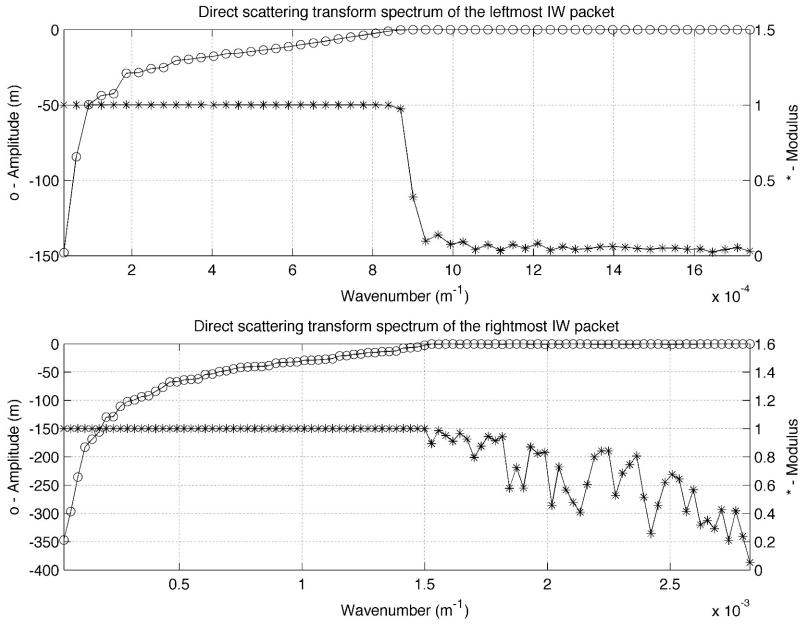


Fig. 9: *Top* and *bottom panels* display the scattering transform spectra of the leftmost (range −150 km to −368 km) and rightmost (range −547 km to −750 km) wave packets of the isopycnal under consideration.

packet, the DST finds 52 solitons traveling on a reference level of −163.9; the Ursell number is 5.718.

The first thing we notice is the similarity in the trend of the nonlinear normal modes' amplitudes (i.e., the rate of increase/decrease of the amplitudes with the wavenumber) in each snapshot. As was the case for the middle internal wave packet we discussed earlier, the largest amplitudes, which decrease quickly (in absolute value) with the wavenumber, are easily seen to be those of the solitary waves visible in the data set. Then, there is a large number of 'hidden' modes whose amplitudes' absolute values decrease approximately linearly with the wavenumber. Furthermore, the spectrum of each wave packet is clearly dominated by solitons, as the amplitudes of the nonlinear normal modes with moduli $m_n < 0.99$ are very small (in absolute value) in comparison with the soliton modes. Again, we emphasize that we cannot be certain whether there are precisely 52 or 28 solitons in the respective internal wave packets. Nonetheless, the DST provides concrete evidence of the nonlinear and *evolving* nature of the packets. Moreover, there is no doubt that solitons are the most prominent part of the spectra, and that their number decreases as the internal wave packets propagate away from the sill.

Furthermore, though for the first wave packet (see top panel of Fig. 9) the non-soliton normal modes are mostly radiation, as their moduli are $m_n \ll 1$, for the middle and rightmost wave packets (see Figs. 5 and bottom panel of 9) we observe more nonlinear normal modes to the right of the 'soliton cutoff' of approximately 1.3×10^{-3} m^{-1}. This correlates with the fact that the Ursell number of this wave packet is the largest of the three — almost twice that of the middle packet and three times that of the leftmost packet. Moreover, this result is consistent with the fact that farther away from the sill the internal waves are, the closer their dynamics are to the KdV (and, eventually, linear) ones.

5 Summary

Data and model studies of internal gravity waves show that their generation and evolution is accompanied by changes in their characteristic spatial and temporal scales. This nonstationary, dispersive behavior arises from nonlinear elements in IW dynamics. In IW studies, dispersion is commonly summarized in amplitude, wavelength, and velocity relationships. Often these are constructed by inspection. In the work described here, objective, analytic tools are employed to investigate the non-stationary behavior of IWs.

In this paper, three methods have been applied to internal wave data generated by Lamb's [2] model designed to simulate IWs observed in the Luzon Strait and South China Sea. They are the following: (1) the discrete Fourier transform, (2) the direct scattering transform and (3) the wavelet transform. The analysis has been applied to linear tide data, to 'short' internal wave data

and to 'long time' segments. Each method yields positive results which in some cases are complementary (DFT, WT) and in some cases unique (DST).

The DST allows for a truly nonlinear analysis of the internal waves and provides a measure of the applicability of the Korteweg-de Vries equation. While the DST does not necessarily give precise quantitative results that can be used for predictive purposes, it provides a 'genuinely nonlinear' decomposition of the data set. In particular, the DST spectra of the snapshots of a wave packet at different stages of its evolution allow us to see the 'nonlinear mode conversions' taking place over time and provides an understanding of solitary wave 'aging' in terms of these modes (notice the decrease in the number of nonlinear oscillation modes for wavenumbers between 0 and $\approx 5 \times 10^{-4}$).

The discrete Fourier transform, the windowed Fourier transform and the wavelet transform res present a continuum of Fourier based approaches to investigate IWs. Each yields the Fourier components of the internal waves and further, the WT (thought of in terms of a refined WFT) provides a view of how these modes change over time. In this regard, we have seen that the WT elucidates the evolving character of internal waves thus providing a time/frequency picture of the evolving dynamics of the internal wave over long periods.

In summary, each technique can be seen to provide positive recognizable details of IW dynamics. It is not uncommon to find that what is obvious to the naked eye cannot be verified by reasonable examination. Thus, while the results described in this paper fall short of a complete quantitative description of IW dynamics, nevertheless, that these methods support and expand on what can be seen 'by eye' is a nontrivial result. Certainly, the entire catalogue of analyses applicable to the investigation of internal waves has not been addressed here. However those described here represent a span of means by which to investigate internal wave dynamics. Moreover, it is reasonable to expect that further work will yield more quantitative results.

Acknowledgements. This research was supported by the Office of Naval Research under PE 62435N, with technical management provided by the Naval Research Laboratory. I.C. acknowledges a fellowship from the ONR/ASSE Naval Research Enterprise Intern Program.

References

1. J. R. Apel: 2003. A new analytical model for internal solitons in the ocean. J. Phys. Oceanogr. **33**, 2247–2269.
2. K. G. Lamb: 1994. Numerical experiments of internal wave generation by strong tidal flow across a finite amplitude bank edge. J. Geophys. Res. **99**(C1), 843–864.
3. A. C. Warn-Varnas, S. A. Chin-Bing, D. B. King, Z. Hallock, and J. A. Hawkins: 2003. Ocean-acoustic solitary wave studies and predictions. Surveys in Geophysics. **24** 39–79.

4. A. Grinsted, J. Moore, and S. Jevrejeva: 2004. Application of the cross wavelet transform and wavelet coherence to geophysical time series, Nonlinear Processes in Geophysics **11**, 561–566.
5. T. F. Duda, J. F. Lynch, J. D. Irish, R. C. Beardsley, S. R. Ramp, and C.-S. Chiu: 2004. Internal tide and nonlinear wave behavior at the Continental Slope in the North China Sea. IEEE J. Ocean Eng. **29**, 1105–1130.
6. S. R. Ramp, 2006. Private communication.
7. S.-Y. Chao, D.-S. Ko, R.-C. Lien, and P.-T. Shaw: 2007. Assessing the West Ridge of Luzon Strait as an internal wave mediator. J. Oceanogr. **63** (No.6), 897–911.
8. Signal Processing Toolbox User's Guide for use with MATLAB. The MathWorks, Inc. (2002).
9. A.R. Osborne, T.L. Burch: 1980. Internal solitons in the Andaman Sea. Science **208**, 451–460.
10. A.R. Osborne: 1994. Automatic algorithm for the numerical inverse scattering transform of the Korteweg–de Vries equation. Math. Comput. Simul. **37**, 431–450.
11. A.R. Osborne, M. Serio, L. Bergamasco, and L. Cavaleri: 1998. Solitons, cnoidal waves and nonlinear interactions in shallow-water ocean surface waves. Physica D **123**, 64–81.
12. A.R. Osborne and E. Segre: 1991. Numerical solutions of the Korteweg–de Vries equation using the periodic scattering transform μ-representation. Physica D **44**, 575–604.
13. A.R. Osborne, E. Segre, G. Boffetta, and L. Calaveri: 1991. Soliton basis states in shallow-water ocean surface waves. Phys. Rev. Lett. **67**, 592–595.
14. A.R. Osborne and M. Petti: 1994. Laboratory-generated, shallow-water surface waves: analysis using the periodic, inverse scattering transform. Phys. Fluids **6**, 1727–1744.
15. A.R. Osborne and L. Bergamasco: The solitons of Zabusky and Kruskal revisited: perspective in terms of the periodic spectral transform. Physica D **18**, 26–46.
16. W.B. Zimmerman and G.W. Haarlemmer: 1999. Internal gravity waves: analysis using the the periodic, inverse scattering transform. Nonlin. Process. Geophys. **6**, 11–26.

AU: Please update ref. [17].

17. I. Christov: 2007. Internal solitary waves in the ocean: analysis using the periodic, inverse scattering transform. Math. Comput. Simul., arXiv:0708.3421, submitted (2007)
18. S. Jevrejeva, J. C. Moore, and A. Grinsted: 2003. Influence of the artic oscillation and El Niño-Southern Oscillation (ENSO) on ice condition in the Baltic Sea: The wavelet approach. J. Geophys. Res. **108**, D21, 4677–4688.
19. P. Kumar and E. Foufoula-Geogiou: 1994. Wavelet analysis in geophysics: an introduction. Wavelets in Geophysics, E. Foufoula-Georgiou and P. Kumar, eds. Academic Press, San Diego. pp. 1–45.
20. L.H. Kantha and C.A. Clayson: 2000. Appendix B: Wavelet Transforms. Numerical Models of Oceans and Oceanic Processes. L.H. Kantha and C.A. Clayson, eds. Academic Press, San Diego. pp. 786–818.
21. C. Torrence, G.P. Compo: 1998. A practical guide to wavelet analysis. Bull. Am. Met. Soc. **79**, 61–78.

Crustal Deformation Models and Time-Frequency Analysis of GPS Data from Deception Island Volcano (South Shetland Islands, Antarctica)

María Eva Ramírez, Manuel Berrocoso, María José González, and A. Fernández

Laboratorio de Astronomía, Geodesia y Cartografía. Departamento de Matemáticas. Facultad de Ciencias. Universidad de Cádiz.
mariaeva.ramirez@uca.es

Abstract. We have applied wavelet techniques to analyze GPS time-series data from REGID geodetic network, deployed at Deception Island Volcano (South Shetland Islands, Antarctica). In the present analysis wavelets are used to detect periodic components and to filter the data. The high frequency components can be associated to the orbital period of the satellites and to local tidal effects, whereas the medium frequencies seem to be related to the weather cycle. The wavelet filtering procedure is based on the SURE estimator, and a considerable reduction in noise is achieved, particularly in the Up component, whose deviation is reduced down to the deviation of the horizontal components before the denoising. An estimation of the displacements in the network for the period 2001/02 – 2005/06 is also included.

1 Introduction and Motivation

The Global Positioning System (GPS) is widely used to study many geoscience problems such as the determination of the motion of the Earth's tectonic plates or volcanic monitoring.
Most of the chapters describing GPS data analysis for evaluation of crustal deformation estimate a single station position over a 24 h period. This processing strategy is suitable when deformation rates are small and vary slowly along the years but a sub-daily position is required for some other applications such as volcanic monitoring. In fact, volcanic activity is usually associated to significant ground deformation, what makes GPS to be considered an ideal technique for both monitoring and studying of active volcanoes.
The usual approach in GPS data processing is limited in the sense that the coordinates of the stations are estimated only once per day, while a much

higher solution rate is obviously needed if a rapidly deformation event must be detected. Moreover, the 24 h sampling rate ignores any variation within a day as well as any periodic component which affects the data with a period shorter than one day.

In this chapter we present a new methodology to analyze GPS time series when a high sampling rate is considered in the data processing. With this sampling, three objectives appear:

- To detect the periodic components in the data that are ignored with processing sessions of 24 h.
- To filter the data in order to decrease the scattering and to remove the detected periodicities.
- To evaluate the displacement in the area of study from the surveyed stations, in order to better understand the pattern of its behaviour.

To tackle these questions we have considered the time-frequency decomposition of the data that the wavelet transformation provides [1].

The analyzed GPS data correspond to the surveying of the geodetic network deployed on Deception Island Volcano (South Shetland Islands, Antarctica) during the last Antarctic campaigns (2003-04 – 2005-06) for monitoring its volcanic activity.

Geodetic studies in Deception Island (Fig. 1) began in the 1950s by the Chilean, Argentinean and British scientists from the bases on the island. These works were focused on updating the existing cartography of the island and they were interrupted at the end of the 60s, when the volcanic eruptions that took place forced the evacuation of the bases. The geodetic and geophysical tasks were interrupted until 1986, when monitoring of the island was reestablished by Argentinean and Spanish researchers. In January 1992, a noticeable increase in seismic activity was detected, with 900 registered events and 4 felt earthquakes. Gravity and magnetic anomalies suggested that the volcano reactivation was due to a 2 km depth magmatic injection at Fumaroles Bay [2]. These evidences started to subside in February 1992. Regarding previous geodetic work, Berrocoso [3] conducted repeated GPS surveying from 1989-90 to 1995-96, obtaining an absolute deformation rate of 4 cm/year and 3.24 cm/year at BARG and FUMA stations respectively, and a value of 2.91 cm/year and 0.89 cm/year at BALL and PEND stations. A subduction process was observed around the island, with values of 1.94 cm/year for BARG and FUMA and 0.94 cm/year and 1.74 cm/year for PEND and BALL respectively (Fig. 2).

At the end of 1998 only a few events were recorded, but this behaviour changed suddenly in the beginning of 1999, with the occurrence of significant seismo-volcanic activity. This crisis included volcano-tectonic (VT) and long period (LP) events together with volcanic tremors. Most of the registered events were localized between Fumaroles and Telephone Bay, some of which were large enough (3–4) to be felt. When the campaign finished, the seismic activity was still high [4].

Although GPS data are available from 1989, in this paper we focused just on those ones collected during the last Antarctic campaigns, due to the improvement achieved in the storage capabilities of the receivers. In particular, we will concentrate on the time-frequency analysis of the GPS time series more than on the interpretation of the detected deformation. Further information about the displacement rate and the deformation on the island and in the South Shetland Islands environment can be found in [5, 6, 7, 8]. In these works a detailed description of the geodetic works on Deception Island can be found and the horizontal deformation models up to the 2002/2003 campaign are presented. The models were obtained by considering 24 h sessions in the processing of the GPS observations but no further analysis of the data was developed. The processing strategy applied in this work, which considers 30 min sessions, allows not only the estimation of the deformation models but also the in depth study of the time series, what constitutes the innovative aspect of this study.
The chapter is organized as follows: the first section includes a brief description of the tectonical setting of the South Shetland Archipelago. The following section deals with the GPS surveying campaigns and the processing strategy. Next section includes the methodology for analyzing the data, where wavelets are applied for two different purposes: to detect the periodic components and to filter the data. The next section includes the application of the explained methodology to the GPS time series. The detected periodicities are discussed and the denoising of the data is described in detail. The last part of the section addresses the estimation of the deformation occurring on the studied area. The deformation pattern estimated from the surveyed stations is shown, and a comparison of the results obtained from the filtered and non-filtered data is exposed. A final discussion and a brief overview of future work follow in the last section of the chapter.

2 Tectonical Setting

Deception Island, located north-west of the Antarctic Peninsula, is situated in the Bransfield Strait marginal basin that separates the South Shetland Islands from the Antarctic Peninsula. It is a horseshoe shaped stratovolcano, whose main volcano-tectonic feature is a central flooded depression related to a spreading centre in the Bransfield Strait. This central caldera has been traditionally described as a collapse caldera, originated after one or more voluminous eruptions [9]. However, other models for Deception Volcano suggested that it was formed progressively by passive normal faulting along nearly orthogonal normal faults that cut across the island according to a regional trend [10]. The Bransfield Strait is a consequence of the rifting and separation of the South Shetland Archipelago and the Antarctic Peninsula, that generated a chain of volcanoes along the Bransfield Sea, three of which emerged: Deception, Pinguin and Bridgeman [2, 11]. The complex structure

of this environment is characterized by the interaction of four small tectonic plates (the Scotia Arc, the Atlantic and Pacific Plates, and the Phoenix Plate) which cause a subduction process from the South Shetland Islands Plate to the Phoenix Plate and an expansion and thermal ascent between the South Shetland Island Plate and the Atlantic Plate along the Bransfield Sea (Fig. 1). Deception Island is also located at the confluence of two major tectonic structures: the SW end of the Bransfield and a southerly extrapolation of the Hero Fracture Zone. In addition to these interesting tectonic features, Deception Island has exhibited a continuous volcanic activity with confirmed eruptions in 1800, 1812, 1842, 1871, 1912, 1956, 1967, 1969, and 1970. The last eruptive process took place from 1967 to 1970 around Telephone Bay and Mont Pond (Fig. 2), along the main fracture in the NNE-SSW direction. It gave rise to a 40 m high cone and an alignment of five craters in the North of the island, causing the collapse of the Chilean Base in Pendulum Cove and the destruction of the British Base in Whaler's Bay due to a lahar action.
Nowadays, the main superficial evidences of the volcanic activity on the island are the presence of fumarolic areas with 100°C and 70°C gaseous emissions at Fumaroles and Whaler's Bay respectively, 100°C hot soils at Hot Hill,

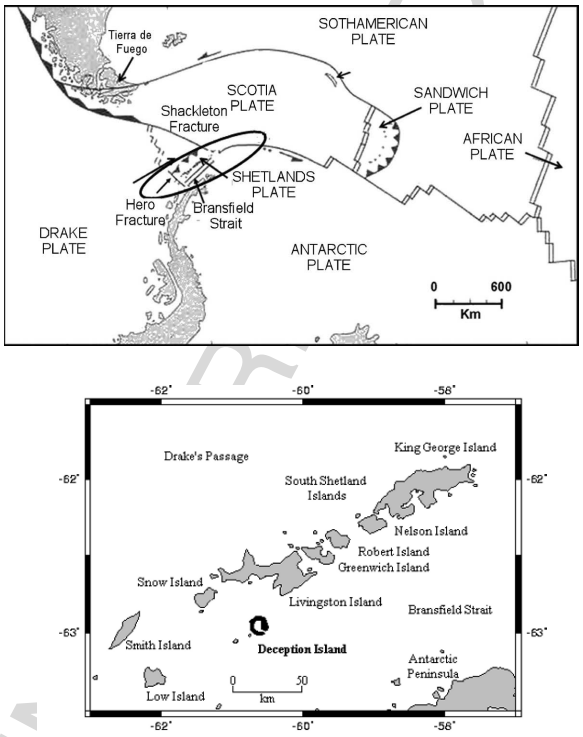


Fig. 1: Tectonic and geographical setting of Deception Island Volcano.

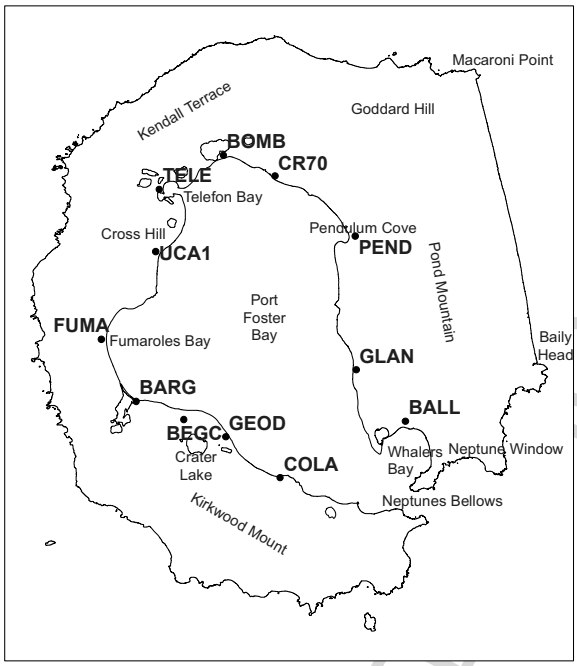


Fig. 2: Distribution of the stations of the geodetic network in Deception Island Volcano. Stations marked with a square are monitored through the whole campaigns, whereas the others are surveyed during some days every year. The locations of the last eruptive process are also indicated in the map.

and 45°C and 65°C thermal springs in Pendulum Cove and Whaler's Bay [12, 13]. A remarkable seismic activity is also registered in some areas of the island, whith a mean of 1000 events detected during the campaigns. The high seismicity reflects a rift expansion, a subduction process and volcanism.
Due to the complex geodynamic characteristics of Deception Island and of the South Shetland environment previously described, the volcanic activity on the island is related to regional tectonics.

3 Data Description and Processing

The analyzed GPS data come from the GPS surveys of the geodetic network in Deception Island Volcano, REGID. At present, the GPS network consists of 12 stations distributed around the inner bay of the island as it is shown in Fig. 2.
The network has been episodically surveyed every Austral summer since 1989. Since the quality of the GPS data improved in the last years, in this study we

have just focused on the data from the last campaigns (2003-04, 2004-05, and 2005-06), with better quality and a larger number of observations. In fact, for the time-frequency analysis we have just considered data from FUMA and PEND stations, at Fumaroles Bay and Pendulum Cove. These stations, together with BEGC station at the Spanish Antarctic Base Gabriel de Castilla, provide a global approach to the deformation occurring on the island and they constitute the fundamental stations of the network being surveyed along the whole campaigns. Table 1 shows the duration of the 2003-04, 2004-05, and 2005-06 campaigns and the number of days each station of the REGID geodetic network was surveyed.
GPS data were processed with BERNESEv4.2 Scientific Software [14], according to the following scheme:

1. Firstly, the absolute coordinates for the reference station BEGC of the REGID geodetic network were obtained from its processing together with BEJC and the IGS station OHI2, at the Chilean Base O'Higgins (150 km away);
2. Coordinates for the rest of the stations of the REGID geodetic network were calculated through radial baselines, setting the fixed station at BEGC as the reference station.

Regarding the configuration parameters of the processing, tropospheric parameters and ambiguities resolution were calculated using the GPSEST subroutine of BERNESE v4.2 software. Tropospheric parameters were estimated hourly from the Saastamoinen tropospheric model as suggested by [15] for GPS data processing in Antarctic regions; ambiguities were solved by applying the quasi ionosphere free (QIF) strategy. Ambiguities for both L1 and

Table 1: Duration of the 2003-04, 2004-05 and 2005-06 campaigns and number of days each station is surveyed.

2003-04 campaign: 12/01/2003–01/01/2004							
Stations (number of surveying days)							
BEGC	(31)	BARG	(6)	FUMA	(31)	PEND	(30)
BALL	(3)	COLA	(6)	GEOD	(9)	UCA1	(3)
TELE	(3)	BOMB	(5)	CR70	(3)	GLAN	(1)
2004-05 campaign: 12/02/2004–02/02/2005							
Stations (number of surveying days)							
BEGC	(71)	BARG	(8)	FUMA	(37)	PEND	(40)
BALL	(0)	COLA	(0)	GEOD	(9)	UCA1	(0)
TELE	(3)	BOMB	(5)	CR70	(3)	GLAN	(0)
2005-06 campaign: 12/17/2005–02/26/2006							
Stations (number of surveying days)							
BEGC	(72)	BARG	(7)	FUMA	(71)	PEND	(71)
BALL	(5)	COLA	(5)	GEOD	(4)	UCA1	(6)
TELE	(6)	BOMB	(5)	CR70	(2)	GLAN	(5)

L2 frequencies were solved after fixing the reference station and estimating ionospheric parameters for every epoch. An iono free solution for every session was obtained. Sessions were 24 h length for obtaining the coordinates of the reference station, and 30 min length for the estimation of the rest of the REGID stations coordinates. The solutions of every session were adjusted by means of the ADDNEQ routine for obtaining the coordinates of the reference station, whereas the solutions for the rest of the stations of the network were not adjusted in order to get a larger number of data in the time series to be analyzed. Precise orbits and pole files were used for the entire procedure.
Once data are processed, the set of the resulting geocentric coordinates (X, Y, Z) are transformed into a local topocentric system (E, N, U), with center in each surveyed station. Therefore, three components are obtained for each station: the East and North horizontal components, and the vertical (or Up) component, being the latter less reliable due to the well-known loss of accuracy of the GPS system in the elevation component (Fig. 3).

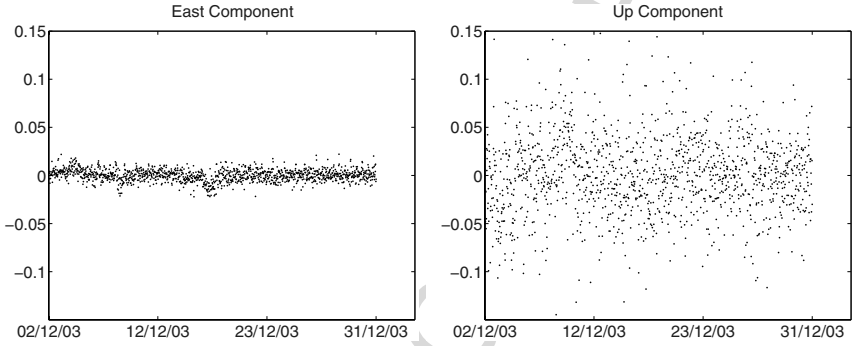


Fig. 3: East and Up components for FUMA station at Fumaroles Bay corresponding to 2003/04 campaign, obtained after the processing of the data with a 30 min sampling rate.

4 Methodology

Geophysical time series are usually generated by processes so complex that their behaviour is difficult to model in the time domain. For many years, Fourier analysis has been the essential tool to study certain characteristics of data in the frequency domain. Nevertheless, these frequency components are not well *time located*. Moreover, they are infinitely disperse in time. The Windowed Fourier Transform allows a restricted partition of the time-frequency plane, considering boxes or windows of fixed shape, but this non-varying shape of the boxes restricts the time location of the frequency content in the data.
The wavelet transform presents an alternative to classical frequency analysis, since the decomposition bases considered for its computation provide a

partition of the time-frequency plane with boxes of varying shape, depending on the frequency resolution we are interested in. This time-frequency transformation allows the decomposition of the data along successive scales, which are related to different frequency resolutions. Therefore, the small scales involve the analysis of the high frequency components of the data, and they correspond to boxes of small time support and wide frequency support in the partition of the time-frequency plane, whereas the greater scales correspond to the lower frequencies, related to boxes with a wide time support and a small support in the frequency domain (Fig. 4). During the last years, several chapters can be found where wavelet techniques are used in the analysis of GPS data. In [16], wavelet decomposition is used to better estimate the secular trend, by rejecting those wavelet coefficients related to the high frequencies, and keeping a smoothed version of the original data. Other studies [17] focus on the filtering of the data to reduce the multipath effect, and also [18] on the detection of sudden changes or occurrence of events in the data.

In this work the wavelet transform was applied to the GPS time series in order to (1) detect the periodic components and (2) to denoise the data to reduce their scattering. In particular, the periodic components of the data will be associated to the scales that concentrate the maximum level of energy in the time-frequency decomposition. Concerning the filtering of the data, the usual wavelet denoising techniques are based on the application of a threshold onto the wavelet coefficients. Those ones that are above the threshold are kept whereas the others are set to zero, and an estimation of the denoised signal is obtained from the filtered coefficients. Nevertheless, this strategy needs to be slightly modified when the data are too noisy. In fact, a high contribution of noise to the signal makes the total energy of the data to be too spread all over the wavelet coefficients which do not exceed the threshold and yields a too smoothed denoised signal. We have considered the *Stein's Unbiased Risk Estimator* for the filtering [19], with the value of the threshold depending not only on the wavelet coefficients but also on a comparison of the signal energy with an estimation of the energy of the noise.

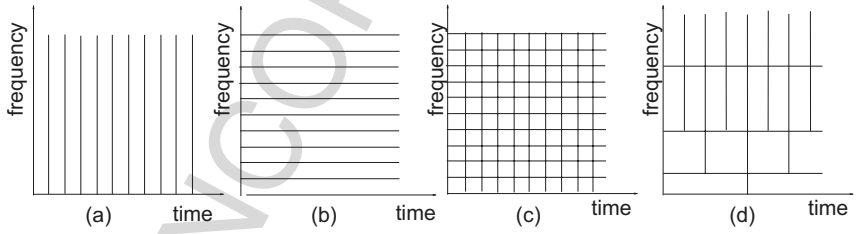


Fig. 4: Decomposition of the time-frequency plane with the Euclidean (a) and Fourier (b) basis; with the windowed Fourier basis (c), and by means of wavelet basis (d).

4.1 Wavelet Background

The wavelet theory relies on the existence of two functions ψ (*mother wavelet* or *wavelet function*) and φ (*father wavelet* or *scale function*) verifying certain properties [1], which provide a decomposition of the data in the time-frequency plane along successive scales. This time-frequency transformation depends on two parameters, the *scale* parameter s, which is related to the frequency, and the *time* parameter u, related to the translation of both functions ψ and φ in the time domain.

The *continuous wavelet transform* of a function $f \in L^2(\mathbb{R})$ is defined as the inner product of f with the translated and dilated version of ψ,

$$Wf(u,s) = \langle f,\, \psi_{s,u} \rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt, \tag{1}$$

where ψ^* denotes the conjugate of ψ. The expresion (1) can be seen as the convolution $f \star \bar{\psi}_s(u)$ of the function f with $\bar{\psi}_s$, where

$$\bar{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi^* \left(\frac{-t}{s} \right). \tag{2}$$

Varying u in time, this convolution provides the frequency components $\{d_{s,u}\}_u$ of the signal associated with the scale s and the time location u, or what is designated as the *details coefficients* of f *at the scale* s *and time* u. A real wavelet transform preserves energy as long as the wavelet function ψ satisfies a weak admissibility condition given by Calderón in [20].

On the other hand, the convolution of f with $\bar{\varphi}_s$, where $\bar{\varphi}_s$ is defined from φ in an analogous way to $\bar{\psi}_s$,

$$f \star \bar{\varphi}_s = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \varphi^* \left(\frac{t-u}{s} \right) dt = \langle f,\, \varphi_{s,u} \rangle, \tag{3}$$

provides a smoothed version of f, given by the *approximation coefficients of* f *at scale* s, $\{a_{s,u}\}_u$.

The time-frequency resolution of the wavelet decomposition depends on the time-frequency spread of the considered wavelet function ψ. In fact, since the time and frequency support are inversely related, as the scale parameter s decreases, the shorter periods (and high frequencies) are captured by the wavelet transform. Conversely, the greater the scale parameter s is, the longer the periods and lower frequencies detected. Nevertheless, this resolution or location in the time-frequency plane is lower bounded by the Heisenberg's Uncertainty Principle [1], which states that the perfect location in both domains simultaneously is not possible.

Particularly important are the *dyadic* versions of both functions ψ and φ, where the scale s is a power of 2, that is, $s = 2^j$, $j \in \mathbb{Z}$, and their integer translations $u = k \in \mathbb{Z}$. This importance lies in the fact that the set $\{\psi_{j,k}\}_{j,\in\mathbb{Z}}$, where $\psi_{j,k}$ is given by

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t-2^j k}{2^j}\right), \tag{4}$$

constitutes an orthonormal base of the space $L^2(\mathbb{R})$ of functions of finite energy. Thus, any function $f \in L^2(\mathbb{R})$ can be expressed in terms of $\{\psi_{j,k}\}_{(j,k)\in\mathbb{Z}^2}$ in a non redundant way,

$$f(t) = \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle \psi_{j,k}(t). \tag{5}$$

The term $\langle f, \psi_{j,k}\rangle$ is defined as the *wavelet coefficient* $d_{j,k}$, so the above expresion (5) can be rewritten as

$$f(t) = \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t). \tag{6}$$

Under these considerations, each j-approximation can be seen as the sum of the $(j+1)$-approximation of a coarser level and the $(j+1)$-details which appear in the scale j but disappear in the approximation at scale $j+1$.

The decomposition determined by the dyadic wavelet can be understood as a discretization of the continuous wavelet transform given by (1). Mallat's algorithm [1] calculates the wavelet coefficients with a cascade of discrete convolutions and subsamplings, what is specially useful to save computation efforts and for certain wavelet applications such as data compression or denoising. Hence, the choice of which of these transformations (continuous or discrete) is better to use depends on the purpose of the data analysis.

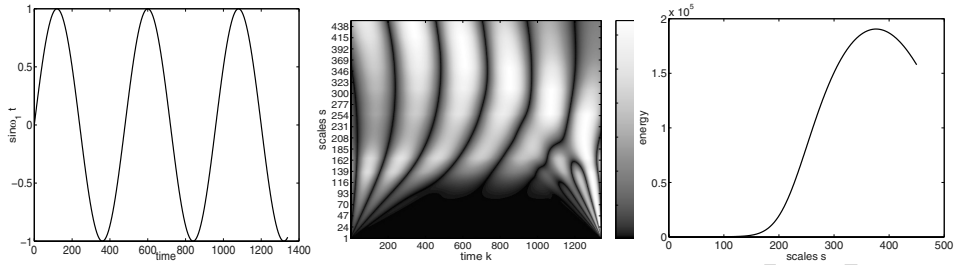
The representation of the wavelet coefficients as a function of scale yields the wavelet spectrum, providing a decomposition of the data in the time-frequency plane. Since every horizontal line in the wavelet spectrum is associated to a frequency component in the data, this decomposition constitutes an useful tool to detect dominant periodic components. Representing the scales of the decomposition versus the corresponding energy of the coefficients, that is, the *scalegram* of the data, the periodic components are identified with the scales whose associated energy reach the maxima in this representation.

In order to validate the detection of periodicities in the data from the maxima of the wavelet scalegram, we have applied the same procedure onto two synthetic sinusoidal signals

$$\begin{aligned} y_1 &= \cos\omega_1 t \\ y_2 &= \cos\omega_1 t + \sin\omega_2 t \end{aligned} \tag{7}$$

where $\omega_1 = 2\pi/864000$ and $\omega_2 = 2\pi/86400$. The time span was set to $[0 : 1800 : 1800 \cdot 1340]$ in order to cover a total of 27 days, and the sampling period was set to $\Delta = 1800$ s to get the same number of data and the same sampling rate as some of the experimental samples.

With these considerations, the wavelet spectrum and the scalegram were calculated. For the first synthetic signal (Fig. 5a) the scalegram (Fig. 5c) reveals



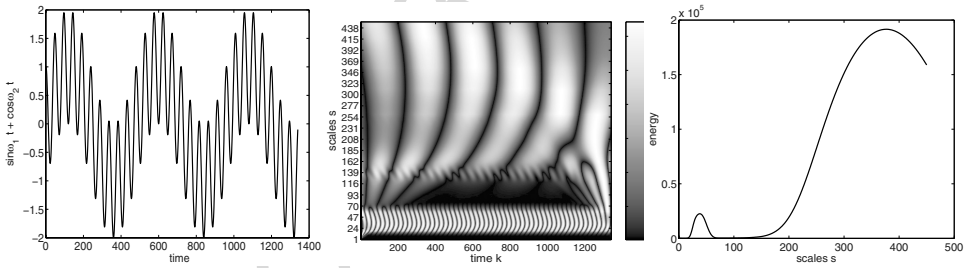
(a) (b) (c)

Fig. 5: Synthetic signal $y_1 = \sin \omega_1 t$, where $\omega_1 = 2 * \pi / T_1$, with $T_1 = 864000$ (a); wavelet spectrum (b) and scalegram (c), whose maximum value is reached for the scale $s = 377$ that captures the oscillation of the sinusoidal signal.

that the maximum value is reached at the scale $s = 377$, which corresponds to a frequency of $f_s = 9.824 \cdot 10^{-7}$, that is, a period of $T \sim 11.7$ days, according to the scale-frequency relation given by

$$f_s = \frac{F_c}{s \cdot \Delta}, \tag{8}$$

where f_s denotes the characteristic frequency for scale s, F_c is the central frequency of the wavelet, related to its maximum oscillation, and Δ is the sampling period. For the second synthetic signal $y_2 = \cos \omega_1 t + \sin \omega_2 t$ (Fig. 6a) it is observed how the wavelet spectrum captures both oscillations (Fig. 6b), which correspond to the local maxima in the scalegram (Fig. 6c). The local maxima are reached at scales $s_1 = 38$ and $s_2 = 377$, which are related to the frequencies $f_{s_1} = 9.746 \cdot 10^{-6}$ and $f_{s_2} = 9.824 \cdot 10^{-7}$, and therefore to periods $T_1 \sim 1.1$ and $T_2 \sim 11.7$ days, respectively.



(a) (b) (c)

Fig. 6: Synthetic signal $y2 = \sin \omega_1 t + \cos \omega_2 t$, where $\omega_1 = 2 * \pi / T_1$ and $\omega_2 = 2 * pi / T_2$, with $T_1 = 864000$ and $T_2 = 86400$, respectivey (a); wavelet spectrum (b) and scalegram (c), whose local maxima are reached for the scales $s = 38$ and $s = 377$ that capture the oscillations of the sinusoidal signal.

4.2 Wavelet Denoising

In practice, experimental data are usually affected by different sources of noise that mask the signal of interest. The experimental data X_n can be written as

$$X_n = f_n + W_n, \tag{9}$$

where f_n is the *pure* magnitude and W_n is a Gaussian noise with σ^2 variance. Most of the wavelet denoising techniques are based on the comparison of the amplitude of wavelet coefficients with a threshold T previously defined. Those coefficients whose amplitudes are above the value T are kept or smoothed, while those ones whose amplitudes are below T are set to zero. Thus, the filtered signal $\tilde{f}$ is estimated by the reconstruction or synthesis from the filtered wavelet coefficients. Since each wavelet coefficient is associated to a time-frequency location, this denoising is adapted to the local regularity of the data. This fact constitutes the main advantage of the wavelet procedure over the usual band pass filtering techniques, where the suppression of certain frequency band affects all the time domain of the signal.
The first question that arises in wavelet-based denoising is the choice of a value for the threshold T. Donoho and Johnstone [21] proved that for the value $T = T_u$

$$T_u = \sigma\sqrt{2\ln N}, \tag{10}$$

where N is the length of the time series and σ is the standard deviation of the noise, the probability of the wavelet coefficients to be above T_u is high. This value of T_u is known as the *universal threshold*. Since the value of σ is unknown, an estimation $\tilde{\sigma}$ for σ is used. A good value for $\tilde{\sigma}$ is given by [1]

$$\tilde{\sigma} = \frac{M_X}{0.6745}, \tag{11}$$

where M_X denotes the median of the wavelet coefficients at the finest scale, that is, $\{d_{1,k}\}_k$ since they are the ones related to the highest frequency components in the data, and therefore, to the noise.
To evaluate the performance of a denoising strategy it is considered a *loss function* which evaluates the norm of the difference between the pure signal f and the estimated $\tilde{f}$ from a threshold T, that is

$$r(f,T) = E\{||f - \tilde{f}||^2\}. \tag{12}$$

The thresholding risk can be reduced by choosing a value of T lower than the one given in (10) and which depends on the data to a greater extent.

SURE Estimator

For thresholding-based denoising the wavelet coefficients $\{d_i\}_{i=1,\ldots,N}$ above the threshold T are smoothed and the ones below T are set to zero. Thus, the risk associated to a signal estimation with a threshold T is given by [1]

$$r(f,T)=\sum_{i=1}^{N}\Phi(d_i),\ \text{with}\ \Phi(x)=\begin{cases}x-\sigma^2, & x\le T\\ \sigma^2+T^2, & x>T\end{cases}. \tag{13}$$

Thus, the value of T must be chosen in order to minimize the expression given in (13). The applied criterium is based on the SURE (*Stein's Unbiased Risk*) estimator, which is briefly described as follows:

1. The wavelet coefficients are arranged according to their modulus in decreasing order:
$$|d_1|\ge|d_2|\ge\ldots\ge|d_N|, \tag{14}$$
where N is the number of wavelet coefficients and $d_i=d_{j,k}$ for certain j and k, and $i=1,2,\ldots,N$.
2. Given a value T of the threshold, let us $\alpha_T\in\mathbb{N}:1\le\alpha_T\le N$ such as
$$|d_1|\ge\ldots|d_{\alpha_T}|\ge T\ge|d_{\alpha_T+1}|\ge\ldots\ge|d_N|. \tag{15}$$
Taken into account the above expression (15), the risk associated to the value T of the threshold is given by
$$\tilde{r}(f,T)=\underbrace{\sum_{k=\alpha_T+1}^{N}|d_k|^2-(N-\alpha_T)\sigma^2}_{(1)}+\underbrace{\alpha_T(\sigma^2+T^2)}_{(2)} \tag{16}$$
where the first summing term (1) is the contribution to the risk of the coefficients whose amplitudes are below T and the second term (2) is related to the coefficients with amplitudes above T.
3. Thus, the expression (16) must be recalculated for each of the N wavelet coefficients and the value of the threshold T will be chosen to be
$$\tilde{T}=|d_\alpha|, \tag{17}$$
for certain $\alpha:1<\alpha<N$ in such a way that (16) is minimum.
4. This algorithm is not always suitable for the filtering of the data. In particular, if data are too noisy, that is, if the energy of the pure signal f is *small* compared to the energy of the noise W, the energy of the data will be too spread among the wavelet coefficients and it can occur that few coefficients are above T, so the reconstructed signal is almost zero since most of the coefficients are set to zero by the filtering procedure.
When this situation occurs, the value for the threshold is taken to be the universal threshold (10). Therefore, the energy of the signal f must be previously compared to a minimum energy level given by [22]
$$\epsilon_N=\sigma^2N^{1/2}(\ln N)^{3/2}. \tag{18}$$

Due to the energy of f is unknown, an estimation must be used. Since

$$E\{||X||^2\} = ||f||^2 + N\sigma^2, \tag{19}$$

an estimation of $||f||^2$ is given by $||f||^2 = ||X||^2 + N\sigma^2$. With this value of ϵ_N, the threshold will be given by

$$T = \begin{cases} \sigma\sqrt{2\ln N}, & \text{if } ||X||^2 - N\sigma^2 \leq \epsilon_N \\ \tilde{T}, & \text{if } ||X||^2 - N\sigma^2 > \epsilon_N \end{cases}, \tag{20}$$

where $\tilde{T}$ is described as in (17).

Threshold Adapted to Scale

Since the number of coefficients decreases as long as the scale increases, it is useful to adapt the value of the threshold to the scale. With this scale-dependence of the threshold, we avoid having a too large value of T for the wavelet coefficients corresponding to a large scale, where the number of coefficients is lower according to Mallat's algorithm subsampling, what would imply that most of the coefficients would be set to zero.
The suggested modification calculates a value T_j for each scale $s = 2^j$, by applying the SURE algorithm previously described onto the wavelet coefficient at level j, that is, onto $\{d_{j,k}\}_k$, and minimizing the expression given in (13).

Choice of the Best Wavelet Basis

Since wavelet denoising is based on the application of a threshold onto the wavelet coefficients, and no good estimation is obtained if the energy of the data is too spread, the best wavelet basis will be the one which concentrates the energy of the data the best, that is, in the minimum number of wavelet coefficients.
Marshall and Olkin proved in [23] that the best basis $\mathcal{B} = \{\psi_n\}_n$ will be the one whose coefficients minimize a *cost function*,

$$\mathcal{C}(f, \mathcal{B}) = \sum_{m=1}^{N} \Phi\left(\frac{\langle X, \psi_m\rangle}{||X||^2}\right), \tag{21}$$

where $\Phi(x)$ is a concave function such as the entropy $\Phi(x) = -x\log x$ or the one derived from the norm $||\cdot||_1$.

5 Application to the GPS Data

5.1 Detection of Periodicities in the Data

The existence of certain periodicities affecting the GPS data is well known, in particular those related to periods of 1 year, 6 months and approximately

400 days. These components appear in GPS data from global and regional networks, with no geographic dependence and in the three components of the surveyed stations. They are associated to the Earth rotational period and the Pole motion [24, 25, 26, 27].

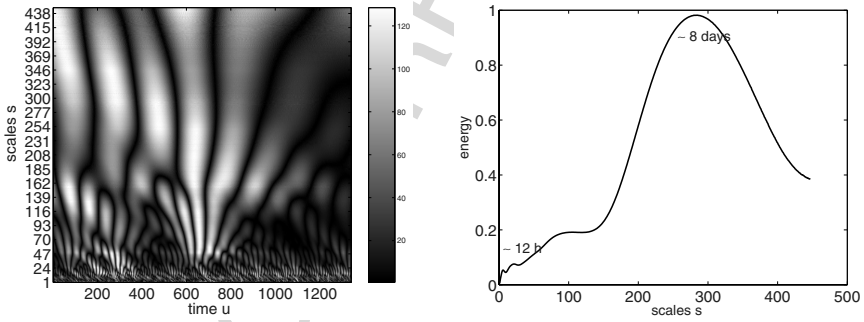
Since our data just cover a period of 3 months at the most, the usual long period components are not detectable. On the contrary, the 30 min solution rate allows the detection of shorter periodicities in the data.

In order to determine the periodic components affecting our GPS data with this sampling rate, the existing gaps and missing data were firstly interpolated to get an uniform sample. The best fitting among an interpolation of 1st, 2nd or 3rd order was chosen. The number of data considered in the interpolation of the gaps depends on the length of the gap according to the following criterium: for gaps shorter than 6 h, the number of points for the interpolation corresponds to the 12 h before and after the gap; for gaps between 6 h and 1 day, the points of the sample cover one day before and after the gap; finally, for gaps greater than 1 day, data corresponding to 3 days before and after the gap are considered in the interpolation.

Once the data were uniformly sampled, the wavelet transform was calculated using MatLab6p5 and the WaveLab8.0 packet by Stanford University, with modifications on some functions according to our needs.

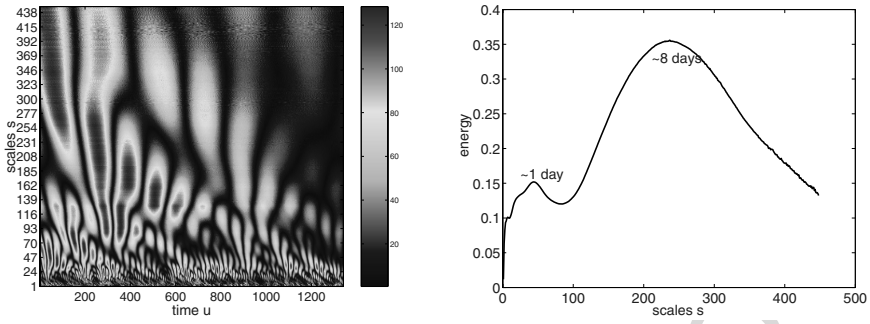
The dominant frequency components are identified with the maxima of the energy of the wavelet transform along the decomposition scales, that is, with the local maxima in the scalegram.

By way of example, Figs. 7 and 8 show the wavelet spectrum and the wavelet scalegram of the East and North component of FUMA station at Fumaroles Bay for the 2003-04 Antarctic campaign, respectively. The detected periodicities for FUMA and PEND stations at Fumaroles Bay and Pendulum



(a) (b)

Fig. 7: Wavelet spectra (a) and scalegram (b) for the East component of FUMA station at Fumaroles Bay and 2003-04 Antarctic campaign. Each time unit u in (a) represents a 30 min interval.



(a) (b)

Fig. 8: Wavelet spectra (a) and scalegram (b) for the North component of FUMA station at Fumaroles Bay and 2003-04 Antarctic campaign. Each time unit u in (a) represents a 30 minutes interval.

Cove respectively for the 2003/04, 2005/05 and 2005/06 campaigns are listed in Table 2.

The causes of some of these periodicities are still unknown. The ones corresponding to the highest frequencies are associated to the satellite orbital period, while medium components are probably due to ionospheric activity, which is particularly important in polar regions. In fact, [28] reveals the existence of some periods of several days which are originated by ionospheric effects. Other possible sources causing the medium frequency components are

Table 2: Detected periodicities for the GPS stations at Fumaroles Bay (FUMA) and Pendulum Cove (PEND) for the 2003/04, 2004/05, and 2005/06 Antarctic campaigns from the analysis of the scalegram of the data.

	FUMA			PEND		
	East	North	Up	East	North	Up
2003/04	12 h 8 d	1 d 8 d	12 h 1 d 4–5 d 12 d	12 h 8–12 d	12 h 6–9 d	12 h 8 d
2004/05	12 h	12 h	12 h 3–5 d 14 d	12 h 1 d 9 d	 1–2 d 6 d	12 h 4–5 d
2005/06	12 h 3 d 6–7 d 27–28 d	12 h 9 d 24–28 d	12 h 25–28 d	12 h 5 d 	12 h 14 d 28 d	1 d 24–28 d

the ones related to the weather cycle, such as the presence of strong winds or the action of some tide effects, typical of marginal seas and inner bays, and different from the usual daily and semi daily solar and lunar tides. The study of some other kind of measurements (e.g. tide gauges measurements, temperature records) is required for the better determining of the sources causing the periodic fluctuation in the data, specially those ones related with medium and long periods.

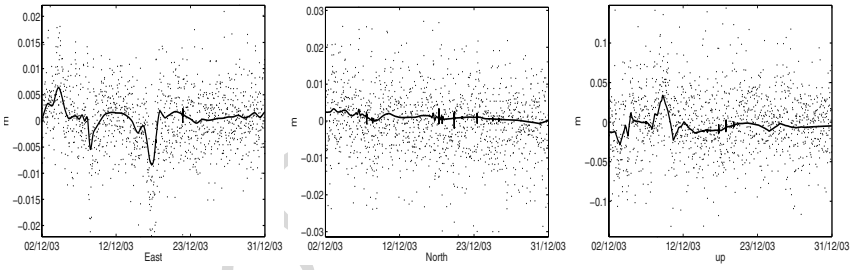
5.2 Data Denoising

The filtering approach described in sect. 4.2 was applied onto the GPS data for FUMA and PEND stations (Fig. 9).
Since the results can vary slightly depending on the considered wavelet function, the denoised signal is estimated as follows:

1. The wavelet decomposition is calculated with several wavelet bases from a set of wavelet functions (*wavelet dictionary*, [1]).
2. The four bases for which the cost criterium given in (21) is lowest are selected. These are the bases that best concentrate the energy of the data.
3. The resulting estimated signal is taken to be the average of the denoised signal from the four wavelet basis determined in step 1.

Table 3 resumes the best bases for each station and component, and the average $\bar{\sigma}^{\mathcal{B}_j}_{comp}$ of the differences between the filtered values obtained with each wavelet basis for each 30 min interval, and the mean value that has been taken as the final solution, that is,

$$\bar{\sigma}^{\mathcal{B}_j}_{comp} = \sum_{i=1}^{N} \frac{x^{\mathcal{B}_j}_{comp}(i) - \bar{x}_{comp}(i)}{N} \tag{22}$$



(a) (b) (c)

Fig. 9: East (a), North (b) and Up (c) component of FUMA station for the GPS campaign 2003/04. The solid line denotes the filtered data obtained by applying the SURE denoising procedure explained in the text.

Table 3: Mean deviation corresponding to the wavelet bases for which the cost function reaches the lowest values. $\mathcal{B}$ denotes the wavelet bases considered in the filtering, and σ_j, $j = E, N, U$ (East, North and Up component) is the mean deviation of the raw data with respect to the filtered data for the corresponding wavelet bases.

2003–04	**FUMA**	**East**	$\mathcal{B}$	bior.4.4	bior5.5	sym5	sym9
			$\bar{\sigma}_E$	–0.05	0.11	0.02	–0.08
		North	$\mathcal{B}$	sym5	sym9	sym7	db3
			$\bar{\sigma}_N$	0.05	–1.2	–1.7	2.8
		Up	$\mathcal{B}$	coif1	coif2	db3	sym3
			$\bar{\sigma}_U$	1.9	3.5	–2.7	–2.7
	PEND	**East**	$\mathcal{B}$	bior5.5	bior4.4	db4	db7
			$\bar{\sigma}_E$	0.4	0.4	–0.03	–0.8
		North	$\mathcal{B}$	coif3	coif1	sym10	sym8
			$\bar{\sigma}_N$	–0.6	1.1	–0.2	–0.2
		Up	$\mathcal{B}$	sym5	sym4	coif4	sym6
			$\bar{\sigma}_U$	–9.8	–1.4	8.8	2.4
2004–05	**FUMA**	**East**	$\mathcal{B}$	coif5	coif1	db7	coif3
			$\bar{\sigma}_E$	–0.4	0.4	–0.06	0.09
		North	$\mathcal{B}$	coif1	sym5	coif5	coif4
			$\bar{\sigma}_N$	–0.5	2.2	–0.8	–0.8
		Up	$\mathcal{B}$	coif4	sym10	sym8	sym5
			$\bar{\sigma}_U$	–2.9	–2.8	0.2	5.6
	PEND	**East**	$\mathcal{B}$	coif1	sym20	db3	sym3
			$\bar{\sigma}_E$	0.8	–0.8	0.01	0.01
		North	$\mathcal{B}$	sym5	sym9	coif4	sym10
			$\bar{\sigma}_N$	–0.8	–0.2	0.4	0.5
		Up	$\mathcal{B}$	sym4	coif2	coif4	sym6
			$\bar{\sigma}_U$	13.3	–5.4	–8.8	–0.9
2005–06	**FUMA**	**East**	$\mathcal{B}$	bior5.5	bior4.4	coif5	rbio1.5
			$\bar{\sigma}_E$	–1	0.2	0.03	–0.05
		North	$\mathcal{B}$	coif1	sym4	coif5	coif2
			$\bar{\sigma}_N$	–0.5	–0.0	0.2	0.3
		Up	$\mathcal{B}$	coif3	coif1	coif2	sym4
			$\bar{\sigma}_U$	–3.9	–2.0	–4.7	6.6
	PEND	**East**	$\mathcal{B}$	bior5.5	rbio6.8	sym4	db6
			$\bar{\sigma}_E$	–0.3	–0.5	1.1	–0.3
		North	$\mathcal{B}$	db3	sym3	coif2	coif1
			$\bar{\sigma}_N$	–0.1	–0.1	0.08	0.2
		Up	$\mathcal{B}$	sym4	coif2	coif1	db3
			$\bar{\sigma}_U$	–6.7	5.8	–1.8	2.7

where N is the sample size, $x_{comp}^{\mathcal{B}_j}(i)$ is the estimated value obtained with the wavelet basis $\mathcal{B}_j$ for $j = 1, \ldots, 4$ and the time interval (or *epoch*) i, and $\bar{x}_{comp}(i)$ is the i-epoch mean value given by

$$\bar{x}_{comp}(i) = \sum_{j=1}^{4} \frac{x_{comp}^{\mathcal{B}_j}(i)}{4} \tag{23}$$

Bases $\mathcal{B}$ given in Table 3 correspond to the following wavelet families: Biorthogonal (*bior*), Symmlet (*sym*), Daubechies (*daub*), Coiflet (*coif*), and Reverse Biorthogonal (*rbio*). The number next to the wavelet name indicates the number of *vanishing moments*, a wavelet property related to the support of the wavelet and to its regularity, in the sense that for a wavelet with p vanishing moments the wavelet coefficient for a p-th order polynomial will be zero [1, 29].

The remarkable point of this method is the great reduction on the deviation of the three components (East, North and Up) of the stations. Table 4 includes the most representative statistical parameters of the raw data and the filtered time series. In particular, it can be observed that the greatest reduction is obtained in the vertical component, what constitutes an important achievement since as it was mentioned before the Up component is the less reliable one since it is more influenced by the effects affecting the GPS signal. In fact, with this denoising procedure the deviation in the Up component is reduced down to the level of the horizontal components before the denoising procedure.

Table 4: Some representative statistical parameters (in mm) of the time series before and after the denoising. $\bar{x}_j$, r_j and σ_j denote the mean, the range and the standard deviation of the j component (j = East, North, Up).

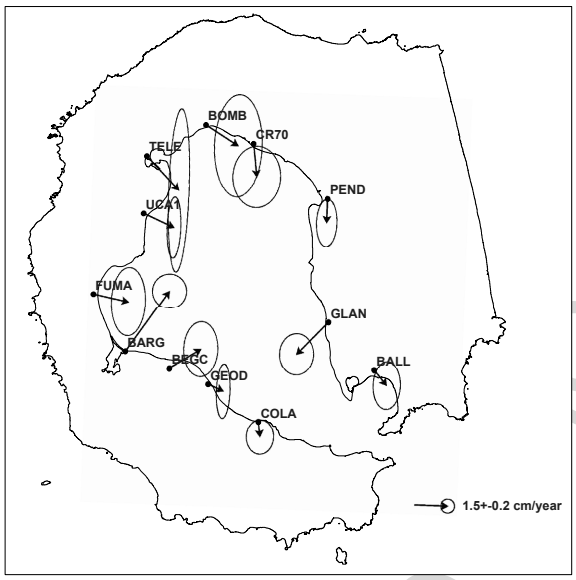
		FUMA				PEND			
Campaign	Component		$\bar{x}_j$	r_j	σ_j	Component		$\bar{x}_j$	r_j	σ_j
2003–04	East	O	0. 2	44. 2	6. 4	East	O	–0. 1	42. 2	5. 7
		F	0. 5	14. 8	2. 2		F	1. 1	4. 3	0. 8
	North	O	–0. 3	62. 3	8. 9	North	O	–0. 1	54. 9	7. 9
		F	1. 0	5. 3	0. 9		F	0. 9	10. 0	0. 7
	Up	O	0. 8	292. 3	40. 5	Up	O	0. 6	248. 7	35. 5
		F	–4. 9	62. 3	8. 6		F	–9. 5	39. 0	5. 2
2004–05	East	O	–1. 3	29. 3	3. 9	East	O	–1. 6	55. 4	6. 5
		F	–0. 6	5. 5	0. 7		F	–0. 6	6. 0	0. 6
	North	O	–0. 3	45. 9	6. 1	North	O	2. 4	59. 2	7. 7
		F	0. 8	10. 6	1. 1		F	2. 2	3. 7	0. 7
	Up	O	–4. 2	201. 9	25. 3	Up	O	–0. 9	240. 2	35. 8
		F	–3. 3	11. 0	1. 9		F	8. 4	49. 4	5. 5
2005–06	East	O	0. 1	28. 1	3. 8	East	O	0. 2	44. 0	5. 7
		F	–0. 1	4. 5	0. 7		F	0. 8	10. 6	1. 4
	North	O	0. 0	40. 4	5. 4	North	O	0. 2	46. 2	6. 1
		F	–0. 2	9. 0	1. 0		F	0. 8	8. 3	0. 6
	Up	O	–0. 1	169. 4	24. 0	Up	O	–0. 4	208. 9	29. 7
		F	5. 1	20. 4	2. 9		F	–9. 1	20. 0	3. 9

5.3 Crustal Deformation Models

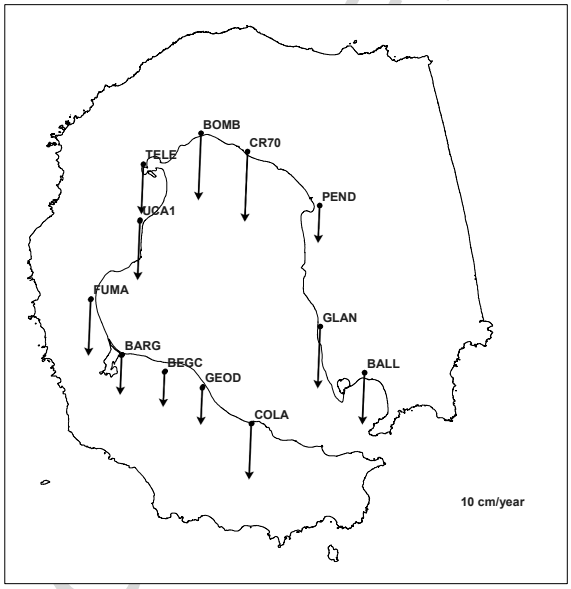
The local deformation of Deception Island was estimated from the filtered data according to the processing procedure exposed in sect. 3.
The variation of the station's coordinates through time provides the global deformation of the island, which includes regional and local tectonic effects. To isolate the local deformation, the displacements relative to BEJC geodetic station at Livingston Island were estimated. In addition, the determination of the displacements of the REGID network on Deception Island with respect to the reference station BEGC provides an inner perspective of the relative deformation on the island. Figures 10–13 include the estimated models, and Table 5 resumes the obtained values for the displacement rates.
The deformation rates estimated from the SURE filtered data were compared to the values obtained by means of a *common-mode* filtering. The common-mode strategy is widely used with this kind of data [30] and it consists in calculating moving averages with a 1-sidereal day length window. We have adapted this procedure, calculating a mean value for every 30 min session within a day, that is, the filtered values are given by

Table 5: Tectonic and volcanic displacement rates for FUMA and PEND stations for the periods 2003/04–2004/05 and 2004/05–2005/06, obtained by filtering the data using common-mode and SURE denoising.

		v, σ (mm/yr)	East	σ_E	North	σ_N	Up	σ_U	$\lvert v_{hor}\rvert$	σ_{hor}
2003/04 – 2004/05		**Absolute tectonic displacements**								
	FUMA	Com-mode	13.0	7	9.5	12	–71.9	33	16.10	13
		SURE Denoising	13.2	4	11.4	3	–75.3	10	17.4	5
	PEND	Com-mode	16.9	12	11.0	18	–7.52	40	20.2	21
		SURE Denoising	16.9	4	10.7	3	–70.0	24	20.0	4
		Volcanic displacements								
	FUMA	Com-mode	6.3	30	–10.8	18	–72.5	29	12.0	34
		SURE Denoising	6.7	3	–8.5	3	–75.9	15	10.0	4
	PEND	Com-mode	10.4	39	–9.3	24	–74.8	36	13.0	45
		SURE Denoising	10.5	4	–9.3	4	–70.8	24	14.0	5
2004/05 – 2005/06		**Absolute tectonic displacements**								
	FUMA	Com-mode	8.2	7	17.4	12	63.7	27	19.2	13
		SURE Denoising	8.2	4	15.5	3	64.5	10	1.5	5
	PEND	Com-mode	23.0	11	28.2	17	71.2	37	36.3	20
		SURE Denoising	23.2	4	28.5	4	68.7	24	36.3	5
		Volcanic displacements								
	FUMA	Com-mode	3.5	25	15.7	15	58.9	24	16.0	3
		SURE Denoising	2	3	–5	3	64.2	15	5.4	4
	PEND	Com-mode	18.4	35	13.6	22	66.2	33	22.8	40
		SURE Denoising	16.1	6	7.8	4	68.2	6	17.8	7

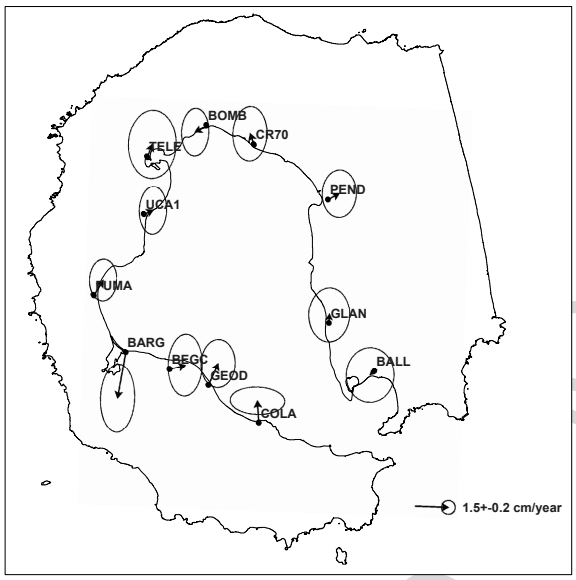


(a)

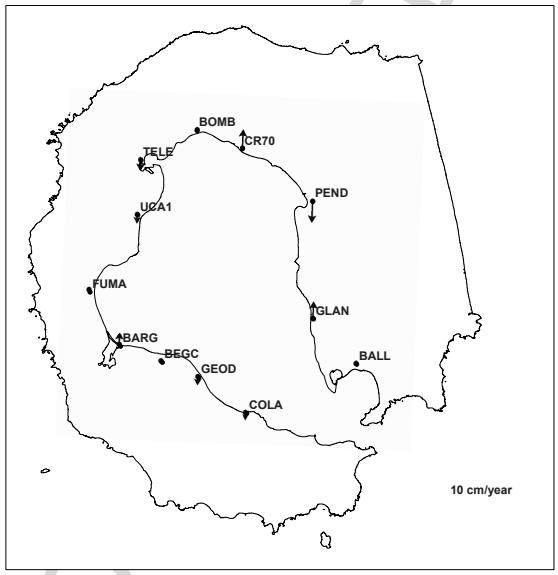


(b)

Fig. 10: Horizontal (top) and vertical (bottom) displacement models related to the local tectonic and volcanic activity of Deception Island for the period 2001/02 – 2002/03.

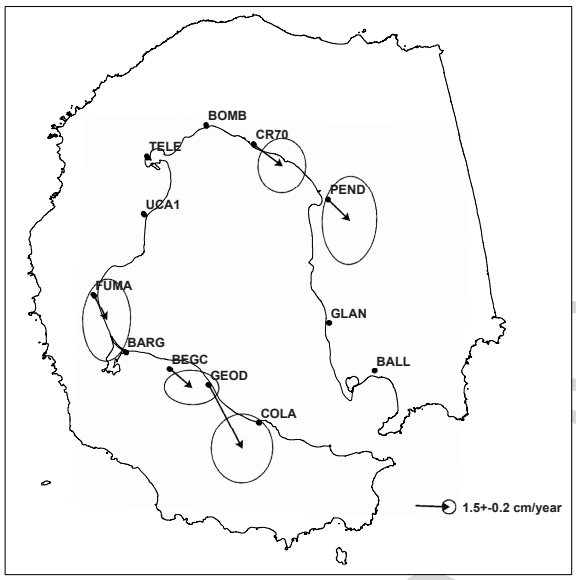


(a)

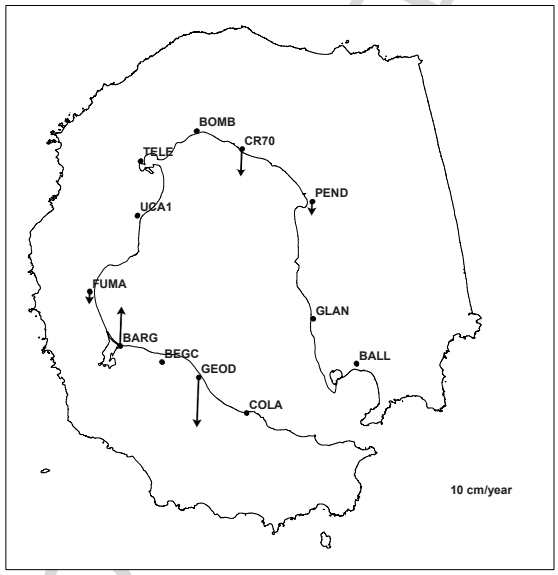


(b)

Fig. 11: Horizontal (top) and vertical (bottom) displacement models related to the local tectonic and volcanic activity of Deception Island for the period 2002/03 – 2003/04.

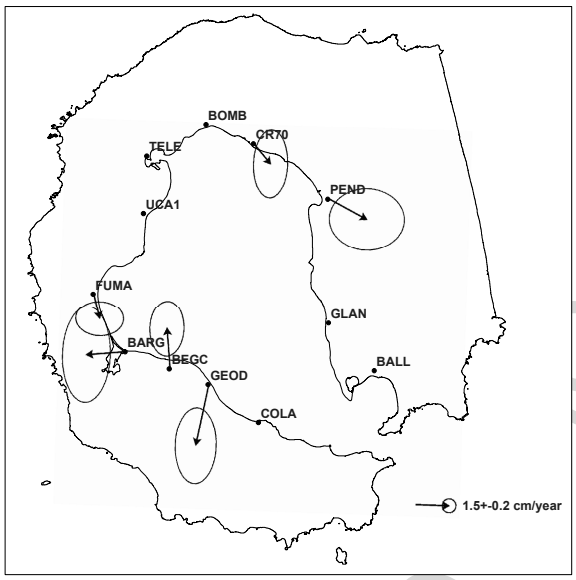


(a)

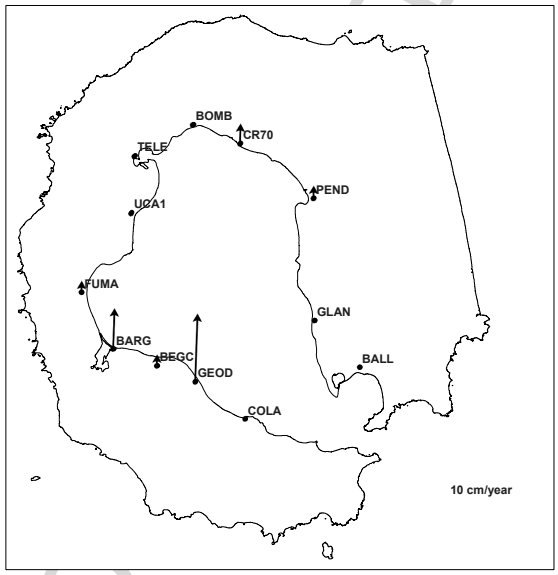


(b)

Fig. 12: Horizontal (top) and vertical (bottom) displacement models related to the local tectonic and volcanic activity of Deception Island for the period 2003/04 – 2004/05.



(a)



(b)

Fig. 13: Horizontal (top) and vertical (bottom) displacement models related to the local tectonic and volcanic activity of Deception Island for the period 2004/05 – 2005/06.

$$\tilde{x}_{ss,d} = x_{ss,d} - \bar{x}_{ss} \tag{24}$$

where $x_{ss,d}$ is the estimated solution for the session $ss = 1, \ldots, 48$ and the observed day d, and

$$\bar{x}_{ss} = \sum_{d=1}^{D} \frac{x_{ss,d}}{D}, \tag{25}$$

with D being the number of surveyed days.
Although the displacement rates obtained by both methods do no differ considerably, with differences being in the order of a few milimeters, the remarkable point of the proposed wavelet methodology is the large reduction in the data deviation, specially in the vertical component, whose standard deviation goes down to the values of the horizontal component before the filtering of the data, as it is shown in Table 5. The calculated displacements agree with the pattern detected for the previous period [7, 31]. In fact, during the period 2001/02 – 2002/03 it was observed a remission of the extensive radial process detected after the volcanic crisis in 1998, which continues in the following period. The deformation pattern for this period is also included in Fig. 10 in order to better interpret the results for the following periods, although GPS data from these campaigns were not considered for the wavelet analysis. No significant displacement were detected afterwards, although it is observed a change in the trend of the surveying stations for the 2003/04 – 2004/05 period, aligned according the Hero Fractures Zone.

6 Conclusions and Outlook

The strategy presented in this paper for the analysis of GPS time series combines relevant information obtained from a multiscale wavelet-based decomposition of the data, with a double objective: (a) the detection of periodic components in GPS data when they are processed with a 30 min sampling rate, and (b) the filtering of the data.
It can be stated that the detected higher frequency components are related to the orbital period of the satellites, while the medium ones seem to be related to more local sources such as weather-related effects; the origin of the longer period components is not still well determined and further research involving the study of other measurements is planned to be done in future Antarctic campaigns, such as tide or temperature records, among others.
On the other hand the denoising strategy has provided very good results, reducing the scatter of the data in the three components. The decrease of the standard deviation of the data yields a reduction in the errors associated to the estimated deformation rates. Particularly remarkable are the results obtained for the Up component: in fact, the error corresponding to the vertical component is one order of magnitude worse than the error corresponding

to the horizontal components before denoising, and it dropps down to the magnitude of the horizontal ones after the wavelet filtering.

Acknowledgements. The realization of this work was possible thanks to financial support to the following projects: "Surveillance and fast monitoring of the volcanic activity of Deception Island" (GEODESIA) (ANT1999-1430-e/HESP), "Geodetic Studies on Deception Island: deformation models, geoid determination and SIMAC", REN 2000-0551-C03-01, "Acquisition of a Scientific Software for GPS data processing" (REN2000-2690-E), "Geodetic monitoring of the volcanic activity of Deception Island" (CGL2004-21547-E/ANT), "Update of the Spanish Cartography for Deception Island" (CGL2004-20408-E/ANT), funded by the Spanish Ministry of Science and Technology through the National Program of Antarctic Research of Natural Resources. We would also like to thank the BIO Las Palmas and BIO Hesperides crew and the members of the Spanish Antarctic Base Gabriel de Castilla for their collaboration during the surveying campaigns.

References

1. S. Mallat: *A wavelet tour of signal processing.* (Academic Press, London 1999) pp 1–637
2. J. Ibáñez, J. Almendros, G. Alguacil, J. Morales, E. Del Pezzo, R. Ortiz: Eventos sísmicos de largo período en Isla Decepción: evidencias de volcanismo activo. *Bol. R. Soc. Esp. Hist. Nat. (Sec. Geol.)* **93**, 1–4:105–112 (1997)
3. M. Berrocoso: *Modelos y formalismos para el tratamiento de observaciones GPS. Aplicación al establecimiento de redes geodésicas y geodinámicas en la Antártida.* (Boletín ROA, Vol 1/97. Ed. Real Instituto y Observatorio de la Armada, San Fernando, Cádiz 1997)
4. J. Ibáñez, J. Almendros, E. Carmona, C. Martínez Arévalo, M. Abril: The recent seismo-volcanic activity at Deception Island volcano. *Deep-Sea Res. II* **50**:1611–1629 (2003)
5. M. Berrocoso, M. E. Ramírez, A. Fernández-Ros: Horizontal Deformation Models for Deception Island (South Shetland Islands, Antarctica). In: *Geodetic Deformation Monitoring: From Geophysical to Engineering Roles*, IAG Vol. 131, ed by F. Sanso and A. Gil (Springer, Berlin Heidelberg New York 2006) pp 217–221
6. M. Berrocoso, A.Fernández-Ros, C. Torrecillas, J. M. Enríquez de Salamanca, M. E. Ramírez, A. Pérez-Peña, M. J. González, R. Páez, Y. Jiménez, A. García, M. Tárraga, F. García: Geodetic Research on Deception Island. In: *Antarctica, Contributions to Global Earth Sciences* (Springer, Berlin Heidelberg New York 2006) pp 391–396
7. A. Fernández-Ros: Displacement models and crustal deformations from GPS observations. Application to Deception Volcano. PhD Thesis, University of Cádiz, Spain (2006)
8. A. Fernández-Ros, M. Berrocoso, M. E. Ramírez: Volcanic deformation models for Deception Island (South Shetland Islands, Antarctica). In: *Antarctica: a*

AU: Please provide the editors nome.

keystone in a changing world. Proceedings for the 10th International Symposium on Antarctic Earth Sciences. Extended abstract 094 (U. S. Geological Survey and The National Academies, California 2007)
9. J. L. Smellie: Lithostratigraphy and volcanic evolution of Deception Island, South Shetland Islands. *Antarc. Sci.* **13** 2:188–209 (2001)
10. J. Martí, J. Vila, J. Rey: Deception Island (Bransfield Strait, Antarctica): An example of volcanic caldera developed by extensional tectonic. *J. Geolog. Soc. Lon.* **32**:253–265 (1994)
11. M. Martini, L. Giannini: Deception Islands (South Shetlands): an area of active Volcanism in Antarctica. *Società Geologica Italiana* **43**:117–122 (1998)
12. A. García and DECVOL Working Group: *A cross-disciplinary study at Deception Island (South shetland Islands, Antarctica). Evaluation of the recent volcanological status.* (Internal Report 2002)
13. J. Ibáñez, E. del Pezzo, J. Almendros, M. La Rocca, G. Alguacil, R. Ortiz, A. García: Seismovolcanic signals at Deception Island Volcano, Antarctica: Wave field analysis and source modelling. *J. Geophys. Res.* **105**:13905–13931 (2000)
14. G. Beutler, H. Bock, E. Brockmann and BERNESE Working Group: *BERNESE Software Version 4.2.* (Ed by U Hungentobler, S Schaer, P Fridez. Astronomical Institute, University of Bern, Bern 2001) pp 1–500
15. M. N. Bouin, C. Vigny: New constraints on Antarctic plate motion and deformation from GPS data. *J. Geophys. Res.* **105**:28279–28293 (2000)
16. K. V. Kumar, L. J. Miyashita K: Secular crustal deformation in central Japan, based on the wavelet analysis of GPS time-series data. *Earth Planets Space* **54**:133–129 (2002)
17. E. M. Souza, J. F. G. Monico: Wavelet Shrinkage: High frequency multipath reduction from GPS relative positioning. *GPS Solut.* **8**:152–159 (2004)
18. C. Ogaja, J. Wang, C. Rizos: Detection of Wind-Induced Response by Wavelet Transformed GPS Solutions. *J. Surv. Engrg.* **129**, 3:99–105 (2003)
19. C. Stein: Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**:1135–1151 (1981)
20. A. P. Calderón: Intermediate spaces and interpolation. *Stud. Math.* **24**:113–190 (1964)
21. D. Donoho, I. Johnstone: Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**:425–455 (1994)
22. D. Donoho, I. Johnsotone: Adapting to unknown smoothness via wavelet shrinkage. *J. American Statist. Assoc.* **90**:1200–1224 (1995)
23. A. W. Marshall, I. Olkin: *Inequalities: Theory of Majorization and its applications.* (Academic Press, Boston 1979) pp 1–569
24. C. Bruyninx, M. Yseboodt: *Frequency analysis of GPS coordinate time from the ROB EUREF Analysis Centre. Technical Report.* (TWG/Status of the EUREF Permanent Network 2001)
25. A. Caporali: Average strain rate in the Italian crust inferred from a permanent GPS network- I. Statistical analysis of the time-series of permanent GPS stations. *Geophys. J. Int.* **155**:241–253 (2003)
26. M. Poutanen, H. Koivula, M. Ollikainen: On the periodicity of GPS time series. In: *Proceedings IAG 2001 Scientific Assembly* (Budapest, Hungary 2001)
27. X. L. Ding, D. W. Zheng, Q. Chen, C. Huang, W. Chen: Analysis of seasonal and interannual variations in the position of permanent GPS tracking stations. In: *Proceedings FIG XXII International Congress* (Washington 2002)

28. D. Altadill: Quasi-periodic oscillations in the high ionosphere related to the planetary waves activity in the medium atmosphere. PhD. Thesis, Ramón Llull University, Barcelona (2001)
29. M. Vetterli, J. Kovacevic: *Wavelets and Subband Coding*. (Ed. Pearson Education, 1995) pp 1–448
30. Y. Bock, R. M. Nikolaidis, P. J. Jonge: Instantaneous geodetic positioning at medium distances with the Global Positioning System. *J. Geophys. Res.* **105**, B12: 28223–28253 (2000)
31. M. E. Ramírez: Crustal deformation models in volcanic areas by means of the wavelet theory. Application to Deception Island Volcano. PhD Thesis, University of Cádiz, Spain (2006)

Describing Seismic Pattern Dynamics by Means of Ising Cellular Automata

Abigail Jiménez[1], Antonio M. Posadas[2], and Kristy F. Tiampo[1]

[1] Department of Earth Sciences Biological and Geological Sciences, University of Western Ontario, London, Canada ajimene@uwo.ca
[2] Department of Applied Physics, University of Almería, Spain aposadas@ual.es

Abstract. This chapter is dedicated to the description and testing of a new method of obtaining Probabilistic Activation Maps of seismic activity. This method is based upon two major concepts: Cellular Automata (CA) and Information Theory. The proposed method can be used in other fields, as long as the spatially extended system is described in terms of a Cellular Automata with two available states, $+1$ and -1, as in the Ising case described here. The crucial point is to obtain the rules of an Ising Cellular Automata that maps one pattern into its future state by means of an entropic principle. We have already applied this technique to the seismicity in two regions: Greece and the Iberian Peninsula. In this chapter, we study other regions to test if the observed behavior holds in general. For this purpose, we will discuss the results for California, Turkey and Western Canada. The Cellular Automaton rules obtained from the correponding catalogs are found to be well described by an Ising scheme. When these rules are applied to the most recent pattern, we obtain a Probabilistic Activation Map, where the probability of surpassing a certain energy (equivalent to a certain magnitude) in the next interval of time is represented, which is a useful information for seismic hazard assessment.

1 Introduction

When looking for a framework that allows for studying nonlinearity and stochasticity at the same time, Information Theory is one of the most natural candidates. Information Theory confronts the problem of constructing models from experimental time series. These models are used to make predictions, but the underlying dynamics is unknown (or known but with a high dimensionality and thus incomputable). Information Theory was described (for the first time) by Shannon [1, 2] and Shannon and Weaver [3]. This formalism was later used by Shaw [4] to study the time series produced by a drop of water that falls from a faucet not properly turned off. He established an alternative way to deal with complex problems in the phase space. The behavior and

evolution of a system in which a series of states is known that occurred at times $T_1, T_2, ..., T_n$, can be characterized by a return map; that is, by representing the state at T_i in the x-axis, and the one at T_{i+1} in the y-axis, and the process goes on until the adequate dimension is obtained. Shaw used the resulting scatter plots and the concept of information based on Shannon's entropy in order to study the knowledge of the future states from the present and the past states. This knowledge can be (among others) characterized by a quantity called mutual information, which will be described later. In a nonlinear, spatially extended system, information can be produced in certain regions, and this information can be observed in other regions of the system some time later. If localized dynamics at point A result in information generation, then with finite-accuracy measurements it will take some time for the information which is observable in A at time t to be observable in point B; the time is, roughly, the diffusive time L^2/D, where L is the spatial separation of A and B, and D is a diffusion coefficient. Information transport can be detected by computation of an information-theoretic quantity, the time-delayed mutual information, between measurements of the system at separate spatial points [5]. Our technique uses this concept in a discrete representation of the system, in the form of a Cellular Automaton (CA). The information transport is carried out by the rules of said CA. Our goal is to maximize the information contained in the neighborhood (A) at a given time about the state in certain point B (the central cell) at a later time.

CA have been proposed as a model for self-organized criticality [6, 7], and have been applied as a simple analogue of earthquake occurrence by many authors [8, 9, 10]. Those CA are proposed as direct models, based on physical hypothesis, and find similar statistics to that found in earthquake catalogs. These models correspond to a new approach in seismology modeling the earthquake activity. From the first models of fault rupture [11], we have moved forward to describing it as a complex system where interactions between the faults play a main role. This is also due to the fact that seismologists have found significant relations between seismicity and critical systems [12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. These findings have lead a new approach to this problem based on statistical mechanics foundations. In this framework, CA models proposed try to capture this critical behavior in a direct way. With our approach, instead, we try to construct a simple CA that best fits the actual sequence of the earthquake occurrence in a seismic catalog. Our technique has already been used in two seismically active regions (Greece and Iberian Peninsula) [22, 23], where we found that an Ising CA model is capable to describe the major behavior of the considered records in a reasonable way. In this contribution, we study earthquake catalogs from three other regions, in order to further validate whether Ising CA models are good candidates for deriving meaningful information about the dynamics of seismic activity. This chapter is organized as follows: In Sect. 2, we describe our method in some detail. In Sect. 3 we present the study regions; Sect. 4 is devoted to our results, and finally, in Sect. 5 we expose our conclusions.

2 Description of the Method

As we already stated, we use a discrete representation of the seismic region. Our dynamical model is a CA which simulates the spatio-temporal evolution of the different seismic patterns obtained from the discretization process. The interactions are independent of time or static. Yet, seismicity might be have non-stationary features, so other types of dynamics are certainly conceivable as well. Our approach deals with an Ising-like behavior in the sense that cells tend to be in the same state as their neighborhood. The corresponding model is fully explained in [23], so here we are pointing out just the principal steps. As a cellular automaton, it is represented by its lattice Z^d of cells, a finite set A of states, a neighborhood set $N \subset Z^d$, and a local rule f updating the cells.

The coarse-graining of the events (Fig. 1), both spatially and temporally, produces a series of lattices and, after a state is assigned (active, $+1$, or quiescent, -1, in this case), a series of patterns is obtained. The dynamics of these patterns is what we want to simulate. With that, we already have chosen the lattice Z^d (square cells in 2D) and the set of states A.

The activation criteria are based on the time series given by the quantity:

$$\epsilon_q(N(\tau)) = \sum_{n=1}^{N(\tau)} \varepsilon_n^q \tag{1}$$

where ε_n is the released energy of the n^{th} event, and $N(\tau)$ is the number of earthquakes in a given interval of time τ, and where the energy is calculated from the relationship between magnitude and energy [24]. If ϵ_q exceeds some certain threshold in the time interval, the cell is considered active $(+1)$ and, otherwise, is quiescent (-1). Note that ϵ_1 is the accumulated energy, $\epsilon_{1/2}$ represents the Benioff strain, $\epsilon_{1/3}$ is a proxy for the accumulated radius of the rupture, and ϵ_0 represents the number of events. When the value of q is lower, it means that small earthquakes have a higher weight in ϵ_q. In general, ϵ_q is a proxy for the stress accumulation in the cell.

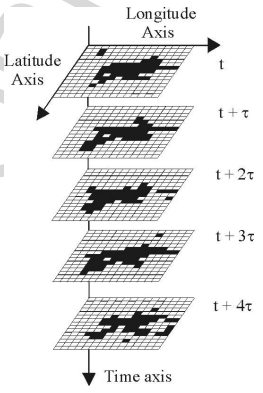


Fig. 1: Coarse-graining.

As we said, in this work we assume an Ising-like framework. This Ising framework is introduced by the way the interactions between the cells are constructed. The rules for updating a cell (f) must follow those of an Ising system. The original Ising model describes the dynamics of a system of coupled spins, where the spin at every site is either up ($+1$) or down (-1). Unless otherwise stated, the interaction is between nearest neighbors only and is given by $-$J if the spins (which we will identify with the state of seismic activation) are parallel and $+$J if the spins are anti-parallel. The total energy can be expressed in the form

$$E = -J \sum_{i,j} S_i S_j - B \sum_i S_i, \tag{2}$$

where $S_i = \pm 1$, J is known as the exchange constant and B is the external (magnetic) field. The analog in the seismic case is the external field, that corresponds to both plate motion and ductile zone under the cells. In a CA representation, the energy is calculated for each site, j being the nearest neighbors, and added. So, we define the state of the cell at a time t by its energy with respect to its neighbors, and the state at a later time, is referred to its activity. Then, we are interested in the maximum transmission of information from the neighborhood to the cell.

At that point, we already have a series of lattice configurations (patterns). In the CA representation, we assume that each cell interacts only with its nearest neighbors, and we can calculate the transition rules directly from these patterns [25] by means of a histogram of occurrences.

In an Ising model, the flip transitions are given by the energy state of the cells, so that a cell in a state has a certain probability of changing depending on the energy of the interactions with its neighborhood and the external field. In our method, these probabilities have to be calculated, yielding an inverse or data assimilation problem. We assume that we have no a priori hypothesis about the nature of the interactions between neighboring sites, nor between the sites and an external field. Therefore, we classify the neighborhoods configuration in terms of its "energetic" state, so that each cell has an associated energy, E_i, given by Eq. (2), with S_i being the central cell's state and S_j the neighboring cells' states, without an external field (which would represent the driving forces, but cannot be calculated), and with the term J set to 1 without loss of generality. The "energetic state" of a cell with respect to its neighborhood is then given by:

$$E_i = -\sum_j S_i S_j. \tag{3}$$

We use a Moore's neighborhood ($N \subset Z^d$), since it is more isotropic [25], and the "energy" can take only discrete values in the interval $E \in [-8, 8]$. No Ising behavior is imposed on the transition probabilities, but they are extracted from the data itself by calculating the distribution of the

neighborhood's states and its influence in the activity or inactivity of the cell in the future.

The transmission of information depends on the number of cells, N, and time interval, t, chosen. By maximizing the time-delayed mutual information, μ_I, between the past and future states we can find the model which contains a higher correlation between them [26]. The expression for μ_I in this particular model is as follows:

$$\mu_I = \sum_{i,j\in\{\pm 1\}} \sum_{k=E_0}^{E_n} p(i;j,k) log_2 \frac{p(i;j,k)}{p(i)p(j,k)} \tag{4}$$

with $p(i;j,k)$ being the joint probability of past and future states, and $p(i)p(j,k)$ a distribution of independent states, i stands for the central cell at time $t+\tau$, and (j,k) relates to the central and its k neighborhood's state at time t, with E_i $(i \in [0,n],\ i \in N)$ representing the possible states. The calculated value of μ_I represents the expected *information gain* when using a model with interacting cells instead of another model where the consecutive states are independent [27]. To find the maximum value of μ_I, a grid search in time steps and number of cells is carried out and, finally, we derive our Cellular Automaton.

After obtaining the transition rules, we can test how well they reproduce the data. Simulations of the future patterns are carried out [28], and real and simulated patterns are compared by means of the correlation function [29] and the Hamming distance [30]. The latter measure gives the number of cells that failed in the prediction, representing the simulation error between two binary patterns in the usual way.

Finally, in our application to seismic data, if the CA rules are applied to the latest pattern, we obtain what we call a Probabilistic Activation Map, with the probability of surpassing certain cumulative ϵ_q (equivalent to certain magnitude) [25, 31, 23]. Since the model takes into account activity $(+1)$ and quiescence (-1), a probability p of becoming active corresponds to a probability $1-p$ of quiescence. To highlight the Ising behavior, we have modified this scale from $[0,1]$ to $[-1,1]$. The maps are slightly smoothed in the corners of the cells, because the spatial extension of the cells that maximize the mutual information is too large, so that the display is more understandable. In a more general case, it will represent the probability of observing a pattern at the next step of the CA.

3 Data Description and Tectonic Setting

In this section, we will describe the three sets of data we are using. The California and Turkey regions have already been extensively studied [32, 33], so they represent two places where a great amount of information has been gathered, which allows a reliable discussion of our results. The Western Canada region is also an important place, where big events can be expected.

- The Southern California catalog: The catalog is maintained by the Southern California Earthquake Center (SCEC) and contains the seismic data for the period 1932–2001. The analyzed area ranges from 32–40° N, and 115–124° W. The magnitude spans from 3.0 to 8.0, where it's complete. The maximum depth is 79 km. The faults in the region are shown in Fig. 2.
- The Turkey catalog: The catalog is maintained by the Kandilli Observatory in Istanbul, and spans the years from 1900 to 2004. The area ranges from 22–46° E and 31–46° N. The magnitude spans the years from 3 to 7.9. It is complete above magnitude 4.5. The tectonic setting for Turkey is shown in Fig. 3.
- The Western Canada catalog: Obtained from the Canadian National Data Centre, data from 1700 to 2004. The area ranges from 120–135° W and 46.7–55.1° N. Maximum depth is 105 km, the magnitude spans from −0.6 to 9. The catalog is complete above magnitude 5. The tectonic setting for Western Canada is shown in Fig. 4.

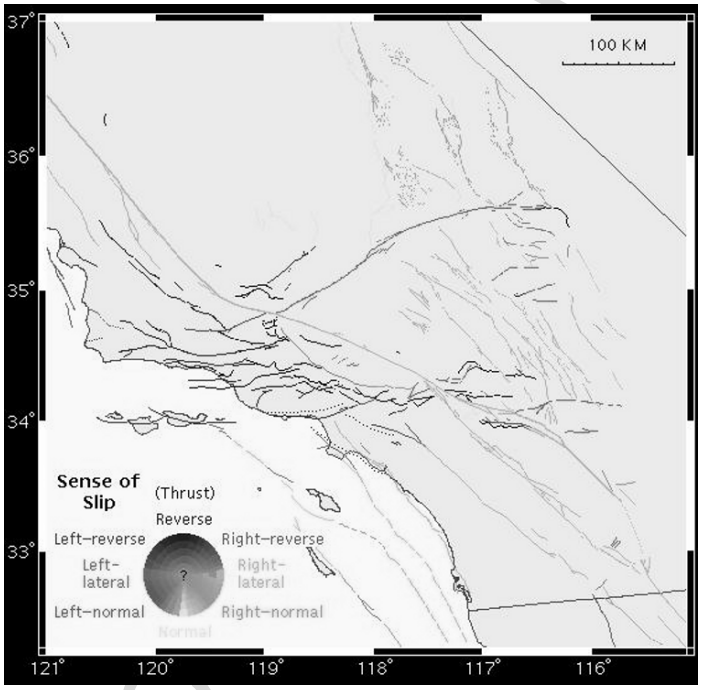


Fig. 2: Sense of slip map for Southern California. Courtesy of John Marquis, created using GMT, with a data set cobbled together from several different sources [34, 35, 36, 37, 38, 39].

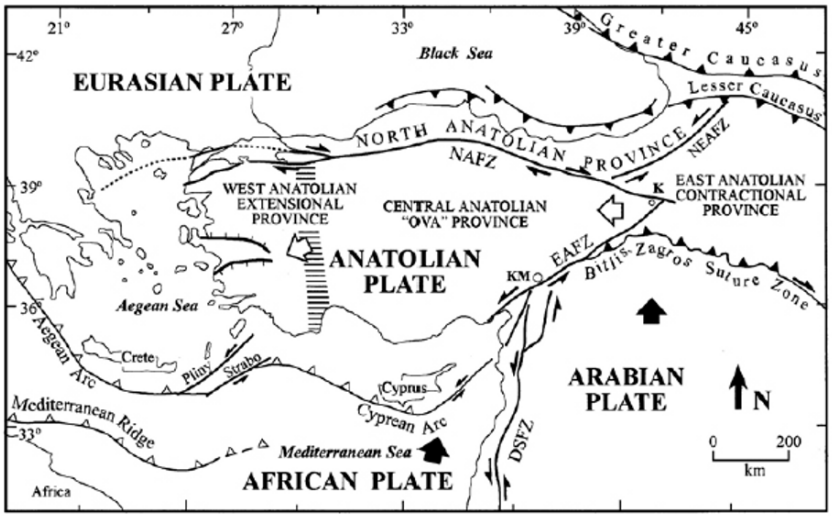


Fig. 3: Tectonic setting in Turkey [40, 41].

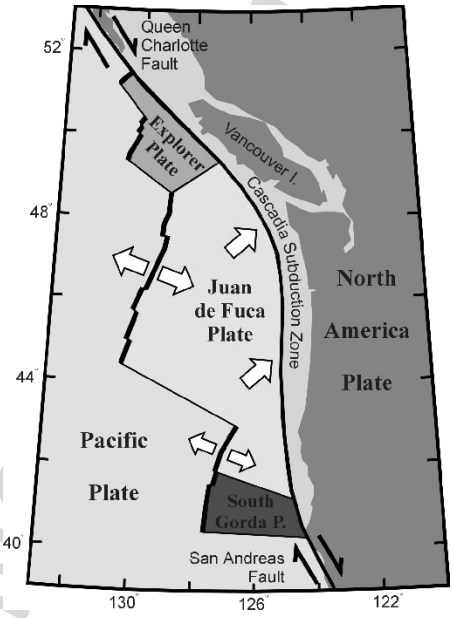


Fig. 4: Tectonic setting in Canada.

4 Results

The results for California are shown in Table 1 for each catalog and each q and magnitude threshold tried for the seismic patterns. The analysis for Turkey is in Table 2; and, finally, in Table 3 we show the results for the Western Canadian catalog. The maximum mutual information is found with a box size of 2 ($\approx 220 \times 220$ km^2) for all the catalogs, except in the case of California and a magnitude threshold of 3, where the mutual information is maximized at 1 ($\approx 110 \times 110$ km^2). For the other catalogs we did not find that, since they are not complete at this magnitude. When we use a box size of 2, the time intervals increase (a lower inter-event time, second row in the tables, where the inter-event time is obtained by dividing the difference between the last and first time in the data set by the time interval) for the maximization of μ_I. After that, the number of intervals decreases with increasing magnitude threshold. A rough estimation of the corresponding diffusion coefficient D [5] gives around 500, 130 and 40 m^2/s for California and magnitude thresholds of 4, 5 and 6, respectively; around 150 and 25 m^2/s in Turkey for magnitude 5 and 6; and lower values for Western Canada, around 30, 20 and 10 m^2/s with magnitudes 5, 6, and 7. As [42] pointed out, those values for California are consistent with the work by [43], and give a realistic diffusion coefficient

Table 1: Results of the maximization of the Ising model with an energy threshold criterion for California (q is the value of q in Eq. (1), m is the magnitude threshold, t and *boxsize* are the number of time intervals and the spatial resolution in degrees that maximize μ_I, the mutual information, the error is the Hamming distance in %, and M is the averaged 'magnetization', or number of active cells).

q	m	t	*boxsize*	μ_I	error (%)	M (%)
1	3	5	1	0.78	8	67
1	4	25	2	0.79	12	56
1	5	7	2	0.72	13	47
1	6	2	2	0.83	4	36
1	7	2	2	0.29	12	12
1/2	3	5	1	0.78	8	67
1/2	4	37	2	0.83	10	55
1/2	5	6	2	0.78	11	59
1/2	6	4	2	0.81	8	39
1/2	7	2	2	0.69	4	20
1/3	3	5	1	0.78	8	67
1/3	4	41	2	0.85	12	56
1/3	5	12	2	0.86	7	56
1/3	6	5	2	0.88	6	50
1/3	7	2	2	0.65	12	38
1/3	8	2	2	0.24	12	12

Table 2: Results of the maximization of the Ising model with an energy threshold criterion for Turkey.

q	m	t	$boxsize$	μ_I	error (%)	M (%)
1	5	9	2	0.54	20	47
1	6	2	2	0.52	13	45
1	7	2	2	0.22	7	14
1/2	5	5	2	0.59	16	63
1/2	6	3	2	0.60	20	44
1/2	7	2	2	0.40	5	18
1/3	5	13	2	0.59	18	49
1/3	6	2	2	0.66	11	60
1/3	7	2	2	0.59	12	35
1/3	8	2	2	0.15	10	7

Table 3: Results of the maximization of the Ising model with an energy threshold criterion for Western Canada.

q	m	t	$boxsize$	μ_I	error (%)	M (%)
1	5	8	2	0.28	9	17
1	6	4	2	0.25	17	16
1	7	2	2	0.12	16	8
1	8	2	2	0.08	3	3
1/2	5	8	2	0.28	9	17
1/2	6	4	2	0.25	17	16
1/2	7	5	2	0.10	7	6
1/2	8	2	2	0.08	3	3
1/3	5	8	2	0.28	9	17
1/3	6	5	2	0.28	11	15
1/3	7	2	2	0.13	25	13
1/3	8	2	2	0.11	9	6

for the region. Turkey seems to have a similar value, but not Western Canada, which appears to be slower.

In general, the mutual information increases with decreasing q as well. The difference is higher (more information can be extracted) for high magnitude cutoffs in the case of $q \neq 0$, and is therefore more useful for seismic hazard assessment. It is also interesting to note the increase in the "magnetization", approaching 50% of active cells (approaching the null magnetization) with a higher mutual information.

The transition probabilities represent the rules for the CA. Although they are stored in tables, they can be better visualized as in Fig. 5. In gray the situation is represented when the cell is initially inactive, and in black, when the

Fig. 5: Rules for the CA. See text for explanation.

cell is initially active. It always happens that there is a value for the "energy" in Eq. (3) when it is more likely to change the cell's initial state (at times an interval oscillates around 50%). That is marked, symbolically, as a vertical line in the figure's bar. The value is not the same for an initially inactive and for an initially active cell. It should be noted that this behavior is consistent with an Ising-like system, where as the "energy" increases, the probability of changing the state is higher. We also point out that this representation can describe almost all the situations we tried, in all the regions. Some regions do not present enough data for the rules, since the threshold magnitudes are too high, and no reliable histogram can be obtained. This is analogous to a ferromagnetic material, where the cells tend to adopt the same state as their neighborhoods. Depending on the difference between the values for changing an initially inactive and an initially active cell, we can interpret that the external field increases ($B < 0$) or decreases ($B > 0$) the probability of changing the state. For California, there is no clear trend in the sense of the external field. If we focus on the results for $q = 1/3$, and other q with low magnitude thresholds, it is positive, so that the plate boundary conditions favor the relaxation of the stresses. However, for high magnitudes and $q = 1$ or $1/2$, the opposite behavior is found. In Turkey, the general trend is $B < 0$, so that the external field is constantly increasing the probability of changing the state. The Western Canada catalog is not clear either, but in this case because of the lack of data.

The fact that for configurations with $E < 0$ (the surrounding activity state is the same as that of the central cell's activity), the state is always reinforced (both active and inactive) is a typical feature of Ising-like behavior. Taking this into account, an active region loads its neighborhood when it releases energy. This is also a feature contained in the Cellular Automata used to simulate the seismicity in the literature [9, 10, 44], that leads to the activation of neighboring areas, if they are close to the rupture point. However, the results obtained here also point out that an active region becomes quiescent because of the neighboring quiescence. This is in accordance with Griffith's principle, in which cells are broken when the release of elastic energy exceeds the surface energy cost [45]. If a cell releases energy and the surrounding areas are not near the rupture point, they will absorb this energy in an elastic way, without becoming active, so that when the initially active cell releases all the exceeding energy, it becomes inactive as well. When $E \geq 0$, the transition to changing the state is less clear, mainly because of the non-uniformity in the stress field. In that case, B is different for each cell, as shown before.

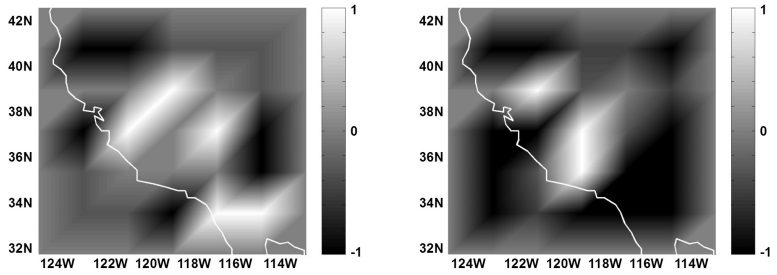


Fig. 6: Probabilistic Activation Maps for the next intervals of time for California with $q = 1$, $m = 6$ (35 years), and $m = 7$ (35 years).

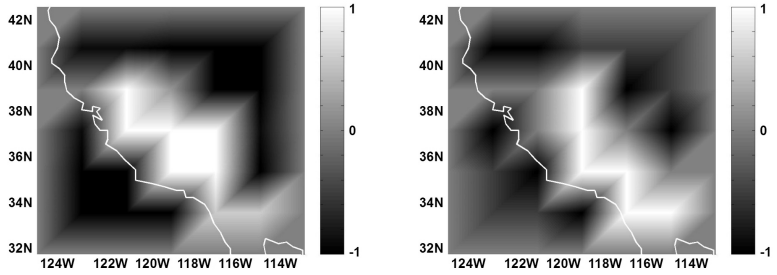


Fig. 7: Probabilistic Activation Maps for the next intervals of time for California with $q = 1/2$, $m = 6$ (17 years), and $m = 7$ (35 years).

Figures 6–8 show the Probabilistic Activation Maps obtained for California, Figs. 9–11 represent the results for Turkey, and Figs. 12–14 give the maps for Western Canada.

As it might be expected, the higher probabilities of occurrence lie on the principal faults: in particular, San Andreas fault is delineated with $q = 1/3$ up to magnitude 7. After that, the most energetic spots are near the Big Bend and near Garlock fault. This feature is seen in all q, so we would expect the bigger earthquakes (around magnitude 7) to occur at those places. In fact, two earthquakes of magnitude 6 have occurred near the Big Bend after our data set.

We find two interesting locations in the Turkey catalog: the earthquakes occurring in the Aegean Sea and the Karliova junction, where the principal faults in the region converge. As it can be seen, the higher probabilities of occurrence for the North Anatolian fault lie at both extremes of the mentioned fault. That result is coincident with the forecast made in 1996 by Stein et al. [33]. They assumed that earthquakes interact, as we also do. However, some earthquakes occurred since that time at the mentioned places, so our maps

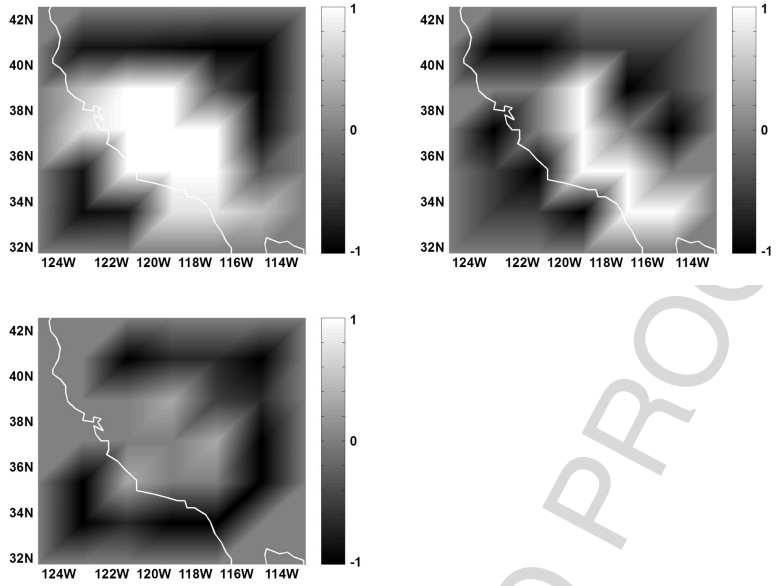


Fig. 8: Probabilistic Activation Maps for the next intervals of time for California with $q = 1/3$, $m = 6$ (14 years), $m = 7$ (35 years), and $m = 8$ (35 years).

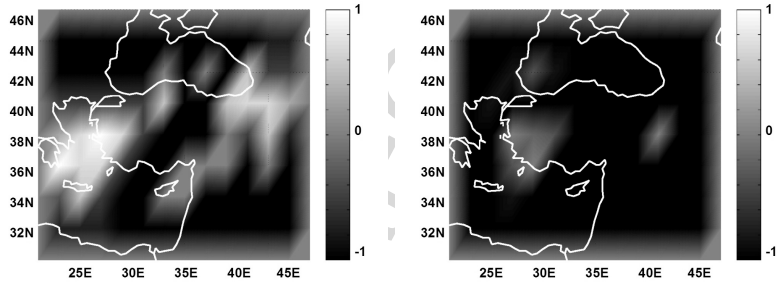


Fig. 9: Probabilistic Activation Maps for the next intervals of time for Turkey with $q = 1$, $m = 6$ (52 years), and $m = 7$ (52 years).

show a continuation of the activity in those regions. Note also that they are hazard maps (in the sense explained before), and not forecasts. Nevertheless, it is interesting to see the coincidence in the locations. As before, as q decreases, the probabilities increase for the higher magnitudes. Only one earthquake of magnitude 6 has occurred after the time covered by our catalog, and it has been located in the Aegean Sea.

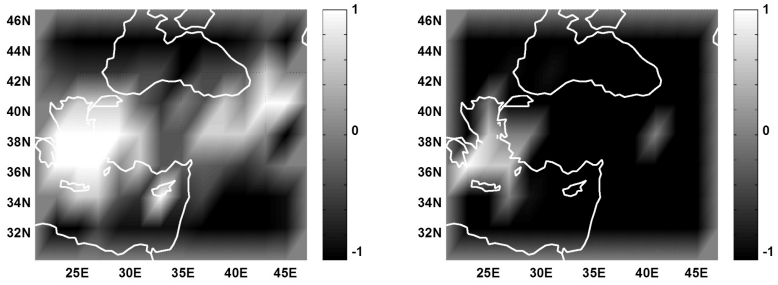


Fig. 10: Probabilistic Activation Maps for the next intervals of time for Turkey with $q = 1/2$, $m = 6$ (35 years), and $m = 7$ (52 years).

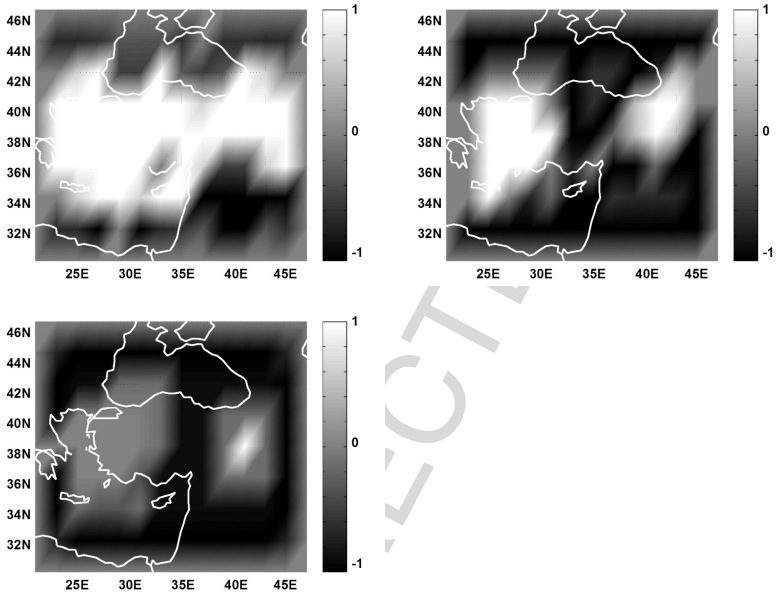


Fig. 11: Probabilistic Activation Maps for the next intervals of time for Turkey with $q = 1/3$, $m = 6$ (52 years), $m = 7$ (52 years), and $m = 8$ (52 years).

In Western Canada we find that the higher energy releases are expected to be related to the Explorer plate, affecting both Vancouver and the Queen Charlotte islands. Around 10 earthquakes of magnitude 5 have occurred in the sea between these two islands after the data finish. By means of a quick calculation, only with a threshold based upon $q = 1/3$ these 10 earthquake of magnitude 5 added up surpass one of magnitude 6. With $q = 1$ and $q = 1/2$, that is not true. The locations are also coincident with the white spot in

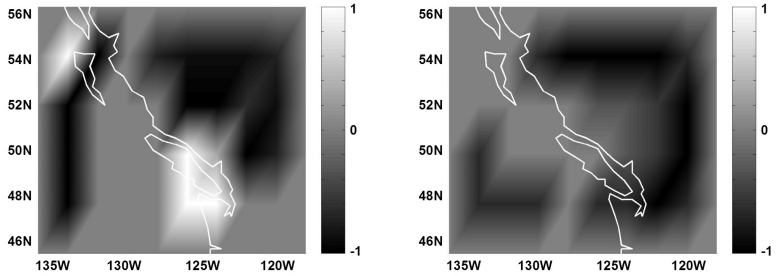


Fig. 12: Probabilistic Activation Maps for the next intervals of time for Western Canada with $q = 1$, $m = 6$ (76 years), and $m = 7$ (152 years).

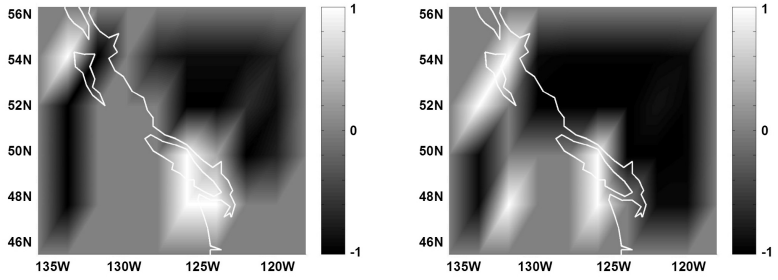


Fig. 13: Probabilistic Activation Maps for the next intervals of time for Western Canada with $q = 1/2$, $m = 6$ (76 years), and $m = 7$ (61 years).

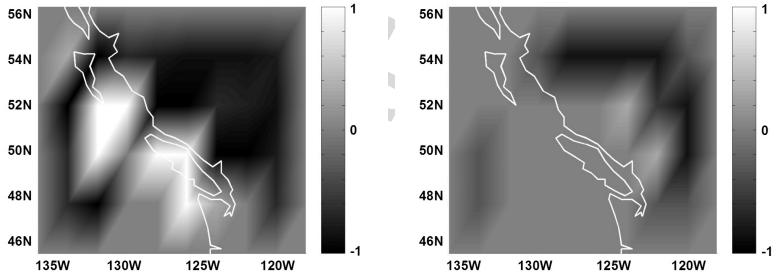


Fig. 14: Probabilistic Activation Maps for the next intervals of time for Western Canada with $q = 1/3$, $m = 6$ (61 years), and $m = 7$ (152 years).

Fig. 14 just under the Queen Charlotte Islands, and that location does not appear with high probability for the other q used. However, since the time intervals involved are long term, we can not decide which map is best, or more helpful for different purposes. Note that the errors change slightly with

the chosen values for q. However, the Hamming distance, as said before, is the usual way to compare two binary patterns.

In all cases, the places where these extreme events are foreseen are consistent with the tectonic setting. Other hazard maps for these regions show the probability of exceedance in 50 years are delineating the principal faults. The important point of our method is the idea that a particular time since the last data needs to be checked. The time-dependence is important in order to obtain the transition probabilities. If we average the energy, we should find the temporal and spatial scales for the averaging. If we use the same as that found in the model, the patterns change, so it would not be time-independent. In a time-independent estimation, the energy (or magnitude) used for the calculations is usually the highest found in each seismogenic zone. That may change in time, depending on the time span of the catalog. So, if we compare the patterns with the highest magnitudes, or add up the energy, we find that they are time-dependent. Note that since the time span is long and the spatial resolution is low, it can not be viewed as a prediction, but as a forecasting, or a Probabilistic Activation Map, but it is not static. It is also noted that there are no attenuation laws and no site responses applied, so the meaning of these Probabilistic Activation Maps are the probabilities of surpassing certain energies (magnitudes) in the different studied areas. So they are understood as seismic hazard maps, rather than an actual forecast.

5 Conclusions

The method proposed in [23] has been used in three catalogs, and the results obtained have shown to be consistent with previous studies. The seismic activity patterns can be translated to an Ising Cellular Automaton model to obtain estimates of the probabilities of surpassing certain energies (magnitudes) at each cell of said CA. The Probabilistic Activation Maps obtained may be useful for time-dependent seismic hazard assessment, although the transition probabilities are static. This methodology is also applicable to other fields, as long as the system is described in terms of a CA, that is, a coarse-graining is carried out, in time, space and state, to describe its main features in a discrete fashion, e.g. in ecological evolution. The evolution of the obtained patterns are given by the rules of the CA that maximize the delayed-mutual information.

The idea from previous results that a lower number of cells usually maximizes the mutual information is reinforced. It reflects the large scale nature of earthquake occurrence. This, joint with the different periods of time for each magnitude threshold, can be roughly explained as a diffusion of the stresses, which are transmitting the information to the whole system.

Our Probabilistic Activation Maps for California mark the Big Bend and Garlock faults as the places where higher magnitude releases might be expected. In Turkey, the higher seismic hazard corresponds to the Aegean Sea

and the Karliova junction, both extremes of the North Anatolian fault. Finally, in Western Canada, we expect the higher seismic activity around Vancouver and the Queen Charlotte Islands, related to the Explorer plate. All these results are consistent with the different tectonic settings.

Finally, we conclude that evolving CA models, although simplified, capture some interesting behavior of the seismic interactions. More work will be done to see if this methodology gives us more insight about regional seismicity and its relationship to statistical mechanics. In particular, we are interested improving the resolution in both space and time, and to see if the fractal nature of seismicity can be used for this purpose. Other possible generalizations of this work are to add up more states for the cells, and to try a Potts model, instead of an Ising one. Finally, the error analysis should be improved, in the sense that should be more sensible to the changes in the parameters, and how different thresholds in the probability maps affect the results.

Acknowledgements. We thank Reik Donner for his helpful comments and corrections. This work was partially supported by the MCYT project CGL2005-05500-C02-02/BTE, the MCYT project CGL2005-04541-C03-03/BTE, and the Research Group 'Geofísica Aplicada' RNM194 (Universidad de Almería, España) belonging to the Junta de Andalucía. The work of KFT was funded by the NSERC and Benfield/ICLR Industrial Research Grant. The work of AJ was funded by a 'Fundación Ramón Areces' Grant.

References

1. Shannon, C.E.: The mathematical theory of communication I. Bell Syst. Tech. J. **27** (1948) 379–423
2. Shannon, C.E.: The mathematical theory of communication II. Bell Syst. Tech. J. **27** (1948) 623–656
3. Shannon, C.E., Weaver, W.: The mathematical theory of communication. University of Illinois Press (1949)
4. Shaw, R.: The dripping faucet as a model chaotic system. The Science Frontier Express Series, Aerial Press, Santa Cruz, California (1984)
5. Vastano, J.A., Swinney, H.L.: Information transport in spatiotemporal systems. Phys. Rev. Lett. (1988) 1773–1776
6. Bak, P., Tang, C., Wiesenfeld, K.: Self-organized criticality: An explanation of 1/f noise. Phys. Rev. Lett. **59** (1987) 381–384
7. Bak, P., Tang, C., Wiesenfeld, K.: Self-organized criticality. Phys. Rev. A **38** (1988) 364–374
8. Ito, K., Matsuzaki, M.: Earthquakes as self-organized critical phenomena. J. Geophys. Res. **95** (1990) 6853–6860
9. Barriere, B., Turcotte, D.L.: A scale-invariant cellular automaton model for distributed seismicity. Geophys. Res. Lett. **18** (1991) 2011–2014
10. Olami, Z., Feder, H.J.S., Christensen, K.: Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys. Rev. Lett. **68** (1992) 1244–1247

11. Reid, H.F.: The mechanics of the California earthquake of april 18, 1906. Report of the State Earthquake Investigative Committee, Washington DC, Carnegie Institute (1910)
12. Hirata, T.: Fractal dimension of faults systems in Japan: fractal structure in rock fracture geometry at various scales. Pure Appl. Geophys. **131** (1989) 157–170
13. Hirata, T., Imoto, M.: Multifractal analysis of spatial distribution of microearthquakes in the Kanto region. Geophys. J. Int. **107** (1991) 155–162
14. Smalley, R.F., Chatelain, J.L., Turcotte, D.L., Prevot, R.: A fractal approach to the clustering of earthquakes: applications to the seismicity of the new hebrides. Bull. Seism. Soc. Amer. **77** (1987) 1368–1381
15. Omori, F.: On the aftershocks of earthquakes. J. Coll. Sci. **7** (1895) 111–200
16. Bufe, C.G., Varnes, D.J.: Predictive modeling of the seismic cycle of the greater San Francisco Bay regiony. J. Geophys. Res. **98** (1993) 9871–9883
17. Sornette, D., Sammis, C.G.: Complex critical exponent from renormalization group theory of earthquakes: Implications for earthquake predictions. J. Phys. I France **5** (1995) 607–619
18. Bowman, D.D., Ouillon, G., Sammis, C.G., Sornette, A., Sornette, D.: An observational test of the critical earthquake concept. J. Geophys. Res. **103** (1998) 24359–24372
19. Huang, Y., Saleur, H., Sammis, C., Sornette, D.: Precursors, aftershocks, criticality and self-organized criticality. Europhys. Lett. **41** (1998) 43–48
20. Bowman, D.D., King, G.C.P.: Accelerating seismicity and stress accumulation before large earthquakes. Geophys. Res. Lett. **28** (2001) 4039–4042
21. Tiampo, K.F., Anghel, M.: Critical point theory and space-time pattern formation in precursory seismicity. Tectonophysics **413** (2006) 1–3
22. Jiménez, A.: Morfodinámica de patrones sísmicos discretos mediante evolución entrópica y autómatas celulares: aproximación estocástica al peligro sísmico de la Península Ibérica y sus zonas sismogenéticas singulares. PhD thesis, University of Almería (2005) (Available at http://www.minas.upm.es/fundacion/jgs/ESP/index.html).
23. Jiménez, A., Tiampo, K.F., Posadas, A.M.: An Ising model for earthquake dynamics. Nonlin. Processes Geophys. **14** (2007) 5–15
24. Gutenberg, B., Richter, C.: Earthquake magnitude, intensity, energy, and acceleration. Bull. Seism. Soc. Am. **46** (1956) 105–145
25. Jiménez, A., Posadas, A.M., Marfil, J.M.: A probabilistic seismic hazard model based on cellular automata and information theory. Nonlinear Processes in Geophysics **12** (2005) 381–396
26. Cover, T., Thomas, J.: Elements of information theory. Wiley and Sons (1991)
27. Daley, D.J., Vere-Jones, D.: Scoring probability forecasts for point processes: the entropy score and information gain. J. Appl. Probab. **41A** (2004) 297–312
28. Posadas, A.M., Hirata, T., Vidal, F., Correig, A.: Spatiotemporal seismicity patterns using mutual information application to southern Iberian peninsula (Spain) earthquakes. Phys. Earth Planet. Inter. **122** (2000) 269–276
29. Vicsek, T.: Fractal growth phenomena. World Scientific (1992)
30. Ryan, M.J., Frater, M.R.: Communications and information systems. Argos Press (2002)
31. Jiménez, A., Posadas, A.M.: A Moore's cellular automaton model to get probabilistic seismic hazard maps for different magnitude releases: A case study for Greece. Tectonophysics **423** (2006) 35–42

32. USGS. http://www.usgs.gov/ (2008)
33. Stein, R.S., Barka, A.A., Dieterich, J.H.: Progressive failure on the North Anatolian fault since 1939 by earthquake stress triggering. Geophys. J. Int. **128** (1997) 594–604
34. Hart, E.W., Bryant, W.A., Kahle, J.E., Manson, M.W., Bortugno, E.J.: Summary report: Fault evaluation program, 1986-1987, Mojave desert region and other areas. Technical report, Depart of Conservation, Division of Mines and Geology (1988)
35. Hart, E.W., Bryant, W.A., Manson, M.W., Kahle, J.E.: Summary report: Fault evaluation program, 1984-1985, southern coast ranges region and other areas. Technical report, Depart of Conservation, Division of Mines and Geology (1986)
36. Hart, E.W., Bryant, W.A., Wills, C.J., Treiman, J.A., Kahle, J.E.: Summary report: Fault evaluation program, 1987-1988, southwestern basin and range region and supplemental areas. Technical report, Depart of Conservation, Division of Mines and Geology (1989)
37. Jennings, C.W.: Fault activity map of California and adjacent areas with location and ages of recent volcanic eruptions. California Geologic Data Map Series, Map No. 6 (1994)
38. Petersen, M.D., Wesnousky, S.G.: Fault slip rates and earthquake histories for active faults in southern California. Bull. Seism. Soc. Am. **84** (1994) 1608–1649
39. Wesnousky, S.G.: Earthquakes, quaternary faults, and seismic hazards in southern California. J. Gophys. Res. **91** (1986) 12587–12631
40. Bozkurt, E.: Neotectonics of Turkey-a synthesis. Geodinamica Acta **14** (2001) 3–30
41. Nesrin, T.: GIS based geothermal potential assessment for Western Anatolia (2006)
42. Takahashi, K., Seno, T.: Diffusion of crustal deformation from disturbances arising at plate boundaries-a case of the detachment beneath the Izu Peninsula, central Honshu, Japan-. Earth Planets Space **57** (2005) 935–941
43. Ida, Y.: Slow-moving deformation pulses along tectonic faults. Phys. Earth Planet. Int. **9** (1974) 328–337
44. Burridge, R., Knopoff, L.: Model and theoretical seismicity. Bull. Seism. Soc. Am. **57** (1967) 341–371
45. Toussaint, R., Pride, S.R.: Interacting damage models mapped onto Ising and percolation models. ArXiv Condensed Matter e-prints (2004)

Part III

Applications in Solar-Terrestrial Physics

Template Analysis of the Hide, Skeldon, Acheson Dynamo

Irene M. Moroz

Mathematical Institute, 24-29 St Giles', Oxford OX1 3LB, UK
moroz@maths.ox.ac.uk

Abstract. Self-exciting dynamos are nonlinear electro-mechanical engineering devices, or naturally-occurring magnetohydrodynamic fluid systems that convert mechanical energy into magnetic energy without the help of permanent magnets. Hide et al. [1] introduced a nonlinear system of three coupled ordinary differential equations to model a self-exciting Faraday disk homopolar dynamo. Since only a small selection of possible behaviours, including two examples of chaotic behaviour, was investigated by them, Moroz [2] performed a more extensive analysis of the dynamo model, including producing bifurcation transition diagrams and generating unstable periodic orbits for the two chaotic examples. We now extend that analysis and use ideas from topology [3] and results from a corresponding analysis of the Lorenz attractor to identify a possible template for the HSA dynamo.

1 Introduction

In 1996 Hide, Acheson and Skeldon [1] (hereafter denoted by HSA) introduced a nonlinear model for a self-exciting Faraday disk dynamo as a simple analogue for the heat storage capacity in the oceans, thought to be a key factor in the dynamical processes underlying the El Niño Southern Oscillation. Self-exciting Faraday disk dynamos, such as the HSA dynamo, are of interest since they contain some of the key ingredients of large-scale naturally occurring magnetohydrodynamic dynamos, while being of considerably lower dimension and therefore more amenable to systematic study.

Since their original paper, there have been many extensions to the original three-mode dynamo ([4, 5, 6, 7, 8, 9, 10, 11] to include such effects as the coupling two or more dynamos together, an azimuthal eddy current, an external magnetic field or a battery, etc. Many of the low order models of this family, have rich ranges of behaviour with irregular reversals a common feature, as well as steady, periodic and coexisting states (due to hysteresis effects). What has been lacking has been a means of distinguishing between these and other models as a prelude to comparing them with the large-scale counterparts. One

possible way is via their spectra of unstable periodic orbits (upos), as well as noting their behaviours when key parameters in the problem vary.

Topological methods have been developed to analyse three-dimensional dissipative dynamical systems in the chaotic regime [3]. Such methods supplement the more conventional approaches such as the calculation of Lyapunov exponents and dimension calculations, although we have also included such calculations in this chapter. The topological approach proceeds by identifying the expansion and contraction mechanisms which are involved in creating the strange attractor. This leads to a branched manifold, also called a template or knot holder, on which the upos are organised in a unique way. Certain topological invariants are computed from a chaotic time series. These invariants, usually determined from the lowest order upos, are the (a) Gauss linking numbers, (b) relative rotation rates and (c) templates themselves. One is then able to determine whether two dynamical systems are equivalent, whether a model accurately represents a physical system etc.

In a recent chapter, Moroz [2] returned to the original HSA study, which only investigated a very small selection of possible parameter values, and produced bifurcation transition diagrams for the two examples of chaotic dynamo behaviour reported by HSA. In addition first return maps were used to obtain unstable periodic orbits for these and other examples, but no attempt was made to identify the branched manifold for the underlying attractor. We rectify this now by reporting the linking number calculations that were used to identify the template, using numerical algorithms developed by Bob Gilmore.

The chapter is organised as follows. In Sect. 2 we review the derivation and the salient features of the HSA dynamo, including the linear stability analysis. In Sect. 3 we review how Koga [12] constructed the Poincaré section for his calculation of upos in the Lorenz equations. Section 4 summarises certain results from [2] and introduces two new cases to be studied here. Section 5 explains the template analysis, following [3] and Sect. 6 uses these ideas, and numerical algorithms for the computation of Gauss integrals, provided by Bob Gilmore, to compute tables of linking numbers for both the Lorenz and the HSA equations. We summarise our results in Sect. 7.

2 The Hide, Skeldon, Acheson Dynamo

We begin by introducing the Hide, Skeldon, Acheson (HSA) dynamo, following the treatment given in [1].

2.1 The HSA Equations

The HSA dynamo is a system of three coupled nonlinear ordinary differential equations for an electrically conducting Faraday disk, connected in series with a coil and a motor via sliding contacts attached to the axle and the rim of the disk. The disk is driven into rotation with angular speed $\Omega(\tau)$ by a steady

applied couple G. In the presence of a magnetic field, an e.m.f. is produced in the rotating disk, and a current $I(\tau)$ flows through the coil and motor.

Applying torque balance to the motor and to the disk respectively gives the following two nonlinear equations:

$$B\dot{\omega} = HI - D\omega, \tag{1a}$$
$$A\dot{\Omega} = G - MI^2 - K\Omega. \tag{1b}$$

Here A is the moment of inertia of the disk, K its coefficient of mechanical friction; B is the moment of inertia of the armature of the motor, D its coefficient of mechanical friction; HI is the torque on the armature, produced by the current; $D\omega$ is the torque due to mechanical friction in the motor; $K\Omega$ is the mechanical friction in the disk; $M/2\pi$ is the mutual inductance between the coil and the rim of the disk, and the dot denotes differentiation with respect to τ. The final equation, for I, comes from identifying the e.m.f.s from the various components of the dynamo and applying Kirchoff's Voltage Law. The e.m.f. generated by the moving disk $MI\Omega$ is balanced by the voltages RI, $L\dot{I}$ and $H\omega$ to give:

$$L\dot{I} = MI\Omega - RI - H\omega, \tag{2}$$

where L is the self-inductance of the system and R is the series resistance.

Introducing dimensionless variables

$$t = \frac{R}{L}\tau, \quad x = \left(\frac{M}{G}\right)^{1/2} I, \quad \tilde{y} = \frac{M}{R}\Omega, \quad z = \frac{RBM^{1/2}}{LHG^{1/2}}\omega, \tag{3}$$

Equations (1) and (2) become

$$\dot{x} = x(\tilde{y} - 1) - \beta z, \tag{4a}$$
$$\dot{\tilde{y}} = \alpha(1 - x^2) - \kappa\tilde{y}, \tag{4b}$$
$$\dot{z} = x - \lambda z, \tag{4c}$$

where the dot now denotes differentiation with respect to t, and

$$\alpha = \frac{GLM}{R^2A}, \quad \beta = \frac{H^2L}{R^2B}, \quad \kappa = \frac{KL}{RA}, \quad \lambda = \frac{DL}{RB}. \tag{5}$$

The four positive parameters appearing in (5) can be interpreted as follows. α is a measure of the applied couple, κ is a measure of the mechanical friction in the disk, β is a measure of the inverse moment of inertia in the armature of the motor and λ its mechanical friction.

2.2 Linear Stability Analysis

To translate the trivial fixed point to the origin we introduce $y(t) = \overline{\alpha} - \tilde{y}(t)$, so that (4) becomes

$$\dot{x} = (\overline{\alpha} - 1)x - xy - \overline{\beta}\lambda z, \tag{6a}$$
$$\dot{y} = \kappa(\overline{\alpha}x^2 - y), \tag{6b}$$
$$\dot{z} = x - \lambda z, \tag{6c}$$

where $\overline{\beta} = \beta/\lambda$ and $\overline{\alpha} = \alpha/\kappa$. This gives

$$\overline{\alpha} = \frac{GM}{RK}, \quad \overline{\beta} = \frac{H^2}{RD}. \tag{7}$$

While $\overline{\alpha}$ is still a measure of the applied couple, G, $\overline{\beta}$ no longer measures the inverse moment of inertia of the armature.

The system (6) possesses three equilibrium solutions:

$$\mathbf{x_0} = (x_0, y_0, z_0) = (0, 0, 0), \tag{8a}$$
$$\mathbf{x_e} = (x_e, y_e, z_e) = (x_e, \overline{\alpha}x_e^2, x_e/\lambda), \tag{8b}$$

where $x_e = \pm[1 - (1 + \overline{\beta})/\overline{\alpha}]^{1/2}$.

The trivial equilibrium $\mathbf{x_0}$ undergoes a pitchfork bifurcation when

$$\overline{\alpha}_s = 1 + \overline{\beta}, \tag{9}$$

and a supercritical Hopf bifurcation on the line

$$\overline{\alpha}_h = 1 + \lambda. \tag{10}$$

The nontrivial equilibria $\mathbf{x_e}$ undergo subcritical Hopf bifurcations on

$$\overline{\alpha}_H = 1 + \frac{3}{2}\overline{\beta} + \frac{\lambda[2\overline{\beta} - (\kappa + \lambda)]}{2(\kappa - \overline{\beta})}, \tag{11}$$

provided

$$\overline{\alpha} + \lambda/2 > 3\overline{\beta}/2 + 1$$

and $\overline{\beta} \neq \kappa$.

All three equilibria undergo a codimension-two double-zero bifurcation at the point

$$(\overline{\beta}, \overline{\alpha}) = (\lambda, 1 + \lambda). \tag{12}$$

2.3 Parameter Regimes

HSA presented time series and phase portraits for two isolated examples of chaotic behaviour:

$$(\alpha, \beta, \kappa, \lambda) = (20.0, 2.0, 1.0, 1.2), \tag{13a}$$
$$(\alpha, \beta, \kappa, \lambda) = (100.0, 1.01, 1.0, 1.0), \tag{13b}$$

the latter having many more oscillations around the non-trivial equilibria before reversals than the former.

This was rectified in [2] where bifurcation transition curves were computed for these two choices of $(\alpha, \kappa, \lambda)$ as functions of β for the whole range in which chaotic behaviour was to be found to be present (see [2], Figs. 2 and 4). In addition bifurcation transition curves were also presented for other choices of κ and λ.

As mentioned in [2], the choice of $\kappa = \lambda$ is degenerate, since (11) and (12) above show that it corresponds to the double-zero bifurcation for the trivial and nontrivial equilibria, coinciding with the vertical asymptote for the subcritical Hopf bifurcation off the nontrivial equilibria. We therefore follow [2] and perturb away from this degeneracy by selecting the same values for λ and κ as for the $\alpha = 20$ problem.

Our linking number calculations and template analysis will therefore test the dependence on α (a nondimensional measure of the applied couple G), as well as on β.

3 The Lorenz Equations

Because of their relevance to the present investigation, we find it convenient to discuss certain aspects of the Lorenz equations [13]. The Lorenz equations

$$\dot{x} = \sigma(y - x), \tag{14a}$$
$$\dot{y} = rx - y - xz, \tag{14b}$$
$$\dot{z} = -bz + xy, \tag{14c}$$

have three equilibrium solutions

$$(x_0, y_0, z_0) = (0, 0, 0), \quad (x_e, y_e, z_e) = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1). \tag{15}$$

A linear stability analysis shows that, while all three equilibria can undergo steady state bifurcations, only (x_e, y_e, z_e) undergo subcritical Hopf bifurcations and there are no double-zero bifurcations, unlike the HSA system.

Koga [12] introduced the change of variable

$$\hat{z} = z - (r - 1),$$

which translates the z nontrivial fixed point to the origin and transforms (14) to

$$\dot{x} = \sigma(y - x), \tag{16a}$$
$$\dot{y} = -y + x(1 - \hat{z}), \tag{16b}$$
$$\dot{\hat{z}} = -b(\hat{z} + r - 1) + xy, \tag{16c}$$

and the equilibria to

$$(x_0, y_0, \hat{z_0}) = (0, 0, 1 - r), \quad (x_e, y_e, \hat{z_e}) = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, 0). \tag{17}$$

He takes the Poincaré section to be

$$\hat{\mathcal{S}} = [(x, y) : \hat{z} = 0, \dot{\hat{z}} > 0, x > 0], \tag{18}$$

and computes upos on this section, presenting a selection of the orbits, together with a table of their periods for the choice $(r, \sigma, b) = (28, 10, 8/3)$.

4 Numerical Investigations

Moroz [2] presented various bifurcation transition curves for $\alpha = 20$ and $\alpha = 100$. The two examples of chaotic dynamics focused on represented examples chosen from the middle of the chaotic regime ($\alpha = 20$) and from near the loss of stability of chaotic behaviour to steady dynamo action (($\alpha = 100$). A selection of unstable periodic orbits were presented but no attempt was made to identify the template of the underlying attractor nor to see how it was affected by different parameter choices. We rectify that here.

We present results for three instances of the $\alpha = 20$ and one of the $\alpha = 100$ cases. In the former, we choose values of β from both extremes of the bifurcation transition diagram, representing the loss of stability to steady states (low values of β) and loss of stability to a stable periodic state (high values of β). In addition, we return to the two main examples of [2]. Our goals in this chapter are to determine the effects of varying β and of varying α on the linking number calculations and the template identification, albeit for a very small selection of parameter values. A complete study is beyond the scope of the present work.

Since the linking numbers and template analysis for the Lorenz attractor will prove key to our analysis, we include pertitent results for the classic choice of $(r, \sigma, b) = (28, 10, 8/3)$ when appropriate.

Following Koga [12] we translate the variable $y(t)$ by its nontrivial equilibrium state, $Y(t) = 1 + \overline{\beta} - y(t)$ so that (6) becomes:

$$\dot{x} = (\overline{\alpha} - 1 - \overline{\alpha} x_e^2)x - xY - \overline{\beta}\lambda z, \tag{19a}$$

$$\dot{Y} = \kappa(\overline{\alpha} x^2 - Y - \overline{\alpha} x_e^2), \tag{19b}$$

$$\dot{z} = x - \lambda z. \tag{19c}$$

and take the Poincaré section to be

$$\mathcal{S} = \{(x, z) : Y = 0, \dot{Y} > 0, x > 0\}. \tag{20}$$

4.1 The Four Examples

We begin by presenting the (x, Y) phase portraits and the $x(t)$ time series for the four examples of interest. In all of our integrations we fixed $\lambda = 1.2$ and $\kappa = 1$. Figure 1 shows the HSA attractor for $\alpha = 20$ and $\beta = 1.25$, close

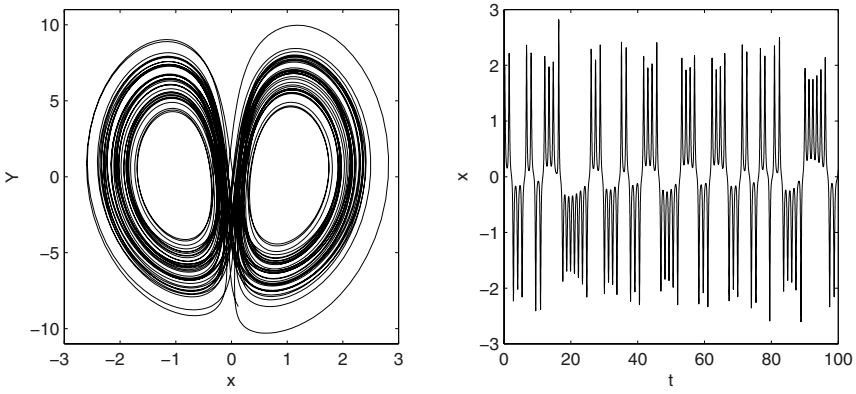


Fig. 1: The phase portrait in the (x, Y)-plane and the time series of $x(t)$ for $\alpha = 20$, $\beta = 1.25$, $\lambda = 1.2$ and $\kappa = 1.0$.

to the loss of stability to steady dynamo behaviour. Figures 2 and 3 show the corresponding plots for $\beta = 2$ (in the middle of the chaotic regime) and $\beta = 2.6$ (close to the loss of chaotic to stable periodic behaviour) respectively. What is evident from these three cases is that the number of oscillations about each of the nontrivial fixed points decreases as β increases. Our fourth example, shown in Fig. 4, is for $\alpha = 100$ and $\beta = 2$ and again represents an example, close to the loss of stability to steady states (see [2]).

We computed the Lyapunov exponents and the Lyapunov dimension for these four cases and compared them with those for the Lorenz equations for their classic parameter choices. We used the Kaplan-Yorke estimate for the

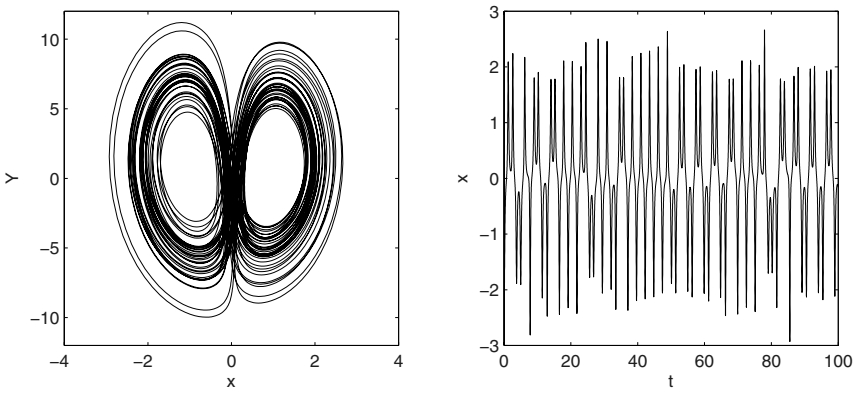


Fig. 2: The phase portrait in the (x, Y)-plane and the time series of $x(t)$ for $\alpha = 20$, $\beta = 2.0$, $\lambda = 1.2$ and $\kappa = 1.0$.

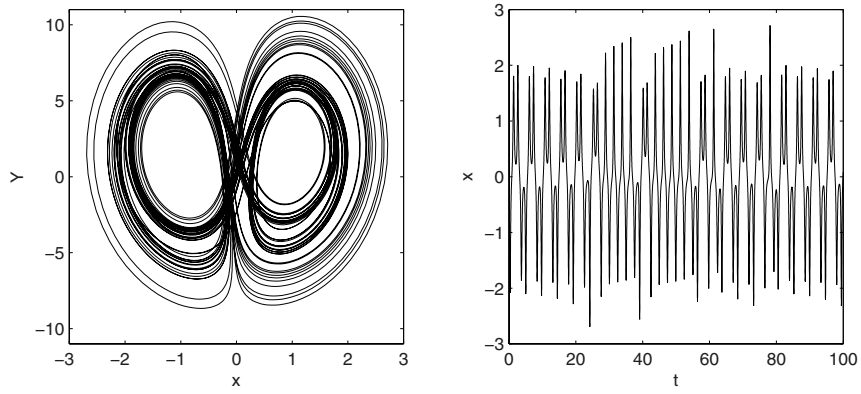


Fig. 3: The phase portrait in the (x, Y)-plane and the time series of $x(t)$ for $\alpha = 20$, $\beta = 2.6$, $\lambda = 1.2$ and $\kappa = 1.0$.

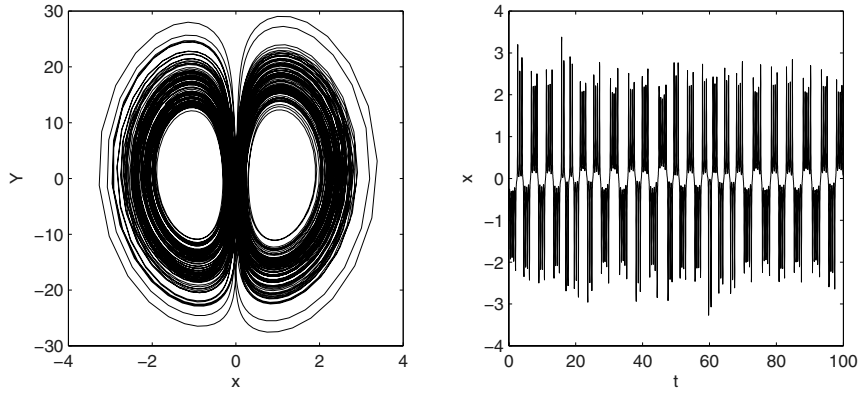


Fig. 4: The phase portrait in the (x, Y)-plane and the time series of $x(t)$ for $\alpha = 100$, $\beta = 2.0$, $\lambda = 1.2$ and $\kappa = 1.0$.

Table 1: Lyapunov exponents and Lyapunov dimension for the four examples of the HSA dynamo considered here.

α	β	Lyapunov Exponents	Lyapunov Dimension
20	1.25	0.4333, 0, −1.0066	2.4305
20	2.0	0.2837, 0, −0.9298	2.3051
20	2.6	0.1850, 0, −0.8639	2.2142
100	2.0	0.4885, 0, −0.8922	2.5475

Lyapunov dimension D_L , where $D_L = 2 + \lambda_1/|\lambda_3|$. The results for the HSA system are shown in Table 1. The magnitudes of the Lyapunov exponents and Lyapunov dimension decrease as β increases. For the Lorenz equations with $(r, \sigma, b) = (28, 10, 8/3)$ we obtained $(0.9076, 0, -14.574)$ with our codes, giving $D_L = 2.0623$. The Lorenz attractor is therefore more strongly contracting than the HSA attractor.

4.2 Unstable Periodic Orbits

Upos were found as close returns on the Poincaré section as follows. The HSA equations were integrated for 60, 000s with a time step of 0.001s. A close return was determined from the condition

$$||\mathbf{Y}_i - \mathbf{Y}_j|| = \sqrt{(x_i - x_j)^2 + (Y_i - Y_j)^2 + (z_i - z_j)^2} < \epsilon,$$

where $\epsilon = 0.005$. Moreover the method of Hénon [14] was used to guarantee that all trajectories landed precisely on the Poincaré section.

Guided by the signs of the nontrivial equilibria, we adopted the protocol of labelling trajectories using symbol sequences R^m if the trajectory cycled m times around the equilibrium $x_e > 0$, and L^n if the trajectory cycled n times around the negative equilibrium $x_e < 0$. Each symbol L or R was taken to correspond to one period. Thus, for example, $R^m L^n$ would be a period-$(m+n)$ orbit, which cycles m times around $x_e > 0$ and n times around $x_e < 0$.

Figure 5 shows a comparison of the histograms of upos for three of the cases for the HSA dynamo with that for the Lorenz equations obtained on their respective Poincaré sections. It is clear from these histograms that the numbers of distinct upos decreases as β increases, and therefore that, for a given choice of parameter values, the HSA dynamo does not contain all possible upos of a given period (unlike the Lorenz equations). This becomes evident in the template analysis below.

Moroz [2] shows a selection of upos for the cases of $(\alpha, \beta) = (20, 2.0)$ and $(\alpha, \beta) = (100, 2.0)$. Here we show a selection of upos for the remaining two cases of $(\alpha, \beta) = (20, 1.25)$ and $(\alpha, \beta) = (20, 2.6)$.

Figure 6 shows examples of the two lowest period upos found for the HSA dynamo when $\alpha = 20$ and $\beta = 1.25$. No examples of the period-2 LR upo was found for this case. Indeed Fig. 1 shows that one would expect upos of higher period to occur. Recall that $\beta = 1.25$ is near to the loss of stability of the oscillatory behaviour to steady dynamo action.

When $\beta = 2.6$, the LR period-2 orbit, with period $2.444s$ predominates (see Fig. 5d). Instead of showing any of the lower order upos, we shall illustrate this case with two different period-10 orbits, which were also used to verify the HSA template. Figure 7 shows examples of a $RL(R^2L^2)^2$ upo with period $12.078s$ and an $(LR)^3L^2R^2$ upo with period $12.378s$.

We can also compare upos found for different values of α. Figure 8 shows the L^3R^3 period-6 orbit for $\alpha = 20$ (upper two panels) compared to that for

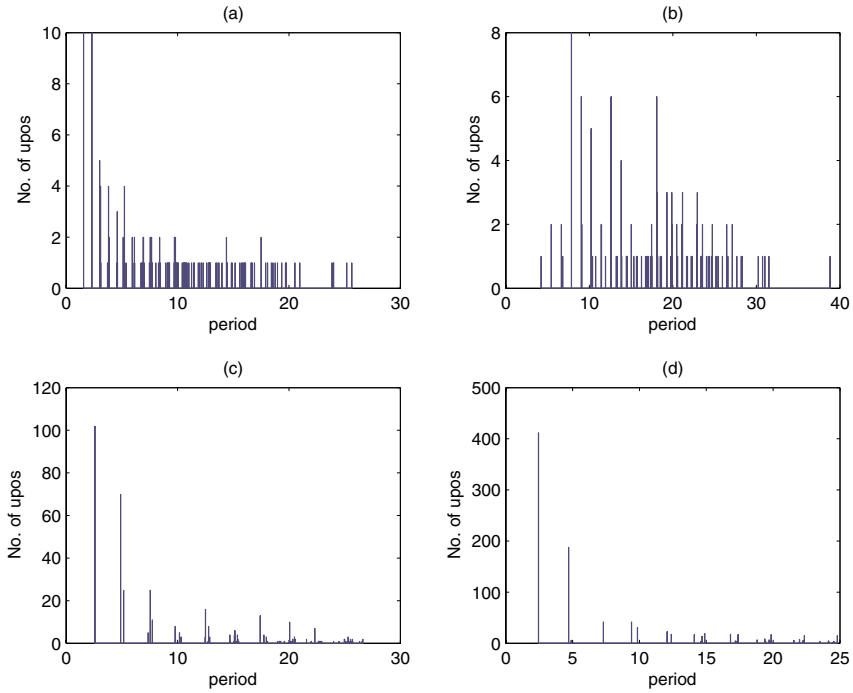


Fig. 5: The histograms of upos for (a) the Lorenz equations for $(r, \sigma, b) = (28.10, 8/3)$, and the HSA dynamo for $\alpha = 20$ and (b) $\beta = 1.25$, (c) $\beta = 2.0$ and (d) $\beta = 2.6$.

$\alpha = 100$ (lower two panels). The former has period $7.353s$, while the latter has period $3.523s$.

5 Template Analysis

For the Lorenz equations at their classic parameter values, examples of upos of all low order periods were found using the method of close returns on the Poincaré section. For the HSA dynamo this is not the case for both of the examples of Fig. 9 in [1], For this reason we calculated the linking numbers for upos of higher periods, in order to verify the validity of our template choice.

The linking number $L(A, B)$ of two upos A and B is defined to be half the number of signed crossings of A and B. It can be computed in two ways: either from a Gauss integral

$$L(A,B) = \frac{1}{4\pi} \oint_A \oint_B \frac{(\mathbf{x_A} - \mathbf{x_B}).(\mathbf{dx_A} \wedge \mathbf{dx_B})}{\|\mathbf{x_A} - \mathbf{x_B}\|^{\mathbf{3}}}, \tag{21}$$

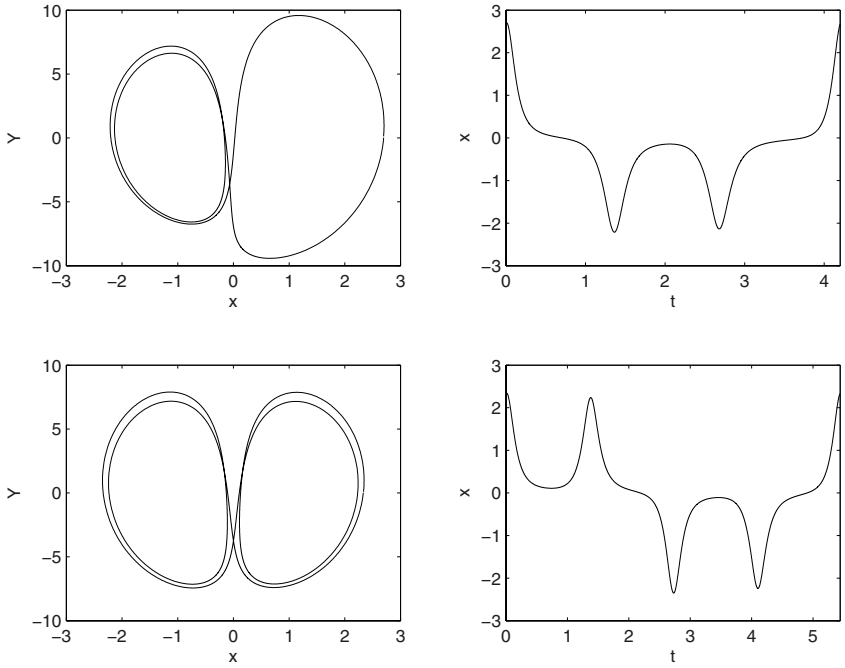


Fig. 6: The (x, Y) phase portraits and $x(t)$ time series of a period-3 L^2R (upper pair) and a period-4 L^2R^2 upo (lower pair) for $\alpha = 20$ and $\beta = 1.25$. The L^2R orbit has a period of $4.207s$, while the L^2R^2 orbit has a period of $5.439s$.

or by projecting the two upos onto a two-dimensional plane and counting half of the number of signed crossings by eye [3]. In our calculations we used both approaches, but primarily a numerical code very kindly supplied by Bob Gilmore to compute the Gauss linking numbers via the integral in (21).

To verify the template and compute the self-linking numbers, we used a second numerical algorithm also supplied by Bob Gilmore. This requires information about the twisting and crossing of branches of the template and is given by the torsion matrix $T(i, j)$. Here $T(i, j)$ is the signed crossings of the ith and jth branches. If $T(i, j) = 0$, the branches do not cross. For the Lorenz system, there are two branches (corresponding to the cycling of the upos around each of two nontrivial equilibrium states) and neither twists, so that $T(i, j) = 0$ for $i, j = 1, 2$. The right hand branch of the template lies in front of the left hand branch, so that the layering information is $(1, -1)$. The HSA system also has two nontrivial equilibria and so two branches, and orbits for the LSA and HSA attractors are describable in terms of two symbols. The algorithm also requires a listing of the upos in terms of period and symbol sequence. The output is a table of linking numbers (including self-linking numbers, whose values cannot be calculated by the Gauss integral (21)).

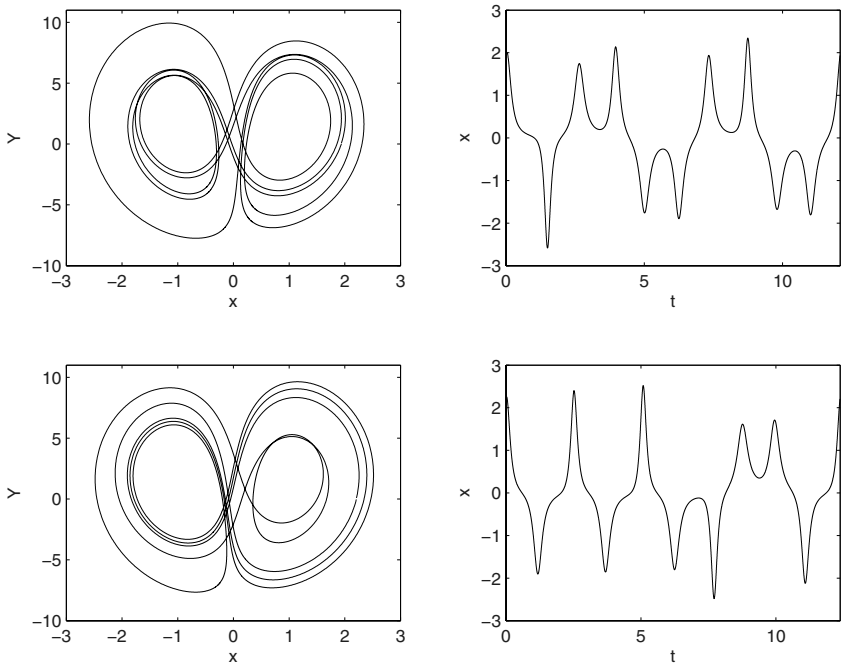


Fig. 7: The (x, Y) phase portraits and $x(t)$ time series of two period-10 upos for $\beta = 2.6$. The upper pair shows a $RL(R^2L^2)^2$ upo with period $12.078s$, while the lower pair shows an $(LR)^3L^2R^2$ upo with period $12.378s$.

6 Identification of the HSA Template

To identify a possible template for the HSA equations, we proceeded as follows. Using the Lorenz equations as a test bed, we calculated linking numbers for all 23 upos in the equations up to, and including, orbits of period 6, together with two period-8 orbits, using the Gauss linking number code. While Gilmore and Letellier [15] have produced tables of linking numbers for the Lorenz equations for all orbits up to period-5, to the best of our knowledge, this is the first time a table for orbits up to period-6 has been computed and presented in the published literature. Because 23 upos are involved, we have found it convenient to include a simplified labelling of the orbits in Table 2, with the linking numbers displayed in Table 3.

The structure of the Lorenz template is already known and the template verification code, referred to in the previous section, was used to compute the full table of linking numbers for all the selected upos, including values for the self-linking numbers. This is shown in Table 3. Linking numbers for distinct upos are found as the off-diagonal entries, while self-linking numbers

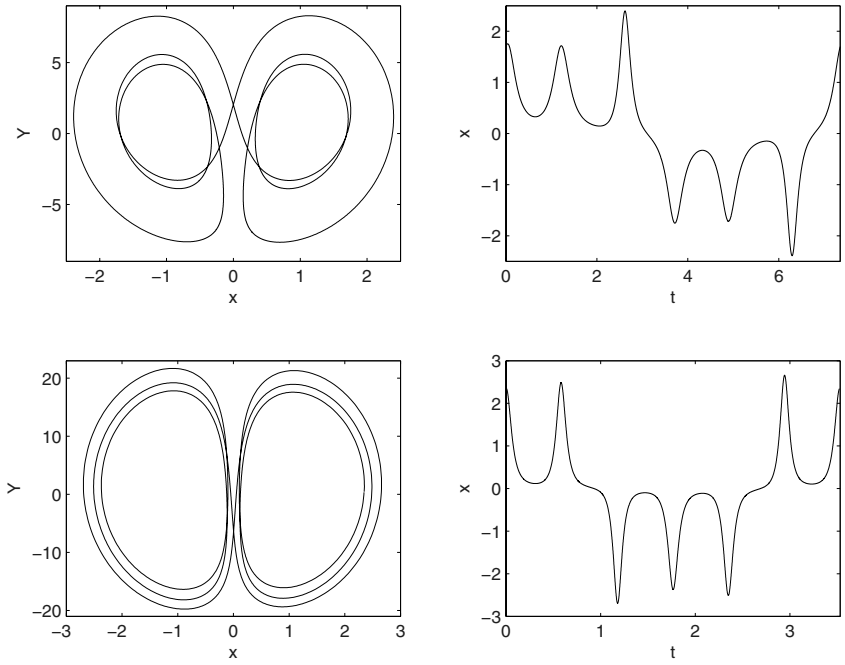


Fig. 8: The (x, Y) phase portraits and $x(t)$ time series of two period-6 L^3R^3 upos for $\beta = 2.0$. The upper pair is for $\alpha = 20$ and has period $7.353s$, while the lower pair is for $\alpha = 100$ with period $3.523s$.

Table 2: Upos of period-6, and two period-8 upos for the Lorenz equations when $(r, \sigma, b) = (28, 10, 8/3)$.

Period	2	3	3	4	4	4
Label	2	3_1	3_2	4_1	4_2	4_3
Periodic Orbit	LR	L^2R	R^2L	L^3R	LR^3	L^2R^2
Period	5	5	5	5	5	5
Label	5_1	5_2	5_3	5_4	5_4	3_6
Periodic Orbit	L^4R	LR^4	L^3R^2	L^2R^3	LRL^2R	RLR^2L
Period	6	6	6	6	6	6
Label	6_1	6_2	6_3	6_4	6_4	6_6
Periodic Orbit	L^5R	LR^5	L^4R^2	L^2R^4	L^3R^3	L^2R^2LR
Period	6	6	6	8	8	
Label	6_7	6_8	6_9	8_1	8_2	
Periodic Orbit	R^2L^2RL	L^3RLR	R^3LRL	$L^2R^2(LR)^2$	$R^2L^2(RL)^2$	

Table 3: Linking numbers for all upos of period-6, and two period-8 upos for the Lorenz equations when $(r, \sigma, b) = (28, 10, 8/3)$. Self-linking numbers are along the diagonal.

orbit	2	3_1	3_2	4_1	4_2	4_3	5_1	5_2	5_3	5_4	5_5	5_6	6_1	6_2	6_3	6_4	6_5	6_6	6_7	6_8	6_9	8_1	8_2
2	**1**	1	1	1	1	2	1	1	2	2	2	2	1	1	2	2	2	3	3	2	2	4	4
3_1	1	**2**	1	2	1	2	2	1	3	2	3	2	2	1	3	2	3	3	3	4	2	4	4
3_2	1	1	**2**	1	2	2	1	2	2	3	2	3	1	2	2	3	3	3	3	2	4	4	4
4_1	1	2	1	**3**	1	2	3	1	3	2	3	2	3	1	4	2	3	3	3	4	2	4	4
4_2	1	1	2	1	**3**	2	1	3	2	3	2	3	1	3	2	4	3	3	3	2	4	4	4
4_3	2	2	2	2	2	**3**	2	2	3	3	4	4	2	2	3	3	4	5	5	4	4	7	7
5_1	1	2	1	3	1	2	**4**	1	3	2	3	2	4	1	4	2	3	3	3	4	2	4	4
5_2	1	1	2	1	3	2	1	**4**	2	3	2	3	1	4	2	4	3	3	3	2	4	4	4
5_3	2	3	2	3	2	3	3	2	**4**	3	5	4	3	2	4	3	4	5	5	5	4	7	7
5_4	2	2	3	2	3	3	2	3	3	**4**	4	5	2	3	3	4	4	5	5	4	5	7	7
5_5	2	3	2	3	2	4	3	2	5	4	**6**	4	3	2	5	4	5	6	6	6	4	8	8
5_6	2	2	3	2	3	4	2	3	4	5	4	**6**	2	3	4	5	5	6	6	4	6	8	8
6_1	1	2	1	3	1	2	4	1	3	2	3	2	**5**	1	4	2	3	3	3	4	2	4	4
6_2	1	1	2	1	3	2	1	4	2	3	2	3	1	**5**	2	4	3	3	3	2	4	4	4
6_3	2	3	2	4	2	3	4	2	4	3	5	4	4	2	**5**	3	4	5	5	6	4	7	7
6_4	2	2	3	2	4	3	2	4	3	4	4	5	2	4	3	**5**	4	5	5	4	6	7	7
6_5	2	3	3	3	3	4	3	3	4	4	5	5	3	3	4	4	**5**	6	6	5	5	8	8
6_6	3	3	3	3	3	5	3	3	5	5	6	6	3	3	5	5	6	**7**	8	6	6	10	11
6_7	3	3	3	3	3	5	3	3	5	5	6	6	3	3	5	5	6	8	**7**	6	6	11	10
6_8	2	4	2	4	2	4	4	2	5	4	6	4	4	2	6	4	5	6	6	**7**	4	8	8
6_9	2	2	4	2	4	4	2	4	4	5	4	6	2	4	4	6	5	6	6	4	**7**	8	8
8_1	4	4	4	4	4	7	4	4	7	7	8	8	4	4	7	7	8	10	11	8	8	**13**	15
8_2	4	4	4	4	4	7	4	4	7	7	8	8	4	4	7	7	8	11	10	8	8	15	**13**

of each upo are shown in bold type along the diagonal. Note the reflectional symmetry about this main diagonal.

We then used the Gauss linking number code to compute linking numbers for upos extracted from the HSA dynamo for the cases of $\alpha = 20$ and $\alpha = 100$ under consideration in this study.

Tables 4–6 show the linking numbers for the upos we used for $\alpha = 20$ and $\beta = 1.25$, $\beta = 2$ and $\beta = 2.6$ respectively, while Table 7 shows those for $\alpha = 100$ and $\beta = 2$. The self-linking numbers are again along the diagonal.

As is evident from the histograms in Fig. 5, unlike the Lorenz system, not all low period upos were found in the HSA system in our time series for the parameter values we considered. We were therefore compelled to use much higher order orbits to verify the template than is usual. That using such high order upos produced no discrepancies in the template verification code, only served to underline the validity of our calculations and guess.

In each of the four examples considered in this chapter, the linking numbers were found to be compatible with the template for the Lorenz equations, and so the same layering of the two torsion-free branches.

Table 4: Linking numbers for some of the upos found in the HSA dynamo for $\alpha = 20$, $\beta = 1.25$, $\lambda = 1.2$, $\kappa = 1$. Self-linking numbers are along the diagonal.

Period	Period Orbits	3 L^2R	3 LR^2	4 L^2R^2	5 LR^4	5 L^4R	6 L^3R^3	7 L^4R^3	7 L^3R^4
3	L^2R	**2**	1	2	1	2	3	3	3
3	LR^2	1	**2**	2	2	1	3	3	3
4	L^2R^2	2	2	**3**	2	2	4	4	4
5	LR^4	1	2	2	**4**	1	3	3	4
5	L^4R	2	1	2	1	**4**	3	4	3
6	L^3R^3	3	3	4	3	3	**5**	5	5
7	L^4R^3	3	3	4	3	4	5	**6**	5
7	L^3R^4	3	3	4	4	3	5	5	**6**

Table 5: Linking numbers for some of the upos found in the HSA dynamo for $\alpha = 20$, $\beta = 2$, $\lambda = 1.2$, $\kappa = 1$. Self-linking numbers are along the diagonal.

Period	Period Orbits	2 LR	4 R^2L^2	6 R^3L^3	6 RLR^2L^2	6 LRL^2R^2	8 $(LR)^2L^2R^2$	8 $(RL)^2R^2L^2$
2	LR	**1**	2	2	3	3	4	4
4	R^2L^2	2	**3**	4	5	5	7	7
6	R^3L^3	2	4	**5**	6	6	8	8
6	RLR^2L^2	3	5	6	**7**	8	11	10
6	LRL^2R^2	3	5	6	8	**7**	10	11
8	$(LR)^2L^2R^2$	4	7	8	11	10	**13**	15
8	$(RL)^2R^2L^2$	4	7	8	10	11	15	**13**

Table 6: Linking numbers for some of the upos found in the HSA dynamo for $\alpha = 20$, $\beta = 2.6$, $\lambda = 1.2$, $\kappa = 1$. Self-linking numbers are along the diagonal.

Period	Period Orbits	2 LR	4 L^2R^2	6 L^2RLR^2	6 R^2LRL^2	8 $L^2R^2(LR)^2$	8 $R^2L^2(RL)^2$	10 $RL(R^2L^2)^2$
2	LR	**1**	2	3	3	4	4	5
4	L^2R^2	2	**3**	5	5	7	7	8
6	L^2RLR^2	3	5	**7**	8	11	10	12
6	R^2LRL^2	3	5	8	**7**	10	11	13
8	$L^2R^2(LR)^2$	4	7	11	10	**13**	15	18
8	$R^2L^2(RL)^2$	4	7	10	11	15	**13**	17
10	$RL(R^2L^2)^2$	5	8	12	13	18	17	**19**

Table 7: Linking numbers for some of the upos found in the HSA dynamo for $\alpha = 100$, $\beta = 2$, $\lambda = 1.2$, $\kappa = 1$. Self-linking numbers are along the diagonal.

Period	Period Orbits	2 LR	6 R^3L^3	6 RLR^2L^2	6 LRL^2R^2	7 L^4R^3	7 L^3R^4	8 L^4R^4	8 $(LR)^2L^2R^2$
2	LR	**1**	2	3	3	2	2	2	4
6	R^3L^3	2	**5**	6	6	5	5	6	8
6	RLR^2L^2	3	6	**7**	8	6	6	6	11
6	LRL^2R^2	3	6	8	**7**	6	6	6	10
7	L^4R^3	2	5	6	6	**6**	5	6	8
7	R^4L^3	2	5	6	6	5	**6**	6	8
8	L^4R^4	2	6	6	6	6	6	**7**	8
8	$(LR)^2L^2R^2$	4	8	11	10	8	8	8	**13**

7 Discussion

In this chapter we have completed the study, begun in [2] into the classification of chaos in the HSA dynamo using unstable periodic orbits. Moroz [2] selected two examples of chaotic behaviour and extracted upos from time series for two values of α, a dimensionless parameter which measures the couple applied to the disk of the dynamo, driving it into rotation.

Bifurcation transition curves produced in [2] indicated that for $\alpha = 20$, the chaotic behaviour occurred for a value of β, mid-way between the transition to steady states (low β) and the loss of stability to stable periodic motion (high β). For $\alpha = 100$, the chosen example was very close to the transition to steady states. No attempt was made by [2] to identify the underlying attractor using topological methods.

Moreover, as well as considering two different values of α, in this chapter, we have also investigated the effects of varying β, choosing two cases which lie close to either the transition to steady or to periodic states. The histograms shown in Fig. 5 indicate that as β increases, the number of distinct upos decreases. Also, for the parameter values investigated, the Lorenz system contains more distinct upos than does the HSA system.

In the present study, we have applied topological ideas, expounded by Gilmore [3], to compute linking numbers for the upos and, thereby, we have identified a possible template on which the orbits lie. This template turns out to be topologically identical to that for the Lorenz equations.

As well as for the HSA system, we have also performed similar calculations of upos and linking numbers for another dynamo model, the Extended Malkus-Robbins dynamo [10, 11], a four-dimensional nonlinear system which reduces to the Lorenz equations when $\beta = 0$. For certain choices of parameter values, the Extended Malkus-Robbins equations are strongly contracting, with a Lyapunov dimension $D_L < 3$. Since the system becomes effectively three-dimensional, the topological methods described here are applicable. Moroz et al. [16] were able to show that the branched manifolds describing the projected

attractors for small and large values of β were reflectionally-symmetric, of Lorenz type with rotation symmetry.

The HSA dynamo is the first of a class of low order dynamo models to have its chaotic behaviour classified using topological methods [7]. Other such models include effects such as an external magnetic field, a battery term, the coupling of two or more dynamo units together, an azimuthal eddy current ([5, 6, 8]). For such and other systems whose Lyapunov dimension $D_L < 3$, the topological approach presented here has the potential to classify the underlying chaotic attractor, and to distinguish between different chaotic attractors.

References

1. R. Hide, A.C. Skeldon and D.J. Acheson: A study of two novel self-exciting single-disk homopolar dynamos:theory. Proc. R. Soc. Lond. A (1996) **452**: 1369-1395
2. I.M. Moroz: The Hide, Skeldon, Acheson dynamo revisited. Proc. R. Soc. Lond. A (2007) **463**: 113-130
3. R. Gilmore: Topological analysis of chaotic dynamical systems. Rev. Mod. Phys. (1998) **70**: 1455-1530
4. R. Hide: The nonlinear differential equations governing a hierarchy of self-exciting coupled Faraday-disk homopolar dynamos. Phys. Earth Plan. Int. (1997) **103**: 281-291
5. I.M. Moroz, R. Hide and A.M. Soward: On self-exciting coupled Faraday disk homopolar dynamos driving series motors. Physica D (1998) **117**: 128-144
6. R. Hide and I.M. Moroz: Effects due to azimuthal eddy currents in the Faraday disk self-exciting homopolar dynamo with series motor: I. Two special cases. Physica D (1999) **134**: 287-301
7. I.M. Moroz: Behaviour of a self-exciting Faraday-disk homopolar dynamo with battery in the presence of an external magnetic field. Int. J. Bif. Chaos (2001a) **11**: 1695-1705
8. I.M. Moroz: Self-exciting Faraday disk homopolar dynamos. Int. J. Bif. Chaos (2001b) **11**: 2961-2975
9. I.M. Moroz: On the behaviour of a self-exciting Faraday disk homopolar dynamo with a variable nonlinear series motor. Int. J. Bif. Chaos (2002) **12**: 2123-2135
10. I.M. Moroz: The extended Malkus-Robbins dynamo as a perturbed Lorenz system. Nonlinear Dynamics (2005) **41**:191-210
11. I.M. Moroz: The Malkus-Robbins dynamo with a linear series motor. Int. J. Bif. Chaos (2003) **13**: 147-161
12. S. Koga: Phase Description Method to Time Averages in the Lorenz System. Prog. Theor. Phys. (1986) **76**: 335-355
13. E.N. Lorenz: Deterministic non-periodic flows. J. Atmos. Sci. (1963) **20**: 130-141
14. M. Hénon: On the numerical computation of Poincaré maps. Physica D (1982) **5**: 412-414
15. R. Gilmore and C. Letellier: *The Symmetry of Chaos*, 1st edn (OUP 2007)
16. I.M. Moroz, C. Letellier and R. Gilmore: When are projections also embeddings? Phys. Rev. E (2007) **75**: Article no. 046201

Methods to Detect Solitons in Geophysical Signals: The Case of the Derivative Nonlinear Schrödinger Equation

Nikolay G. Mazur[1], Viacheslav A. Pilipenko[2], and Karl-Heinz Glassmeier[3]

[1] Institute of the Physics of the Earth, Russian Academy of Sciences, Moscow, Russia n.g.mazur@mtu-net.ru
[2] Space Research Institute, Russian Academy of Sciences, Moscow, Russia pilipenko_va@mail.ru
[3] Institut für Geophysik und extraterrestrische Physik, Technische Universität Braunschweig, Germany kh.glassmeier@tu-bs.de

Abstract. Methods to detect solitons and determine their parameters are considered. The first simple observational test for soliton identification is based on the determination of statistical relationships between amplitude, duration, and carrying frequency of the detected signals, and their comparison with the relevant relationships from the soliton theory. The second method is based on the solution of the direct scattering problem for the relevant nonlinear equations. As an example the Derivative Nonlinear Schrödinger (DNLS) equation has been considered. The integral reflection coefficient, which should rapidly drop when a signal is close to the N-soliton profile, has been used as a soliton detector. Application of this technique to numerically simulated signals shows that it is more efficient than the standard Fourier transform and can be used as a practical tool for the analysis of outputs from nonlinear systems.

1 Introduction: Solitons in Geophysical Media

Nonlinear waves and solitons are frequently observed in all geophysical media: the solar corona [1], interplanetary space [2], Earth's magnetosphere [3, 4], topside ionosphere [5], atmosphere [6, 7], and Earth's crust [8]. In a nonlinear medium a disturbance with finite amplitude commonly evolves to the soliton state [9, 10]. The modern theory predicts and has mathematical tools to describe N-soliton structures and soliton turbulence gas [11, 12]. The detection of the soliton component and determination of its properties demands elaboration of special nonlinear methods of signal analysis. Standard methods of spectral analysis based on the Fourier Transform (FT) fit well the detection of linear waves and determination of their properties, but they are not very effective for the examination of highly structured space plasma turbulence.

The simplest approach is based on the determination of the statistical relationships between amplitudes, duration, velocity, etc. of the observed signal ensemble. Then, the comparison with the theoretically predicted relationships for a given soliton class may be used as a simple observational test for its identification [3]. In this paper, we provide the necessary basic relationships for some types of solitons, and indicate the validity and limitations of these relationships.

However, the above simple statistical method of soliton identification requires an analysis of substantial number of signals desirably under the same external conditions. Thus, the approach that can be applied to a case study is highly desirable.

The idea of "nonlinear Fourier analysis" of observational time series, based on the numerical solution of the direct scattering transform (ST) associated with Korteveg–de Vries equation, was suggested and implemented by Osborne et al. [13, 14]. This approach was applied to the analysis of ocean surface waves. Later on Hada et al. [15] suggested to apply the ST associated with the Derivative Nonlinear Schrödinger equation to a complex time series.

The approach of [15] is further developed in this paper. We have built an effective numerical algorithm to implement the ST. Below we give a short description of this algorithm, comprising calculations of discrete data of the scattering problem (otherwise, soliton parameters), and apply this technique to numerically simulated signals.

2 Statistical Method for Soliton Detection

A linear wave packet in a dispersive medium decays upon propagation due to the packet spreading. If dissipation is weak, the energy conservation law predicts the following relationship between the observed amplitude A and duration T as follows $A^2T \simeq$ const. However, in any realistic geophysical system the amplitude of generated wave packets may vary in a wide range, and this relationship has no practical sense. Contrary to linear signals, the soliton amplitude A is not a free parameter, but it is intrinsically related to other signal parameters, such as duration T, nonlinear component of velocity V, carrying frequency ω, etc. The statistical relationships between them may be used as a simple observational test for soliton identification [3].

To apply adequately this method, the basic relationships, as well as their limitations, are to be taken into account. From many evolutionary equations we consider the following two main model equations.

2.1 Nonlinear Schrödinger Equation (NLS)

The NLS equation

$$i\partial_t\psi + \lambda\partial_{xx}\psi = \nu|\psi|^2\psi \tag{1}$$

is quite universal. It is commonly used to describe weakly nonlinear and dispersive wave packets. The medium nonlinearity produces higher harmonics which results in the slow (on the carrying wave scale) spatial-temporal change of the packet envelope.

In the case of modulation instability, according to the Lighthill condition $\lambda\nu < 0$, the wave breaks with time into isolated wave packets. The envelopes of these packets are the solitons of the Eq. (1) (e.g., [16]):

$$\psi = a\sqrt{-2\lambda/\nu}\,\mathrm{sech}\left[a(x-2bt)\right]\exp\left[ibx + i\lambda(a^2-b^2)t\right].$$

This expression shows that the soliton amplitude $A = a\sqrt{-2\lambda/\nu}$ and its spatial scale $L = a^{-1}$ are not independent variables, as for linear waves, but they are coupled by the relationship $(AL)^2 = -2\lambda\nu^{-1}$. However, usually the observed parameter is not the spatial scale L, but the soliton duration $T = L/V$, where V is the soliton velocity relative to the registration site. Thus

$$(AT)^2 = -2\lambda\nu^{-1}V^{-2}. \tag{2}$$

The relationship (2) can be used as a criterion for soliton identification. When the background medium is motionless, the velocity V equals the group velocity, $V \simeq V_g = 2b$. In the case of fast medium flow relative to a detector with the velocity U which much exceeds V_g (e.g., solar wind), one should suppose that $V \simeq U$.

The right-hand part of Eq. (2) depends in general case on the carrying wave frequency ω. However, the dependence (2) of the product AT on ω for various waves, though described by the same NLS equation, is not universal, but depends on the medium property, namely, on the wave dispersion law and nonlinearity coefficient. For example, for the quasi-longitudinal magnetohydrodynamic wave propagation in a plasma $AT \propto \omega^{-1/2}$, whereas for the quasi-perpendicular fast magnetosonic wave propagation $AT \propto \omega^{-1}$ [3]. Further, for waves at the surface of deep water [16] $AT = g\omega^{-3}$, where g is the gravitational acceleration.

2.2 Derivative Nonlinear Schrödinger Equation

The DNLS equation

$$\partial_{t'}b + i\delta\partial_{x'x'}b + \alpha\partial_{x'}(b|b|^2) = 0 \tag{3}$$

is not so universal as the NLS equation. It was used foremost for the description of weakly nonlinear dispersive Alfvén waves in the case of quasiparallel propagation [17]. Moreover, the DNLS equation was applied to describe Alfven solitary structures in magnetized dusty plasmas [18]. The elliptically polarized field of the Alfvén wave can be presented in the complex form as

$b(x,t) = b_y + ib_z$. The coefficients of the Eq. (3) for the dispersive Alfvén wave are as follows

$$\alpha = \frac{V_A}{4(1-\beta)}, \quad \delta = \frac{V_A^2}{2\Omega_i}\,, \tag{4}$$

where V_A is the Alfvén velocity, $\beta = V_s^2/V_A^2$, V_s is the sound velocity, Ω_i is the ion gyrofrequency. The parameter δ is related to the ion inertia dispersion length: $2\delta/V_A = V_A/\Omega_i$.

The variables x' and t' are related to the original physical variables, coordinate x and time t, in the laboratory coordinate system, by the relationships

$$x' = \varepsilon^2(x - V_A t), \quad t' = \varepsilon^4 t, \tag{5}$$

where $\varepsilon = A \equiv \max|\mathbf{B}|/B_0$ is the small amplitude of the magnetic field disturbance. This change of variables is used upon derivation of Eq. (3) by the perturbation technique [19].

In order to avoid additional formula complications we reduce the DNLS equation to the following normalized form

$$\partial_\tau \psi + \partial_\xi(\psi|\psi|^2) + i\partial_{\xi\xi}\psi = 0. \tag{6}$$

This reduction uses the following change of scales:

$$\tau = 2\Omega_i t', \quad \xi = 2\Omega_i V_A^{-1} x', \quad \psi = \frac{1}{2}(1-\beta)^{-1/2} b. \tag{7}$$

We consider the case of rapid decrease of solution of the Eq. (6) at $|x| \to \infty$. One-soliton solution of (6) for this case has the form [15]

$$\psi_{sol}(\xi,\tau;\lambda) = a(X;\lambda)e^{i\theta}, \tag{8}$$

where

$$\begin{gathered} a(X;\lambda)^2 = \frac{8\lambda_i^2}{|\lambda|\cosh(4\lambda_i X) - \lambda_r}, \quad X = \xi - 4\lambda_r\tau, \\ \theta = -2\lambda_r\xi - 4\left(\lambda_i^2 - \lambda_r^2\right)\tau + \tilde{\theta}, \quad \tilde{\theta}(X;\lambda) = 3\arctan\left[\frac{\lambda_i \tanh(2\lambda_i X)}{|\lambda| - \lambda_r}\right]. \end{gathered} \tag{9}$$

This solution depends on the complex parameter $\lambda = \lambda_r + i\lambda_i$ ($\lambda_i > 0$). It is determined to within an initial soliton location at the initial moment, and the initial phase.

One may see from the above formulas that the structure of DNLS solitons is more complicated than the structure of NLS solitons. A feature of the DNLS solitons is the nonuniform variation of phase: θ is not a linear function of either x, or t. Hence, the frequency and wavenumber of the soliton packet may be determined only locally as $\omega = -\partial\theta/\partial t$ and $k = \partial\theta/\partial x$.

The physical parameters of the DNLS solitons and the relationships between them can be found by reverting to the physical variables x and t with

the help of the coordinate transforms (5) and (7). As a result, the following expressions for the variables in (8) and (9) are obtained:

$$X(x,t) = \frac{2\Omega_i}{V_A}\varepsilon^2 \left[x - V_A\left(1 + 4\varepsilon^2\lambda_r\right)t\right], \tag{10}$$

$$\theta(x,t) = -\frac{4\lambda_r\Omega_i}{V_A}\varepsilon^2\left[x - V_A\left(1 + 2\varepsilon^2\frac{\lambda_r^2 - \lambda_i^2}{\lambda_r}\right)t\right] + \tilde{\theta}(X;\lambda), \tag{11}$$

where ε is the soliton amplitude. The magnetic field of the soliton is expressed via ψ_{sol} as follows

$$B_{sol}(x,t) = \varepsilon b_{sol}(x,t) = 2\varepsilon B_0\sqrt{1-\beta}\psi_{sol}(x,t;\lambda). \tag{12}$$

Here ψ_{sol} has been normalized by the condition

$$4(1-\beta)\max|\psi_{sol}|^2 = 1, \quad \text{i.e.} \quad 32(1-\beta)\left(|\lambda| + \lambda_r\right) = 1. \tag{13}$$

From (10) and (11) the group velocity of the soliton packet follows

$$V = V_A\left(1 + 4\varepsilon^2\lambda_r\right). \tag{14}$$

The spatial scale and time duration are

$$L = V_A(8\varepsilon^2\Omega_i\lambda_i)^{-1}, \quad T = L/V_A = (8\varepsilon^2\Omega_i\lambda_i)^{-1}. \tag{15}$$

The local frequency is determined by

$$\omega = -\partial_t\theta = \omega_\infty + 2\varepsilon^2\Omega_i\partial_X\tilde{\theta} = \omega_\infty + \frac{12\varepsilon^2\Omega_i h}{1 + (1 + h^2\lambda_i^{-2})\sinh^2(2\lambda_i X)}, \tag{16}$$

where $h = [32(1-\beta)]^{-1}$, and $\omega_\infty = -4\varepsilon^2\lambda_r\Omega_i$. The maximum of local frequency $\omega_{\max} = \omega_\infty + 12\varepsilon^2\Omega_i h > 0$ is reached in the center of the soliton $X = 0$. In the above formulas for L, T, and ω the small correction terms of higher order in amplitude have been omitted for breivity.

The specific for DNLS solitons nonuniform phase variation means that the local frequency within the soliton packet grows from the limiting value ω_∞ to the maximum ω_{max} and returns back to ω_∞. Under $\lambda_r > 0$ the limit ω_∞ and maximum ω_{max} have opposite signs, that is the phase changes nonmonotonically.

It is necessary to mention that the physical values V, L, T, ω_∞, and ω_{max} are determined by two independent real parameters only, because owing to the normalization (13) the real and imaginary parts of the parameter λ are coupled as follows: $\lambda_i^2 = h^2 - 2h\lambda_r$.

The DNLS solitons can be visualized as a wave packet envelope of some high-frequency signal only in the limiting case $\lambda_i \to \infty$. Really, in this case $\omega_{\max} \approx \omega_\infty$, because $\lambda_r \approx -(2h)^{-1}\lambda_i^2 \to -\infty$, so $(\omega_{\max} - \omega_\infty)\omega_\infty^{-1} =$

$-3h\lambda_r^{-1} \to 0$. The frequency (16) practically does not depend on x or t, $\omega = 4\varepsilon^2\Omega_i|\lambda_r|$. We find from (15)

$$(\varepsilon T)^2 = (64\varepsilon^2\Omega_i^2\lambda_i^2)^{-1} = (128h\varepsilon^2\Omega_i^2|\lambda_r|)^{-1} = (32h\Omega_i\omega)^{-1},$$

i.e. the relationship between $\varepsilon = A$, T, and ω:

$$(AT)^2 = (1-\beta)(\Omega_i\omega)^{-1}.$$

The above relationship in the quasi-monochromatic limit $\omega T \to \infty$ is simple enough for practical use. However, the above simple method of soliton identification requires an analysis of a statistically significant number of signals under the same external conditions. An alternative method described below can be applied to a single event.

3 Integral Reflection Coefficient as a Soliton Detector

We demonstrate the proposed method using as an example DNLS solitons. For consistency with [20] we transfer from variables in the normalized equation (6) as follows: $\psi(\xi,\tau) \to b(x,t)$. The exact solution of the DNLS equation (6) may be reduced to the solution of a linear problem with a well elaborated algorithm. This algorithm is based on the solution of the direct and inverse scattering problems for the auxiliary linear system [20]:

$$\begin{aligned} \partial_x v_1 &= -i\lambda v_1 + \sqrt{\lambda} b v_2 \ , \\ \partial_x v_2 &= \ i\lambda v_2 + \sqrt{\lambda} b^* v_1 \ , \end{aligned} \tag{17}$$

where λ is the spectral parameter. If the function $b(x,t)$ evolves in time according to the DNLS equation (6), then the functions v_1, v_2 satisfy also another linear system:

$$\begin{aligned} \partial_t v_1 &= Av_1 + Bv_2 \ , \\ \partial_t v_2 &= Cv_1 - Av_2 \ , \end{aligned} \tag{18}$$

where $B = -\sqrt{\lambda}\left(2\lambda b + b^* b^2 + i\partial_x b\right)$, $C = -\sqrt{\lambda}\left(2\lambda b^* + (b^*)^2 b - i\partial_x b^*\right)$, and $A = i\left(2\lambda^2 + \lambda b^* b\right)$. In other words, the compatibility condition for these linear systems for all values of λ is just the DNLS equation.

Any solution of the system (17) can be decomposed over either of two function bases $\varphi, \bar{\varphi}$ or $\psi, \bar{\psi}$ (Jost functions). These function bases are characterized by their asymptotic behavior at $x \to -\infty$

$$\varphi \sim \begin{pmatrix} 1 \\ 0 \end{pmatrix} \exp(-i\lambda x), \quad \bar{\varphi} \sim \begin{pmatrix} 0 \\ -1 \end{pmatrix} \exp(i\lambda x), \tag{19}$$

or at $x \to +\infty$

$$\psi \sim \begin{pmatrix} 0 \\ 1 \end{pmatrix} \exp(i\lambda x), \quad \bar{\psi} \sim \begin{pmatrix} 1 \\ 0 \end{pmatrix} \exp(-i\lambda x). \tag{20}$$

The scattering coefficients $s_1(\lambda)$ and $s_2(\lambda)$ are determined by the decomposition (similar to the combination of incident and reflected waves)

$$\varphi = s_1\bar{\psi} + s_2\psi\,. \tag{21}$$

The ratio $r(\lambda) = s_2/s_1$ for real λ is the reflection coefficient. Zeros of the function $s_1(\lambda)$ in the upper half-plane of the complex variable λ are the discrete eigenvalues of the spectral problem (17); for such $\lambda = \lambda_n$ in view of (21) we have

$$\varphi(x;\lambda_n) = s_2(\lambda_n)\psi(x;\lambda_n)\,.$$

From above it follows, with account for (19) and (20), that function $\varphi(x;\lambda_n)$ decays exponentially at both sides, which in quantum mechanics corresponds to a "bound state ".

The function of real argument $r(\lambda)$, the eigenvalues λ_n, and the normalization coefficients $c_n = i\lambda_n^{-1/2}s_2(\lambda_n)/s_1'(\lambda_n)$ constitute the scattering dataset. If $b(x,t)$ from (17) evolves in accordance with (6), then the scattering data owing to (18) vary in a simple way as:

$$r(\lambda,t) = r(\lambda,0)\exp(4i\lambda^2 t), \qquad \lambda_n(t) = \lambda_n(0), \qquad c_n(t) = c_n(0)\exp(4i\lambda_n^2 t).$$

This allows to find $b(x,t)$ by solving first the direct scattering problem for the given initial condition $b(x,0)$, and, second, the inverse scattering problem for any t.

The observations provide the complex function $b(x)$ as a time series $x_0, x_1, \ldots, x_M$ in $M+1$ points. The onset and end of the examined interval, x_0 and x_M are to be chosen in such a way to make $|b(x)|$ sufficiently small, e.g., less than 10^{-4}. For the numerical solution of the direct scattering problem it is necessary to find the coefficients $s_1(\lambda)$ and $s_2(\lambda)$ for real λ, and to find all roots of the equation $s_1(\lambda) = 0$ in the upper half-plane. In accord with the definition of the coefficients (21), the initial condition for the differential equation system (17) at $x = x_0$ is to be taken as $\varphi_1(x_0;\lambda) = \exp(-i\lambda x_0)$, and $\varphi_2(x_0;\lambda) = 0$ (see (19)). Then, the Eq. (17) are numerically integrated from x_0 to x_M. After that, $s_1(\lambda)$ and $s_2(\lambda)$ can be found with the use of (20) and (21):

$$s_1(\lambda) = \varphi_1(x_M;\lambda)\exp(i\lambda x_M), \qquad s_2(\lambda) = \varphi_2(x_M;\lambda)\exp(-i\lambda x_M)\,.$$

Finally, for each real λ the reflection coefficient $r(\lambda) = s_2(\lambda)/s_1(\lambda)$ is calculated.

The method of calculation relies upon the fact that the discrete eigenvalues are zeros of the function $s_1(\lambda)$ of complex spectral parameter. The following two-stage algorithm for the effective numerical calculations has been elaborated. At the first stage, the values $\mathrm{Re}\, s_1(\lambda)$ and $\mathrm{Im}\, s_1(\lambda)$ on the real axis are to be calculated and the points where these values are equal to zero are to be found. Then, starting from the points obtained, the contours of zero

level $\mathrm{Re}\, s_1(\lambda) = 0$ and $\mathrm{Im}\, s_1(\lambda) = 0$ are calculated in the upper half-plane. Calculations continue until those contours intersect, as illustrated in Fig. 4.

As a detector of DNLS solitons the integral reflection coefficient R is used, that drastically decreases (theoretically to zero) when a signal under analysis is close to the N-soliton profile $b_N(x,t)$. This integral reflection coefficient is introduced with the formula

$$R = \int_{-\infty}^{\infty} |r(\lambda)|\, d\lambda.$$

Any N-soliton spatial profile is a reflectionless potential and for such a profile this value R is exactly zero, whereas it is positive for any other distribution $b(x)$. Any changes in the spatial scale (that is changes of sampling step δ over x in a numerical values of the profile under study) result in a change of the quantity R. For an exact N-soliton profile these variations of scale yield a sharp minimum of the function $R(\delta)$ for a correct step δ. Thus, the integral reflection coefficient is very sensitive to the change of the linear scale. Therefore, if the analysis of an experimentally detected spatial profile with the use of the scale variation method gives a dependence $R(\delta)$ with an evident minimum, this may imply an occurrence of a substantial soliton component in this disturbance. As a by-product of this method, a correct value of the step $\delta = \delta_s$ is determined. Using the determined value of δ one can calculate the discrete data of the scattering problem and retrieve a pure soliton part of the disturbance under study, with the help of known formulas for the N-soliton solution.

Let us consider the situation when the complex function $b(x) = b_y(x) + ib_z(x)$ is measured in a fixed observation site $x^{(0)}$, whereas a wave disturbance $b(x - Vt)$ propagates along it with unknown velocity V. It is supposed that soliton has been formed outside the observational region and does not undergo nonlinear deviations upon propagation through this region. Thus, we have in the point $x^{(0)}$ the time series $b^{(0)}(t_j) = b(x^{(0)} - Vt_j)$, $j = 0, 1, \ldots, M$, with time sampling rate $t_{j+1} - t_j = \Delta t$. Under unknown velocity V the spatial structure $b(x)$ can be determined only disregarding the spatial step $\delta = V\Delta t$. Therefore, the method in lieu of time series $b^{(0)}(t_j)$ analyzes the spatial profile $b(x;\delta)$, determined in the points $x_j = x_0 + j\delta$, with unknown step δ. The calculated integral reflection coefficient R as a function of δ tends to zero under certain $\delta = \delta_s$, if the wave $b(x - Vt)$ is the exact N-soliton profile. Then, the step δ_s and wave velocity can be found as $V = \delta_s/\Delta t$.

The example of the $R(\delta)$ dependence, calculated for a given one-soliton profile corresponding to the eigenvalue $\lambda = 0.3 + 0.5i$ is shown in Fig. 1 (curve A). This plot shows that at $\delta = \delta_s$ the function $R(\delta_s)$ indeed has a sharp minimum (near zero), whereas to both sides away from this point the function $R(\delta)$ rapidly grows.

When the complex eigenvalue $\lambda = \lambda_r + i\lambda_i$ has been determined the physical parameters of searched soliton such as velocity V, characteristic length

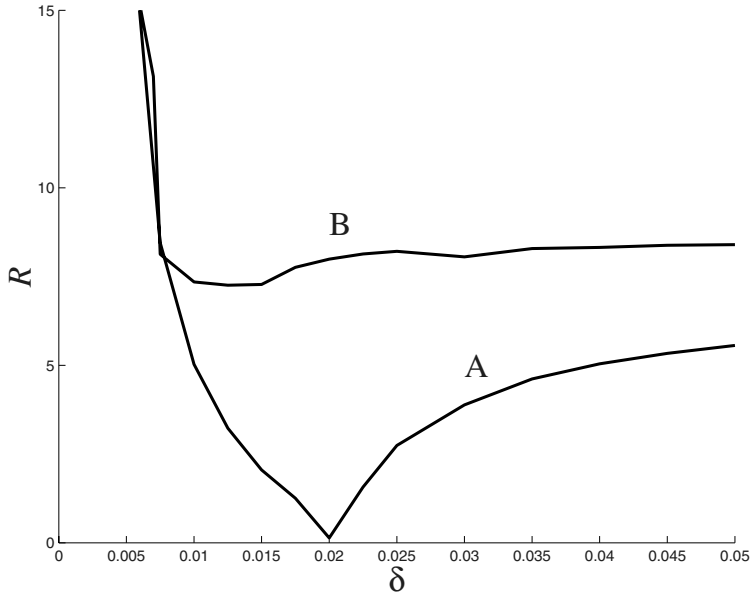


Fig. 1: Integrated reflection coefficient $R(\delta)$ for the exact one-soliton profile (8) with parameter $\lambda = 0.3{+}0.5i$ (A) and its imitation with the localized wave packet $b_{imit}(x)$ (22) (B).

L, duration T, and local frequency can be found with the help of explicit formulas (14), (15) and (16).

4 Discrimination of solitons from "linear" Wave Packets

Here we examine how well the proposed method of time series analysis can discriminate between actual soliton and a similar localized wave packet. As a test we use the soliton $b_{sol}(x;\lambda)$ with eigenvalue $\lambda = 0.3 + 0.5i$. The result of calculation (curve A in Fig. 1) shows that the dependence $R(\delta)$ has an evident minimum, reaching zero under the same sampling step $\delta = \delta_s$ as the raw function $b_{sol}(x;\lambda)$ was determined. In due course, the linear wave packet may be described by the imitating function of the following form

$$b_{imit}(x) = A \exp\left[-\gamma(x - x_0)^2\right] \left[\cos(\theta - \theta_c) + i \sin(\theta - \theta_s)\right], \quad \theta = k(x - x_0). \tag{22}$$

This function depends on many parameters, which enabled us to choose a function $b_{imit}(x)$ with nearly the same waveform as the exemplary soliton $b_{sol}(x;\lambda)$. The values of the parameters used for calculation are: $A = 3.0$, $\gamma = 1.5$, $k = 2.4$, $x_0 = -0.27$, $\theta_1 = 1.04$, $\theta_2 = 0.54$. The waveforms of "soliton" $b_{sol}(x)$ and "linear" $b_{imit}(x)$ functions are shown in Fig. 2.

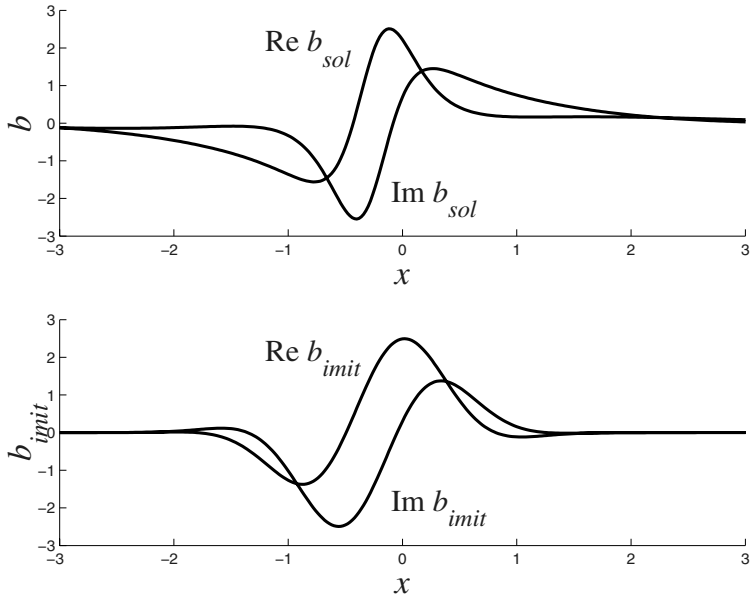


Fig. 2: Comparison of soliton $b_{sol}(x)$ and "linear" signal $b_{imit}(x)$, which have been used for the calculation of integral reflection coefficient $R(\delta)$ in Fig. 1.

The calculation of the integral reflection coefficient $R(\delta)$ for the function $b_{imit}(x)$ has provided the following result (Fig. 1). The plot $R(\delta)$ for the soliton-imitating signal (22) (curve B) is essentially different from the corresponding plot for actual soliton (curve A). Instead of a minimum at $\delta = \delta_s$, the coefficient $R(\delta)$ has at δ near δ_s a plateau at rather high level. Thus, the proposed function $R(\delta)$ for characterizing the soliton-like nature of the signal has turned out to be very sensitive to a signal deviation from a soliton waveform.

5 Influence of High-Frequency Noise on the Soliton Detector

To validate and show the robustness of the proposed technique to the possible occurrence of high-frequency noise in data, this method has been applied to the testing signal, consisting of a soliton $b_{sol}(x)$ with parameter $\lambda = 0.3+0.5i$, and high-frequency interference signal $b_{pert}(x) = \alpha \exp(2i\pi x)$ with amplitude $\alpha = 0.3$ (which provide noise/signal ratio ~ 11 %):

$$b(x) = b_{sol}(x) + b_{pert}(x). \tag{23}$$

The estimated reflection coefficient $R(\delta)$ for this "noisy"soliton is shown in Fig. 3 (curve B). The comparison with pure soliton (curve A) shows that

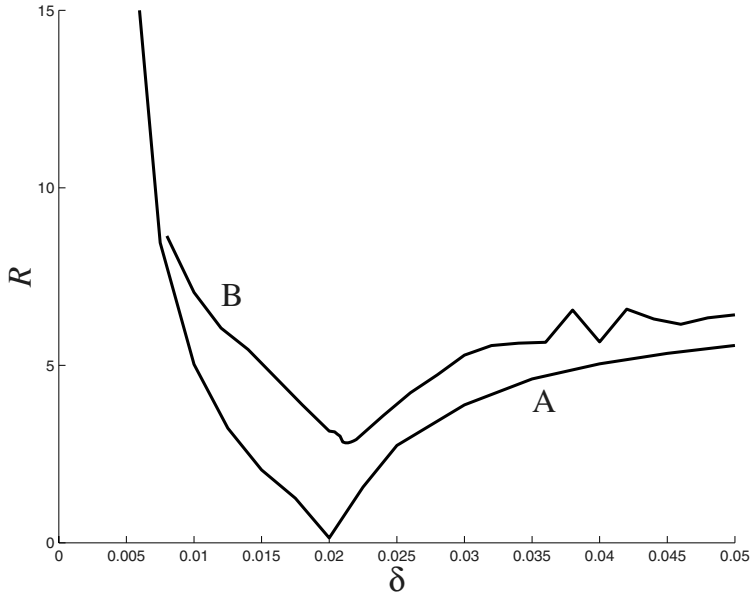


Fig. 3: Comparison of the reflection coefficient $R(\delta)$ for the pure soliton (A) and the soliton with high-frequency harmonic noise (B).

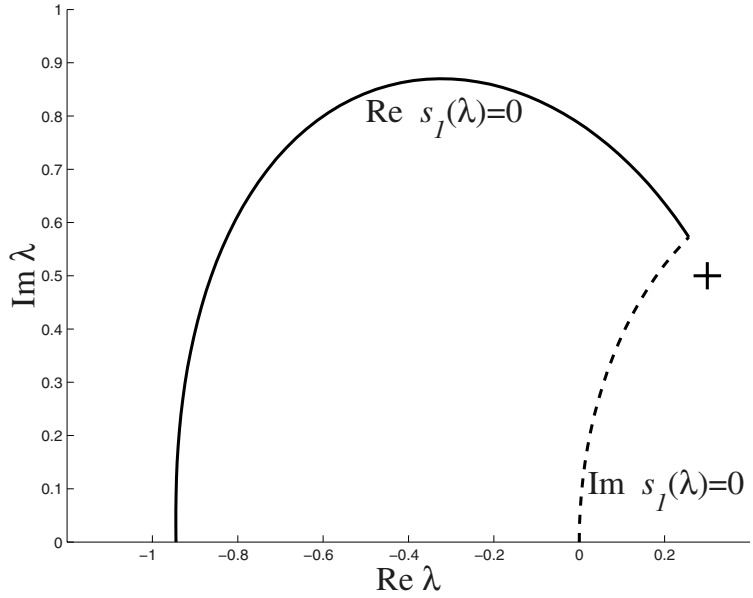


Fig. 4: The example of the eigenvalue λ finding for one-soliton profile, perturbed by high-frequency harmonic interference. The value λ for the exact one-soliton profile is marked by cross.

despite the noise occurrence the coefficient R still has an evident minimum in the vicinity of $\delta = \delta_s$.

The calculation of eigenvalues λ for a "noisy"soliton (23) has been made for the sampling scale δ_s with the use of the algorithm described above. Figure 4 shows the zero-level lines $\mathrm{Re}\, s_1(\lambda) = 0$ and $\mathrm{Im}\, s_1(\lambda) = 0$; their intersection is the searched complex eigenvalue λ. For a relatively weak noise amplitude $\alpha = 0.3$ a perturbed eigenvalue does not shift far from a nominal eigenvalue for pure soliton $b_{sol}(x)$ (marked by a cross).

The sensitivity of the method to the noise amplitude α is demonstrated by the results of the analysis of several "noisy"soliton profiles (23) (Fig. 5). The calculations of $R(\delta)$ and λ show that under even a relatively high level of interference signal $\alpha = 0.45$, when the $R(\delta)$ minimum near δ_s is rather unclear, the eigenvalue λ nevertheless does not shift too far from "actual" unperturbed value $\lambda = 0.3{+}0.5i$. Thus, the proposed method is sufficiently robust regarding high-frequency noise.

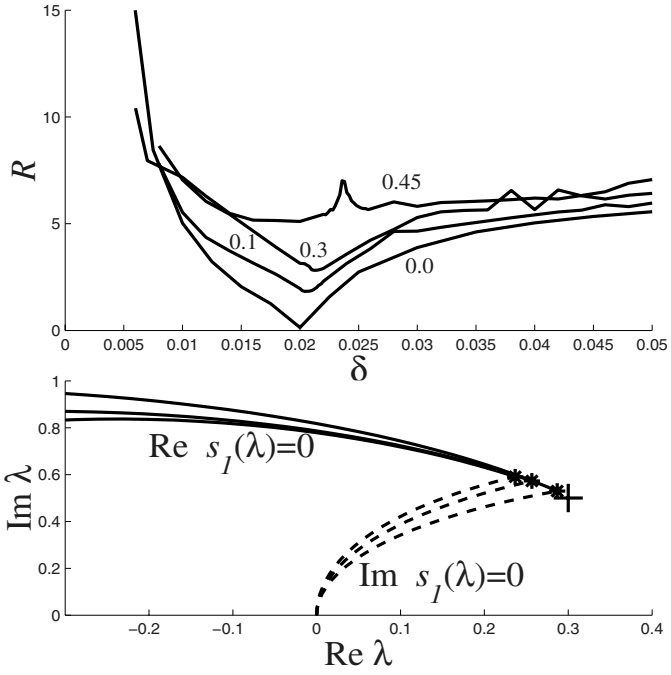


Fig. 5: The dependence of the results of the perturbed soliton profile (23) analysis on noise amplitude α (indicated near relevant curves) and the results of the eigenvalue λ calculation for the same noise levels (asterisks at bottom plot). The value λ for the exact one-soliton profile is marked by cross.

6 Possible Applications and Further Studies

The method of integral reflection coefficient can be applied, strictly speaking, to any soliton solutions of the integrable nonlinear wave equations. The localized solutions of more general non-integrable nonlinear wave equations, which may be called "soliton-like"or "coherent structures", cannot be directly found by this technique. However, analysis of the observational data with ST can provide information on how strongly viscosity and other complicating factors distort in reality the soliton structure of non-linear waves.

We have elaborated this technique for DNLS because this equation describes a wide range of nonlinear phenomena in space plasma. The feature of the DNLS soliton is the consistent variations of both amplitude and phase, so the identification of this type of solitons by only its amplitude wave form (e.g., by least square approximation) is not sufficient.

The suggested method is not limited by the treatment of one-soliton profile $b(x)$, but may be applied to the analysis of more complicated events with several interacting solitons. Figures 6 and 7 show examples for the $N = 3$ case. The function $b_{sol}(x)$, shown in Fig. 6, corresponds to the interaction of three

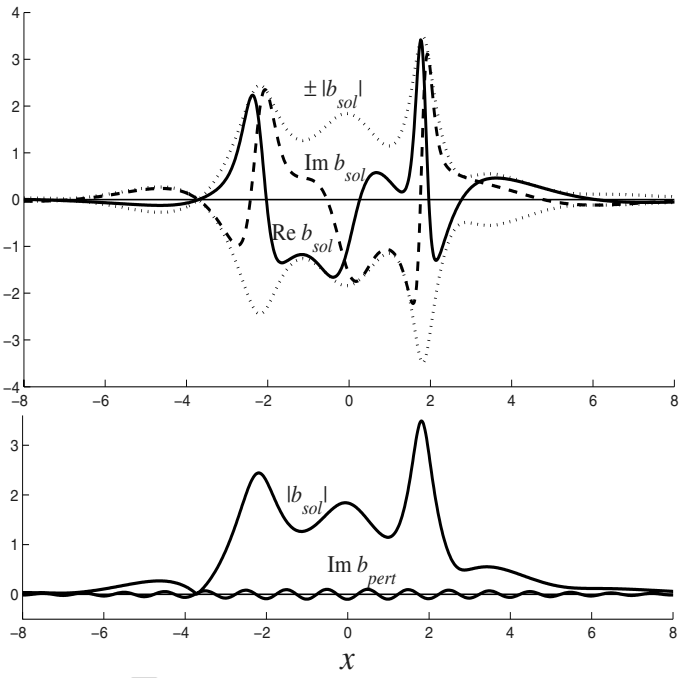


Fig. 6: The function $b_{sol}(x)$, describing the interaction of 3 near solitons with parameters $\lambda_1 = 0.4 + 0.3i$, $\lambda_2 = 0.5i$, and $\lambda_3 = -0.4 + 0.4i$ (upper panel); comparison of the noise with amplitude $\alpha = 0.1$ and modulus of the function $b_{sol}(x)$ (bottom panel).

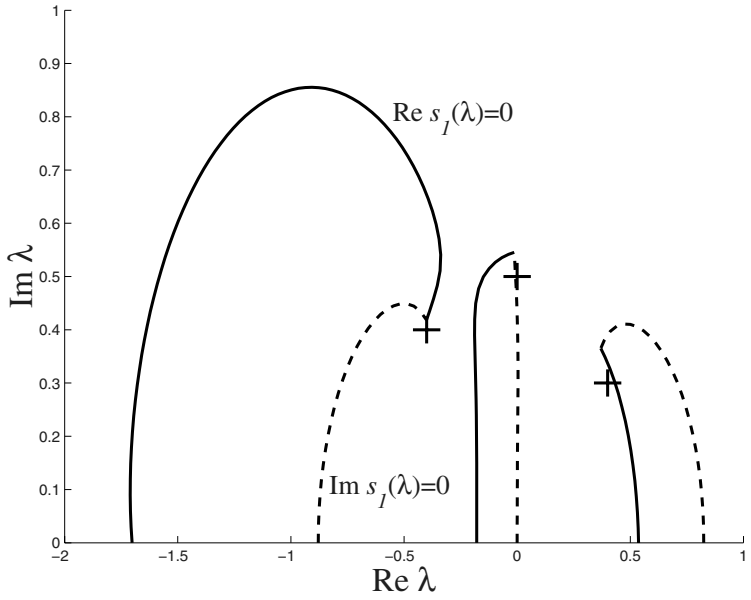


Fig. 7: Calculation of the soliton component for the 3-soliton profile with high-frequency interference, shown in Fig.6. The noise amplitude is $\alpha = 0.1$.

near-by solitons with parameters $\lambda_1 = 0.4 + 0.3i$, $\lambda_2 = 0.5i$, and $\lambda_3 = -0.4 + 0.4i$. The bottom panel shows the interference signal with amplitude $\alpha = 0.1$ and absolute value of function $b_{sol}(x)$. Figure 7 presents the calculation results of the soliton parameters. One can see, that even in the case of complicated profile composed from several interacting solitons and high-frequency noise the soliton parameters can be reliably retrieved.

The DNLS solitons are balanced nonlinear objects which are robust in respect to disturbances and noises. Therefore the retrieval of soliton component from a signal has a fundamental importance. At the same time, the component of nonlinear wave corresponding to the continuous spectrum ("radiation") is apt to dispersion and hardly distinguishable from noise. Therefore, we have not tried to retrieve the radiation component.

7 Conclusion

When the observational conditions enable one to acquire a statistically significant number of relationships between basic parameters (amplitude, duration, etc.) of the expected soliton signals the simple test for soliton identification, based on their comparison with the relevant relationships from soliton theory, can be applied. For a case study, the method to detect soliton and determine

its parameters, based on the scattering transform, may be recommended. We have constructed the algorithm of numerical solution of the direct scattering problem for the linear system (17). This system is associated in the inverse scattering techniques with the DNLS equation (6).

The integral reflection coefficient which steeply drops (theoretically to zero) when a signal is close to the N-soliton waveform has been used as a detector of the DNLS solitons. Application of this technique to numerically simulated signals showed that it is more efficient than standard FT and can be developed into a practical tool for the analysis of outputs from nonlinear systems. The application of the proposed technique to modeling signals shows its superiority over the standard FT. The technique effectively discriminates N-soliton solution ($N = 1 - 5$) from non-soliton isolated disturbances (Gaussian packets). Examples of the Fourier spectra of solitons given in [15] show that standard spectral analysis practically does not evidence the occurrence of the soliton structure in time series. However, a wave envelope that seems complicated to the FT may be a superposition of just a few solitons, easily retrievable with the proposed method.

This approach seems promising for the analysis of nonlinear signals in space physics, often detected in the solar wind, magnetosheath, auroral region, etc. This method enables one to determine from single-point observations the basic parameters of soliton component of a disturbance, such as velocity, amplitude, duration, etc. A similar approach after a minor modification can be applied for the detection of solitons described by other integrable nonlinear equations [21]. As a next step, we will apply the developed technique to the data of satellite observations of electromagnetic disturbances in the near-Earth environment.

Acknowledgements. Useful comments of E.N. Fedorov and all reviewers are appreciated. This study is supported by the INTAS grant 05-1000008-7978 (VAP, KHG), and grant 07-05-00185a from the Russian Fund for Basic Research (NGM).

References

1. K. Stasiewicz: Heating of the solar corona by dissipative Alfven solitons, Phys. Rev. Lett., **96**, 175003, doi:10.1103 /PhysRevLett.96.175003 (2006).
2. C.R. Ovenden, N.A. Shah, S.J. Schwartz: Alfven solitons in solar wind, J. Geophys. Res., **88**, 6095–6101 (1983).
3. A.V. Guglielmi, N.M. Bondarenko, V.N. Repin: Solitary waves in near-Earth environment, Doklady AN SSSR, **240**, 47–50 (1978).
4. V.L. Patel, B. Dasgupta: Theory and observations of Alfven solitons in the finite beta magnetospheric plasma, Physica, **D27**, 387–398 (1987).
5. O.A. Pokhotelov, D.O. Pokhotelov, M.B. Gokhberg, F.Z. Feygin, L. Stenflo, P.K. Shukla: Alfven solitons in the Earth's ionosphere and magnetosphere, J. Geophys. Res., **101**, 7913–7916 (1996).

6. E.N. Pelinovsky, N.N. Romanova: Nonlinear stationary waves in the atmosphere, Physics of the atmosphere and ocean, **13**, 1169–1182 (1977).
7. V.I. Petviashvili, O.A Pokhotelov.: *Solitary Waves in Plasmas and in the Atmosphere*. (London: Gordon and Breach Sci. Publ., 1992).
8. F. Lund: Interpretation of the precursor to the 1960 Great Chilean Earthquake as a seismic solitary wave, Pure Appl. Geophys., **121**, 17–26 (1983).
9. C.I. Christov, V.P. Nartov: *Random Point Functions and Large-Scale Turbulence*. (Nauka-Siberian Branch, Novosibirsk, 1992).
10. C.I. Christov, V.P. Nartov: On a bifurcation emerging from the stochastic solution in a variational problem connected with plane Poiseuille flow, Doklady Acad. Sci. USSR, **277**, 825–828 (1984).
11. A.V. Gurevich, N.G. Mazur, K.P. Zybin: Statistical limit in a completely integrable system with deterministic initial conditions, J. Experim. Theor. Phys., **90**, 695–713 (2000).
12. N.G. Mazur, V.V. Geogdzhaev, A.V. Gurevich, K.P. Zybin: A statistical limit in the solution of the nonlinear Schrödinger equation under deterministic initial conditions, J. Experim. Theor. Phys., **94**, 834–851 (2002).
13. A.R. Osborne: "Nonlinear Fourier analysis" in: *Nonlinear Topics in Ocean Physics* (Elsevier, Amsterdam, 1991) pp. 669–699.
14. A.R. Osborne: Nonlinear Fourier analysis for the infinite-interval Korteweg-de Vries equation I: An algorithm for the direct scattering transform, J. Computational Physics, **94**, 284–313 (1991).
15. T. Hada, R.J. Hamilton, C.F. Kennel: The soliton transform and a possible application to nonlinear Alfven waves in space, Geophys. Res. Lett., **20**, 779–782 (1993).
16. M. Ablowitz, H. Segur: *Solitons and the Inverse Scattering Transform*. (SIAM, Philadelphia, 1981).
17. C.F. Kennel, B. Buti, T. Hada, R. Pellat: Nonlinear, dispersive, elliptically polarized Alfven waves, Physics of Fluids, **31**, 1949–1961 (1988).
18. A.A. Mamun: Alfven solitary structures and their instabilities in a magnetized dusty plasma, Physica Scripta, **60**, 365–369 (1999).
19. S.R. Spangler: Kinetic effects on Alfven wave nonlinearity. II. The modified nonlinear wave equation, Phys. Fluids B, **2**, 407–418 (1990).
20. D.J. Kaup, A.J. Newell: An exact solution for a derivative nonlinear Schrödinger equation, J. Math. Phys., **19**, 798–801 (1978).
21. V.I. Karpman: *Nonlinear waves in dispersive media*. (Moscow, Nauka, 1973) 233pp.

Detecting Oscillations Hidden in Noise: Common Cycles in Atmospheric, Geomagnetic and Solar Data

Milan Paluš[1] and Dagmar Novotná[2]

[1] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic mp@cs.cas.cz
[2] Institute of Atmospheric Physics, Academy of Sciences of the Czech Republic, Boční II/1401, 141 31 Prague 4, Czech Republic nov@ufa.cas.cz

Abstract. In this chapter we present a nonlinear enhancement of a linear method, the singular system analysis (SSA), which can identify potentially predictable or relatively regular processes, such as cycles and oscillations, in a background of colored noise. The first step in the distinction of a signal from noise is a linear transformation of the data provided by the SSA. In the second step, the dynamics of the SSA modes is quantified in a general, nonlinear way, so that dynamical modes are identified which are more regular, or better predictable than linearly filtered noise. A number of oscillatory modes are identified in data reflecting solar and geomagnetic activity and climate variability, some of them sharing common periods.

Keywords: signal detection, statistical testing, Monte Carlo SSA, sunspots, geomagnetic activity, NAO, air temperature, solar-terrestrial relations

1 Introduction

The quest for uncovering physical mechanisms underlying experimental data in order to understand, model, and predict complex, possibly nonlinear processes, such as those studied in geophysics, in many cases starts with an attempt to identify trends, oscillatory processes and/or other potentially deterministic signals in a noisy environment. The distinction of a relatively regular part of the total variability of a complex natural process can be a key for understanding not only such a process itself, but also interactions with other processes or phenomena, if they posses, for instance, oscillations on a similar temporal scale.

Singular system (or singular spectrum) analysis (SSA) [1, 2, 3] in its original form (closely related to the principal component analysis or Karhunen-Loève decomposition) is a method for identification and distinction of important information in multivariate data from noise. It is based on an orthogonal decomposition of a covariance matrix of multivariate data under study. The SSA provides an orthogonal basis onto which the data can be transformed, making thus individual data components ("modes") linearly independent. Each of the orthogonal modes (projections of the original data onto the new orthogonal basis vectors) is characterized by its variance, which is given by the related eigenvalue of the covariance matrix. Here, we will deal with a univariate version of SSA (which, however, can be generalised into a multivariate version, see, e.g. [4]) in which the analyzed data is a univariate time series and the decomposed matrix is a time-lag covariance matrix, i.e., instead of several components of multivariate data, a time series and its time-lagged versions are considered. This type of SSA application, which has frequently been used especially in the field of meteorology and climatology [5, 6, 7, 8, 9], can provide a decomposition of the studied time series into orthogonal components (modes) with different dynamical properties. Thus, "interesting" phenomena such as slow modes (trends) and regular or irregular oscillations (if present in the data) can be identified and retrieved from the background of noise and/or other "uninteresting" non-specified processes.

In the traditional SSA, the distinction of "interesting" components (signal) from noise is based on finding a threshold (jump-down) to a "noise floor" in a sequence of eigenvalues given in descending order. This approach might be problematic if the signal-to-noise ratio is not sufficiently large, or the noise present in the data is not white but "colored". For such cases, statistical approaches utilizing the Monte Carlo simulation techniques have been proposed [6, 10] for reliable signal/noise separation. The particular case of Monte Carlo SSA (MCSSA) which considers the "red" noise, usually present in geophysical data, has been introduced by Allen & Smith [11]. In this chapter, we present and apply an extension of the Monte Carlo singular system analysis based on evaluating and testing the regularity of dynamics of the SSA modes. In our approach, we retain the decomposition exploiting the linear covariance structure of the data, however, in the testing (detection) part of the method, we evaluate the regularity of dynamics of the SSA modes using a measure of general, i.e., nonlinear dependence. The latter gives a clue in inferring whether the studied data contain a component which is more regular and predictable, in a general, nonlinear sense, than linearly filtered noise. Attempts to generalize SSA-like approach to accounting for nonlinear dependence structures are also known [12, 13, 14, 15], however, are not considered in this chapter.

2 Monte Carlo Singular System Analysis and its Enhancement

2.1 The Basic Univariate Singular System Analysis

Let a univariate time series $\{y(i)\}$, $i = 1, \ldots, N_0$, be a realization of a stochastic process $\{Y(i)\}$ which is stationary and ergodic. A map into a space of n-dimensional vectors $\mathbf{x}(i)$ with components $x^k(i)$, where $k = 1, \ldots, n$, is given as

$$x^k(i) = y(i + k - 1). \tag{1}$$

The sequence of the vectors $\mathbf{x}(i)$, $i = 1, \ldots, N = N_0 - (n - 1)$, is usually referred to as the $n \times N$ trajectory matrix $\mathbf{X} = \{x_i^k\}$, the number n of the constructed components is called the embedding dimension, or the length of the (embedding) window. Suppose that the n-dimensional time series (the trajectory matrix $\mathbf{X}$) results from a linear combination of m different dynamical modes, $m < n$. Then, in an ideal case, the rank of the trajectory matrix $\mathbf{X}$ is $\mathrm{rank}(\mathbf{X}) = m$, and $\mathbf{X}$ can be transformed into a matrix with only m nontrivial linearly independent components. In the univariate SSA, it is supposed that this procedure decomposes the original series $\{y(i)\}$ into a sum of several components and noise. Exceptional care must be taken when the trajectory matrix $\mathbf{X}$ is constructed from a time series possibly containing short-range correlated or nonlinear signals such as chaotic signals. The emergence of additional, linearly independent modes when the lags used in construction of the trajectory matrix are larger than the correlation length of such a signal has been discussed in [16].

Instead of the $n \times N$ matrix $\mathbf{X}$, it is more convenient to decompose the symmetric $n \times n$ matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$, since $\mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{C})$. The elements of the covariance matrix $\mathbf{C}$ are

$$c_{kl} = \frac{1}{N} \sum_{i=1}^{N} x^k(i) x^l(i), \tag{2}$$

where $1/N$ is the proper normalization and the components $x^k(i)$, $i = 1, \ldots, N$, are supposed to have zero mean. The symmetric matrix $\mathbf{C}$ can be decomposed as

$$\mathbf{C} = \mathbf{V}\Sigma\mathbf{V}^T, \tag{3}$$

where $\mathbf{V} = \{v_{ij}\}$ is an $n \times n$ orthonormal matrix, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, σ_i are non-negative eigenvalues of $\mathbf{C}$, by convention given in descending order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. If $\mathrm{rank}(\mathbf{C}) = m < n$, then

$$\sigma_1 \geq \cdots \geq \sigma_m > \sigma_{m+1} = \cdots = \sigma_n = 0. \tag{4}$$

In the presence of noise, however, all eigenvalues are positive, and the relation (4) takes the following form [17]:

$$\sigma_1 \geq \cdots \geq \sigma_m \gg \sigma_{m+1} \geq \cdots \geq \sigma_n > 0. \tag{5}$$

Then, the modes ξ_i^k

$$\xi_i^k = \sum_{l=1}^{n} v_{lk} x_i^l, \tag{6}$$

for $k = 1, \ldots, m$ are considered as the "signal" part, and the modes ξ_i^k, $k = m+1, \ldots, n$, are considered as the noise part of the original time series. The "signal" modes can be used to reconstruct the denoised signal $\tilde{x}_l^k$ as

$$\tilde{x}_i^k = \sum_{l=1}^{m} v_{kl} \xi_i^l. \tag{7}$$

Of course, the original time series x_i^k can be reconstructed back from the modes as

$$x_i^k = \sum_{l=1}^{n} v_{kl} \xi_i^l. \tag{8}$$

In the latter relation – decomposition (8), the modes ξ_i^k can also be interpreted as time-dependent coefficients and the orthogonal vectors $\mathbf{v}_k = \{v_{kl}\}$ as basis functions, usually called the empirical orthogonal functions (EOF's).

2.2 Monte Carlo Singular System Analysis

The clear signal/noise distinction based on the eigenvalues $\sigma_1, \sigma_2, \ldots, \sigma_n$ can only be obtained in particularly idealized situation when the signal/noise ratio is large enough and the background consists of white noise. In many geophysical processes, however, so-called "red" noise with power spectrum of the $1/f^\alpha$ (power-law) type is present [11]. Its SSA eigenspectrum also has the $1/f^\alpha$ character [18], i.e., the eigenspectrum of red noise is equivalent to a coarsely discretized power spectrum, where the number of frequency bins is given by the embedding dimension n. The eigenvalues related to the slow modes are much larger than the eigenvalues of the modes related to higher frequencies. Thus, in the classical SSA approach applied to red noise, the eigenvalues of the slow modes might incorrectly be interpreted as a (nontrivial) signal, or, on the other hand, a nontrivial signal embedded in red noise might be neglected if its variance is smaller than the slow-mode eigenvalues of the background red noise. Therefore, Allen & Smith [11] proposed comparing the SSA spectrum of the analyzed signal with the SSA spectrum of a red-noise model fitted to the studied data. Such a red-noise process can be modeled by using an AR(1) model (autoregressive model of the first order):

$$u(i) - \hat{u} = \alpha(u(i-1) - \hat{u}) + \gamma z(i), \tag{9}$$

where $\hat{u}$ is the process mean, α and γ are process parameters, and $z(i)$ is Gaussian white noise with zero mean and unit variance.

In order to correctly detect a signal in red noise, we will apply the following approach, inspired by Allen & Smith [11]:

First, the eigenvalues are plotted not according to their values, but according to a frequency associated with a particular mode (EOF), i.e., the eigenspectrum in this form becomes a sort of a (coarsely) discretized power spectrum in general, not only in the case of red noise (when the eigenspectra have naturally this form, as mentioned above).
Second, the eigenspectrum obtained from the studied data set is compared, in a frequency-by-frequency way, with the eigenspectra obtained from a set of realizations of an appropriate noise model (such as the AR(1) model (9)), i.e., an eigenvalue related to a particular frequency bin obtained from the data is compared with a range of eigenvalues related to the same frequency bin, obtained from the set of so-called surrogate data, i.e., the data artificially generated according to the chosen noise model (null hypothesis) [11, 19, 20, 21]. Allen & Smith [11] also discuss other relevant approaches how to compare the eigenvalues from the tested data and the surrogates.

The detection of a nontrivial signal in an experimental time series becomes a statistical test in which the null hypothesis that the experimental data were generated by a chosen noise model is tested. When (an) eigenvalue(s) associated with some frequency bin(s) differ(s) with a statistical significance from the range(s) of related noise model eigenvalues, then one can infer that the studied data cannot be fully explained by the considered null hypothesis (noise model) and could contain an additional (nontrivial) signal. This is a rough sketch of the approach, for which we will use the term Monte Carlo SSA (MCSSA), as coined by Allen & Smith [11] (see [11] where also a detailed account of the MCSSA approach with analyses of various levels of null hypotheses is given), although the same term was earlier used for other SSA methods, which considered a white noise background [6, 10].

2.3 Enhanced MCSSA: Testing Dynamics of the SSA Modes

The MCSSA described above is a sophisticated technique. However, it still assumes a very simple model, i.e., that the signal of interest has been linearly added to a specified background noise. Therefore the variance in the frequency band, characteristic for the searched signal, is significantly larger than the typical variance in this frequency band obtained from the considered noise model. If the studied signal has a more complicated origin, e.g., when an oscillatory mode is embedded into a background process without significantly increasing variance in a particular frequency band, the standard MCSSA can fail. In order to be able to detect any interesting dynamical mode independent of its (relative) variance, Paluš & Novotná [22] have proposed to test also dynamical properties of the SSA modes against the modes obtained from surrogate data. From this idea, the question arises how we can characterize dynamics in a simple, computationally effective way.

Consider a complex, dynamic process evolving in time. A series of measurements done on such a system in consecutive instants of time $t = 1, 2, \ldots$ is usually called a time series $\{y(t)\}$. Consider further that the temporal evolution of the studied system is not completely random, i.e., that the state of the system at time t in some way depends on the state in which the system was at time $t - \tau$. The strength of such a dependence per unit time delay τ, or, inversely, a rate at which the system "forgets" information about its previous states, can be an important quantitative characterization of temporal complexity in the system's evolution. The time series $\{y(t)\}$, which is a record of (a part of) the system's temporal evolution, can be considered as a realization of a stochastic process, i.e., a sequence of stochastic variables. Uncertainty in a stochastic variable is measured by its entropy. The rate with which the stochastic process "produces" uncertainty is measured by its entropy rate. The concept of entropy rates is common to the theory of stochastic processes as well as to information theory where the entropy rates are used to characterize information production by information sources [23].

Alternatively, the time series $\{y(t)\}$ can be considered as a projection of a trajectory of a dynamical system, evolving in some measurable state space. A. N. Kolmogorov, who introduced the theoretical concept of classification of dynamical systems by information rates, was inspired by information theory and generalized the notion of the entropy of an information source [24]. The Kolmogorov-Sinai entropy (KSE) [24, 25, 26] is a topological invariant, suitable for classification of dynamical systems or their states, and is related to the sum of the system's positive Lyapunov exponents (LE) according to the theorem of Pesin [27].

Thus, the concept of entropy rates is common to theories based on philosophically opposite assumptions (randomness vs. determinism) and is ideally applicable for a characterization of complex geophysical processes, where possible deterministic rules are always accompanied by random influences. However, possibilities to compute the exact entropy rates from experimental data are limited to a few exceptional cases. Therefore Paluš [28] has proposed "coarse-grained entropy rates" (CERs) instead. The CERs are relative measures of regularity and predictability of analyzed time series and are based on coarse-grained estimates of information-theoretic functionals. In the simplest case, applied here, we use the so-called mutual information. The mutual information $I(X;Y)$ of two random variables X and Y is given by $I(X;Y) = H(X) + H(Y) - H(X,Y)$, where the entropies $H(X)$, $H(Y)$, $H(X,Y)$ are defined in the usual Shannonian sense [23]:

Let X and Y be random variables with sets of values Ξ and Υ, respectively, probability distribution functions (PDF) $p(x)$, $p(y)$, and a joint PDF $p(x,y)$. The entropy $H(X)$ of a single variable, say X, is defined as

$$H(X) = -\sum_{x \in \Xi} p(x) \log p(x), \tag{10}$$

and the joint entropy $H(X,Y)$ of X and Y is

$$H(X,Y) = -\sum_{x\in\Xi}\sum_{y\in\Upsilon} p(x,y)\log p(x,y). \tag{11}$$

The mutual information $I(X;Y)$ then can be expressed as

$$I(X;Y) = \sum_{x\in\Xi}\sum_{y\in\Upsilon} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{12}$$

A detailed account on relations between entropy rates and information-theoretic functionals is given in [28, 29]. For a time series $\{x(t)\}$, considered as a realization of a stationary and ergodic stochastic process $\{X(t)\}$, $t = 1,2,3,\ldots$, we compute the mutual information $I(x;x_\tau)$ as a function of time lag τ. We mark $x(t)$ as x and $x(t+\tau)$ as x_τ. For defining the simplest form of CER let us find τ_{max} such that for $\tau' \geq \tau_{max}$, $I(x;x_{\tau'}) \approx 0$ for the analysed datasets. Then, we define the norm of the mutual information

$$||I(x;x_\tau)|| = \frac{\Delta\tau}{\tau_{max}-\tau_{min}+\Delta\tau}\sum_{\tau=\tau_{min}}^{\tau_{max}} I(x;x_\tau) \tag{13}$$

with $\tau_{min} = \Delta\tau = 1$ sample as a usual choice. The CER h^1 is then defined as

$$h^1 = I(x,x_{\tau_0}) - ||I(x;x_\tau)||. \tag{14}$$

It has been shown that the CER h^1 provides the same classification of states of chaotic systems as the exact KSE [28]. Since usually $\tau_0 = 0$ and $I(x;x) = H(X)$ which is given by the marginal probability distribution $p(x)$, the sole quantitative descriptor of the underlying dynamics is the mutual information norm (13) which we will call the regularity index. Since the mutual information $I(x;x_\tau)$ measures the average amount of information contained in the process $\{X\}$ about its future τ time units ahead, the regularity index $||I(x;x_\tau)||$ gives an average measure of predictability of the studied signal and is inversely related to the signal's entropy rate, i.e., to the rate at which the system (or process) producing the studied signal "forgets" information about its previous states.

There are plenty of approaches to estimate the mutual information $I(x;x_\tau)$ [30]. If we are not interested in an exact value, but rather in a relative comparison of values obtained from the tested data and from the surrogate set, a simple box-counting approach based on marginal equiquantization [21, 28, 29] is satisfactory. The latter means that the marginal boxes (bins) are not defined equidistantly, but in a such a way that there is approximately the same number of data points in each marginal bin.

2.4 Implementation of the Enhanced MCSSA

We realize the enhanced MC SSA as follows:

1. The studied time series undergoes SSA as briefly described above or in [31], i.e., using an embedding window of length n, the $n \times n$ lag-correlation matrix $\mathbf{C}$ is decomposed using the SVDCMP routine [32]. In the eigenspectrum, the position of each eigenvalue on the abscissa is given by the dominant frequency associated with the related EOF, i.e., detected in the related mode. That is, the studied time series is projected onto the particular EOF, the power spectrum of the projection (mode) is estimated, and the frequency bin with the highest power is identified. This spectral coordinate is mapped onto one of the n frequency bins, which equidistantly divide the abscissa of the eigenspectrum.
2. An AR(1) model is fitted to the series under study, and the residuals are computed.
3. The surrogate data are generated using the above AR(1) model, where "scrambled" (randomly permutated in temporal order) residuals are used as innovations, i.e., the noise term $\gamma z(i)$ in (9) .
4. Each realization of the surrogates undergoes SSA as described in step 1. Then, the eigenvalues for the whole surrogate set are sorted in each frequency bin, and the values for the 2.5th and 97.5th percentiles are found. In the eigenspectra, the 95% range of the surrogates' eigenvalue distribution is illustrated by a horizontal bar between the above percentile values.
5. For each frequency bin, the eigenvalue obtained from the studied data is compared with the range of the surrogate eigenvalues. If an eigenvalue lies above the range given by the above percentiles, the null hypothesis of the AR(1) process is rejected, i.e., there is a probability $p < 0.05$ that such an eigenvalue as observed can emerge from the background of the null noise model.
6. For each SSA mode (a projection of the data onto a particular EOF), the regularity index is computed, as well as for each SSA mode for all the realizations of surrogate data. The regularity indices are processed and statistically tested in the same way as the eigenvalues. The regularity index is based on mutual information obtained by a simple box-counting approach with marginal equiquantization [21, 28, 29].

Performing MCSSA using the embedding window of the length n, there are n eigenvalues in the eigenspectrum, and n statistical tests are done. Therefore, in general, the problem of simultaneous statistical inference should be considered (see [21] and references therein). However, in many relevant applications we are interested in a detection of a signal in a specific frequency band (and not in rejecting the null hypothesis by a digression from the surrogate range by an eigenvalue or a regularity index in *any* frequency band), therefore we will not discuss this topic here.

Rejecting the null hypothesis of the AR(1) (or another appropriate) noise model, one can infer that there is "something more" in the data than a realization of the null hypothesis (noise) model. The rejection based on the eigenvalues indicates a different covariance structure than the noise model used. The rejection based on the regularity index indicates that the studied data contains a dynamically interesting signal with higher regularity and predictability than a mode obtained by linear filtration of the considered noise model.

3 Numerical Examples

3.1 A Signal in AR(1) Background

For an example of the application of the presented approach, let us consider numerically generated data data – a periodic signal with randomly variable amplitude (Fig. 1a) mixed with a realization of an AR(1) process with a strong slow component (Fig. 1b). The used noise model is defined as $x_i = 0.933x_{i-1} + \xi_i$, where ξ_i are Gaussian deviates with zero mean and unit variance. The signal to noise ratios (i.e., the ratios of the respective standard deviations of

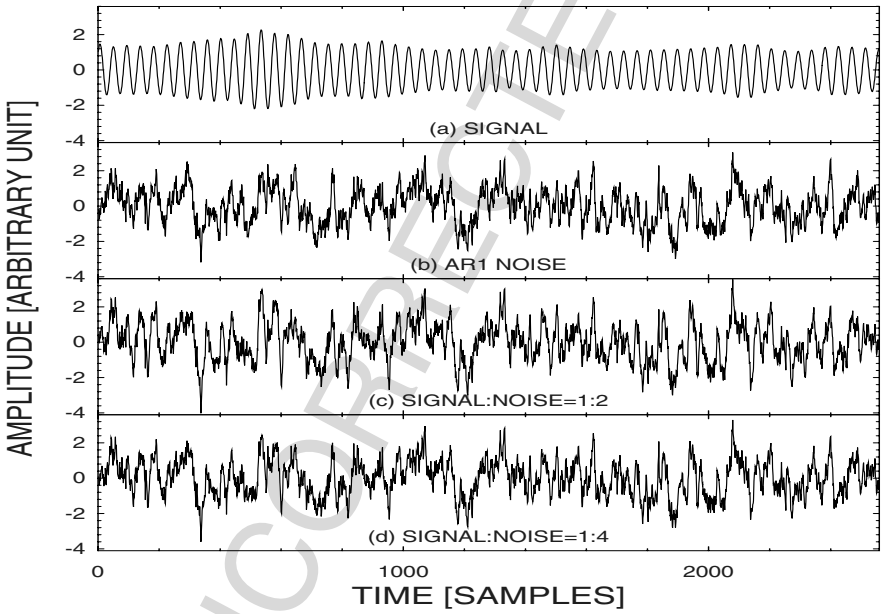


Fig. 1: Numerically generated test data: (a) A periodic signal with randomly variable amplitude was mixed with (b) a realization of an AR(1) process with a strong slow component, obtaining the signal to noise ratio 1:2 (c), and 1:4 (d). Adapted from Paluš & Novotná [31].

signal and noise component) obtained by mixing the signals were 1:2 (Fig. 1c), and 1:4 (Fig. 1d). The latter two series are analyzed by the presented method.

The eigenspectrum of the time series consisting of the signal (Fig. 1a) and the AR(1) noise (Fig. 1b) in the ratio 1:2 (Fig. 1c) is presented in Fig. 2a, where logarithms of the eigenvalues are plotted as the bursts ("LOG POWER"). The series is considered as unknown experimental data, so that an AR(1) model is fitted on the data and the surrogates are generated as described above. The vertical bars in the eigenspectrum represent the surrogate eigenvalue ranges from 2.5th to 97.5th percentiles, which were obtained from 1500 surrogate realizations (here, as well as in the following example). The eigenvalues of the AR(1) surrogates uniformly fill all the n frequency bins (here, as well as in the following example, $n = 100$), while in the case of

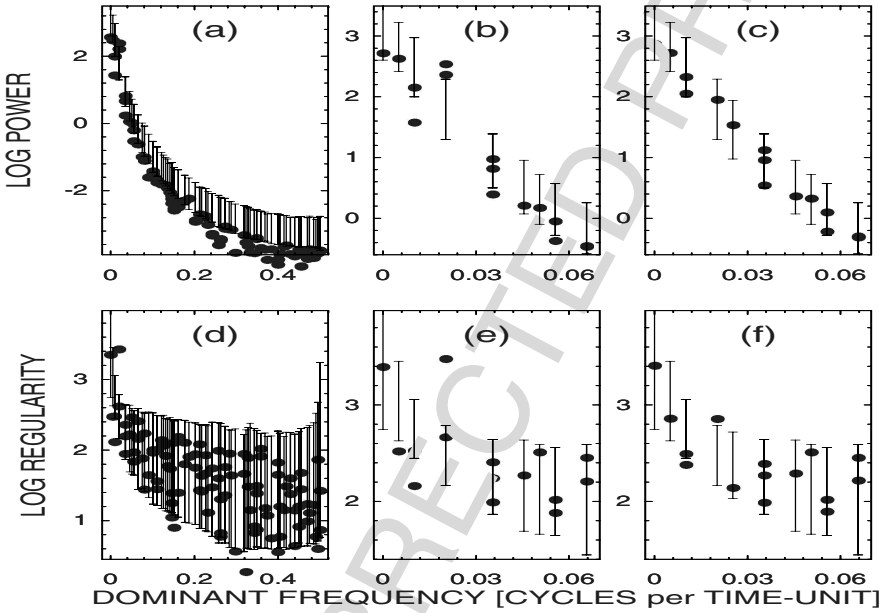


Fig. 2: The standard – eigenvalue based (a–c) and the enhanced – regularity index based (d–f) MCSSA analysis of the numerical data, presented in Fig. 1. (a) The full eigenspectrum and (b) the low-frequency part of the eigenspectrum – logarithms of eigenvalues ("LOG POWER") plotted according to the dominant frequency associated with particular modes, for the signal to noise ratio 1:2. (c) Low frequency part of the eigenspectrum for the signal to noise ratio 1:4. (d) The regularity spectrum and (e) its low frequency part for the signal to noise ratio 1:2. (f) Low frequency part of the regularity spectrum for the signal to noise ratio 1:4. Bursts – eigenvalues or regularity indices for the analysed data; bars – 95% of the surrogate eigenvalues or regularity index distribution, i.e., the bar is drawn from the 2.5th to the 97.5th percentiles of the surrogate eigenvalues/regularity indices distribution. Adapted from Paluš & Novotná [31].

the test data, some bins are empty, others contain one, two, or more eigenvalues. We plot the surrogate bars only in those positions, in which (an) eigenvalue(s) of the analyzed data exist(s). Note the $1/f^{\alpha}$ character of the surrogate eigenspectrum, i.e., the eigenvalues plotted against the dominant frequency associated with the related modes are monotonously decreasing in a $1/f^{\alpha}$ way. The low-frequency part of the eigenspectrum from Fig. 2a is enlarged in Fig. 2b. The two data eigenvalues related to the frequency 0.02 (cycles per time unit) are clearly outside the range of those from the surrogates, i.e., they are statistically significant, the null hypothesis is rejected, and a signal not consistent with the null hypothesis is detected. A close look to the significant modes shows that they are related to the embedded signal from Fig. 1a, in particular, one of the modes contains the signal together with some noise of similar frequencies, and the other include an oscillatory mode shifted by $\pi/2$ relatively to the former one. Note that the simple SSA based on the mutual comparison of the data eigenvalues could be misleading, since the AR(1) noise itself "produces" two or three eigenvalues which are larger than the two eigenvalues related to the signal embedded in the noise.

The same analysis applied to the series possessing the signal/noise ratio 1:4 (Fig. 2c), however, fails to detect the embedded signal — all eigenvalues obtained from the test data are well confined between the 2.5th and 97.5th percentiles of the surrogate eigenvalues distributions. Applying the test based on the regularity index to the mixture with the signal to noise ratio 1:2 (Fig. 2d,e), for one data eigenvalue, the regularity index has been found significantly higher than the related surrogate indices. It was obtained from the mode related to the frequency bin 0.02, as in the case of the significant eigenvalues in Fig. 2a,b. This is the mode which contains the embedded signal (Fig. 1a) together with some noise of similar frequencies. The orthogonal mode, related to the same frequency bin, which has a variance comparable to the former one (Fig. 2a,b), has its regularity index close to the 97.5th percentile of the surrogate regularity indices distribution. With other words, if a (nearly) periodic signal is embedded in a (colored) noise background, the SSA approach, in principle, is able to extract this signal together with some noise of neighboring frequencies, and produces an orthogonal "ghost" mode which has a comparable variance. However, its dynamical properties are closer to those of the modes obtained from the pure noise (null model), as measured by the regularity index (13). Nevertheless, the regularity index used as a test statistic in the MCSSA manner is able to detect the embedded signal with a high statistical significance in this case (signal:noise = 1:2), as well as in the case of the signal to noise ratio 1:4 (Fig. 2f), when the standard (variance-based) MCSSA failed (Fig. 2c). In the latter case, the orthogonal "ghost" mode did not appear, and the regularity index of the signal mode was lower than in the previous case, since the mode contains larger portion of the isospectral noise. However, the signal mode regularity index is still safely above the surrogate bar, i.e., significant with $p < 0.05$ (Fig. 2f).

3.2 A Signal in Multifractal Background

As a more complex example we "embed" the test signal (Fig. 1a) into a realization of a multifractal process (Fig. 3b) generated by a log-normal random cascade on a wavelet dyadic tree [33] using the discrete wavelet transform [32]. Using wavelet decomposition, we embed the most significant part of the signal (Fig. 1a) related to a particular wavelet scale – this wavelet-filtered signal is illustrated in Fig. 3a. The mixing is done in the space of wavelet coefficients. In the first case (in Fig. 3 referred to as "signal added to multifractal"), the standard deviation (SD) of the signal wavelet coefficients is twice the SD of the wavelet coefficients of the multifractal signal in the related scale (Fig. 3c), i.e., the added signal deviates from the covariance structure of the "noise" (multifractal) process. In the second case, we adjusted the SD of both sets of wavelet coefficients to 50% of the SD of the wavelet coefficients of the original multifractal signal in the associated scale (Fig. 3d), so that the total variance in this scale (frequency band) does not exceed the corresponding variance of the "clean" multifractal. Then, it is not surprising, that the variance-(eigenvalues)-based MCSSA test, using the AR(1) surrogate data (Fig. 4a,b), clearly distinguishes the signal from the multifractal background in the first case (Fig. 4a) including its orthogonal "ghosts", while in the second case, no

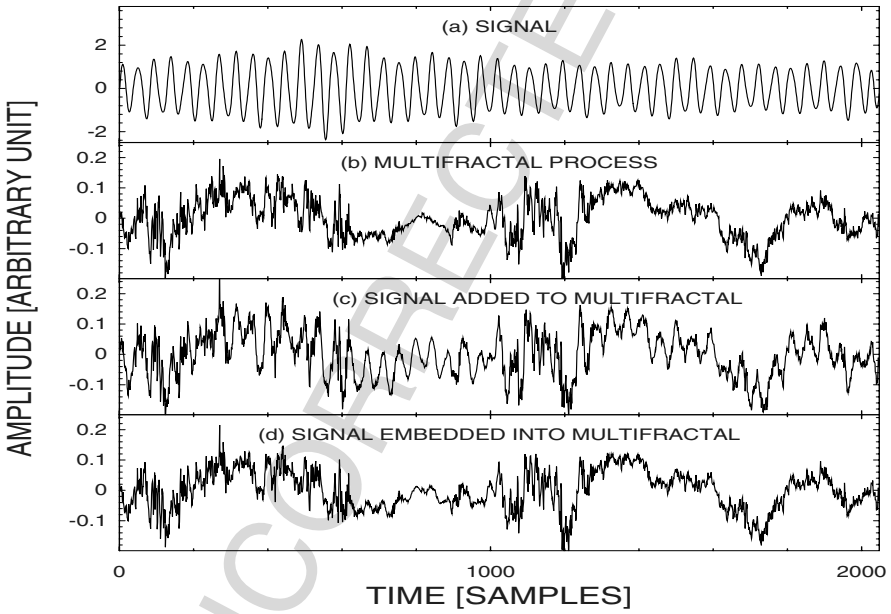


Fig. 3: Numerically generated test data: (a) The wavelet filtered signal from Fig. 1a was embedded into (b) a realization of a multifractal process, obtaining the ratio of related wavelet coefficients 1:2 (c), and 0.5:0.5 (d). Adapted from Paluš & Novotná [31].

eigenvalue is outside the AR(1) surrogate range, but the slow trend mode (Fig. 4b). The AR(1) process is unable to correctly mimic the multifractal process - the slow mode (the zero frequency bin) scores as a significant trend over the AR(1) surrogate range, while the variance on subsequent frequencies is overestimated (Fig. 4a,b). On the other hand, even the AR(1) surrogate model is able to detect the added signal in the first case (Fig. 4a). If we use realizations of the same multifractal process as the surrogate data, the signal is detected in the first case (not presented, just compare the bursts on frequency 0.02 in Fig. 4a and the related surrogate bar in Fig. 4c), while in the second case, the eigenvalues-based MCSSA neglects the signal embedded into the multifractal "noise" – all the data mixture eigenvalues (bursts) are inside the multifractal surrogate bars (Fig. 4c). In the MCSSA tests using the regularity index, the embedded signal is safely detected together with its orthogonal "ghosts" and higher harmonics not only in the first case (Fig. 4d), but also in the second case, either using AR(1) (Fig. 4e) or the multifractal surrogate data (Fig. 4f), when it is, from the point of view of the covariance structure, indistinguishably embedded into the multifractal process.

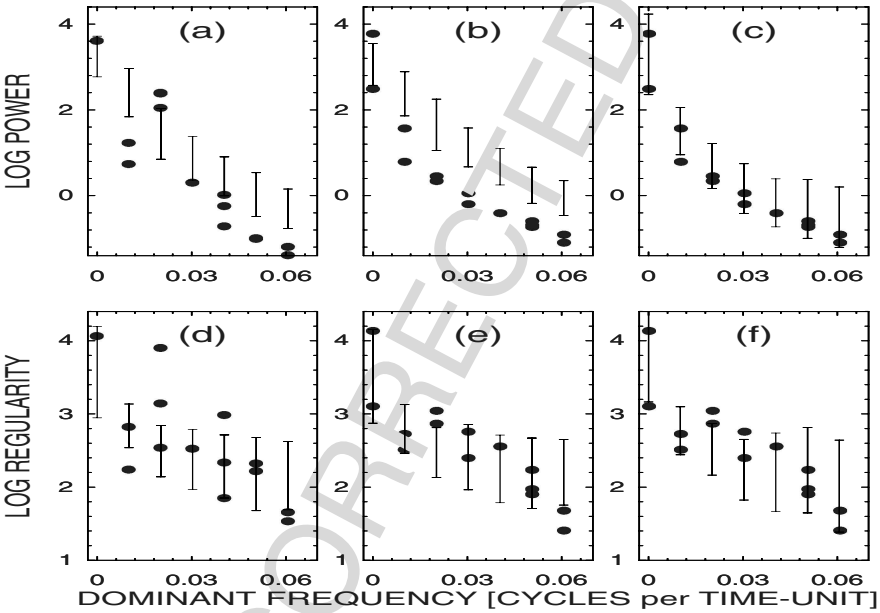


Fig. 4: The low frequency parts of the MCSSA eigenspectra (a–c) and regularity spectra (d–f) for the signal embedded into a multifractal process with wavelet coefficient ratio 1:2 (a,d) and 0.5:0.5 (b,c,e,f). Bursts – eigenvalues or regularity indices for the analysed data; bars – 95% of the surrogate eigenvalues or regularity index distribution obtained from the AR(1) (a,b,d,e) and the multifractal (c,f) surrogate data. Adapted from Paluš & Novotná [31].

4 Detection of Irregular Oscillations in Geophysical Data

Temperature measurements are among the longest available instrumental data characterizing the long term evolution of the atmosphere and climate in a particular location. For instance, the data from the Prague–Klementinum station are available since 1775. On the other hand, large-scale circulation patterns reflect a more global view on the atmospheric dynamics. The North Atlantic Oscillation (NAO) is a dominant pattern of atmospheric circulation variability in the extratropical Northern Hemisphere, accounting for about 60% of the total sea-level pressure variance. The NAO has a strong effect on European weather conditions, influencing meteorological variables including the temperature [34]. The NAO – temperature relationship, however, is not straightforward and its mechanism is not yet fully understood.

The possible influence of the solar variability on the climate change has been a subject of research for many years, however, there are still open questions and unsolved problems (for reviews, see e.g. [35, 36, 37]). Probably the longest historical record of the solar variability are the well-known sunspot numbers. After the sunspot numbers, aa index, the time series of the geomagnetic activity provides the longest data set of solar proxies which goes back to 1868 [38]. Since there are no direct measurements of solar irradiance available until the beginning of the 1980s, the data of geomagnetic variations are used for an additional study of solar activity, especially of irradiance.

It might be interesting if the atmospheric data, both the local and global, and the geomagnetic and solar data possess any common, repeating variability pattern such as cycles or oscillatory modes. The enhanced MCSSA can give an answer to such a question.

4.1 The Data

The NAO index is traditionally defined as the normalized pressure difference between the Azores and Iceland. The NAO data used here and their description are available at `http://www.cru.uea.ac.uk/cru/data/nao.htm`.

Monthly average near-surface air temperature time series from ten European stations were used (see [31] for details), obtained from the Carbon Dioxide Information Analysis Center Internet server (`ftp://cdiac.esd.ornl.gov/pub/ndp041`) as well as a time series from the Prague–Klementinum station from the period 1781 – 2002. The long-term monthly averages were subtracted from the data, so that the annual cycle was effectively filtered out.

The aa-index is defined by the average, for each 3-hour period, of the maximum of magnetic elements from two near-antipodal mid-latitude stations in Australia (Melbourne) and England (Greenwich). The data spanning the period 1868–2005 were obtained from World Data Centre for Solar-Terrestrial Physics, Chilton, `http://www.ukssdc.ac.uk/data/wdcc1/wdc_menu.html`.

The monthly sunspot data, spanning the period 1749–2006, has been obtained from the SIDC-team, Royal Observatory of Belgium, Ringlaan 4, 1180 Brussels, Belgium, `http://sidc.oma.be/DATA/monthssn.dat`.

4.2 The Results

Figure 5 presents the results from the enhanced MCSSA for the considered monthly NAO index and the monthly average near-surface air temperature time series from Prague (Prague–Klementinum station) and Berlin, obtained using the embedding dimension $n = 480$ months. In the standard MCSSA, the only eigenvalue undoubtedly distinct from the surrogate range is the trend (zero frequency) mode in the temperature (Fig. 5b,c). Further, there are two modes at the frequency 0.0104 just above the surrogate bar in the Prague temperature and NAO test (Fig. 5a,b). These results, however, are still "on the edge" of significance and are not very convincing. In the case of Berlin,

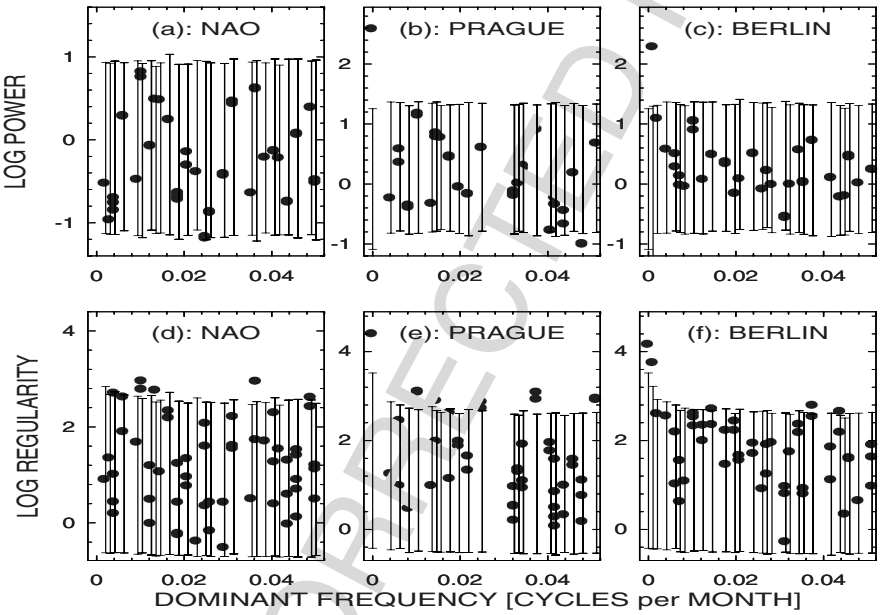


Fig. 5: Enhanced MCSSA analysis of the monthly NAO index (a,d) and monthly average near-surface air temperature series from Prague–Klementinum (b,e) and Berlin (c,f). Low-frequency parts of eigenspectra – logarithms of eigenvalues ("LOG POWER") (a,b,c) and regularity index spectra (d,e,f). Bursts – eigenvalues or regularity indices for the analysed data; bars – 95% of the surrogate eigenvalues or regularity index distribution, i.e., the bar is drawn from the 2.5th to the 97.5th percentiles of the surrogate eigenvalues/regularity indices distribution. The datasets span the period 1824–2002, the embedding dimension $n = 480$ months was used.

the eigenvalues of the modes at the frequency 0.0104 are confined within the surrogate range (Fig. 5c).

A quite different picture is obtained from the analyses based on the regularity index (Figs. 5d,e,f). Several oscillatory modes have been detected with a high statistical significance. The distinction of the regularity indices of these modes from the related surrogate ranges is clear and even the simultaneous statistical inference cannot jeopardize the significance of the results. The significant modes in the NAO are located at the frequencies (in cycles per month) 0.004, 0.006, 0.0104, 0.014, 0.037 and 0.049, corresponding to the periods of 240, 160, 96, 73, 27 and 20 months, respectively. Besides the zero frequency (trend) mode, the significant modes in the Prague temperature are located at the frequencies 0.0104, 0.014, 0.016, 0.018, 0.025, 0.037 and 0.051, corresponding to the periods of 96, 68, 64, 56, 40, 27 and 20 months, respectively. In the case of Berlin, there are some differences, namely the modes with the periods 20, 40, 56 and 64 months are missing, while modes with periods 23, 29 and 58 months, as well as a slow mode next to the zero frequency mode appeared. The significant modes with the periods 27, 68 and 96 months were detected in both the records.

The modes with a period of 8 years were extracted and analysed in [31], their mean frequency was estimated with higher precision as 7.8 years. Besides the latter modes (and the trend mode in the temperature), the highest regularity index was obtained for the modes with a period of 27 months (frequency 0.037). This frequency lies within the range of the quasi-biennial oscillations (QBO). The behavior of these modes was studied in some detail in [39].

The results of the enhanced MCSSA analysis of the aa index are presented in Fig. 6. In the standard (eigenvalue) analysis (Fig. 6a), we can see significant modes representing the trend, i.e., the zero frequency mode, and a mode with a frequency of 0.0073 which corresponds to the period of 136 months, i.e. to the 11-year solar activity cycle. The analysis based on the regularity index (Fig. 6b) confirms the previous two modes and adds two more ones on frequencies of 0.0104 and 0.016, corresponding to periods of 96 and 64 months.

The mode with the period of 96 months or 8 years has been detected in all the above analyzed data sets, i.e., in the near-surface air temperature, the NAO index, and the geomagnetic aa index. The time series of the modes extracted using the SSA, i.e., by projecting the input data on the particular EOF, are presented in Fig. 7a,c,e. When the modes are extracted using SSA, there is an uncertainty of timing of the modes given by the embedding window, and a part of the data equal to the embedding window is lost. We positioned the SSA modes on the time axis by maximizing the cross-correlation between the mode and the original data. This approach, however, does not always give unambiguous results. Therefore, Paluš & Novotná [39] studied the possible relationships of the QBO modes from the temperature and NAO index not only using the SSA-extracted modes, but also using modes extracted from the data by means of complex continuous wavelet transform (CCWT) [40]. Here we compare the modes with the period 96 months extracted by SSA

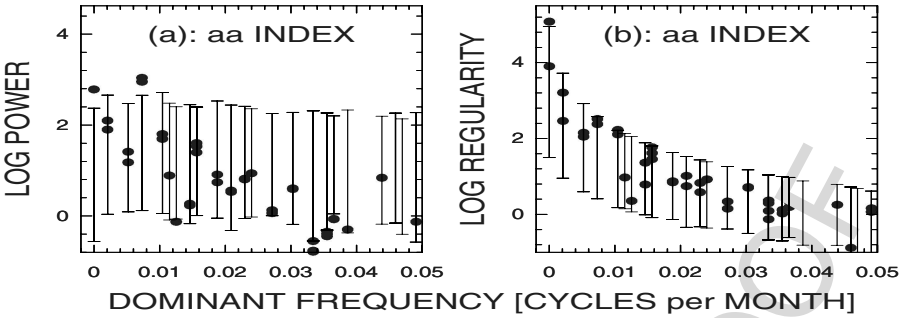


Fig. 6: Enhanced MCSSA analysis of the monthly aa index. The low-frequency part of the eigenspectrum – logarithms of eigenvalues ("LOG POWER") (a) and the regularity index spectrum (b). Bursts – eigenvalues or regularity indices for the analysed data; bars – 95% of the surrogate eigenvalues or regularity index distribution, i.e., the bar is drawn from the 2.5th to the 97.5th percentiles of the surrogate eigenvalues/regularity indices distribution. The dataset spans the period 1868–2005, the embedding dimension $n = 480$ months was used.

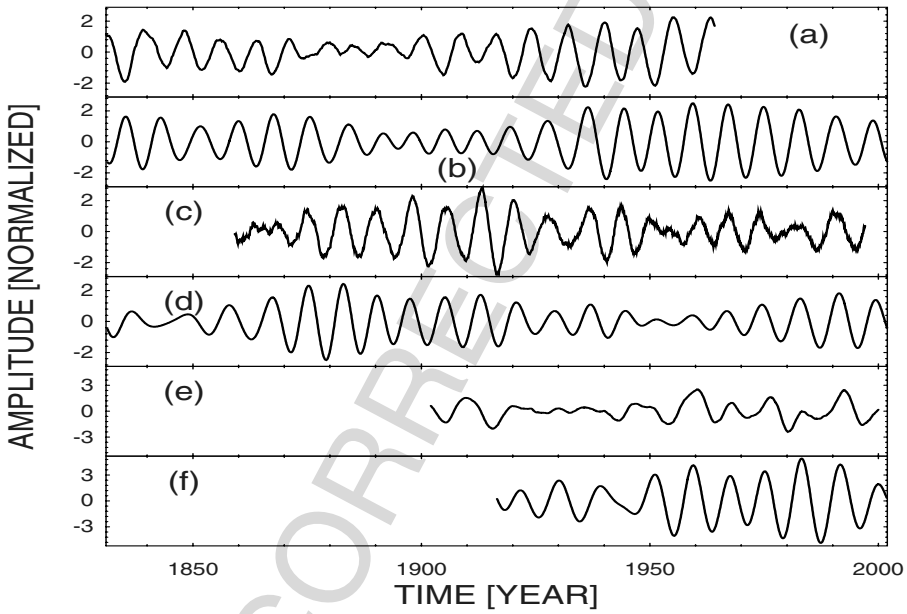


Fig. 7: The oscillatory modes with the approximately 8-year period extracted by using SSA (a,c,e) and CCWT (b,d,f) from the near-surface air temperature (a,b), the NAO index (c,d), and the aa index (e,f).

(Fig. 7a,c,e) with the modes obtained by using CCWT with the central wavelet frequency set to the period of 96 months (Fig. 7b,d,f). The SSA mode and the wavelet mode, obtained from the Prague temperature data (Fig. 7a,b, respectively) are shifted by π (a half of the period), otherwise their agreement is very good. The timing of the SSA and CCWT modes from the NAO index (Fig. 7c,d, respectively) is consistent, however, the wavelet transform performs stronger smoothing. In the aa index, the CCWT mode is smoother and slightly shifted in time in comparison with the SSA mode (Fig. 7f,e, respectively).

Analyzing the monthly sunspot data, the only clear significance in both the eigenspectrum (Fig. 8a) and the regularity index spectrum is the mode with a period of 136 months. The long-term trend at the zero-frequency mode lies at the edge of significance (Fig. 8a). After removal of the 136 month mode and subsequent analysis of the data residuals, the zero frequency mode

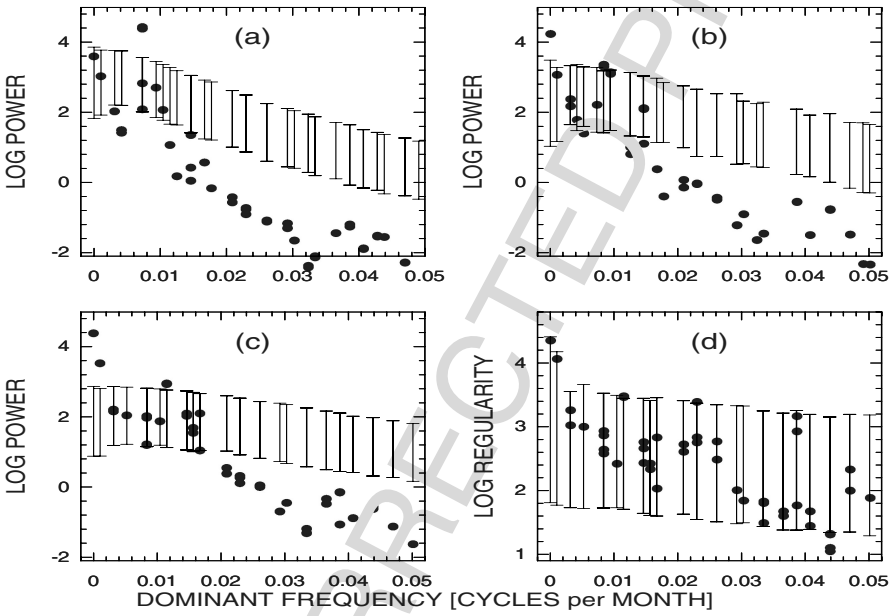


Fig. 8: Enhanced MCSSA analysis of the monthly sunspot data. Low-frequency parts of eigenspectra – logarithms of eigenvalues ("LOG POWER") for the raw sunspot data (a), the sunspot data after removal of the mode with the period 136 months (b), and for the sunspot data after removal of the modes with the periods 136, 120 and 106 months (c). (d): Low-frequency part of the regularity index spectrum for the sunspot data after removal of the modes with the periods 136, 120 and 106 months. Bursts – eigenvalues or regularity indices for the analysed data; bars – 95% of the surrogate eigenvalues or regularity index distribution, i.e., the bar is drawn from the 2.5th to the 97.5th percentiles of the surrogate eigenvalues/regularity indices distribution. The dataset spans the period 1749–2006, the embedding dimension $n = 480$ months was used.

becomes highly significant and another slow mode, with a period about 80 years emerges. Two new significant modes, related to the 11-year solar cycle appear on the frequency bins following the frequency bin of the previously defined mode with the period 136 months. Their periods are 120 and 106 months (Fig. 8b). After removal of all three modes (i.e., the modes with the periods 136, 120 and 106 months) which can be considered as a decomposition of the 11-year cycle, the standard MCSSA analysis of the sunspot data residuals uncovers another interesting oscillatory mode in the frequency bin corresponding to a period of 7.4 years (Fig. 8b). The enhanced MCSSA analysis of the sunspot data residuals confirms all the modes from the standard MCSSA (zero frequency and period 80 and 7.4 years) and adds two new significant modes with the periods of 43.5 and 26 months (Fig. 8d).

It is important to note that the frequency or period accuracy of the SSA approach is limited by the number of frequency bins given by the embedding dimension. The accuracy of the frequency (or the period) of a particular mode can be increased after the extraction of this mode from the original data and its subsequent spectral or autocorrelation analysis, as Paluš & Novotná [22, 31] have done for the temperature mode. On the other hand, oscillatory modes from natural processes are never strictly periodic and their frequency is variable. We illustrate this variability by presenting histograms of instantaneous

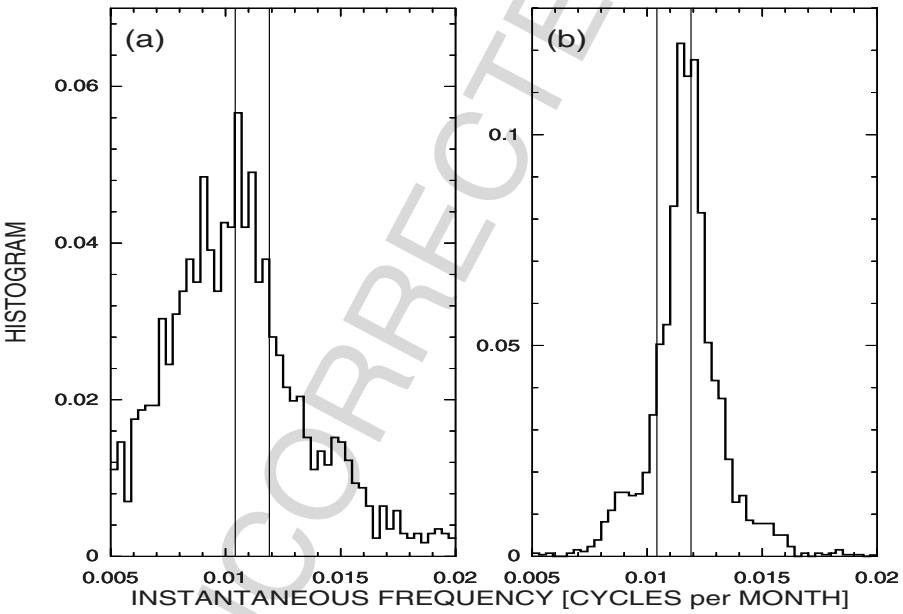


Fig. 9: Histograms of the instantaneous frequencies of the 7.8 yr temperature mode (a) and the 7.4 yr sunspot mode (b). The thin vertical lines mark the frequencies corresponding to the period of 8 and 7 years, reading from the left to the right side.

frequencies of the two close modes – the mode with the period 7.8 yr from the Prague temperature (Fig. 9a), and the period 7.4 yr mode obtained from the sunspot data residuals after modes related to the 11 yr cycle have been previously removed (Fig. 9b). The instantaneous frequencies were obtained by differentiation of the instantaneous phases [41, 42]. The latter can easily be computed by applying the analytic signal approach to the two orthogonal (shifted by $\pi/2$) components of each oscillatory mode, see Refs. [39, 43] for details. Thus the presented histograms are not necessarily equivalent to the power (Fourier) spectra, but they better reflect possibly nonstationary fluctuations of the frequencies of the modes. We can see that the most probable period of the sunspot mode is 7.4 years, with the slight tendency to higher frequencies (Fig. 9b), while in the case of the temperature mode, the most probable period is 7.8 years, with considerable weight on slower frequencies (Fig. 9a). There is, however, a great deal of common frequencies of the two modes, giving thus the possibility of interactions during some time intervals.

Considering both the available accuracy and the natural variability of the frequency of the detected oscillatory modes, the periods given here should be understood as limited accuracy estimates of average periods of particular modes.

The common occurrence of the oscillatory modes with the periods of approximately 11, 5.5, and 2.2 years and in the range 7–8 years in the sunspot numbers, the aa index, the near-surface air temperature and the NAO index is summarized in Table 1.

Table 1: Occurrence of the most significant oscillatory modes with periods of approximately 11, 7 – 8, 5.5 and 2.2 years in the sunspot numbers, the aa index, the average near-surface air temperature and the NAO index.

Source	Period [years]			
data	≈ 11	7 – 8	≈ 5.5	≈ 2.2
sunspots	+	+	–	+
aa	+	+	+	–
T	–	+	+	+
NAO	–	+	–	+

We can see that the modes with a period in the range 7–8 years have been detected in all the analysed datasets. These modes, obtained from the near-surface air temperature, from the NAO index and the geomagnetic aa index have already been presented in Fig. 7, the related modes from the sunspot data are illustrated in Fig. 10. Again, we can compare the mode extracted by SSA in the natural EOF base (Fig. 10a) with the modes obtained by CCWT with the Morlet basis [40], using two close central wavelet frequencies corresponding to the periods 8 yr (Fig. 10b) and 7.4 yr (Fig. 10c). We can see that the wavelet extracted modes have a more limited frequency range and

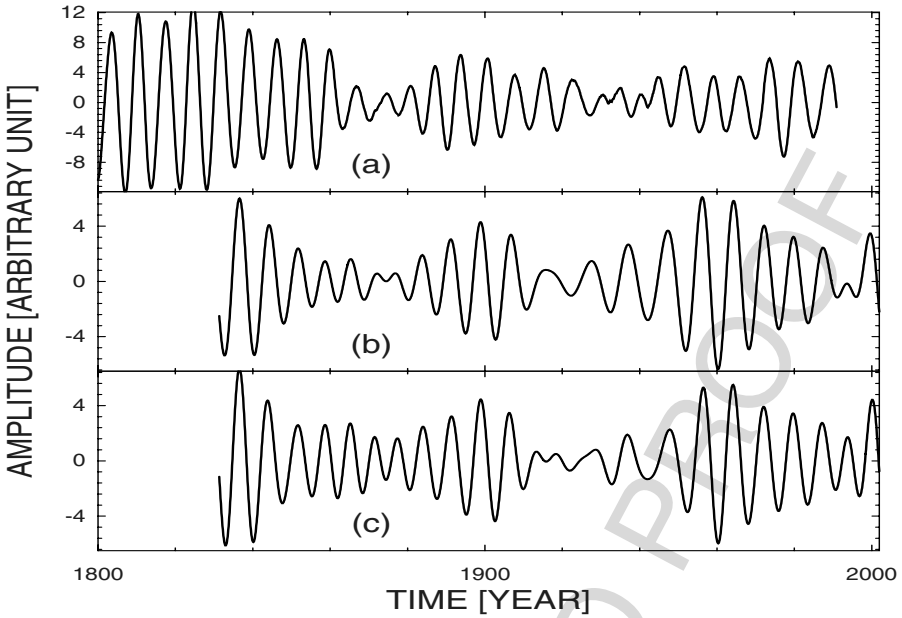


Fig. 10: The oscillatory mode with the approximately 7.4 yr period obtained from the sunspot data residuals after previously removed modes related to the 11 yr cycle, extracted by using SSA (a) and CCWT with the central wavelet frequency corresponding to periods 8 yr (b) and 7.4 yr (c).

the wavelets with different central frequency are able to better fit the mode shapes in different temporal segments dominated by different frequencies.

5 Discussion and Conclusion

The Monte Carlo Singular System Analysis has been extended by evaluating and testing the regularity of the dynamics of the SSA modes against the colored noise null hypothesis in addition to the test based on variance (eigenvalues). The nonlinear approach to the measurement of regularity and predictability of the dynamics, based on a coarse-grained estimate of the mutual information, gives a possibility to detect dynamical modes which are more regular than those obtained by decomposition of colored noise. Using numerical examples, we have demonstrated that such an enhanced MCSSA test is more sensitive in detection of oscillatory modes hidden in a noisy background. There are, however, some facts about accuracy and consistency of the results which should not be neglected. Already in the previous section, we have discussed the accuracy of the estimation of the period of detected oscillatory modes. We have stated that we are only able to provide a limited accuracy estimate of an average period or frequency, since the frequency of

oscillatory modes in the studied natural phenomena is variable. One should keep this fact in mind in comparisons of results found in the literature. Not only frequency, but also the relative variance and the regularity of the oscillatory modes is variable. Due to this nonstationary behaviour, any conclusion about the existence and significance of a mode is dependent on the temporal range of analysed data. Obtained eigenvalues and regularity indices give an average quantification of the relative variance and regularity, respectively, for the analysed time span of the data. It is possible that in some data segments, the results can change. Thus it is reasonable to combine the MCSSA analysis with a wavelet analysis, using the latter one as an exploratory tool and the former one as a hypothesis testing tool.

Another important question is that of the relevance of the used null hypothesis. While in many cases the simple AR(1) process seems to work satisfactorily, for instance, in the case of the sunspot numbers, it is not generally appropriate. In this case, the AR(1) process does not fit the long-range dependence in the data, but the short-range correlation inside the 11yr cycle. As a consequence, the covariance structures of the data and the null noise model are not consistent (see Fig. 8a,b,c where the surrogate bars overestimate the data eigenvalues). The situation is improved after removal of the modes related to the 11yr cycle, and especially, in the case of the regularity test, the null hypothesis seems to be consistent with the noise part of the data (Fig. 8d). In the further development of the MCSSA, it is desirable to consider also more sophisticated null hypotheses including long-range correlated, fractal and multifractal models, since such properties have been observed in geophysical data, especially in the long-term air temperature records [44, 45].

The enhanced MCSSA has been applied to records of monthly average near-surface air temperature from several European locations, to the monthly NAO index, as well as to the monthly aa index and the sunspot numbers. A number of significant oscillatory modes have been detected in all the different source data, some of them with common periods (Table 1). While the 11yr solar cycle is shared by the solar and geomagnetic data, the quasi-biennial mode is present in the atmospheric data and also in the solar data. The mode with the period in the range 7–8 years is present in all the analysed data, i.e., in the atmospheric temperatures, in the NAO index, in the aa index and in the sunspot numbers.

It is interesting to note that the oscillatory mode with a period of 7.8 years has been detected in the NAO, in the Arctic Oscillation (AO), in the Uppsala winter near-surface air temperature, as well as in the Baltic Sea ice annual maximum extent by Jevrejeva and Moore [46]. Applying MCSAA on the winter NAO index, Gámiz-Fortis et al. [47] detected oscillations with the period 7.7 years. Moron et al. [48] have observed oscillatory modes with the period about 7.5 years in the global sea surface temperatures. Da Costa and de Verdiere [49] have detected oscillations with the period 7.7 years in interactions of the sea surface temperature and the sea level pressure. Unal and Ghil [50] and Jevrejeva et al. [51] observed oscillations with periods 7–8.5

years in a number of sea level records. Feliks and Ghil [52] report the significant oscillatory mode with the 7.8 year period in the Nile River record, Jerusalem precipitation, tree rings and in the NAO index. Our first application of the enhanced MCSSA [22] yielded the observation of the mode with the period 7.8 years in near-surface air temperature from several European locations. Recently, the enhanced MCSSA analyses of the temperature data were refined and the analysis of the NAO index was added [31]. In the present work the number of processes containing the oscillatory mode with the period in the range 7–8 years was extended by the geomagnetic activity aa index and the sunspot numbers.

These findings give a solid basis for further research of relations among the dynamics reflected in the analysed data and thus between the solar and geomagnetic activity and the climate variability. The existence of oscillatory modes open the possibility to apply the recently developed synchronization analysis [53, 54] which already has found successful applications in studies of relations between atmospheric phenomena. Maraun & Kurths [55] discovered epochs of phase coherence between El Niño/Southern Oscillation and Indian monsoon, while Paluš & Novotná [39] demonstrated phase synchronization or phase coherence between the above mentioned QBO modes extracted from the temperature and the NAO index. The analysis of instantaneous phases of oscillatory processes allows to detect very weak interactions [53] and also causality relations if one oscillatory process drives the other one [56, 57]. In such analysis, Mokhov & Smirnov [58] have demonstrated that the NAO interacts with (or is influenced by) the other main global atmospheric oscillatory process – the El Niño Southern Oscillation. We believe that the synchronization analysis will help uncovering the mechanisms of the tropospheric responses to the solar and geomagnetic activity and contribute to a better understanding of the solar-terrestrial relations and their role in climate change.

Acknowledgements. The authors would like to thank the editor, R. Donner, and two anonymous referees for numerous comments and suggestions which helped to improve this chapter. This study was supported by the Grant Agency of the Academy of Sciences of the Czech Republic projects No. IAA3042401 and IAA300420805, and in part by the Institutional Research Plans AV0Z10300504 and AV0Z30420517.

References

1. J.B. Elsner, A.A. Tsonis, *Singular Spectrum Analysis. A New Tool in Time Series Analysis*. Springer, Berlin (1996)
2. N. Golyandina, V. Nekrutkin, A. Zhigljavsky, *Analysis of Time Series Structure. SSA and Related Techniques*, Chapman & Hall/CRC, Boca Raton (2001)

3. M.R. Ghil, M. Allen, M.D. Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi, P. Yiou, Advanced spectral methods for climatic time series. *Rev. Geophys.*, 40, 3-11–3-13 (2002)
4. M.R. Allen, A.W. Robertson, Distinguishing modulated oscillations from coloured noise in multivariate datasets. *Climate Dynamics*, 12, 775–784 (1996)
5. R. Vautard, M. Ghil, Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica D*, 35, 395–424 (1989)
6. M. Ghil, R. Vautard, Interdecadal oscillations and the warming trend in global temperature time series. *Nature*, 350(6316), 324–327 (1991)
7. C.L. Keppenne, M. Ghil, Adaptive filtering and the Southern Oscillation Index. *J. Geophys. Res.*, 97, 20449–20454 (1992)
8. P. Yiou, M. Ghil, J. Jouyel, D. Paillard, R. Vautard, Nonlinear variability of the climatic system, from singular and power spectra of Late Quarternary records. *Clim. Dyn.*, 9, 371–389 (1994)
9. M.R. Allen, L.A. Smith, Investigatings the origins and significance of low-frequency modes of climate variability. *Geophys. Res. Lett.*, 21, 883–886 (1994)
10. R. Vautard, P. Yiou, M. Ghil, Singular spectrum analysis: a toolkit for short noisy chaotic signals. *Physica D*, 58, 95–126 (1992)
11. M.R. Allen, L.A. Smith, Monte Carlo SSA: Detecting irregular oscillation in the presence of colored noise. *J. Climate*, 9(12), 3373–3404 (1996)
12. W. W. Hsieh, Nonlinear multivariate and time series analysis by neural network methods. *Rev. Geophys.*, 42, RG1003 (2004)
13. S. S. P. Rattan, W.W. Hsieh, Nonlinear complex principal component analysis of the tropical Pacific interannual wind variability. *Geophys. Res. Lett.*, 31(21), L21201 (2004)
14. W. W. Hsieh, Nonlinear principal component analysis of noisy data. *Neural Networks*, 20, 434–443 (2007)
15. W. W. Hsieh, A. J. Cannon, Towards robust nonlinear multivariate analysis by neural network methods. This volume.
16. M. Paluš, I. Dvořák, Singular-value decomposition in attractor reconstruction: pitfalls and precautions. *Physica D* 55, 221–234 (1992).
17. D.S. Broomhead, G.P. King, Extracting qualitative dynamics from experimental data. *Physica D*, 20, 217–236 (1986)
18. J.B. Gao, Y. Cao, J.-M. Lee, Principal component analysis of $1/f^{\alpha}$ noise. *Phys. Lett. A*, 314, 392–400 (2003)
19. L.A. Smith, Identification and prediction of low-dimensional dynamics. *Physica D*, 58, 50–76 (1992)
20. J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J.D. Farmer, Testing for nonlinearity in time series: the method of surrogate data. *Physica D* 58, 77–94 (1992)
21. M. Paluš, Testing for nonlinearity using redundancies: Quantitative and qualitative aspects. *Physica D*, 80, 186–205 (1995)
22. M. Paluš, D. Novotná, Detecting modes with nontrivial dynamics embedded in colored noise: Enhanced Monte Carlo SSA and the case of climate oscillations. *Phys. Lett. A*, 248, 191–202 (1998)
23. T.M. Cover, J.A. Thomas, *Elements of Information Theory*. J. Wiley & Sons, New York (1991)
24. Ya.G. Sinai, *Introduction to Ergodic Theory*. Princeton University Press, Princeton (1976)

25. I.P. Cornfeld, S.V. Fomin, Ya.G. Sinai, *Ergodic Theory*. Springer, New York (1982)
26. K. Petersen, *Ergodic Theory*, Cambridge University Press, Cambridge (1983)
27. Ya.B. Pesin, Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Math. Surveys*, 32, 55–114 (1977)
28. M. Paluš, Coarse-grained entropy rates for characterization of complex time series. *Physica D*, 93, 64–77 (1996)
29. M. Paluš, Kolmogorov entropy from time series using information-theoretic functionals. *Neural Network World*, 7(3), 269–292 (1997) (`http://www.cs.cas.cz/~mp/papers/rd1a.pdf`)
30. K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.*, 441, 1–46 (2007).
31. M. Paluš, D. Novotná, Enhanced Monte Carlo Singular System Analysis and detection of period 7.8 years oscillatory modes in the monthly NAO index and temperature records. *Nonlin. Proc. Geophys.*, 11, 721–729 (2004)
32. W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*. Cambridge Univ. Press, Cambridge (1986)
33. A. Arneodo, E. Bacry, J.F. Muzy, Random cascades on wavelet dyadic trees. *J. Math. Phys.*, 39(8), 4142–4164 (1998)
34. J.W. Hurrell, Y. Kushnir, M. Visbeck, Climate - The North Atlantic oscillation. *Science*, 291(5504), 603 (2001)
35. D. Rind, The Sun's Role in Climate Variations. *Science*, 296, 673–677 (2002)
36. E. Bard, M. Frank, Climate change and Solar variability: What's new under the sun? *Earth Planet. Sci. Lett.*, 248(1–2), 1–14 (2006)
37. R.P. Kane, Sun-Earth relation: Historical development and present status- A brief review. *Advances in Space Research*, 35(5), 866–881 (2005)
38. P.N. Mayaud, The aa indices: a 100year series characterizing the magnetic activity. *J. Geophys. Res.*, 77(34), 6870–6874 (1972)
39. M. Paluš, D. Novotná, Quasi-Biennial Oscillations extracted from the monthly NAO index and temperature records are phase-synchronized. *Nonlin. Proc. Geophys.* 13, 287–296 (2006)
40. C. Torrence, G.P. Compo, A practical guide to wavelet analysis. *Bull. Amer. Meteorological Society*, 79(1), 61–78 (1998)
41. M. Paluš, D. Novotná, Sunspot cycle: a driven nonlinear oscillator? *Phys. Rev. Lett.*, 83, 3406–3409 (1999)
42. M. Paluš, J. Kurths, U. Schwarz, N. Seehafer, D. Novotná, I. Charvátová, The solar activity cycle is weakly synchronized with the solar inertial motion. *Phys. Lett. A*, 365, 412–428 (2007)
43. M. Paluš, D. Novotná, P. Tichavský, Shifts of seasons at the European mid-latitudes: Natural fluctuations correlated with the North Atlantic Oscillation. *Geophys. Res. Lett.*, 32, L12805 (2005)
44. R. B. Govindan, D. Vjushin, S. Brenner, A. Bunde, S. Havlin, H.-J. Schellnhuber, Long-range correlations and trends in global climate models: Comparison with real data. *Physica A*, 294(1–2), 239–248 (2001)
45. R. B. Govindan, A. Bunde, S. Havlin, Volatility in atmospheric temperature variability. *Physica A*, 318(3–4) 529–536 (2003)
46. S. Jevrejeva, J. C. Moore, Singular Spectrum Analysis of Baltic Sea ice conditions and large-scale atmospheric patterns since 1708. *Geophys. Res. Lett*, 28(23), 4503–4506 (2001)

47. S.R. Gámiz-Fortis, D. Pozo-Vázques, M.J. Esteban-Parra, Y. Castro-Díez, Spectral characteristics and predictability of the NAO assessed through Singular Spectral Analysis, *J. Geophys. Res.*, 107(D23), 4685 (2002)
48. V. Moron, R. Vautard, M. Ghil, Trends, interdecadal and interannual oscillations in global sea surface temperatures. *Clim. Dyn.*, 14, 545 – 569 (1998)
49. E. D. da Costa, A. C. de Verdiere, The 7.7 year North Atlantic Oscillation. *Q. J. R. Meteorol. Soc.*, 128, 797 – 817 (2002)
50. Y. S. Unal, M. Ghil, Interannual and interdecadal oscillation patterns in sea level. *Climate Dynamics* 11, 255–278 (1995)
51. S. Jevrejeva, A. Grinsted, J. C. Moore, S. Holgate, Nonlinear trends and multiyear cycles in sea level records. *J. Geophys. Res.* 111, C09012 (2006)
52. Y. Feliks, M. Ghil, Interannual, synchronized oscillations over the Norh Atlantic, Eastern Mediterranean and Ethiopian Plateau. *Geophysical Research Abstracts*, 9, 05600 (2007)
53. A. Pikovsky, M. Rosenblum, J. Kurths, *Synchronization. A Universal Concept in Nonlinear Sciences*. Cambridge University Press, Cambridge (2001)
54. M. Paluš, Detecting phase synchronization in noisy systems. *Phys. Lett. A*, 235, 341–351 (1997)
55. D. Maraun, J. Kurths, Epochs of phase coherence between El Niño/Southern Oscillation and Indian monsoon. *Geophys. Res. Lett.*, 32(15), L15709 (2005)
56. M.G. Rosenblum, A.S. Pikovsky, Detecting direction of coupling in interacting oscillators. *Phys. Rev. E*, 64, 045202(R) (2001)
57. M. Paluš, A. Stefanovska, Direction of coupling from phases of interacting oscillators: An information-theoretic approach. *Phys. Rev. E*, 67, 055201(R) (2003)
58. I.I. Mokhov, D.A. Smirnov, El Niño-Southern Oscillation drives North Atlantic Oscillation as revealed with nonlinear techniques from climatic indices. *Geophys. Res. Lett.*, 33, L03708 (2006)

Phase Coherence Analysis of Decadal-Scale Sunspot Activity on Both Solar Hemispheres

Reik Donner

Institute for Transport and Economics, Dresden University of Technology,
Andreas-Schubert-Str. 23, 01062 Dresden, Germany,
e-mail: donner@vwi.tu-dresden.de

Abstract. Coherent or synchronous motion of oscillatory components is a feature of many geoscientific systems. In this work, we review and compare different possible approaches to detect and quantify the phase coherence between time series of oscillatory systems. In particular, methods originated in the theory of phase synchronisation phenomena and the concept of recurrence plots are considered. As a particular example, the sunspot activity on both solar hemispheres and the corresponding phenomenon of north-south asymmetry are studied. It is shown that this asymmetry can be understood in terms of a different "phase diffusion" of two coupled chaotic oscillators, which do however evolve coherently in time. The statistical reliability and implications of this result are discussed. Apart from the particular problem of sunspot activity, the methods described in this chapter may be used to study a variety of other phenomena in geoscientific systems, for example, the coherent motion of certain atmospheric oscillation patterns.

Keywords: Decadal-scale variability, phase coherence analysis, wavelet analysis, sunspots, north-south asymmetry

1 Introduction

Many solar and geo-physical processes are characterised by coherent oscillatory components in their dynamics. In particular, the solar activity on decadal time scales is clearly dominated by the so-called Schwabe cycle with an average period of about 11 years, which can be observed in terms of indicators like sunspot numbers, flare activity, or total solar flux. Any of these "sunspot cycles" is accompanied by a reversal of the polarity of the solar magnetic field, which means that the magnetic cycle of the Sun is dominated by a roughly 22-years period (Hale cycle). Detailed analyses of recent observations additionally indicate the presence of other distinct periodicities in the solar activity, ranging from short periods [1] to long-term components [2, 3, 4] like

the Gleissberg (period of about 80–100 years), Suess/de Vries (210 years), and Hallstatt (2300 years) cycles.

In general, it is known that on longer time scales, the quasi-regular Schwabe cycle is modulated by long-term fluctuations which affect both its amplitude and frequency. Observational data on sunspot numbers which record these variations are continuously available since the mid of the 19th century, and have been roughly reconstructed from distinct historical observations as well as climatological sources with a remarkably high temporal resolution for the last millenium. These reconstructions show that there were distinct periods of rather weak solar activity [5, 6], known as the Dalton (approx. 1790–1820 AD), Maunder (1645–1715 AD), Spörer (1420/50–1550/70 AD), Wolf (1280–1350 AD), and Oort (1040–1080 AD) minima. Most of these minima have been associated with certain climatic conditions on the Earth, for example, very cold winters in Europe during the so-called "little ice age" (showing its first climatic minimum at about 1650 AD) which coincides well with the Maunder and Dalton minima. Even for the time before 1000 AD, distinct historical sources [7] allow to determine time intervals of extraordinarily strong solar activity. Moreover, the observation that low solar activity is accompanied by a reduced net irradiation on the Earth surface and has therefore signatures in the climate system has motivated reconstructions based on high-resolution climate archives like tree rings or sediment as well as ice core records, which give indirect information about the activity during the past millenia. On even longer time scales, the abundance of certain cosmic isotopes (for example, ^{10}Be) in ice cores can be used to trace variations of the solar activity.

The variations of solar activity are known to trigger not only the long-term climate change itself. It is also known that there is a distinct feedback with the geomagnetic field, which itself influences the climate on large time scales. For the decadal-scale variability of the Sun (i.e., the "sunspot" cycle), various authors have reported that its signatures can be found in different parts of the climate system, from the lower troposphere [8] to the stratosphere [9], surface temperatures [10, 11, 12] and the precipitation activity and resulting lake-level standings in Central Africa [13]. Recently, it has been suggested that the hemispheric asymmetry of decadal-scale solar activity may have a certain importance for the Earth's atmospheric circulation [14]. In contrast to these findings, Moore et al. [15] have shown that major atmospheric oscillation patterns like the Quasi-Biennial Oscillation (QBO), the Arctic Oscillation (AO), and the El Niño Southern Oscillation (ENSO) are not directly linked to the Schwabe cycle. In addition, recent studies have proven that the present global warming cannot be attributed to a gradual increase of solar activity, which follows from the too small amplitude of solar activity variations [16] and an opposite direction of both signals [17] during the last decades.

In this chapter, some techniques will be reviewed that allow to trace the dynamic signatures of oscillatory variability on the Sun in different indicators as well as the Earth's climate system. The key feature to be studied is the

phase coherence of distinct oscillatory signals. It has to be mentioned that efforts have been made to also perform such studies without pronounced periodic signals (for example, in the case of the relationship between the El Niño activity and the strength of the Indian monsoon [18, 19, 20]), however, these efforts have not yet been fully convincing. In Sect. 2 of this chapter, some methods are introduced that have been recently suggested for investigating phase synchronisation or, more general, phase coherence phenomena in coupled oscillating systems. In Sect. 3, the dynamic features of the solar activity are studied in some detail, with a special emphasis on its decadal-scale variability. The asymmetry of both solar hemispheres is discussed in Sect. 4, whereas all results are summarised and discussed in Sect. 5.

2 Phase Coherence Analysis

During the past about 20 years, there has been an increasing interest in the study of synchronisation phenomena between coupled oscillatory systems in nature and society [21]. In general, the term synchronisation refers to a *process* of mutual adjustment of oscillations of two distinct, but coupled systems, leading from a non-coherent to a coherent motion of the oscillators, which may be periodic, quasi-periodic, or chaotic. Even without explicit oscillations, one may understand the emergence of a coherent motion of two coupled systems as a synchronisation phenomenon (generalised synchronisation). If the coupling between the systems is not bivariate, it is however more reasonable to speak of a locking instead of synchronisation.

Among the different types of synchronisation phenomena, the emergence of phase synchronisation, i.e., a coherent motion with a fixed ratio of average frequencies, is particularly relevant for the understanding of many situations. However, for a synchronisation phenomenon in the strict sense, a clear distinction between two coupled systems is required. If this is not the case (for example, when studying two observables of the same system), one should rather speak of phase coherence between the considered oscillations. In the context of time series analysis, this may lead to an identification problem: If there is no sufficient information about the particular structure of an observed system, a phase-coherent motion can be interpreted in terms of two coupled self-sustained oscillators (i.e., phase synchronisation), two components or observables of one oscillatory system, one observable viewed through two (nonlinear) observation functions, etc.

In the following, different approaches to quantify the degree of phase coherence or phase synchronisation based on time series analysis will be reviewed. For the presented methods, sufficiently stationary conditions are assumed, i.e., temporal variations in the presence or degree of phase coherence are not explicitly considered. In order to study instationary phase coherence, the presented approaches can be applied in terms of a piecewise analysis of the observations,

given a sufficiently high temporal resolution compared to the time-scale on which the corresponding changes occur.

2.1 Phase Definition and Phase Coherence

The classical approach of studying the coherent motion of coupled oscillatory systems considers the spectral coherence, i.e., the existence of oscillations with the same frequency in two or more time series. However, this approach requires stationary conditions, i.e., the presumed frequency must contribute at all times with equal strength. In real-world systems (in particular, in the geosciences, but also in physiological systems), this assumption is usually violated, which calls for time-sensitive generalisations like wavelet coherence. In contrast to these "frequency coherence" concepts, phase coherence (or phase synchronisation) analysis refers to the *instantaneous* frequency of oscillations (i.e., the time derivative of a suitably defined phase variable), which may change with time. This generalisation allows to study also the joint behaviour of rather complex systems like chaotic oscillators.

According to the above mentioned paradigm of instantaneous frequencies, traditional phase coherence analysis is based on the proper definition of phases. For this problem, there is a variety of different approaches:

- Poincaré sections of the (possibly embedded) time series may be used to define points in time that correspond to fixed phase values $2k\pi$ with $k \in \mathbb{N}$ [21]. Between these points, the phase variable is defined via interpolation, for example, using piecewise linear functions.
- If there is a pronounced oscillation (whose frequency and amplitude may still vary with time), the analytic signal approach has become a standard way of defining a phase variable. In this framework, the Hilbert transform

$$Y(t) = \frac{1}{\pi} \mathcal{P.V.} \int_{-\infty}^{\infty} \frac{X(\tau)}{t-\tau} \, d\tau \tag{1}$$

  is used to continue a scalar signal X into the complex plane ($Z = X + iY$) [22], where a phase variable can be easily assigned as

$$\phi(t) = \arctan\left(\frac{Y(t)}{X(t)}\right). \tag{2}$$

- In order to make the analytic signal approach being applicable to noisy or slightly non-coherent oscillators as well, it has been shown that a geometric phase definition based on the local derivatives of the time series X and its Hilbert transform Y might be helpful *(curvature method)* [23, 24, 25]

$$\phi(t) = \arctan\left(\frac{\dot{Y}(t)}{\dot{X}(t)}\right). \tag{3}$$

- As a generalisation of the frequency coherence approach, one may use a Wavelet transform [26, 27, 28, 29, 30, 31] or other methods of time-frequency analysis like the Gabor transform [32] to select the (temporally variable) strength of oscillations on a fixed reference frequency in terms of both amplitude and phase. If the frequency of the dominating oscillation pattern varies significantly itself, one may also consider these variations using the phases belonging to the frequency with the largest local oscillation amplitude. As a further alternative, coherent oscillatory modes with possibly variable frequencies can be selected using empirical mode decomposition [33, 34], bandpass filtering in the Fourier space [35], independent component analysis [36] or other suitable methods, before being used for defining a phase variable by the analytic signal approach.

Having thus defined the phase variables for two time series, one may compare the joint evolution of the phases of both systems. In general, the phases increase with an average rate corresponding to the average frequency of the systems. If this average frequency is the same in both systems, their phase difference $\delta\phi(t) = \phi_2(t) - \phi_1(t)$ is bounded. In a more general way, one may define $m : n$ phase synchronisation if the average frequencies of both systems have the ratio $m : n$ with m and n being integers. Remaining phase differences correspond to different fluctuations of the instantaneous frequencies around their average values, corresponding to the so-called *phase diffusion*. This term is motivated by the fact that for chaotic oscillators, the dynamics of the phase after substracting a linear increase according to the average frequency may resemble a stochastic diffusion process [21].

According to these general considerations, measures of phase coherence may be defined based on either the statistical properties of the phase differences or the joint evolution of the phases. Whereas the latter approach is realised in terms of the mutual information between the phase variables [37], suitable statistical properties of the phase differences are their standard deviation, normalised Shannon entropy [38], or the Rayleigh measure *(mean resultant length)*

$$r = \frac{1}{N}\left|\sum_{i=1}^{N} e^{i\delta\phi(t_i)}\right|, \tag{4}$$

where N is the total length of the considered time series. As the last quantity has a direct interpretation in terms of directional statistics, it is convenient to use it as a corresponding measure in all further analyses. It has to be mentioned that the power of the different statistics may be very different. In order to statistically test for the presence of phase synchronisation, bootstrap approaches have been proposed [39]. However, although (frequency) and phase coherence methods are usually sensitive, it has been shown that they are not specific in distinguishing between coupled self-sustained oscillators and time series being connected by a certain transfer function [40], which is closely related to the above mentioned identification problem.

2.2 Recurrence Plots: Phase Coherence Analysis Without Phases

As an alternative to the "traditional" phase synchronisation analysis described above, one may use a topological approach which is based on the concept of *recurrence plots* [41, 42]. This method has originally been designed as a tool for visualising the correlation pattern within a single time series comparing the values at all times t_i with all observations at other times t_j. A simple graphical representation is obtained by comparing the difference between the distances of every pair of values and a prescribed threshold value ϵ, which is then binarily encoded according to the corresponding order relationship. Mathematically, this leads to the so-called *recurrence matrix*

$$R_{ij} = \Theta(\epsilon - ||X_i - X_j||) \tag{5}$$

(where $\Theta(\cdot)$ is the Heaviside function, $X_i = X(t_i)$, and $||\cdot||$ some suitable norm), which depends on the specific choice of ϵ. Note that in order to enhance the meaning of non-zero entries as representatives of a similar dynamics, in Eq. (5), the time series X is often replaced by a suitably embedded version of itself. The graphical visualisation of R in terms of a black-white structure is called a recurrence plot. As a particularly remarkable feature, it follows from the definition that $R_{ii} = 1$ independent of ϵ. The corresponding main diagonal structure in the plot is called the *line of identity (LOI)*. Besides the intuitive interpretation of the emerging structures in a recurrence plot in terms of a similar dynamics, statistics on the distributions of continuous vertical and diagonal structures allow to estimate a variety of dynamic invariants [42].

The concept of recurrence plots may be extended to study the joint evolution of two coupled systems. If X and Y are two time series reflecting the dynamics of the same physical quantity, *cross-recurrence plots* are defined as [43]

$$CR_{ij} = \Theta(\epsilon - ||X_i - Y_j||). \tag{6}$$

In contrast to univariate recurrence plots, the line of identity is not anymore present in cross-recurrence plots. However, if the dynamics of X and Y is topologically similar, other continuous structures emerge as both considered times t_i and t_j increase. The corresponding structure next to the main diagonal, the *line of synchronisation (LOS)* [44], can be understood as a representative of the phase difference between the two systems.

Recurrence plots may be used to detect different types of synchronised dynamics. In particular, signatures of phase synchronisation may be isolated without explicitly defining a phase variable [45, 46]. For this purpose, one considers the diagonal-wise recurrence rate

$$\hat{p}^X(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \Theta(\epsilon - ||X_i - X_{i+\tau}||), \tag{7}$$

which may be understood as a *generalised auto-correlation function* of the time series X [47, 48]. One may convince oneself that the presence of phase

synchronisation between two time series X and Y requires that the resulting generalised auto-correlation functions show the same behaviour. According to this, the correlation coefficient between these two functions,

$$CPR = \left\langle \bar{p}^X(\tau)\bar{p}^Y(\tau) \right\rangle_\tau \tag{8}$$

(where $\bar{p}$ stands for the function p standardised to zero mean and unit variance), can be used as an indictor of phase synchronisation. In order to test the significance of this measure, the concept of *twin surrogates* is used. Here, one of the original time series is used to successively generate independent, topologically consistent replications of the time series within which random points are replaced by others with the same local neighborhood, i.e., the same column in the recurrence plot [49]. Bootstrapping one of the systems in this way and computing the CPR indices with respect to the original time series of the other system for every realisation, one approaches a distribution of index values. Comparing the CPR value obtained from the original two systems with this distribution, one may conclude about the statistical significance of CPR and, hence, about the presence of phase synchronisation. Note that a significant difference requires the presence of remaining phase differences between both systems, i.e., a different "phase diffusion". If this is not the case (for example, in the case of complete synchronisation), it is not anymore possible to use this approach to test for the presence of phase synchronisation.

2.3 Example: Two Coupled Roessler Oscillators

To illustrate the performance of different aproaches to phase synchronisation analysis, we review the paradigmatic example of two chaotic Roessler oscillators that are diffusively coupled via their y-coordinates [25, 45, 50]:

$$\dot{x}_{1,2} = -\omega_{1,2}y_{1,2} - z_{1,2}, \tag{9}$$

$$\dot{y}_{1,2} = \omega_{1,2}x_{1,2} + ay_{1,2} + \mu(y_{2,1} - y_{1,2}), \tag{10}$$

$$\dot{z}_{1,2} = b + z_{1,2}(x_{1,2} - c). \tag{11}$$

The Roessler system is known to show chaotic oscillations on a relatively broad frequency band. Moreover, there are two distinct time scales of the dynamics, which correspond to the short-term fluctuations of the oscillation amplitude (with high values of entropy and fractal dimension) and the long-term phase diffusion (with lower values of entropy and fractal dimension) [51]. As it will be shown later, this superposition of short- and long-term dynamics can be seen as an analogy of the sunspot activity which varies in a roughly similar way.

Tuning the parameter a of the Roessler system to a certain range of values, there is a transition from coherent oscillations (i.e., oscillations around a well-defined origin in the x-y plane) to the non-coherent Funnel regime. As in earlier studies [25], the following parameters have been chosen as an illustrative example for the performance of different methods for detecting and

quantifying phase coherence: $\omega_1 = 0.98$, $\omega_2 = 1.02$, $b = 0.1$, and $c = 8.5$. For the parameter a, values of 0.16 (standard phase-coherent Roessler system) and 0.2925 (non-coherent Funnel regime) have been used.

In the phase-coherent regime, the standard Hilbert transform approach is well suited to define a meaningful phase variable. In order to illustrate the performance of the wavelet-based approach, a coupling strength of $\mu = 0.05$ has been applied, which corresponds to a phase synchronised system [45]. In Fig. 1, the results of a scale-resolved phase synchronisation analysis are shown

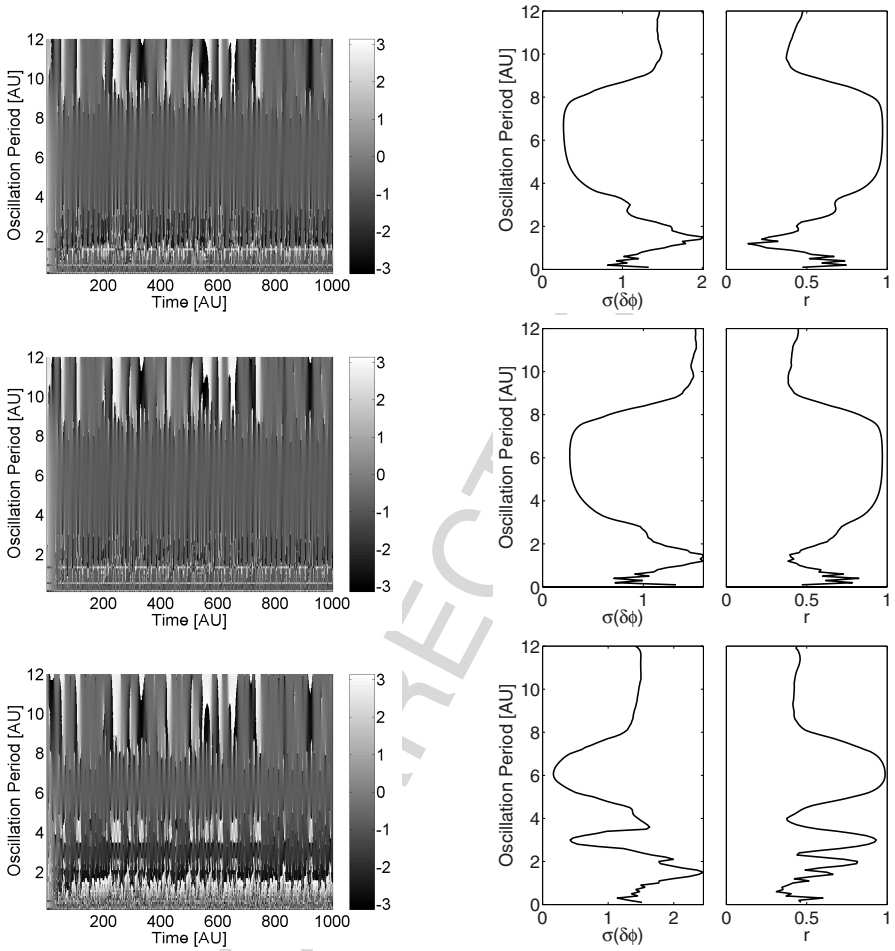


Fig. 1: Left panels Normalised phase differences (within the interval $[-\pi, \pi]$) between the phases of the two Roessler systems (coupled with $\mu = 0.05$) estimated with a complex Morlet wavelet on different scales for the respective x, y, and z components (from top to bottom). Right panels: Resulting standard deviations $\sigma(\delta\phi)$ (left) and mean resultant lengths r (right) of the phase differences as a function of the considered reference period.

for all three components of both systems. One may see that there is a broad continuum of frequencies on which phase synchronisation may be detected. In the case of the z-components, this range is split into two frequency bands of coherent motion in the phase variables.

Considering the transition from non-synchronised to phase synchronised and completely synchronised conditions (see Fig. 2), one may observe that the applied methodology is only capable to identify phase synchronisation.

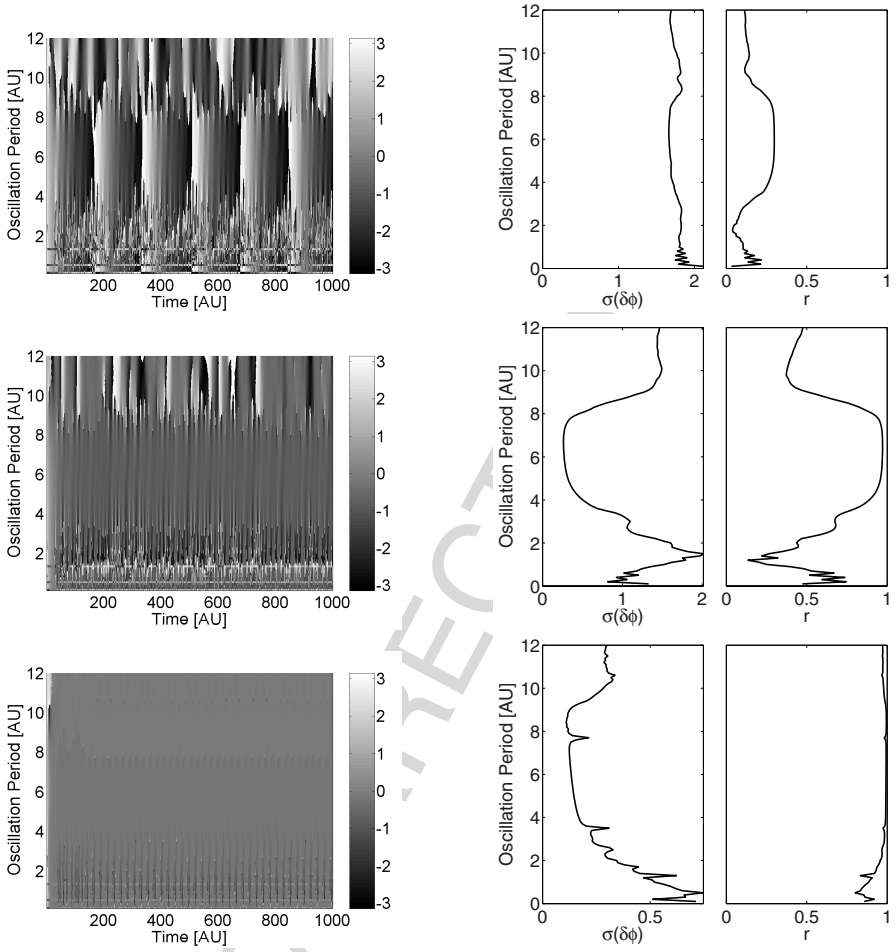


Fig. 2: Left panels Normalised phase differences between the phases of the x variables of the two coupled Roessler systems (estimated with a complex Morlet wavelet on different scales) for coupling strengths of $\mu = 0.02$ (no synchronisation), $\mu = 0.05$ (phase synchronisation), and $\mu = 0.2$ (complete synchronisation) (from top to bottom). Right panels: Resulting standard deviations $\sigma(\delta\phi)$ (left) and mean resultant lengths r (right) as a function of the considered reference period.

Indeed, this is also true for other methods of phase synchronisation analysis. For example, Romano [50] reported that the recurrence plot based CPR index shows values of 0.115 for $\mu = 0.02$ (non-synchronised systems) and 0.998 for $\mu = 0.05$ (phase synchronisation). The transition between low and high values of this index was found to be relatively sharp for the considered parameters. Looking in some more detail at the involved frequencies as resolved by the wavelet based method, it turns out that whereas the phase coherence of the two chaotic attractors is restricted to a certain frequency band, a corresponding coherence is found for a much broader range of reference periods in the case of complete synchronisation.

In contrast to the regime of coherent chaotic oscillations, in the Funnel regime, a much wider range of periodicities is present in the wavelet spectrograms. Applying the wavelet based approach to phase synchronisation, it turns out that the frequency-dependence of the phase synchronisation index r can hardly be distinguished from the signature of complete synchronisation in the case of coherent oscillations (see Fig. 3). This result shows that the width of a

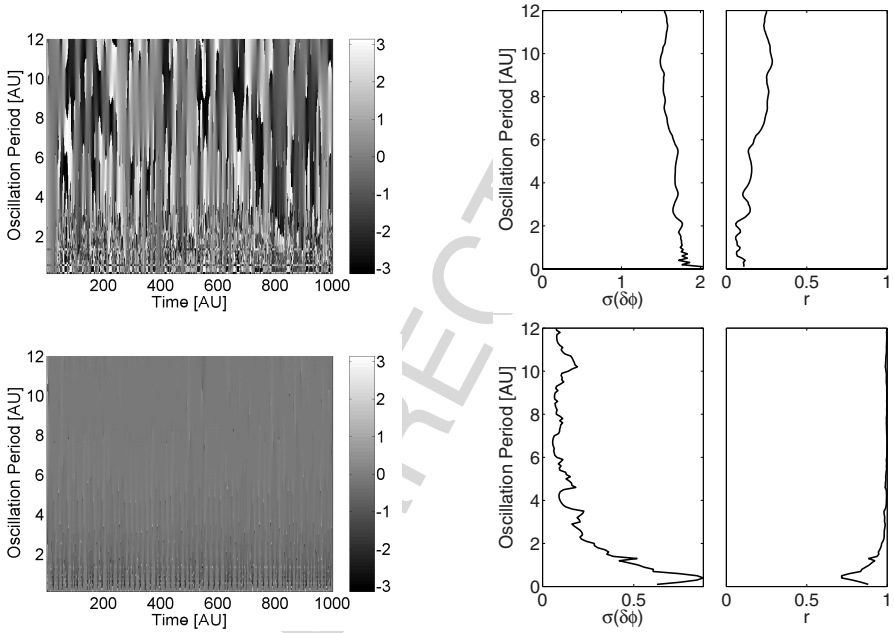


Fig. 3: Left panels Normalised phase differences between the phases of the x variables of the two coupled Roessler systems in the Funnel regime (estimated with a complex Morlet wavelet on different scales) for coupling strengths of $\mu = 0.05$ (no synchronisation, top) and $\mu = 0.2$ (phase synchronisation, bottom). Right panels: Resulting standard deviations $\sigma(\delta\phi)$ (left) and mean resultant lengths r (right) as a function of the considered reference period.

possible coherent frequency range can hardly be considered as an indicator for the transition from phase to complete synchronisation of chaotic oscillators.

3 Decadal-Scale Solar Activity

3.1 Description of the Data

The decadal-scale variability of solar activity can be observed in a variety of different observables, including sunspot numbers and areas, the 10.7-cm radio flux, and the total solar irradiation arriving at the Earth. This study will exclusively focus on sunspot observations. As different measures, the total sunspot areas A (in units of millionths of a hemisphere), their values A^n and A^s for the northern and southern hemisphere of the Sun, and the group and international (Wolf, Zurich) sunspot numbers are considered. These quantities have the advantage that observational data are available which cover a sufficiently large amount of time (i.e., several solar activity cycles).

Whereas the sunspot areas are absolute quantities, sunspot numbers are relative index values. If g and n are the numbers of identified sunspot groups and individual spots, respectively, the international sunspot number R_I (or R_Z) is defined as

$$R_I = k(10g + n), \tag{12}$$

where k is a correction factor for the considered observer. In a similar way, but with a slightly inconsistent normalisation, the american sunspot numbers R_A can be considered [52]. As a more objective measure, Hoyt and Schatten [53] introduced the group sunspot number R_G using data from different observers $i = 1, \ldots, N$ as

$$R_G = 12.08 \, \frac{1}{N} \sum_{i=1}^{N} k_i \; g_i. \tag{13}$$

The above definition allowed to extend the time series of relative sunspot numbers back until 1610. For this purpose, in addition to the observations made at the Zurich observatory, other references have been taken into account. By the corresponding extension of the sunspot time series, studies based on direct observations of the solar activity during the Maunder minimum became possible [54, 55, 56]. In addition to the numbers given for the entire Sun, hemispherically resolved values of the international sunspot numbers (hereafter called R_I^n and R_I^s, respectively) are available since 1992. An extended catalog for the time interval 1945–2004 has been provided by Temmer et al. [57, 58], combining observations from two Austrian and Slovakian observatories that have been normalised to be consistent with the values of R_I. For convenience, in this contribution, the relative sunspot numbers from this catalogue will be referred to as R^n and R^s, respectively.

The main statistical characteristics of the different relative sunspot numbers, sunspot areas, and the 10.7-cm radio flux have been reported to be

Table 1: Time coverage of the sunspot data used in this study.

Quantity	Coverage (daily)	Coverage (monthly)	Source
R_I	1849(1818)-present	1749-present	[A]
$R_I^{n,s}$	1992–2006	1992–2006	[A]
R_A	1944-present	1944-present	[A]
R_G	1796(1610)–1995	1610–1995	[A]
$R^{n,s}$	1945–2004	1945–2004	[B]
A	—	1874-present	[C]
$A^{n,s}$	—	1874-present	[C]

[A] http://www.ngdc.noaa.gov/stp/SOLAR/ftpsunspotnumber.html
[B] http://cdsweb.u-strasbg.fr/cgi-bin/qcat?J/A+A/447/735
[C] http://solarscience.msfc.nasa.gov/greenwch.shtml

consistent with each other [4, 59]. In the following, recent results on the linear as well as non-linear statistical properties of sunspot time series will be reviewed and further extended. In particular, it will be examined on which scales and up to which degree the aforementioned observables are *actually* phase-coherent.

3.2 Signatures of Low-Dimensional Chaos

In his seminal work on auto-regressive spectral estimation, Yule [60] described the time series of relative sunspot numbers as a disturbed harmonic process, considering the perturbation being an auto-regressive process of second order. In later work, the presence of exclusively linear-stochastic behaviour in the sunspot record has been excluded by surrogate data testing [61]. Figure 4 shows the significance of the time irreversibility (Q) statistics for the monthly record of the international sunspot number. Here, a significance value of S means that the normalised cubed difference $Q(\tau) = \left\langle (X_{t+\tau} - X_t)^3 \right\rangle / \left\langle (X_{t+\tau} - X_t)^2 \right\rangle$ of the original time series lies outside of the S-fold standard deviation of the corresponding values obtained from a set of CAAFT surrogates [62]. The presented results for the time series of annual averages are in excellent agreement with the findings of Theiler et al. [61]. If however the correlation dimension is used as discriminating statistics, Theiler *et al.* found that the sunspot data are indiscernible from the AAFT surrogates. This underlines the importance of a proper choice of the statistics in testing against linear-stochastic dynamics.

It has to be pointed out that the significance of the test using the Q statistics shows almost no dependence on the resolution of the time series, in particular, moving average filtering has no significant effect unless variations are smoothed on a scale of several years. This suggests that the observed signature of nonlinearity is not an effect of short-term fluctuations, but rather of

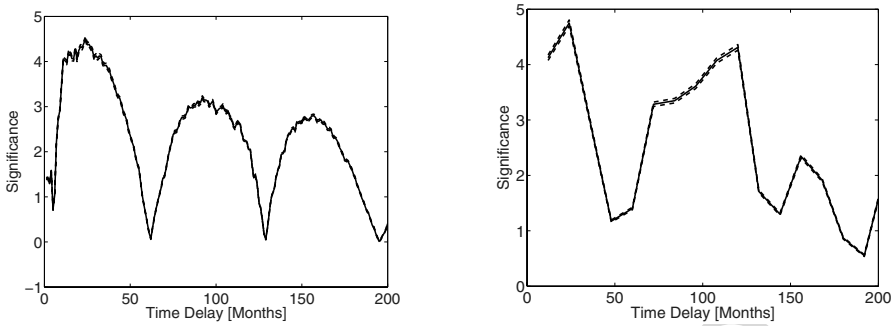


Fig. 4: Significance $S(\tau) = |(Q^{orig}(\tau) - \mu_Q(\tau))|/\sigma_Q(\tau)$ (where $Q^{orig}(\tau)$ is the normalised cubed difference of the original time series, and $\mu_Q(\tau)$ and $\sigma_Q(\tau)$ are the mean and standard deviation of the same statistics computed over $N_{surr} = 1000$ realisations of the CAAFT surrogate algorithm (left panel) and of the Barnes model after transforming its data to a Gaussian distribution (right panel)) of the test against linear-stochastic behaviour in the monthly international sunspot numbers. Dashed lines very close to the main solid curve indicate error estimates of the significance given by $\Delta S(\tau) = \sqrt{(1 + S^2(\tau)/2)}/N_{surr}$ [61]. Note that the Barnes model relates to the annual mean sunspot numbers, which leads to a remarkably lower temporal resolution in the plot.

the decadal-scale variability dominating the record. This hypothesis is further underlined by the fact that the quantity $(Q^{orig} - \mu_Q)/\sigma_Q$ (whose absolute value is just the significance index considered here) shows (as a function of the time shift τ) a periodicity of about 11 years, which corresponds to the average period of the solar Schwabe cycle.

It has to be underlined that the violation of the hypothesis of a stationary linear-stochastic process may be explained by nonstationarities, nonlinear stochastic models, or deterministic chaos. A more detailed discrimination between these alternatives requires refined statistical techniques and is beyond the scope of this work. Considering the hypothesis of a nonlinear stochastic process, Barnes et al. [63] modelled the variability of annual mean sunspot numbers by a narrowband Gaussian stochastic process, which was defined by a nonlinear mapping $Y_n = Z_n^2 + \alpha(Z_{n-1}^2 - Z_{n-2}^2)^2$ where Z_n was assumed to be an ARMA[2,2] process. Realisations of this model show variability patterns that resemble those of the sunspot activity with quasi-regular cycles and superimposed long-term amplitude variations. However, a more detailed analysis reveals that this model is very likely not capable to reproduce the sunspot activity time series of the last 300 years, which can be shown by either considering the corresponding amplitude-frequency relationship [64] or the time irreversibility measure Q as a discriminating statistics for both original (annual) sunspot numbers and realisations of the Barnes model. In particular, the qualitative behaviour of the corresponding significance levels resembles

strongly that obtained with CAAFT surrogates (cf. Fig. 4). As a potential improvement, Tsai and Chan [65] found that a fourth-order threshold autoregressive (NLCAR[4]) model may explain the yearly sunspot record much better than the Barnes model.

The above mentioned approaches to modelling the dynamics of the sunspot activity have been based on the assumption of a process that can be exclusively described by stochastic fluctuations. With the development of the theory of nonlinear dynamic systems, increasing evidence was found that at least the decadal-scale solar activity cycle is however better described by deterministic chaotic processes. Various authors considered measures like fractal or correlation dimension, Lyapunov exponents, or entropies of the sunspot time series to characterise the complexity and predictability of the underlying attractor, e.g. [66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83]. Paluš and co-workers provided statistical evidence that the corresponding process can be understood as a driven nonlinear oscillator, which is weakly synchronised with the solar internal motion [84].

Watari [85] showed that a suitable decomposition into components which vary on different time scales allows a separation of periodic, chaotic, and random components. His results are in qualitative agreement with the findings of Qin [86], who reported that low-dimensional chaos can be found on time scales of 8 years or longer, whereas the behaviour on shorter time scales has to be described by high-dimensional chaos or stochastic processes. In order to further validate these findings, some additional analyses have been carried out. For this purpose, the full record of monthly international sunspot numbers has been subjected to a wavelet decomposition in order to extract components that vary on periods of 1–240 months. As a next step, the recurrence matrices of the resulting time series have been computed and used for estimating different nonlinear quantitative measures, which are based on statistics of the lengths of continuous diagonal as well as vertical structures in the recurrence plots.

In Fig. 5, some of the results are shown. All measures indicate that there is a broad range of reference periods between about 5 and 15 years, on which the recurrence plots show strong dynamic regularities suggesting mainly deterministic processes. For shorter as well as longer time scales, the degree of determinism is significantly smaller, which supports the results of Qin [86]. Following [82] where it was pointed out that the z-component of a phase-coherent Roessler oscillator resembles the dynamics of the sunspot activity, the same computations have also been carried for the system studied in Sect. 2.3. Whereas the existence of a broad frequency band with a high degree of determinism is indeed similar for both systems, the Roessler attractor also shows a very high degree of determinism on short time scales, which differs from the behaviour of the sunspot number time series. Hence, one may conclude that the short-term behaviour (one month to some years) of the observed solar activity is actually dominated by a dynamics that cannot be described by a low-dimensional chaotic attractor.

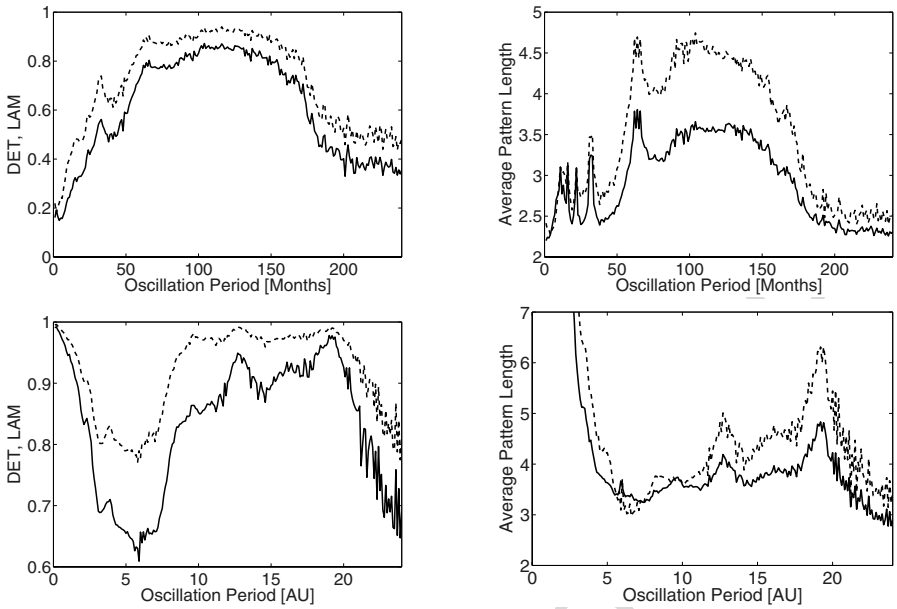


Fig. 5: Upper panels Left: Degree of determinism (solid) and laminarity (dashed) of the wavelet filtered monthly international sunspot numbers R_I as a function of the considered reference period. Right: Corresponding values for the average lengths of diagonal (solid) and vertical (dashed) structures in the recurrence plots. Lower panels: The same quantities computed for the z-component of the first Roessler system from Sect. 2.3 in the phase synchronised coherent state ($\mu = 0.05$). High values indicate a more deterministic dynamics of the system.

3.3 Phase Coherence of Different Sunspot Observables

In order to check the consistency between the daily values of the international sunspot numbers and the more extended Austrian-Slovakian composite catalogue, Temmer et al. [58] studied the corresponding scatter plots. Although a remaining scatter can be observed in the daily data, a linear correlation coefficient of 0.99 indicates that both quantities actually coincide very well. However, comparing the scatter between the monthly sunspot areas and the respective sunspot numbers on both hemispheres, it turns out that the scatter is much larger (see Fig. 6). In particular, the linear correlation coefficient between sunspot areas and numbers is "only" 0.968 (0.860 and 0.824 for northern and southern hemisphere, respectively) in the case of the international sunspot numbers (with only 15 years of joint coverage) and 0.974 (0.958 and 0.956 for both hemispheres) in the case of the extended catalogue (60 years). Although these coincidences appear to be quite reliable, they leave some space for possible inconsistencies.

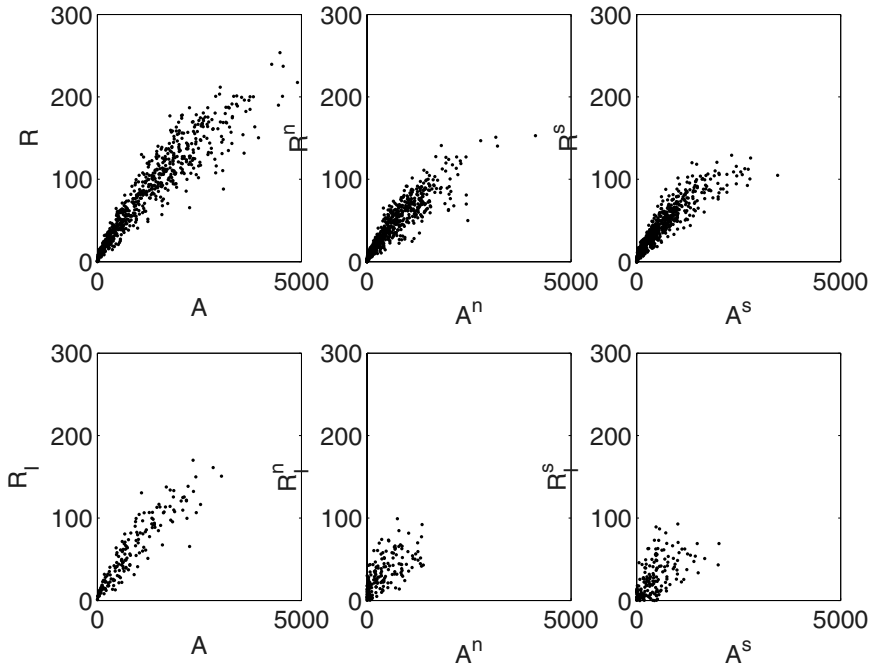


Fig. 6: From left to right: Comparison of the total, northern, and southern hemispheric sunspot areas (x-axis) and sunspot numbers (y-axis) for the extended catalogue of [58] (top, 1945–2004) and the international sunspot number (bottom, 1992–2006).

In order to examine whether there are differences in the joint dynamics, the behaviour of the corresponding phases has been compared for different reference frequencies, using the wavelet decomposition as described in Sect. 2. The results presented in Fig. 7 clearly demonstrate that the dynamics of the different observables at time scales within a range of about 8–14 years can be considered phase-coherent. In this interval, the phases of sunspot numbers and areas do very well coincide with each other (see Fig. 8), with a maximum phase shift of about 3 months over the last 60 years (which means less than 3% of the average duration of a sunspot cycle). A more detailed inspection reveals that within this time interval, there has been a gradual change from conditions where the phase variations of the sunspot areas occur earlier than those of the sunspot numbers towards the opposite conditions. Looking on longer time scales using the continuous record of the international sunspot number, it turns out that this successive trend indeed set in around 1950. Considering the history of the last about 130 years, one finds that before 1890 and since about 1970, the decadal-scale variations of the sunspot numbers

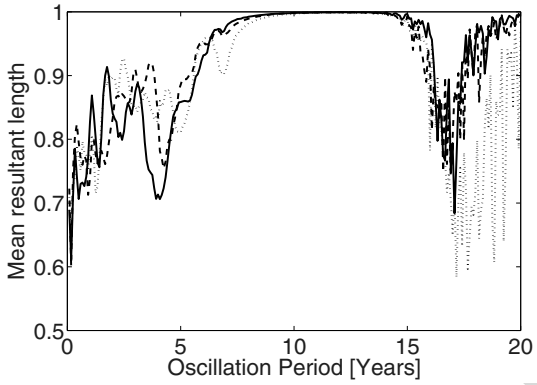


Fig. 7: Frequency-dependent phase coherence between the monthly sunspot numbers from the Austrian-Slovakian composite catalogue and the sunspot areas for the entire Sun (solid lines) and the values for the northern (dashed) and southern (dotted) hemispheres.

occur earlier than those of the sunspot areas (however, these numbers depend slighty on the considered time scales).

In contrast to the longer periods, on shorter time scales, there are remarkable deviations in the phases of sunspot numbers and areas (see Fig. 7). In particular, one has to conclude that the relationship between both types of observables cannot be exclusively described by a monotonous (possibly nonlinear) transformation, but involves irregular short-term contributions. The presented analysis does not allow to reveal whether the relative scatter between sunspot numbers and areas due to fluctuations on smaller time scales is mainly an effect of "observational noise" (which would be smoothed out when going to longer time scales) or of dynamically relevant deterministic or stochastic processes that act on short scales in both time and space.

4 The North-South Asymmetry of Solar Activity

Since observational records with a sufficient time coverage have become available, there has been an increasing interest in the spatial structure of sunspot activity. Among the first to study the corresponding hemispheric asymmetry, Newton and Milsom [87] introduced a very simple asymmetry index,

$$NA = \frac{N - S}{N + S}, \tag{14}$$

where N and S are the values of the respective observables (i.e., sunspot areas or numbers) on the northern and southern hemisphere of the Sun. They found

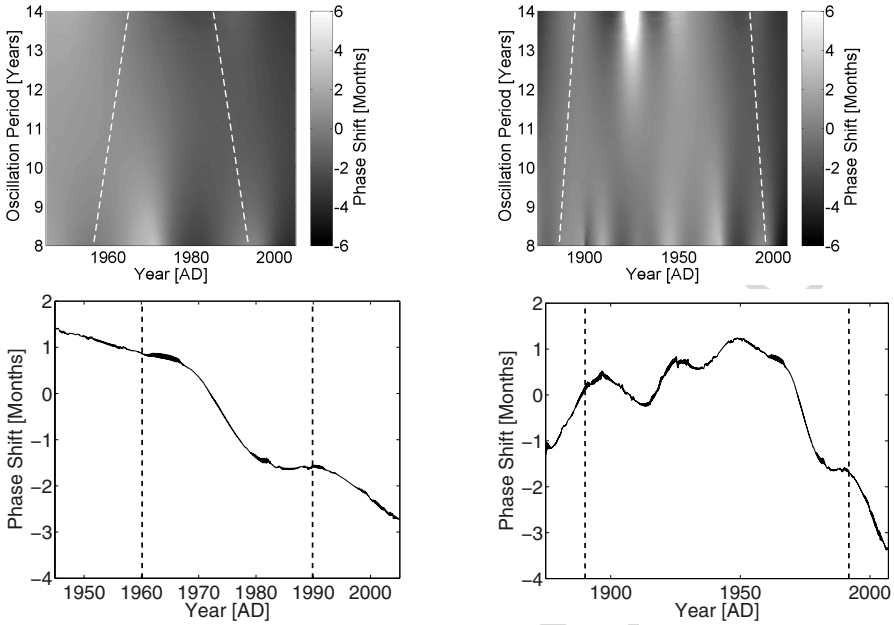


Fig. 8: Estimated phase shift between monthly sunspot areas and sunspot numbers from the Austrian-Slovakian composite catalogue (left panel, 1945–2004) and the international sunspot numbers (right panel, 1875–2006) for different reference frequencies. The upper panels represent the results for all reference time scales between 8 and 14 years, whereas in the lower panels, only the phase shift on a scale of $T = 10.75$ years is displayed. All plotted curves have been subjected to a one-year moving average filter for smoothing. Dashed lines correspond to the cone of influence (with a width of $\sqrt{2}T$ on a scale T for a Morlet wavelet [110]), outside which the results of wavelet analysis are biased due to boundary effects.

that the asymmetry "changes from cycle to cycle, which although not random, appear to have no definite period". In addition to their results, Waldmeier [88] provided evidence that the phase shift between the activity on both hemispheres is another important parameter, which may have a certain influence on the values of NA. The interrelationship between the variations of the north-south asymmetry and major solar flares has been addressed by a number of authors [89, 90, 91, 92]. More detailed statistical studies on the long-term variability were subsequently published and subjected to intensive discussions [93, 94, 95, 96, 97, 98, 99, 100, 101, 102]. Whereas some authors focussed on the question whether there are significant periodicities inside the asymmetry index [103, 104, 105, 106, 107, 108], increasing efforts have been made to study whether the variations of the asymmetry can be understood as a chaotic process [105, 109]. Ballester et al. concluded that the results of

the corresponding studies may be statistically more reliable, if the absolute asymmetry

$$AA = N - S \tag{15}$$

is considered instead of its normalised variant NA.

Most of the aforementioned studies have considered the sunspot areas, whereas less results have been obtained based on sunspot numbers due to the worse time coverage of the corresponding hemispherically resolved records [58, 93, 100, 109]. In this section, the meaning of phase shifts between northern and southern hemispheric activity (i.e., the corresponding sunspot areas or numbers) for the north-south asymmetry will be further examined, following the corresponding ideas originally pointed out by Waldmeier. The approaches that are used in the following potentially yield new measures of the north-south asymmetry, which do not directly compare to NA or AA in that they are based on a fundamentally different assumption, namely the importance of phase differences between the oscillations on both hemispheres. A thorough comparison of the variability of both traditional and phase-based quantities will be subject to future studies.

4.1 Scale-Resolved Phase Coherence Analysis

In a recent paper [111], the wavelet-based approach to phase coherence analysis has been applied to the full record of monthly values of hemispheric sunspot areas between 1874 and 2006. It has been shown that in a range of about 8–14 years, the variations of sunspot areas on both solar hemispheres allow the definition of proper phase variables. Figure 9 underlines that this result holds for both sunspot areas and sunspot numbers. The observed interval of reference periods on which phase coherence is found matches well the results of the previous section on the phase coherence between different observables as well as the presence of low-dimensional chaos [86].

Within the coherent range, the phase variables and their corresponding differences vary on rather long time-scales. In particular, it has been shown that before about 1925 and since about 1965, the sunspot activity in the northern hemisphere occurred earlier than that in the southern hemisphere, whereas in the meantime, the opposite conditions were present.

In Fig. 10, the phase difference time series for the hemispheric sunspot areas and numbers are shown. In particular, the behaviour on a reference time scale of $T = 10.75$ years is shown, for which the drift of the phase variables is minimised (average frequency) [111]. The corresponding results for the sunspot areas have already been presented in [111], however, the remaining short-term fluctuations used here are significantly smaller. The reason for this is that in [111], the phases have been obtained via a Hilbert transform of the real part of the wavelet filtered series (which may involve some numerical errors), whereas in this contribution, both real and imaginary parts from the wavelet decomposition have been directly used for computing the phases.

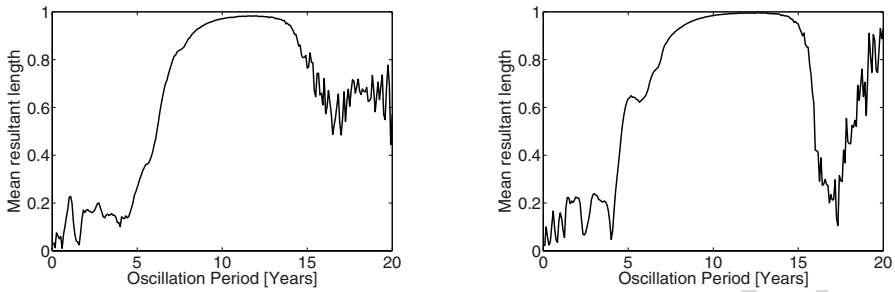


Fig. 9: Frequency-dependent phase coherence between the sunspot areas $A^{n,s}$ (left panel) and between the sunspot numbers from the composite catalogue $R^{n,s}$ (right panel) on both solar hemispheres, quantified by the mean resultant length (4).

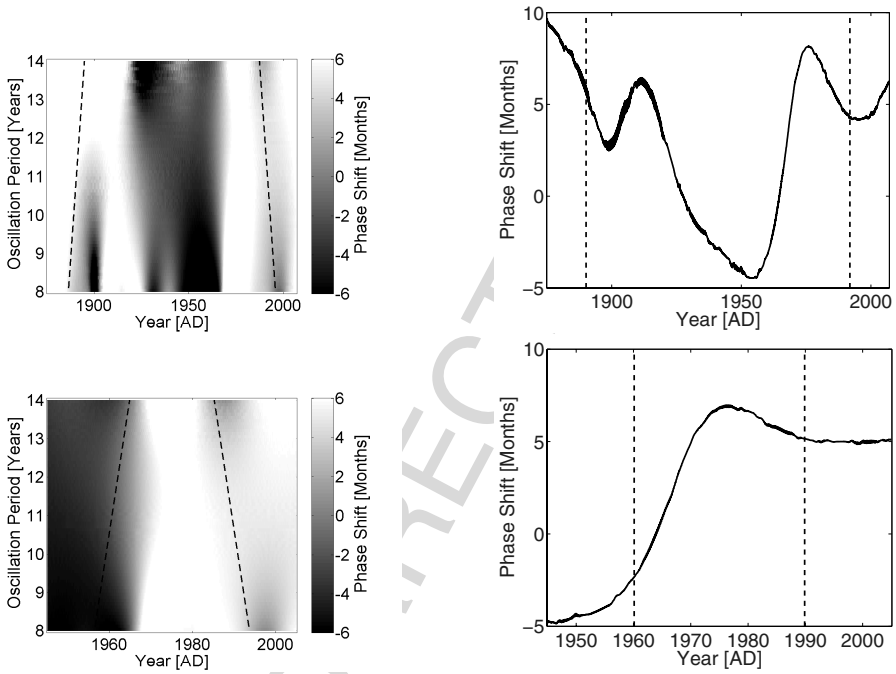


Fig. 10: Left: Phase differences between the sunspot areas $A^{n,s}$ (upper panel) and sunspot numbers $R^{n,s}$ (lower panel) for time scales between 8 and 14 years estimated with a complex Morlet wavelet. Right: Respective phase differences on a scale of $T = 10.75$ years. Note the different scalings on the time (x) axis. Dashed lines correspond to the cone of influence.

However, the underlying pattern is fully equivalent. In order to exclude possibly erroneous results at the boundaries of the record, the corresponding cones of influence [110] have been computed, which give a rough estimate of the part of the wavelet-filtered time series that is statistically reliable.

Examining the inferred phase shift variability in some more detail, the same qualitative behaviour is found for sunspot numbers and areas. As a particularly relevant feature, there is a steep increase of the phase difference between about 1960 and 1975, where conditions with the activity occuring significantly earlier on the southern hemisphere than on the northern one are replaced by the opposite behaviour within only about one solar cycle.

4.2 Topological Phase Coherence Analysis

The wavelet-based method discussed above requires the explicit selection of a distinct reference frequency. As a potential alternative, Zolotova and Ponyavin [112, 113] suggested to use the line of synchronisation (LOS) [44] in the cross-recurrence plot as an indicator of the phase shift of sunspot activity on both hemispheres. Indeed, the presence of such a continuous structure can be interpreted as a signature of time-scale synchronisation, i.e., a coherent dynamics of the two considered time series if the relative time scales are adjusted in a corresponding way. Some successful applications of this method have been demonstrated with respect to palaeoclimatic time series [44, 114, 115], where the corresponding problem of adjusting age-depth models to different sedimentary or ice core records is very typical [116, 117, 118].

In [111], it has already been argued that in the case of the sunspot activity records, the LOS approach might however have been misleading. One major point of criticism is that the algorithm for finding an "optimal path" through a cross-recurrence plot is not very robust. This is also one particular reason why the LOS technique has not yet become standard in palaeoclimatology, where still traditional methods of "sequence slotting" [116, 117] are used. In particular, using *unthresholded recurrence plots* (distance plots) $CR^u_{ij} = ||X_i - Y_j||$ instead of cross-recurrence plots (i.e., a matrix of pairwise distances between all observations in both considered time series) would be a much more natural approach, for which powerful dynamic programming algorithms are available [119]. Moreover, the choice of threshold values ϵ and the specific norm may have an influence on the estimated LOS.[1]

Besides the rather weak robustness against perturbations of the data, there is the conceptual problem that unlike claimed by Zolotova and Ponyavin [112], the LOS pattern is not necessarily a representative of the phase shift, as the

[1] It has to be mentioned that there is another drawback of the path finding algorithm for estimating the LOS [44, 42] which has been implemented in the Matlab CRP toolbox used in this study as well as in the work of Zolotova and Ponyavin [112, 113]: There is a clear preference towards horizontal steps compared to vertical ones [120] which may lead to systematic trends in the estimated LOS.

definition of cross-recurrence plots implies a dependence on both phase and amplitude of the considered time series. However, there are some possible modifications of the cross recurrence plot method that may contribute to a solution of this problem:

1. **Relative recurrences.** Instead of using the standard definition of cross-recurrence plots, one may consider normalised distances, for example, by setting

$$\overline{CR}_{ij} := \begin{cases} \Theta\left(\epsilon - \frac{||X_i - Y_j||}{||X_i|| + ||Y_j||}\right), & ||X_i|| + ||Y_j|| > 0 \\ 1, & \text{else} \end{cases} . \tag{16}$$

   Here, X_i and Y_j correspond to the values of the particularly considered observable (here, sunspot numbers or areas) in the northern and southern solar hemisphere at times t_i and t_j, respectively. The resulting matrix $\overline{CR}$ might be referred to as *relative cross-recurrence matrix*, and its graphical representation to as a *relative cross-recurrence plot*. The term *relative recurrence* is motivated by the fact that the ratio in the argument of the Heaviside function may only have values between 0 and 1, which yields a normalised range for the possible threshold values ϵ. As an unthresholded version, one may also define *relative distance matrices* in a similar way as

$$\overline{CR}^{u}_{ij} := \begin{cases} \frac{||X_i - Y_j||}{||X_i|| + ||Y_j||}, & ||X_i|| + ||Y_j|| > 0 \\ 0, & \text{else} \end{cases} . \tag{17}$$

2. **Dynamic encoding and symbolic recurrences.** A second possibility to make the results of the LOS method becoming more robust against small fluctuations of the time series values is applying a coarse-graining (equivalent to a symbolic encoding) to the record before computing the recurrence matrix. This coarse-graining may be static (i.e., based on the order relationships of the observations with respect to a set of predefined threshold values), dynamic (i.e., either based on the order relationships of subsequent subsets of observations or a static encoding of the difference filtered time series), or a mixed form. Generalising the recently proposed concepts of order patterns recurrence plots [121, 122, 123] and ordinal recurrence plots [42, 122], these approaches may be understood as *symbolic recurrence plots*. In order to make the results as sensitive to phase signals as possible, the use of a dynamic encoding appears to be most promising. In particular, the concept of order patterns recurrence plots might be especially suited for this purpose.

It will be a subject of further studies whether the two mentioned approaches are indeed capable to derive meaningful phase shifts between oscillatory signals from time series. In order to test for this, more robust and reliable algorithms for the estimation of the LOS are required, which are currently not yet available.

4.3 Significance Test

In order to prove the relevance of the derived phase difference time series, the concept of "natural surrogates" [124] has been proposed as a potential basis [111]. Here, a signal is used as a surrogate of the considered time series that represents the same or some essentially similar dynamical system. In the case of the sunspot areas, the consideration of records of the international sunspot number from earlier time intervals has been suggested. The results of Sect. 3.3 may be considered as a validation of this approach, as the phase shift between the total sunspot areas and numbers on both hemispheres is much smaller than that between the records on northern and southern hemisphere. The fact that the resulting phase difference time series of the original data lies outside the confidence levels obtained from the natural surrogates [111] demonstrates that the joint phase diffusion of the time series from both hemispheres deviates remarkably from the corresponding phase shift with respect to a structurally equivalent independent signal. In order to test whether phase coherence (in the sense of synchronisation) is actually present in the hemispheric sunspot areas and the corresponding sunspot numbers, a sophisticated statistical test would have to be applied to one of the resulting phase coherence measures. A corresponding bootstrap approach to quantifying the significance of the mean resultant length (4) has been recently proposed by Allefeld and Kurths [39].

According to the above considerations, the derived phase shift pattern is actually significant and may be considered as a proxy for the north-south asymmetry of solar activity. The fact that the phase difference time series deviates from the series computed using natural surrogates from earlier time intervals is most likely an effect of a particular type of nonstationarity of the system, which is related to a gradual change of the average frequency of oscillations. On the one hand, results have been reported that the solar activity of the last decades has been unusually strong [125, 126]. However, these findings have mainly been related to the magnitude of solar activity. On the other hand, Duhau [127, 128] interpreted pronounced increases of the solar activity between 1923 and 1949 and after 1993 as signatures of *phase catastrophes* associated with a breakdown of centennial scale oscillations (Gleissberg cycle) of the solar activity. It has been pointed out that similar dynamics can be regarded as a feature of many chaotic oscillators. The question whether the presence of such a speculative phase catastrophe can be considered being a reason for the results of the significance test has to be further investigated in future studies, possibly considering other types of surrogates.

5 Discussion

This chapter has summarised some recently developed approaches for testing the phase coherence of different oscillatory modes in the dynamics of complex systems. In particular, it has been argued that for inferring phase difference

patterns, at the present state of research, the use of methods based on an explicit phase definition (e.g., via wavelet or Hilbert transforms) is superior to purely topological methods like the LOS method. However, the latter class of approaches offers a high potential for future developments, which is underlined by the successful introduction of synchronisation indices based on recurrence plots (see Sect. 2.2).

As a particular application, the classical problem of decadal-scale oscillations of solar activity has been considered. It has been demonstrated that phase coherence between both solar hemispheres is present on time scales between about 8 and 14 years, whereas on shorter scales, irregular components contribute differently to the respective dynamics. With respect to the identification problem discussed in Sect. 2, it has to be underlined that further development of empirically motivated physical models is necessary in order to attribute this coherent behaviour to phase synchronisation of two distinct oscillatory components or two observations of the same oscillatory system.

The long-term variability of the phase difference between the activity on both solar hemispheres has been derived for the last about 130 years. It has been shown that the inferred pattern does not significantly depend on the choice of a reference frequency or of the sunspot areas or numbers as the considered observables. Very likely, the inferred phase shift is a major contributor to the north-south asymmetry of solar activity, which supports suggestions going back to Waldmeier 50 years ago [88]. Referring to the most recent literature, one has to mention that the presented results are not in accordance with the findings of Zolotova and Ponyavin [112, 113] who reported a much more irregular phase variability pattern with much larger phase differences. Following the arguments given in Sect. 4.2, the results of the wavelet-based approach used in this study are more reliable that those of the mentioned references, where the very unstable LOS method has been used. In addition, one may criticise that Zolotova and Ponyavin attributed the presence of a corresponding different phase dynamics on both solar hemispheres to the presence of "phase asynchronisation" [112, 113, 129]: Apart from the fact that the use of synchronisation terms is rather doubtful in the considered problem, the concept of synchronisation refers to a process rather than a state [21], i.e., the term "asynchronisation" is physically meaningless.

In this contribution, wavelet decomposition has been used to derive coherent oscillatory signals for which a phase variable can be defined. A potential alternative would be the consideration of empirical mode decomposition [113] or similar methods, that go beyond the requirement of a fixed reference frequency of oscillations. However, it has to be noted that in such case, the particular physical meaning of the inferred modes has to be carefully examined before using them for further analyses.

Finally, one has to mention that the north-south asymmetry of solar activity has traditionally been attributed to different amplitudes of the corresponding observables. In this work as well as a number of subsequent contributions [111, 112, 113, 129], first attempts have been made to use temporally

varying phase differences between decadal-scale oscillations as an alternative way of quantitatively describing this asymmetry. Using the information on phase shifts, one may adjust the corresponding time series to equal phases for considering the effect of different amplitudes separately from the phase dynamics. A corresponding data-adaptive redefinition of asymmetry indices and their thorough analysis is outlined to future studies.

Acknowledgements. This work has been financially supported by the Helmholtz foundation virtual institute "Pole-Equator-Pole", the Japanese Society for the Promotion of Science (JSPS) (project no. PE 06066), and the German Research Foundation. Discussions with M. Thiel, M.C. Romano, and N. Marwan are gratefully acknowledged. For the investigations on recurrence plots and recurrence quantification analysis, the MATLAB cross-recurrence plot toolbox by Norbert Marwan has been used, which is available as a part of the TOCSY toolbox at http://tocsy.agnld.uni-potsdam.de.

References

1. E. Forgács-Dajka, T. Borkovits, Searching for mid-term variations in different aspects of solar activity – looking for probable common origins and studying temporal variations of magnetic polarities. *Mon. Not. R. Astron. Soc.,* 374, 282–291 (2007)
2. P. Frick, D. Galyagin, D.V. Hoyt, E. Nesme-Ribes, K.H. Schatten, D. Sokoloff, V. Zakharov, Wavelet analysis of solar activity recorded by sunspot groups. *Astron. Astrophys.,* 328, 670–681 (1997)
3. M.G. Ogurtsov, Yu.A. Nagovitsyn, G.E. Kocharov, H. Jungner, Long-period cycles of the Sun's activity recorded in direct solar data and proxies. *Solar Phys.,* 211, 371–394 (2003)
4. H.H. Faria, E. Echer, N.R. Rigozo, L.E.A. Vieira, D.J.R. Nordemann, A. Prestes, A comparison of the spectral characteristics of the Wolf sunspot number (R_Z) and group sunspot number (R_G). *Solar Phys.,* 223, 305–318 (2004)
5. M. Waldmeier, *The sunspot activity in the years 1610–1960.* Schulthess, Zurich (1961)
6. C.P. Sonett, M.S. Giampapa, M.S. Matthews (eds.), *The Sun in Time,* University of Arizona Press, Tucson (1992)
7. A.D. Wittmann, Z.T. Xu, A catalogue of sunspot observations from 165 BC to AD 1684. *Astron. Astrophys. Suppl. Ser.,* 70, 83–94 (1987)
8. K. Coughlin, K.K. Tung, Eleven-year solar cycle signal throughout the lower atmosphere. *J. Geophys. Res.* 109, D21105 (2004)
9. K.T. Coughlin, K.K. Tung, 11-Year solar cycle in the stratosphere extracted by the empirical mode decomposition method. *Adv. Space Res.,* 34, 323–329 (2004)
10. D.I. Ponyavin, Solar cycle signal in geomagnetic activity and climate. *Solar Phys.,* 224, 465–471 (2004)
11. D.I. Ponyavin, T.V. Barliaeva, N.V. Zolotova, Hypersensitivity of climate response to solar activity output during the last 60 years. *Mem. S.A.It.,* 76, 1026–1029 (2005)

12. C.D. Camp, K.K. Tung, Surface warming by the solar cycle as revealed by the composite mean difference projection. *Geophys. Res. Lett.,* 34, L14703 (2007)
13. J.C. Stager, A. Ruzmaikin, D. Conway, P. Verburg, P.J. Mason, Sunspots, El Niño, and the levels of Lake Victoria, East Africa. *J. Geophys. Res.,* 112, D15106 (2007)
14. K. Georgieva, B. Kirov, P. Tonev, V. Guineva, D. Atanasov, Long-term variations in the correlation between NAO and solar activity: the importance of north-south solar activity asymmetry for atmospheric circulation. *Adv. Space Res.,* 40, 1152–1166 (2007)
15. J. Moore, A. Grinsted, S. Jevrejeva, Is there evidence for sunspot forcing of climate at multi-year and decadal periods? *Geophys. Res. Lett.,* 33, L17705 (2006)
16. P. Foukal, C. Fröhlich, H. Spruit, T.M.L. Wigley, Variations in solar luminosity and their effect on the Earth's climate. *Nature,* 443, 161–166 (2006)
17. M. Lockwood, C. Fröhlich, Recent oppositely directed trends in solar climate forcings and the global mean surface air temperature. *Proc. R. Soc. A,* 463, 2447–2460 (2007)
18. C. Torrence, P.J. Webster, Interdecadal Changes in the ENSO-Monsoon System. *J. Climate,* 12, 2679–2690 (1999)
19. D. Maraun, J. Kurths, Epochs of Phase Coherence between El Niño/Southern Oscillation and Indian Monsoon. *Geophys. Res. Lett.,* 32, L15709 (2005)
20. R.K. Tiwari, S. Sri Lakshmi, Signature of ENSO and NAO signals in the Indian subcontinent monsoon. In: Y.-T. Chen (ed.), *Advances in Geosciences – Proceedings of the Asia-Oceania Geosciences Society (AOGS) Annual Meeting 2005, Singapore*, 22 (2005)
21. A. Pikovsky, M.G. Rosenblum, J. Kurths, *Synchronization – A Universal Concept in Nonlinear Sciences,* Cambridge University Press, Cambridge (2001)
22. M.G. Rosenblum, A.S. Pikovsky, J. Kurths, Phase Synchronization of Chaotic Oscillators. *Phys. Rev. Lett.,* 76, 1804–1807 (1996)
23. J.Y. Chen, K.W. Wong, H.Y. Zheng, J.W. Shuai, Intermittent phase synchronization of coupled spatiotemporal chaotic systems. *Phys. Rev. E,* 64, 016202 (2001)
24. J.Y. Chen, K.W. Wong, J.W. Shuai, Properties of phase locking with weak phase-coherent attractors. *Phys. Lett. A,* 285, 312–318 (2001)
25. G.V. Osipov, B. Hu, C. Zhou, M.V. Ivanchenko, J. Kurths, Three Types of Transitions to Phase Synchronization in Coupled Chaotic Oscillators. *Phys. Rev. Lett.,* 91, 024101 (2003)
26. A. Bandrivskyy, A. Bernjak, P. McClintock, A. Stefanovska, Wavelet Phase Coherence Analysis: Application to Skin Temperature and Blood Flow. *Cardiovasc. Engin.,* 4, 89–93 (2004)
27. A.A. Koronovskii, A.E. Hramov, Chaotic Phase Synchronization Studied by Means of Continuous Wavelet Transform. *Techn. Phys. Lett.,* 30, 587–590 (2004)
28. A.E. Hramov, A.A. Koronovskii, An approach to chaotic synchronization. *Chaos,* 14, 603–610 (2004)
29. A.A. Koronovskii, M.K. Kurovskaya, A.E. Hramov, Relationship between Phase Synchronization of Chaotic Oscillators and Time Scale Synchronization. *Techn. Phys. Lett.,* 31, 847–850 (2005)

30. C. Allefeld, S. Frisch, Phase Synchronization Analysis of Event-Related Potentials in Language Processing. In: S. Boccaletti, B.J. Gluckman, J. Kurths, L.M. Pecora, R. Meucci, O. Yordanov (eds.), *Experimental Chaos: 8th Experimental Chaos Conference*, AIP Conf. Proc. 742, Springer, New York, 106–111 (2004)
31. C. Allefeld, S. Frisch, M. Schlesewsky, Detection of early cognitive processing by event-related physe synchronization analysis. *Neuroreport,* 16, 13–16 (2005)
32. B. Schack, S. Weiss, Quantification of phase synchronization phenomena and their importance for verbal memory processes. *Biolog. Cybern.,* 92, 275–287 (2005)
33. M. Chavez, C. Adam, V. Navarro, S. Boccaletti, J. Martinerie, On the intrinsic time scales involved in synchronization: A data-driven approach. *Chaos,* 15, 023904 (2005)
34. A. Goska, A. Krawiecki, Analysis of phase synchronization of coupled chaotic oscillators with empirical mode decomposition. *Phys. Rev. E,* 74, 046217 (2006)
35. D.J. DeShazer, R. Breban, E. Ott, R. Roy, Detecting Phase Synchronization in a Chaotic Laser Array. *Phys. Rev. Lett.,* 87, 044101 (2001)
36. F.C. Meinecke, A. Ziehe, J. Kurths, K.-R. Müller, Measuring Phase Synchronization of Superimposed Signals. *Phys. Rev. Lett.,* 94, 084102 (2005)
37. M. Paluš, Detecting phase synchronization in noisy systems. *Phys. Lett. A,* 235, 341–351 (1997)
38. P. Tass, M.G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkmann, A. Schnitzler, H.-J. Freund, Detection of $n : m$ Phase Locking from Noisy Data: Application to Magnetoencephalography. *Phys. Rev. Lett.,* 81, 3291–3294 (1998)
39. C. Allefeld, J. Kurths, Testing for phase synchronization. *Int. J. Bifurcation Chaos,* 14, 405–416 (2004)
40. M. Winterhalder, B. Schelter, J. Kurths, A. Schulze-Bonhage, J. Timmer, Sensitivity and specificity of coherence and phase synchronization analysis. *Phys. Lett. A,* 356, 26–34 (2006)
41. J.-P. Eckmann, S. Oliffson Kamphorst, D. Ruelle, Recurrence plots of dynamic systems. *Europhys. Lett.,* 4, 973–977 (1987)
42. N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems. *Phys. Rep.,* 438, 237–329 (2007)
43. J.P. Zbilut, A. Giuliani, C.L. Webber Jr., Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Phys. Lett. A,* 246, 122–128 (1998)
44. N. Marwan, M. Thiel, N.R. Nowaczyk, Cross Recurrence Plot Based Synchronization of Time Series. *Nonlin. Proc. Geophys.,* 9, 325–331 (2002)
45. M.C. Romano, M. Thiel, J. Kurths, I.Z. Kiss, J.L. Hudson, Detection of synchronization for non-phase-coherent and non-stationary data. *Europhys. Lett.,* 71, 466–472 (2005)
46. M.C. Romano, M. Thiel, J. Kurths, M. Rolfs, R. Engbert, R. Kliegl, Synchronization Analysis and Recurrence in Complex Systems. In: B. Schelter, M. Winterhalder, J. Timmer (eds.), *Handbook of Time Series Analysis,* Wiley-VCH, Weinheim, 231–264 (2006)
47. R. Donner, Interdependences between daily European temperature records: Correlation or phase synchronisation? In: P. Marquié (ed.), *Nonlinear Dynamics of Electronic Systems (NDES 2006),* Université de Bourgogne, Dijon, France, 26–29 (2006)

48. R. Donner, Spatial Correlations of River Runoffs in a Catchment. In: J. Kropp, H.-J. Schellnhuber (eds.), *Correlations and Extremes in Climate and Hydrology,* Springer, Berlin (2008), subm.
49. M. Thiel, M.C. Romano, J. Kurths, M. Rolfs, R. Kliegl, Twin surrogates to test for complex synchronisation. *Europhys. Lett.,* 75, 535–541 (2006)
50. M.C. Romano, *Synchronization Analysis by Means of Recurrences in Phase Space.* PhD thesis, University of Potsdam (2004)
51. M. Thiel, *Recurrences: Exploiting Naturally Occurring Analogues.* PhD thesis, University of Potsdam (2004)
52. C.H. Hossfield, A History of the Zurich and American Relative Sunspot Number Indices. *JAAVSO,* 30, 48–53 (2001)
53. D.V. Hoyt, K.H. Schatten, Group Sunspot Numbers: A New Solar Activity Reconstruction. *Solar Phys.,* 179, 189–219 (1998)
54. J.E. Beckman, T.J. Mahoney, The Maunder Minimum and Climate Change: Have Historical Records Aided Current Research? In: U. Grothkopf, H. Andernach, S. Stevens-Rayburn, M. Gomez (eds.), *Library and Information Services in Astronomy III,* ASP Conference Series, 153, 212–217 (1998)
55. I.G. Usoskin, K. Mursula, G.A. Kovaltsov, Cyclic behaviour of sunspot activity during the Maunder minimum. *Astron. Astrophys.,* 354, L33–L36 (2000)
56. H. Miyahara, D. Sokoloff, I.G. Usoskin, The solar cycle at the Maunder minimum epoch. In: W.-H. Ip, M. Duldig (eds.), *Advances in Geosciences, Vol. 2,* 1–20 (2006)
57. M. Temmer, *Solar Activity Patterns - Hemisphere-Related Studies,* PhD thesis, University of Graz (2004)
58. M. Temmer, J. Rybák, P. Bendík, A. Veronig, F. Vogler, W. Otruba, W. Pötzi, A. Hanslmeier, Hemispheric sunspot numbers R_n and R_s from 1945–2004: Catalogue and N-S asymmetry analysis for solar cycles 18–23. *Astron. Astrophys.,* 447, 735–743 (2006)
59. D.H. Hathaway, R.M. Wilson, E.J. Reichmann, Group sunspot numbers: Sunspot cycle characteristics. *Solar Phys.,* 211, 357–370 (2002)
60. G.U. Yule, On a Method of Investigating Periodicities in Disturbed Series, with special reference to Wolfer's Sunspot Numbers. *Phil. Trans. R. Soc. Lond. A,* 226, 267–298 (1927)
61. J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, J.D. Farmer, Testing for nonlinearity in time series: the method of surrogate data. *Physica D,* 58, 77–94 (1992)
62. D. Kugiumtzis, Surrogate data test for nonlinearity including nonmonotonic transforms. *Phys. Rev. E,* 62, R25–R28 (2000)
63. J.A. Barnes, H.H. Sargent III, P.V. Tryon, Sunspot cycle simulation using random noise. In: R.O. Pepin, J.A. Eddy, R.B. Merrill (eds.), *The Ancient Sun,* 159–163 (1980)
64. M. Paluš, D. Novotná, Sunspot Cycle: A Driven Nonlinear Oscillator? *Phys. Rev. Lett.,* 83, 3406–3409 (1999)
65. H. Tsai, K.S. Chan, A note on testing for nonlinearity with partially observed time series. *Biometrika,* 89, 245–250 (2002)
66. N.O. Weiss, Periodicity and aperiodicity in solar magnetic activity. *Phil. Trans. R. Soc. Lond. A,* 330, 617–625 (1990)
67. J. Kurths, A.A. Ruzmaikin, On forecasting the sunspot numbers. *Solar Phys.,* 126, 407–410 (1990)

68. V.M. Ostryakov, I.G. Usoskin, On the dimension of the solar attractor. *Solar Phys.,* 127, 405–412 (1990)
69. J. Feynman, S.B. Gabriel, Period and phase of the 88-year solar cycle and the Maunder minimum: Evidence for a chaotic Sun. *Solar Phys.,* 127, 393–403 (1990)
70. M.D. Mundt, W.B. Maguire II, R.R.P. Chase, Chaos in the sunspot cycle: Analysis and prediction. *J. Geophys. Res. A,* 96, 1705–1716 (1991)
71. C.P. Price, D. Prichard, E.A. Hogenson, Do the Sunspot Numbers Form a "Chaotic" Set? *J. Geophys. Res. A,* 97, 19,113–19,120 (1992)
72. A. Ruzmaikin, J. Feynman, V. Kosacheva, On long-term dynamics of the solar cycle. In: K.L. Harvey (ed.), *The Solar Cycle,* ASP Conf. Ser., 27, 547–556 (1992)
73. M. Carbonell, R. Oliver, J.L. Ballester, A search for chaotic behaviour in solar activity. *Astron. Astrophys.,* 290, 983–994 (1994)
74. M.N. Kremliovsky, Can we understand time scales of solar activity? *Solar Phys.,* 151, 351–370 (1994)
75. M.N. Kremliovsky, Limits of predictability of solar activity. *Solar Phys.,* 159, 371–380 (1995)
76. K. Jinno, S. Xu, R. Berndtsson, A. Kawamura, M. Matsumoto, Prediction of sunspots using reconstructed chaotic system equations. *J. Geophys. Res. A,* 100, 14, 773–14,782 (1995)
77. J.M. Polygiannakis, X. Moussas, C.P. Sonett, A nonlinear RLC solar cycle model. *Solar Phys.,* 163, 193–203 (1996)
78. P. Hoyng, Is the solar cycle timed by a clock? *Solar Phys.,* 169, 253–264 (1996)
79. T. Serre, E. Nesme-Ribes, Nonlinear analysis of solar cycles. *Astron. Astrophys.,* 360, 319–330 (2000)
80. I.G. Usoskin, K. Mursula, G.A. Kovaltsov, Regular and random components of sunspot activity during active Sun and great minima: Model simulation. *Proc. 1st Solar & Space Weather Euroconference "The Solar Cycle and Terrestrial Climate,* ESA, Santa Cruz de Tenerife, 447–450 (2000)
81. N. Jevtić, J.S. Schweitzer, C.J. Cellucci, Nonlinear time series analysis of northern and southern solar hemisphere daily sunspot numbers in search of short-term chaotic behavior. *Astron. Astrophys.,* 379, 611–615 (2001)
82. C. Letellier, L.A. Aguirre, J. Maquet, R. Gilmore, Evidence for low dimensional chaos in sunspot cycles. *Astron. Astrophys.,* 449, 379–387 (2006)
83. Q.X. Li, K.J. Li, Low Dimensional Chaos from the Group Sunspot Numbers. *Chin. J. Astron. Astrophys.,* 7, 435–440 (2007)
84. M. Paluš, J. Kurths, U. Schwarz, N. Seehafer, D. Novotná, I. Charvátová, The solar activity cycle is weakly synchronized with the solar inertial motion. *Phys. Lett. A,* 365, 421–428 (2007)
85. S. Watari, Separation of periodic, chaotic, and random components in solar activity. *Solar Phys.,* 168, 413–422 (1996)
86. Z. Qin, The transitional time scale from stochastic to chaotic behavior for solar activity. *Solar Phys.,* 178, 423–431 (1998)
87. H.W. Newton, A.S. Milsom, Note on the observed differences in spottedness of the Sun's northern and southern hemispheres. *Monthly Not. R. Astron. Soc.,* 115, 398–404 (1955)
88. M. Waldmeier, Der lange Sonnenzyklus. *Zeitschr. Astrophys.,* 43, 149–160 (1957)

89. B. Bell, A Long-Term North-South Asymmetry in the Location of Solar Sources of Great Geomagnetic Storms. *Smithsonian Contrib. Astrophys.,* 5, 187–194 (1962)
90. J.G. Wolbach, On the Unequal Spottedness of the Two Solar Hemispheres. *Smithsonian Contrib. Astrophys.,* 5, 195–202 (1962)
91. B. Bell, J.G. Wolbach, On Short-Period Relations Between North-South Asymmetry in Spottedness and in Great-Storm Sources. *Smithsonian Contrib. Astrophys.,* 5, 203–208 (1962)
92. J.-R. Roy, The north-south distribution of major solar flare events, sunspot magnetic classes and sunspot areas. *Solar Phys.,* 52, 53–61 (1977)
93. D.B. Swinson, H. Koyama, T. Saito, Long-term variations in north-south asymmetry of solar activity. *Solar Phys.,* 106, 35–42 (1986)
94. W. Yi, The north-south asymmetry of sunspot distribution. *J. Roy. Astron. Soc. Can.,* 86, 89–98 (1992)
95. V.K. Verma, On the north-south asymmetry of solar activity cycles. *Astrophys. J.,* 403, 797–800 (1993)
96. V.K. Verma, Periodic Variation of the North-South Asymmetry of Solar Activity Phenomena. *J. Astrophys. Astr.,* 21, 173–176 (2000)
97. K.J. Li, J.X. Wang, S.Y. Xiong, H.F. Liang, H.S. Yun, X.M. Gu, Regularity of the north-south asymmetry of solar activity. *Astron. Astrophys.,* 383, 648–652 (2002)
98. K.J. Li, X.H. Liu, H.S. Yun, S.Y. Xiong, H.F. Liang, H.Z. Zhao, L.S. Zhan, X.M. Gu, Asymmetrical Distribution of Sunspot Groups in the Solar Hemispheres. *Publ. Astron. Soc. Japan,* 54, 629–633 (2002)
99. E.S. Vernova, K. Mursula, M.I. Tyasto, D.G. Baranov, A new pattern for the north-south asymmetry of sunspots. *Solar Phys.,* 205, 371–382 (2002)
100. M. Temmer, A. Veronig, A. Hanslmeier, Hemispheric Sunspot Numbers R_n and R_s: Catalogue and N-S asymmetry analysis. *Astron. Astrophys.,* 390, 707–715 (2002)
101. S.I. Zharkov, V.V. Zharkova, Statistical analysis of the sunspot area and magnetic flux variations in 1996–2005 extracted from the Solar Feature Catalogue. *Adv. Space Res.,* 38, 868–875 (2006)
102. M. Carbonell, J. Terradas, R. Oliver, J.L. Ballester, The statistical significance of the North-South asymmetry of solar activity revisited. *Astron. Astrophys.,* in press, arXiv:0709.1901 [astro-ph]
103. G. Vizoso, J.L. Ballester, The north-south asymmetry of sunspots. *Astron. Astrophys.,* 229, 540–546 (1990)
104. V.K. Verma, The distribution of the north-south asymmetry for the various activity cycles. In: K.L. Harvey (ed.), *The Solar Cycle,* ASP Conf. Ser., 27, 429–435 (1992)
105. M. Carbonell, R. Oliver, J.L. Ballester, On the asymmetry of solar activity. *Astron. Astrophys.,* 274, 497–504 (1993)
106. R. Oliver, J.L. Ballester, The north-south asymmetry of sunspot areas during solar cycle 23. *Solar Phys.,* 152, 481–485 (1994)
107. R. Knaack, J.O. Stenflo, S.V. Berdyugina, Periodic oscillations in the north-south asymmetry of the solar magnetic field. *Astron. Astrophys.,* 418, L17–L20 (2004)
108. J.L. Ballester, R. Oliver, M. Carbonell, The periodic behaviour of the North-South asymmetry of sunspot areas revisited. *Astron. Astrophys.,* 431, L5–L8 (2005)

109. S. Watari, Chaotic behavior of the north-south asymmetry of sunspots? *Solar Phys.,* 163, 259–266 (1996)
110. C. Torrence, G.P. Compo, A Practical Guide to Wavelet Analysis. *Bull. Am. Met. Soc.,* 79, 61–78 (1998)
111. R. Donner, M. Thiel, Scale-resolved phase coherence analysis of hemispheric sunspot activity: A new look at the north-south asymmetry. *Astron. Astrophys.,* 475, L33–L36 (2007)
112. N.V. Zolotova, D.I. Ponyavin, Phase asynchrony of the north-south sunspot activity. *Astron. Astrophys.,* 449, L1–L4 (2006)
113. N.V. Zolotova, D.I. Ponyavin, Synchronization in Sunspot Indices in the Two Hemispheres. *Solar Phys.,* 243, 193–203 (2007)
114. M.H. Trauth, B. Bookhagen, N. Marwan, M.R. Strecker, Multiple landslide clusters record Quaternary climate changes in the northwestern Argentina Andes. *Palaeogeogr. Palaeoclim. Palaeoecol.,* 194, 109–121 (2003)
115. N. Marwan, M.H. Trauth, M. Vuille, J. Kurths, Comparing modern and Pleistocene ENSO-like influences in NW Argentina using nonlinear time series analysis methods. *Clim. Dyn.,* 21, 317–326 (2003)
116. R. Thompson, R.M. Clark, Sequence slotting for stratigraphic correlation between cores: theory and practice. *J. Paleolimnol.,* 2, 173–184 (1989)
117. L.E. Lisiecki, P.A. Lisiecki, Application of dynamic programming to the correlation of paleoclimate records. *Paleoceanogr.,* 17, 1049 (2002)
118. R.J. Telford, E. Heegaard, H.J.B. Birks, All age-depth models are wrong: but how badly? *Quat. Sci. Rev.,* 23, 1–5 (2004)
119. H. Sakoe, S. Chiba, Dynamic Programming Algorithms Optimization for Spoken Word Recognition, *IEEE Transact. Acoust. Speech Signal Proc.,* 26, 43–49 (1978)
120. N. Marwan, M. Thiel, M.C. Romano, priv. comm.
121. A. Groth, Visualization of coupling in time series by order recurrence plots. *Phys. Rev. E,* 72, 046220 (2005)
122. A. Groth, *Analyse der Wiederkehr in dynamischen Systemen auf einer Ordinalskala.* PhD thesis, University of Greifswald (2006)
123. N. Marwan, A. Groth, J. Kurths, Quantification of order patterns recurrence plots of event related potentials. *Chaos Complex. Lett.,* 2, 301–314 (2007)
124. P. van Leeuwen, D. Geue, S. Lange, D. Cysarz, H. Bettermann, D.H.W. Grönemeyer, Is there evidence of fetal-maternal heart rate synchronization? *BMC Physiol.,* 3, 2 (2003)
125. I.G. Usoskin, S.K. Solanki, M. Schüssler, K. Mursula, K. Alanko, Millennium-Scale Sunspot Number Reconstruction: Evidence for an Unusually Active Sun since the 1940s. *Phys. Rev. Lett.,* 91, 211101 (2003)
126. S.K. Solanki, I.G. Usoskin, B. Kromer, M. Schüssler, J. Beer, Unusual activity of the Sun during recent decades compared to the previous 11,000 years. *Nature,* 431, 1084–1087 (2004)
127. S. Duhau, An early prediction of maximum sunspot number in solar cycle 24. *Solar Phys.,* 213, 203–212 (2003)
128. C. de Jager, Solar forcing of climate. 1: Solar variability. *Space Sci. Rev.,* 120, 197–241 (2005)
129. N.V. Zolotova, D.I. Ponyavin, Was the unusual solar cycle at the end of the XVIII century a result of phase asynchronization? *Astron. Astrophys.,* 470, L17–L20 (2007)