Valliappa Lakshmanan · Eric Gilleland
Amy McGovern · Martin Tingley *Editors*

Machine Learning and Data Mining Approaches to Climate Science

Proceedings of the 4th International Workshop on Climate Informatics

Springer

Machine Learning and Data Mining Approaches to Climate Science

Valliappa Lakshmanan • Eric Gilleland
Amy McGovern • Martin Tingley
Editors

Machine Learning and Data Mining Approaches to Climate Science

Proceedings of the 4th International Workshop on Climate Informatics

Springer

Editors
Valliappa Lakshmanan
The Climate Corporation
Seattle, WA, USA

Amy McGovern
Computer Science
University of Oklahoma
Norman, OK, USA

Eric Gilleland
Research Applications Laboratory
National Center for Atmospheric Research
Boulder, CO, USA

Martin Tingley
Meteorology and Statistics
Pennsylvania State University
University Park, PA, USA

ISBN 978-3-319-17219-4 ISBN 978-3-319-17220-0 (eBook)
DOI 10.1007/978-3-319-17220-0

Library of Congress Control Number: 2015944106

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

The threat of climate change makes it crucial to improve our understanding of the climate system. However, the volume and diversity of climate data from satellites, environmental sensors, and climate models can make the use of traditional analysis tools impractical and necessitate the need to carry out knowledge discovery from data. Machine learning has made significant impacts in fields ranging from web search to bioinformatics, and the impact of machine learning on climate science could be as profound (Monteleoni et al. 2013). However, because the goal of machine learning in climate science is to improve our understanding of the climate system, it is necessary to employ techniques that go beyond simply taking advantage of co-occurrence and, instead, enable increased understanding.

The Climate Informatics workshop series seeks to build collaborative relationships between researchers from statistics, machine learning, and data mining and researchers in climate science. Because climate models and observed datasets are increasing in complexity and volume, and because the nature of our changing climate is an urgent area of discovery, there are many opportunities for such partnerships. The series was cofounded by Claire Monteleoni and Gavin Schmidt and the first workshop held in August 2011 at the New York Academy of Sciences, New York, NY. Since then, the workshop has been held yearly at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, with logistical support from NCAR's Mathematics Applied to Geosciences (IMAGe) led by Doug Nychka.

The 4th International Workshop on Climate Informatics was sponsored by the National Science Foundation, The Climate Corporation, Oak Ridge Associated Universities, and NCAR and held over 2 days, on September 25 and 26, 2014, in Boulder, CO. The workshop drew 74 participants from universities, government laboratories, and industry. There were 43 posters presented at the workshop, as well as four invited talks. The editors selected and reviewed the 22 chapters in this volume to represent the state of the field and provide indications of where new advances will come from.

It has been heartening to see collaborations fostered in previous years bear fruit in the form of presentations in later years. For researchers in either field (machine learning or climate science) looking for a new subspecialty in which to make an impact, Climate Informatics presents a great opportunity. We hope that this book will spark new ideas and foster new collaborations and encourage interested readers to join us in Boulder for the 5th International Workshop on Climate Informatics.

Seattle, WA, USA — Valliappa Lakshmanan
Boulder, CO, USA — Eric Gilleland
Norman, OK, USA — Amy McGovern
State College, PA, USA — Martin Tingley
February 2015

Reference

Monteleoni C, Schmidt GA, Alexander F, Niculescu-Mizil A, Steinhaeuser K, Tippett M, Banerjee A, Blumenthal MB, Ganguly AR, Smerdon JE, Tedesco M (2013) Climate Informatics. In: Yu T, Chawla N, Simoff S (eds) Computational intelligent data analysis for sustainable development; data mining and knowledge discovery series. CRC Press, Taylor & Francis Group, Boca Raton. Chapter 4, pp 81–126

Contents

Part I
Machine Learning Methods

Chapter 1
Combining Analog Method and Ensemble Data Assimilation: Application to the Lorenz-63 Chaotic System

Pierre Tandeo, Pierre Ailliot, Juan Ruiz, Alexis Hannart, Bertrand Chapron, Anne Cuzol, Valérie Monbet, Robert Easton, and Ronan Fablet

Abstract Nowadays, ocean and atmosphere sciences face a deluge of data from space, in situ monitoring as well as numerical simulations. The availability of these different data sources offers new opportunities, still largely underexploited, to improve the understanding, modeling, and reconstruction of geophysical dynamics. The classical way to reconstruct the space-time variations of a geophysical system from observations relies on data assimilation methods using multiple runs of the known dynamical model. This classical framework may have severe limitations including its computational cost, the lack of adequacy of the model with observed data, and modeling uncertainties. In this paper, we explore an alternative approach

P. Tandeo (✉) • R. Fablet
Télécom Bretagne, Plouzané, France
e-mail: pierre.tandeo@telecom-bretagne.eu; ronan.fablet@telecom-bretagne.eu

P. Ailliot
Université de Bretagne Occidentale, Brest, France
e-mail: pierre.ailliot@univ-brest.fr

J. Ruiz • A. Hannart
National Scientific and Technical Research Council, Universidad de Buenos Aires, Buenos Aires, Argentina
e-mail: jruiz@cima.fcen.uba.ar; alexis.hannart@cima.fcen.uba.ar

B. Chapron
Ifremer, Issy-les-Moulineaux, Ifremer, Brest, France
e-mail: bertrand.chapron@ifremer.fr

A. Cuzol
Université de Bretagne Sud, Lorient, France
e-mail: anne.cuzol@univ-ubs.fr

V. Monbet
Université de Rennes I, Rennes, France
e-mail: valerie.monbet@univ-rennes1.fr

R. Easton
University of Colorado, Boulder, CO, USA
e-mail: robert.easton@colorado.edu

V. Lakshmanan et al. (eds.), *Machine Learning and Data Mining Approaches to Climate Science*, DOI 10.1007/978-3-319-17220-0_1

and develop a fully data-driven framework, which combines machine learning and statistical sampling to simulate the dynamics of complex system. As a proof concept, we address the assimilation of the chaotic Lorenz-63 model. We demonstrate that a nonparametric sampler from a catalog of historical datasets, namely, a nearest neighbor or analog sampler, combined with a classical stochastic data assimilation scheme, the ensemble Kalman filter and smoother, reaches state-of-the-art performances, without online evaluations of the physical model.

Keywords Data-driven modeling • Data assimilation • Stochastic filtering • Nonparametric sampling • Analog method • Lorenz-63 model

1.1 Introduction

Understanding and estimating the space-time evolution of geophysical systems constitute a challenge in geosciences. For an efficient restitution of geophysical fields, classical approaches typically combine a physical model based on fluid dynamics equations and remote sensing data or in situ observations. These approaches are generally referred to as data assimilation methods and stated as inverse problems for dynamical processes (see, e.g., Evensen 2009 and reference therein). Two main categories of data assimilation approaches may be distinguished: variational assimilation methods, which resort to the gradient-based minimization of a variational cost function and rely on the computation of the adjoint of the dynamical model (Lorenc et al. 2000), and stochastic data assimilation schemes, which involve Monte Carlo strategies and are particularly appealing for their modeling flexibility (Bertino et al. 2003). These stochastic methods iterate the generation of a representative set of scenarios (hereinafter referred to members), whose consistency is evaluated with respect to the available observations. To reach good estimation performance, this number of members must be high enough to explore the state space of the physical model.

Different limitations can occur in the stochastic data assimilation approaches presented above. Firstly, it generally involves intensive computations for practical applications since the physical model needs to be run with different initial conditions at each time step in order to generate the members. Moreover, intensive modeling efforts are needed to take into account fine-scale effects. Regional geophysical models are typical examples (Ruiz et al. 2010). Secondly, dissimilarities often occur between model outputs and observations. For instance, it can be the case when combining high-resolution model forecasts with high-resolution satellite or radar images. Thirdly, the dynamical model is not necessarily well known, and parameterizations may be highly uncertain. This is particularly the case in subgrid-scale processes, taking into account local and highly nonlinear effects (Lott and Miller 1997). These different examples tend to show that multiple evaluations of an explicit physical model are computationally demanding, and model uncertainties can produce dissimilarities between forecasts and observations.

As an alternative, the amount of observation and simulation data has grown very quickly in the last decades. The availability of such historical datasets strongly advocates for exploring implicit data-driven schemes to build realistic statistical simulations of the dynamics for data assimilation issues. Satellite sequence images are typical examples. When the spatiotemporal sampling and the amount of historical remote sensing data are sufficient, we may able to learn dynamical operators to construct relevant statistical forecasts with a good consistency with satellite observations. Such implicit data-driven schemes may also provide fast implementation alternatives as well as flexible strategies to deal with the abovementioned modeling uncertainties. In this case, historical simulated data with different parameterizations, initial conditions, and forcing terms may provide various scenarios to explore larger state spaces.

In this paper, we aim at demonstrating a proof of concept of such data-driven strategies to reconstruct complex dynamics from partial noisy observations. The feasibility of our data assimilation method is illustrated on the classical chaotic Lorenz-63 model (Lorenz 1963). The paper is organized as follows. In Sect. 1.2, we propose to use a nonparametric sampler, based on the analog (or nearest neighbors) method, to generate the forecast members (Delle Monache et al. 2013). Then, we use the ensemble Kalman recursions to combine these members with the observations (Evensen 2009). In Sect. 1.3, we numerically evaluate the methodology on the Lorenz-63 model such as various previous works (see, e.g., Pham 2001, Hoteit et al. 2008). We further discuss and summarize the key results of our investigations in Sect. 1.4.

1.2 Combining Machine Learning and Stochastic Filtering Methods

Data assimilation for dynamical systems is generally stated according to the following state space model (see, e.g., Bertino et al. 2003):

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathscr{M}(\mathbf{x}(t), \boldsymbol{\eta}(t)) \tag{1.1}$$

$$\mathbf{y}(t) = \mathscr{H}(\mathbf{x}(t), \boldsymbol{\epsilon}(t)) \quad . \tag{1.2}$$

The dynamical model given in Eq. (1.1) describes the evolution of the true physical process $\mathbf{x}(t)$. It includes a random perturbation $\boldsymbol{\eta}(t)$ which accounts for the various sources of uncertainties (e.g., boundary conditions, forcing terms, physical parameterization, etc.). As an illustration, $\mathscr{M}$ refers in the next sections to the Lorenz-63 dynamical model, in which the state of the system $\mathbf{x}$ is a three-dimensional vector (x, y, z). The observation model given in Eq. (1.2) links the observation $\mathbf{y}(t)$ to the true state at the same time t. It also includes a random noise $\boldsymbol{\epsilon}(t)$ which models observation error and uncertainties, change of support (i.e., downscaling/upscaling effects), and so on.

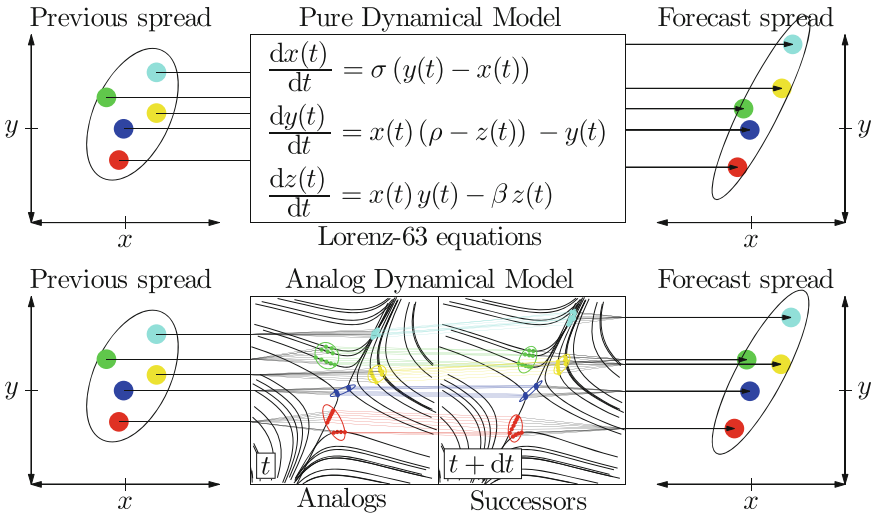

Fig. 1.1 Sketch of the forecast step in stochastic data assimilation schemes using pure (*top*) and analog (*bottom*) dynamical models. As an example, we consider the three-dimensional Lorenz-63 chaotic model. For visualization convenience, we only represent the x-y plane, centered at the origin. We track five statistical members with the variability depicted by ellipsoids accounting for the covariance structure

The key originality of the methodology proposed in this paper consists in using a nonparametric statistical sampling within a classical ensemble Kalman framework. As described in Fig. 1.1 (top), the classical approach exploits an explicit knowledge of the pure dynamical model (PDM) to propagate the ensemble members from a given time step to the next one. By contrast, we assume here that a representative catalog of examples of the time evolution of the state is available. This catalog is used to build an analog dynamical model (ADM) to simulate $\mathscr{M}$ and the associated error $\boldsymbol{\eta}$ given in Eq. (1.1). We proceed as follows. Let us denote by $\mathbf{x}(t)$ the state at time t. Its analogs or nearest neighbors are the samples in the catalog which are the closest to $\mathbf{x}(t)$. Such nearest neighbor schemes are among the state-of-the-art machine learning strategies (Friedman et al. 1977). In the geoscience literature, we talk about analog methods (see, e.g., Lorenz 1963 or Van den Dool 2006). They were initially devised for weather prediction, but applications to downscaling issues (Timbal et al. 2003) or climate reconstructions (Schenk and Zorita 2012; Yiou et al. 2013) were also proposed. As described in Fig. 1.1 (bottom), for each member at a given time, we use the successors of its analogs to generate possible forecast states at time $t+\mathrm{d}t$. The variability of the selected successors also provides a characterization of the forecast error, namely, here, its covariance. From a methodological point of view, analog techniques provide nonparametric representations. They are associated with computationally efficient implementations and prove highly flexible to account

for nonlinear and chaotic patterns as soon as the catalog of observed situations is rich enough to describe all possible state dynamics (Lorenz 1969).

Then, this nonparametric data-driven sampling of the state dynamics is plugged into a classical ensemble data assimilation method. It leads to the estimation of the filtering or smoothing probabilities of the state-space model given in Eqs. (1.1)–(1.2). It might be noted that previous works have analyzed the convergence of these estimated probabilities to the true ones, when the size of the catalog tends to infinity (Monbet et al. 2008). Here, we exploit the low-computational ensemble Kalman recursions (see Evensen 2009 for more details), but other stochastic methods could be used such as particle filters.

1.3 Application to the Lorenz-63 Chaotic System

In this section, we perform a simulation study to assess the assimilation performance of the proposed method on the classical Lorenz-63 model. This model has been extensively used in the literature on data assimilation (see, e.g., Miller et al. 1994, Anderson and Anderson 1999 or Van Leeuwen 1999). From a methodological point of view, it is particularly interesting due to its simplicity (in terms of dimensionality and computational cost) and its chaotic behavior. We first describe how we generate the catalog (Sect. 1.3.1) and detail how we implement the analog dynamical model in a classical stochastic filtering (Sect. 1.3.2). We then evaluate assimilation performance with respect to classical state-of-the-art data assimilation techniques (Sect. 1.3.3).

1.3.1 Synthetic Data

We generate three different datasets (true state, noisy observations, and catalog) using the exact Lorenz-63 differential equations given in Fig. 1.1 (top) with the classical parameters $\rho = 28$, $\sigma = 10$, $\beta = 8/3$ and the time step $\mathrm{d}t = 0.01$. From a random initial condition and after 500 time steps, the trajectory converges to the attractor, and we append the associated data to our datasets as follows. At each time t, the corresponding Lorenz trajectory is given by the variables x, y, and z. We store the three variables in the true state vector $\mathbf{x}(t)$. Then, we randomly generate the observations $\mathbf{y}(t)$ as the sum of the state vector and of independent Gaussian white noises with variance 2. To generate the catalog, we use another random initial condition, and after 500 time steps, we start to append the consecutive state vectors $\mathbf{z}(t)$ (the analogs) and $\mathbf{z}(t + \mathrm{d}t)$ (the successors) in the catalog. Examples of the samples stored in this catalog are given in Table 1.1.

Table 1.1 Samples of the catalog used in the ADM presented in Fig. 1.1 (bottom) to simulate realistic Lorenz-63 trajectories with a time step $\mathrm{d}t = 0.01$

$\mathbf{z}(t) \rightarrow$ Analogs	$\mathbf{z}(t + dt) \rightarrow$ Successors
$(-0.3268, +3.2644, +25.5134)$	$(+0.0131, +3.2278, +24.8371)$
$(+0.0131, +3.2278, +24.8371)$	$(+0.3177, +3.2017, +24.1889)$
$\vdots$	$\vdots$
$(-2.7587, -4.5007, +19.1790)$	$(-2.9344, -4.7112, +18.8037)$
$(-2.9344, -4.7112, +18.8037)$	$(-3.1147, -4.9464, +18.4530)$

1.3.2 The Analog Ensemble Kalman Filter and Smoother

As stressed in Sect. 1.2, the key feature of the proposed approach is to build a nonparametric sampler of the dynamics (ADM). For the considered application to Lorenz-63 dynamics, we resort to a first-order autoregressive process between $\mathbf{z}(t)$ and $\mathbf{z}(t + \mathrm{d}t)$ with $\mathrm{d}t = 0.01$ (see Sprott 2003, chapter 10, for similar applications in other chaotic models). We consider the first ten analogs (or the first ten nearest neighbors) of a given state within the built catalog of simulated Lorenz-63 trajectories presented in Table 1.1. Note that we here consider an exhaustive search within the entire catalog. This ADM is plugged into classical ensemble Kalman recursions. We implement both the ensemble Kalman filter (EnKF) and smoother (EnKS). Whereas EnKF only exploits the available observation up to the current state (i.e., past and current observations), EnKS exploits the entire observation series (i.e., both past, present, and future observations with respect to the current state). We implement the EnKF and EnKS with 100 members, value sufficiently important to correctly estimate the covariances. In the next results, we perform numerical experiments to assess the performance of the proposed approach. We vary both the time steps of the observations and the size of the catalog and analyze the impact on assimilation performance. We carry out a comparative evaluation with respect to reference assimilation models using a parametric autoregressive process and the pure dynamical Lorenz-63 equations (PDM). For each experiment, we display the ensemble mean and the 95 % confidence interval (transparent error area) of the assimilated states issued from the Gaussian smoothing probabilities estimated by the EnKS.

1.3.3 Evaluation of Assimilation Performance

We first analyze assimilation performance for noisy observations sampled at different time rates (noted as $\mathrm{d}t_{\mathrm{obs}}$), from 0.01 to 0.40. Considering the analogy between the Lorenz-63 and atmospheric time scales, note that $\mathrm{d}t_{\mathrm{obs}} = 0.08$ is equivalent to a 6 h variability in the atmosphere. As an illustration of the complexity of Lorenz-63 dynamics, we report in Fig. 1.2 (left column) the scatter cloud of two

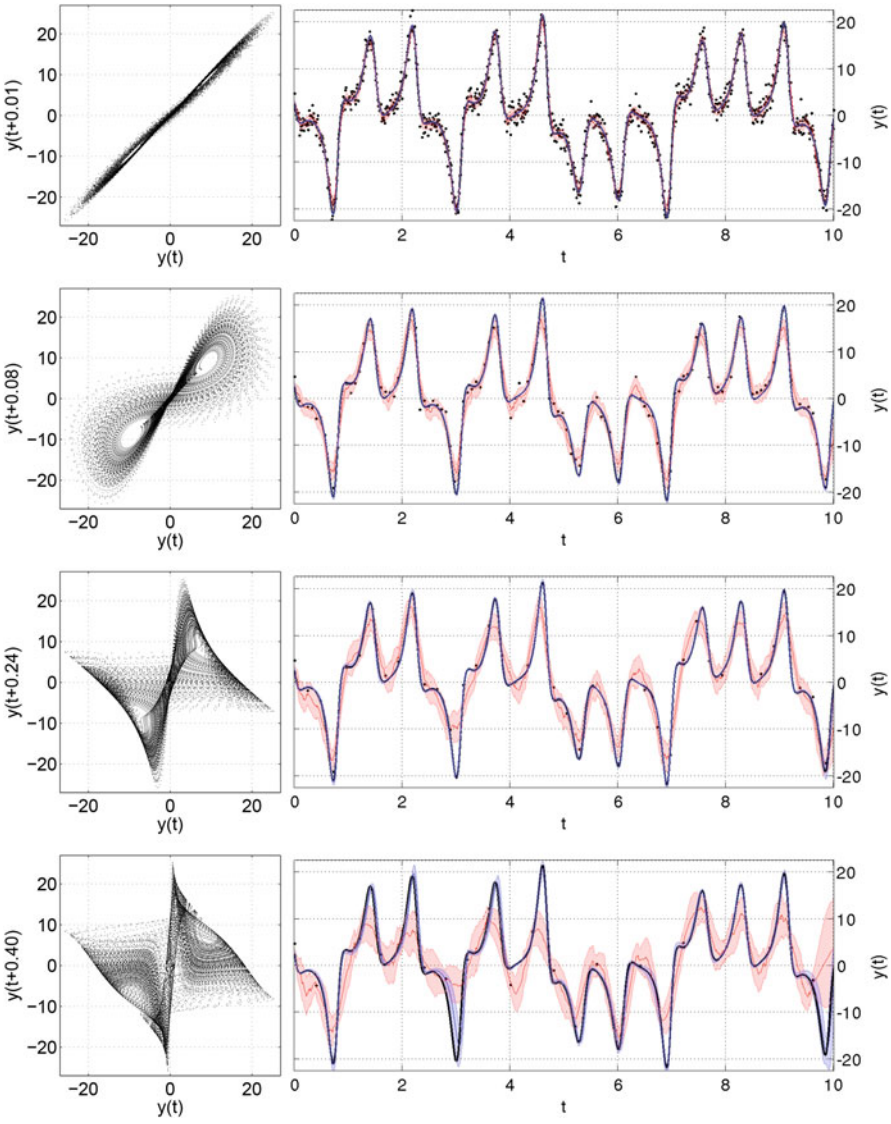

Fig. 1.2 The *left column* displays the scatter plot between two consecutive values of the Lorenz-63 second variable y. In the *right column*, the noisy observations and true states of the Lorenz-63 are respectively represented with *black dots* and *black curves*. We also display the smoothed mean estimate and the 95 % confidence interval of the assimilation of the noisy observations using a simple linear and parametric AR(1) model (*red*) and the proposed nonparametric ADM (*blue*). Experiments are carried out for different sampling rates between consecutive observations, from 0.01 to 0.40 (*top* to *bottom*)

consecutive values of the second Lorenz-63 variable y in the catalog. Whereas we observe a linear-like pattern for the fine sampling rate of 0.01 (first row), all other sampling rates clearly exhibit nonlinear patterns, which can hardly be captured by a linear dynamical model. For each time step setting, we also compare in Fig. 1.2 (right column) the observations (black dots), the true state (black curves), and the assimilation results using different dynamical models. Two results are reported: the nonparametric ADM presented in Sect. 1.3.2 (blue curves) and the parametric first-order linear autoregressive AR(1) model (red curves). For very small sampling rates between consecutive observations, a simple linear AR(1) dynamical model proves sufficient to assimilate the state of the system. But, as soon as the sampling rate becomes greater (from 0.08), such an AR(1) model can no longer drive the assimilation to relevant states. By contrast, the proposed ADM does not suffer from these limitations and show weak effects of the sampling rates on the quality of the assimilated states.

We also compare the performance of the proposed nonparametric ADM to the classical EnKS assimilation using the PDM, i.e., allowing online evaluations of the Lorenz-63 equations. We perform different simulations varying the time sampling rate between two consecutive observations $\mathrm{d}t_{\mathrm{obs}} = \{0.01, 0.08, 0.24, 0.40\}$ and the size of the catalog $n = \{10^3, 10^4, 10^5, 10^6\}$. For each experiment, we compute the root mean square error (RMSE) between the true and estimated smoothed states of the Lorenz-63 trajectories. These RMSE are computed over 10^5 time steps. To solve the differential equations of the Lorenz-63 model in the PDM, we use the explicit (4,5) Runge-Kutta integrating method (cf. Dormand and Prince 1980). Figure 1.3 summarizes the results. As benchmark curves, in dashed lines, we plot the results of the classical EnKS using the PDM. In solid lines, we report the results of the proposed EnKS using ADM. We observe a decrease of the error when the size n of

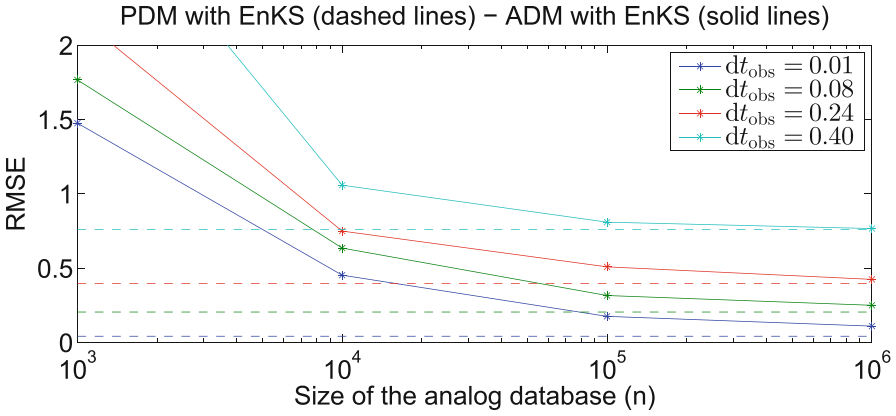

Fig. 1.3 root mean square error (RMSE) for the three variables of the Lorenz-63 model as a function of the size of the catalog (n) and the time sampling rate between consecutive observations ($\mathrm{d}t_{\mathrm{obs}}$). *Dashed* and *solid lines* refer respectively to the reanalysis (smoothed estimates) for the classical EnKS using PDM and the proposed EnKS using ADM (see Fig. 1.1 for the difference between the two approaches)

the catalog increases (x-axis in log scale). It also shows that the difference in RMSE between the two kinds of reanalysis (with and without an explicit knowledge of the Lorenz-63 equations) decreases when the time sampling rate (and thus the forecast error) between two consecutive observations $\mathrm{d}t_{\mathrm{obs}}$ increases (colors in legend). Overall, for a catalog of 10^6 samples, we report RMSE difference below 0.05 for sampling rates equal or greater than 0.08.

1.4 Conclusion and Perspectives

In this paper, we show that the statistical combination of Monte Carlo filters and analog procedures is able to retrieve the chaotic behavior of the Lorenz-63 model when the size of the catalog is sufficiently important. The proposed methodology may be a relevant alternative to the classical data assimilation schemes when (i) large observational or model-simulated databases of the process are available and (ii) physical models are computationally demanding and/or modeling uncertainties are important. The data-driven methodology proposed in this paper is a relatively low-cost procedure, which directly samples new ensembles from previously observed or simulated data, and potentially allows for an exploration of more scenarios.

Our future work will particularly investigate the application of the proposed methodology to archives of in situ measurements, remote sensing observations, and model-simulated data for the multi-source reconstruction of geophysical parameters at the surface of the ocean. The methodology seems particularly appealing for such surface oceanographic studies for three reasons: (i) the low dimensionality of the state in comparison with atmosphere and a 3D spatial grid, (ii) the less chaotic behavior of the dynamics due to the water viscosity and (iii) the amount oceanographic data at the surface of the ocean. Indeed, in the last two decades, satellite and in situ measurements have provided a wealth of information with high spatial and temporal resolutions.

Future work will also address methodological aspects, especially regarding the search procedures for the analogs and the construction of the catalog. In this Lorenz-63 example, a small part of the trajectory is really chaotic (zone close to the origin, between the two attractors), and most of the time, a simple autoregressive process is able to produce relevant forecasts in non-chaotic regions. An effort is therefore needed to evaluate the complexity of the trajectory, what may, for instance, rely on Lyapunov exponent (see Sprott 2003, chapter 10), and carefully select the samples indexed in the catalog upon their representativeness of the underlying chaotic dynamics. Another important aspect is the size of the sampled trajectories between analogs and successors in the catalog. In this paper, we use a very small time lag ($\mathrm{d}t = 0.01$), but other strategies can be used, e.g., sampling successors with the same time lag than consecutive observations ($\mathrm{d}t_{\mathrm{obs}}$). A last methodological aspect concerns the filtering methods. In such low-cost emulation of the dynamical model, particle filters and smoothers may allow more flexibility to take into account non-Gaussian assumptions.

Acknowledgements This work was supported by both EMOCEAN project funded by the "Agence Nationale de la Recherche" and a "Futur et Ruptures" postdoctoral grant from Institute Mines-Télécom.

References

Anderson JL, Anderson SL (1999) A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Mon Weather Rev 127(12):2741–2758

Bertino L, Evensen G, Wackernagel H (2003) Sequential data assimilation techniques in oceanography. Int Stat Rev 71(2):223–241

Delle Monache L, Eckel FA, Rife DL, Nagarajan B, Searight K (2013) Probabilistic weather prediction with an analog ensemble. Mon Weather Rev 141(10):3498–3516

Dormand JR, Prince PJ (1980) A family of embedded Runge-Kutta formulae. J Comput Appl Math 6(1):19–26

Evensen (2009) Data assimilation: the ensemble Kalman filter. Springer, Berlin

Friedman JH, Bentley JL, Finkel RA (1977) An algorithm for finding best matches in logarithmic expected time. ACM Trans Math Softw (TOMS) 3(3):209–226

Hoteit I, Pham DT, Triantafyllou G, Korres G (2008) A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. Mon Weather Rev 136(1):317–334

Lorenc AC, Ballard SP, Bell RS, Ingleby NB, Andrews PLF, Barker DM, Bray JR, Clayton AM, Dalby T, Li D, Payne TJ, Saunders FW (2000) The Met. Office global three-dimensional variational data assimilation scheme. Q J R Meteorol Soc 126(570):2991–3012

Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20(2):130–141

Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. J Atmos Sci 26(4):636–646

Lott F, Miller MJ (1997) A new subgrid-scale orographic drag parametrization: its formulation and testing. Q J R Meteorol Soc 123(537):101–127

Miller RN, Ghil M, Gauthiez F (1994) Advanced data assimilation in strongly nonlinear dynamical systems. J Atmos Sci 51(8):1037–1056

Monbet V, Ailliot P, Marteau P-F (2008) L1-convergence of smoothing densities in non-parametric state space models. Stat Inference Stoch Process 11(3):311–325

Pham DT (2001) Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon Weather Rev 129(5):1194–1207

Ruiz J, Saulo C, Nogués-Paegle J (2010) WRF model sensitivity to choice of parameterization over South America: validation against surface variables. Mon Weather Rev 138(8):3342–3355

Schenk F, Zorita E (2012) Reconstruction of high resolution atmospheric fields for Northern Europe using analog-upscaling. Clim Past Discuss 8(2):819–868

Sprott JC (2003) Chaos and time-series analysis. Oxford University Press, Oxford

Timbal B, Dufour A, McAvaney B (2003) An estimate of future climate change for western France using a statistical downscaling technique. Clim Dyn 20(7–8):807–823

Van den Dool H (2006) Empirical methods in short-term climate prediction. Oxford University Press, Oxford

Van Leeuwen PJ (1999) Nonlinear data assimilation in geosciences: an extremely efficient particle filter. Q J R Meteorol Soc 136(653):1991–1999

Yiou P, Salameh T, Drobinski P, Menut L, Vautard R, Vrac M (2013) Ensemble reconstruction of the atmospheric column from surface pressure using analogues. Clim Dyn 41(5–6):1333–1344

Chapter 2
Machine Learning Methods for ENSO Analysis and Prediction

Carlos H.R. Lima, Upmanu Lall, Tony Jebara, and Anthony G. Barnston

Abstract The El Niño-Southern Oscillation (ENSO) plays a vital role in the interannual variability of the global climate. In order to reduce its adverse impacts on society, many statistical and dynamical models have been used to predict its future states. However, most of these models present a limited forecast skill for lead times beyond 6 months. In this paper, we present and discuss results from previous work and describe the University of Brasilia/Columbia Water Center (UNB/CWC) ENSO forecast model, which has been recently developed and incorporated into the ENSO Prediction Plume provided by the International Research Institute for Climate and Society. The model is based on a nonlinear method of dimensionality reduction and on a regularized least squares regression. This model is shown to have a skill similar to or better than other ENSO forecast models, particularly for longer lead times. Many dynamical and statistical models predicted a strong El Niño event in 2014. The UNB/CWC model did not, consistent with the subsequent observations. The model's ENSO predictions for 2014 are presented and discussed.

Keywords Dimensionality reduction • Nonlinear • Regularized least squares

C.H.R. Lima (✉)
Civil and Environmental Engineering, University of Brasilia, Brasilia, Brazil
e-mail: chrlima@unb.br

U. Lall
Earth and Environmental Engineering, Columbia University, New York, NY, USA
e-mail: ula2@columbia.edu

T. Jebara
Computer Science, Columbia University, New York, NY, USA
e-mail: jebara@cs.columbia.edu

A.G. Barnston
International Research Institute for Climate and Society, The Earth Institute of Columbia University, New York, NY, USA
e-mail: tonyb@iri.columbia.edu

V. Lakshmanan et al. (eds.), *Machine Learning and Data Mining Approaches to Climate Science*, DOI 10.1007/978-3-319-17220-0_2

2.1 Introduction

The term El Niño-Southern Oscillation (ENSO) refers to a coupled ocean-atmosphere phenomenon that takes place along the Tropical Pacific Ocean and consists of anomalies in the sea surface temperature (SST) and sea level pressure (SLP) across the entire Pacific basin. Positive anomalies (warm events) in the eastern Tropical Pacific SST are associated with a reduction in the SLP gradient across the basin, and this event is called El Niño. It has a periodicity of about 4–6 years (Diaz and Markgraf 2000) and is accompanied by changes in the atmospheric circulation in the equatorial region, most notably in the Walker circulation cells, which in turn affect rainfall and temperature patterns across the globe. The opposite phase of El Niño is called La Niña (ENSO cold events) and consists of negative anomalies in the SST in the central and eastern part of the equatorial Pacific basin and an enhancement of the cross-basin SLP gradient and consequently in the trade winds. We refer the reader to Diaz and Markgraf (2000) for further details on ENSO variability and its impacts on climate and society.

A recent review (Barnston et al. 2012) of the skill of 12 dynamical and 8 statistical ENSO models for real-time forecasts during 2002–2011 shows an average correlation skill of 0.42 at a 6-month lead time, which is lower than the average correlation skill (0.65) for the 1981–2010 period obtained from the same models and lead time but in a hindcast design. Barnston et al. (2012) suggest that the difference in the skills is explained by the design of the forecasts (real time vs. hindcast) as well as by the lower ENSO variability during 2002–2011, which makes forecasts more challenging. Barnston et al. (2012) emphasize that predictions at lead times greater than 6 months continue to lack skill.

For predicting ENSO indices, statistical models have used gridded SST, wind and SLP fields, and, more recently, ocean subsurface temperature data (Drosdowsky 2006). Principal component analysis (PCA) has been widely applied to identify the key modes of variability in such data and for reducing the dimensionality of the predictors in forecasting models. A regression model that uses the leading modes is then used to predict an ENSO index. However, since PCA is based on the eigenvalue decomposition of the covariance (or correlation) matrix of the input data, it considers only the linear dependence structure. In high-dimensional spaces, where variables are nonlinearly correlated, PCA may need a large number of principal components to approximate the main modes of spatiotemporal variability of such systems.

In this paper, we extend previous work (Lima et al. 2009) and describe the University of Brasilia/Columbia Water Center (UNB/CWC) ENSO forecast model, which has been recently developed and incorporated into the ENSO Prediction Plume provided by the International Research Institute for Climate and Society (IRI). We apply a nonlinear method of dimensionality reduction developed by the machine learning community (Weinberger and Saul 2006) to identify the spatiotemporal variability of the depth of the 20°C isotherm (D_{20}) along the Tropical Pacific Ocean, which is a proxy for the thermocline and a carrier of the long-lead

ENSO signal (Drosdowsky 2006). The leading modes of variability of the Tropical Pacific thermocline data are obtained by this method and used as predictors in a regression model for operational ENSO forecasts at different lead times. We use the top three modes at different lags to predict ENSO through a regularized least squares regression model. The rest of this paper is organized as follows. In Sect. 2.2, we present the climate dataset. The mathematical details of the forecast model are presented in Sect. 2.3. Some features of the spatial modes of the D_{20} field and the model skills for cross-validated ENSO forecasts are offered in Sect. 2.4, which is followed by a summary of the paper.

2.2 Climate Dataset

As a proxy for the Tropical Pacific thermocline and heat content, we use the National Oceanic and Atmospheric Administration (NOAA)/National Centers for Environmental Prediction (NCEP) thermocline depth at 20 °C (D_{20}), which is derived from a global ocean data assimilation system (GODAS) (Behringer and Xue 2004). Our focus here is on the Pacific D_{20} bounded by the region 26°N–28°S and 122°E–77°W. The dataset starts in January 1980 and is updated regularly. It consists of 26,243 data points located in an equally spaced grid cell with resolution 1/3 degree by 1/3 degree. As a representative of ENSO events (Barnston et al. 1997), we use the NCEP NINO3.4 index defined as the monthly mean SST anomalies averaged over the area 5°N–5°S and 170°W–120°W. Both datasets are provided by IRI at http://iridl.ldeo.columbia.edu/SOURCES/.

2.3 Technical Approach

2.3.1 Nonlinear Dimensionality Reduction

Nonlinear methods of dimensionality reduction are usually derived by first mapping the original dataset that lies on a nonlinear space (or manifold) onto a linear space (the feature space) and second by applying PCA on the projected input data. A common method is kernel principal component analysis, which was first introduced by Schölkopf et al. (1998) and uses the concept of kernels to map the original dataset onto a linear feature space. Mathematically, let $\mathbf{X^T}$ be a $N \times M$ centered matrix of inputs. Here, $\mathbf{X^T}$ refers to the transpose matrix of $\mathbf{X}$, and N and M are the number of months and grid points of the D_{20} data used in the analysis, respectively. Using the concept of singular value decomposition factorization $\mathbf{X^T} = \mathbf{U\Sigma V^T}$, the $L \times N$ matrix $\mathbf{Y}$ of the projection of the data matrix $\mathbf{X}$ onto the first L eigenvectors is given by:

$$\mathbf{Y} = \Sigma \mathbf{V^T} \tag{2.1}$$

where $\mathbf{V}$ is the $N \times L$ matrix of eigenvectors of the Gram matrix $\mathbf{G} = \mathbf{X}\mathbf{X}^{\mathbf{T}}$ corresponding to the top L eigenvalues and Σ is the diagonal matrix of *square roots* of the top L eigenvalues of $\mathbf{G}$.

Consider now a nonlinear function Φ defined by any nonlinear basis function (e.g., $\Phi(\mathbf{x}_i) = \mathbf{x}_i^2$) that maps each point of the input data to the feature space $\mathscr{H}$. The idea here is to apply PCA in the space defined by $\Phi(\mathbf{X})$ rather than $\mathbf{X}$, in order to obtain a set of low-dimensional vectors that accounts for the maximum variance in the new space $\mathscr{H}$. The leading modes can be obtained in a manner similar to PCA:

$$\Phi(\mathbf{X})^{\mathbf{T}} = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}} \tag{2.2}$$

where $\mathbf{U}$ has the eigenvectors of $\Phi(\mathbf{X})^{\mathbf{T}}\Phi(\mathbf{X})$, $\mathbf{V}$ the eigenvectors of $\mathbf{K} = \Phi(\mathbf{X})\Phi(\mathbf{X})^{\mathbf{T}}$, and Σ is the diagonal matrix of *square roots* of the eigenvalues of $\mathbf{K}$.

Using the so-called *kernel trick*, the elements of the N by N Gram matrix K are obtained without the need to compute $\Phi(\mathbf{x})$ explicitly. The principal modes of $\mathbf{X}$ are obtained as in Eq. (2.1), but substituting the Gram matrix $\mathbf{G}$ by the kernel function $\mathbf{K}$.

Instead of defining a function Φ, Weinberger and Saul (2006) proposed to maximize the trace (sum of the eigenvalues) of the kernel matrix $\mathbf{K}$ by exploring choices of kernel values between pairs of inputs that still preserve the distances between nearby points in the original space. This method, known as maximum variance unfolding (MVU), can be defined through the following optimization problem:

Maximize trace($\mathbf{K}$) s.t.:

$$\mathbf{K} \succeq 0; \tag{2.3}$$

$$\sum_{ij} K_{ij} = 0; \tag{2.4}$$

$$K_{ii} + K_{jj} - K_{ij} - K_{ji} = G_{ii} + G_{jj} - G_{ij} - G_{ji}, \quad \forall i,j \text{ where } \eta_{ij} = 1, \tag{2.5}$$

where η_{ij} is 1 if i and j are k-nearest neighbors of each other and 0 otherwise. More details about the optimization problem can be seen in Weinberger and Saul (2006). The leading modes $\mathbf{Y}$ of $\mathbf{X}$ in the new space $\mathscr{H}$ are obtained as in Eq. (2.1) but substituting $\mathbf{G}$ by $\mathbf{K}$.

2.3.2 Forecast Model

The forecast model for the NINO3.4 index for a lead time τ can be written as:

$$F(t+\tau) = \beta_{0,\tau,t} + \beta_{1,\tau,t} \cdot O(t) + \sum_{l=t-24}^{t} \beta_{2,\tau,l} \cdot Y_1(l) + \beta_{3,\tau,l} \cdot Y_2(l) + \beta_{4,\tau,l} \cdot Y_3(l) + \epsilon_\tau(t), \tag{2.6}$$