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 offer 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, modeling uncertainties. In this paper, we explore an alternative approach 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

Pierre Tandeo
Télécom Bretagne, e-mail: pierre.tandeo@telecom-bretagne.eu

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

Juan Ruiz
National Scientific and Technical Research Council, e-mail: jruiz@cima.fcen.uba.ar

Alexis Hannart
National Scientific and Technical Research Council, e-mail: alexis.hannart@cima.fcen.uba.ar

Bertrand Chapron
Ifremer, e-mail: bertrand.chapron@ifremer.fr

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

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

Robert Easton
University of Colorado, e-mail: robert.easton@colorado.edu

Ronan Fablet
Télécom Bretagne, e-mail: ronan.fablet@telecom-bretagne.eu

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, reach state-of-the-art performances, without online evaluations of the physical model.

Key words: Data assimilation, Stochastic filtering, Nonparametric sampling, Analog method, Lorenz-63 model

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., [1] 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 ([3]), and stochastic data assimilation schemes, which involve Monte Carlo strategies and are particularly appealing for their modeling flexibility ([4]). 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 ([18]). 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 ([19]). These different examples tend to show that multiple evaluations of an explicit physical model is 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 advocate for exploring implicit data-driven schemes to build realistic statistical simu-

lations of the dynamics for data assimilation issues. Satellite sequence images are typical examples. When the spatio-temporal sampling and the amount of historical remote sensing data is 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 above mentioned 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 ([6]). The paper is organized as follows. In Section 2, we propose to use a nonparametric sampler, based on the analog (or nearest neighbors) method, to generate the forecast members ([5]). Then, we use the ensemble Kalman recursions to combine these members with the observations ([1]). In Section 3, we numerically evaluate the methodology on the Lorenz-63 model such as various previous works (see e.g., [21], [20]). We further discuss and summarize the key results of our investigations in Section 4.

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. [4]):

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathscr{M}(\mathbf{x}(t), \eta(t)) \tag{1}$$

$$\mathbf{y}(t) = \mathscr{H}(\mathbf{x}(t), \varepsilon(t)) \quad . \tag{2}$$

The dynamical model given in Eq. (1) describes the evolution of the true physical process $\mathbf{x}(t)$. It includes a random perturbation $\eta(t)$ 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 3-dimensional vector (x,y,z). The observation model given in Eq. (2) links the observation $\mathbf{y}(t)$ to the true state at the same time t. It also includes a random noise $\varepsilon(t)$ which models observation error and uncertainties, change of support (i.e. downscaling/upscaling effects) and so on.

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 (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 η given in Eq. (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 ([2]). In the geoscience literature, we talk about analog methods (see e.g. [6] or [11]). They were initially devised for weather prediction, but applications to downscaling issues ([13]) or climate reconstructions ([14], [15]) were also proposed. As described in Fig. 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 ([12]).

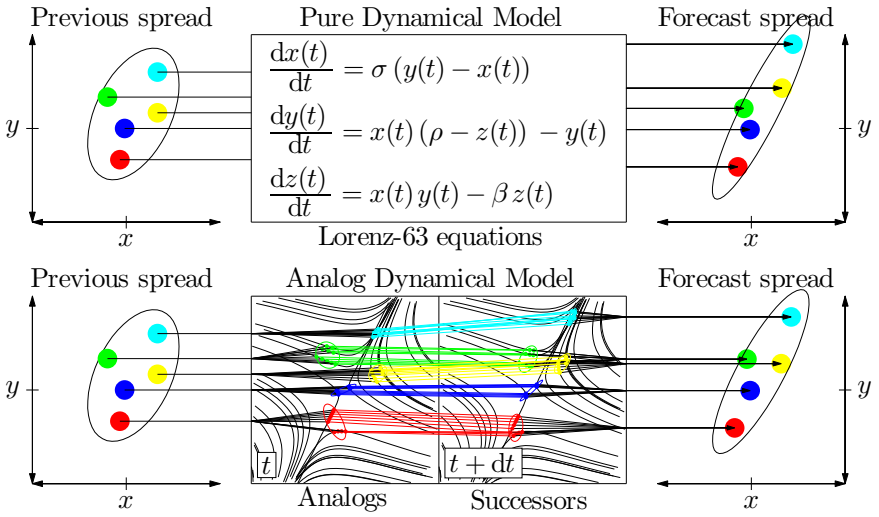


Fig. 1 Sketch of the forecast step in stochastic data assimilation schemes using pure (top) and analog (bottom) dynamical models. As an example, we consider the 3-dimensional Lorenz-63 chaotic model. For visualization convenience, we only represent the xy-plane, centered at the origin. We track five statistical members with the variability depicted by ellipsoids accounting for the covariance structure.

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 Eq. (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

[7]. Here, we exploit the low-computational ensemble Kalman recursions (see [1] for more details) but other stochastic methods could be used such as particle filters.

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., [16], [9] or [8]). 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. 3.1) and detail how we implement the analog dynamical model in a classical stochastic filtering (Sect. 3.2). We then evaluate assimilation performance with respect to classical state-of-the-art data assimilation techniques (Sect. 3.3).

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 (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 states 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.

Table 1 Samples of the catalog used in the ADM presented in Fig. 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
$(-0.3268, +3.2644, +25.5134)$	$(+0.0131, +3.2278, +24.8371)$
$(+0.0131, +3.2278, +24.8371)$	$(+0.3177, +3.2017, +24.1889)$
$\vdots$	$\vdots$
$(-2.7587, -4.5007, +19.1790)$	$(-2.9344, -4.7112, +18.8037)$
$(-2.9344, -4.7112, +18.8037)$	$(-3.1147, -4.9464, +18.4530)$

3.2 The Analog Ensemble Kalman Filter and Smoother

As stressed in Sect. 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 [17], chapter 10, for similar applications in other chaotic models). We consider the first 10 analogs (or the first 10-nearest neighbors) of a given state within the built catalog of simulated Lorenz-63 trajectories presented in Table 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 exploit 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.

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 hours variability in the atmosphere. As an illustration of the complexity of Lorenz-63 dynamics, we report in Fig. 2 (left column) the scatter cloud of two 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. 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. 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 limita-

tions and show weak effects of the sampling rates on the quality of the assimilated states.

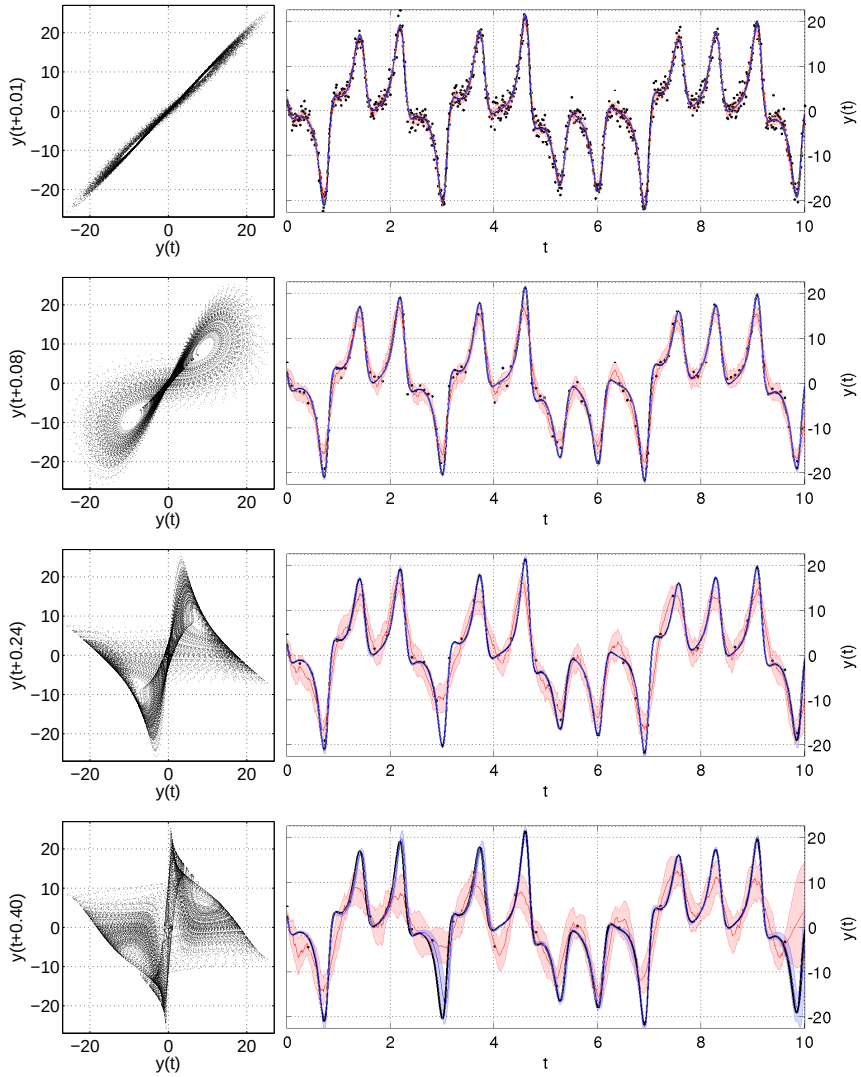


Fig. 2 The left column displays the scatters plot between two consecutive values of the Lorenz-63 second variable y. In the right column, the noisy observations and true states of the Lorenz-63 are respectively represented with black dots and black curves. We also display the smoothed mean estimate and 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).

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. [10]). Fig. 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 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.

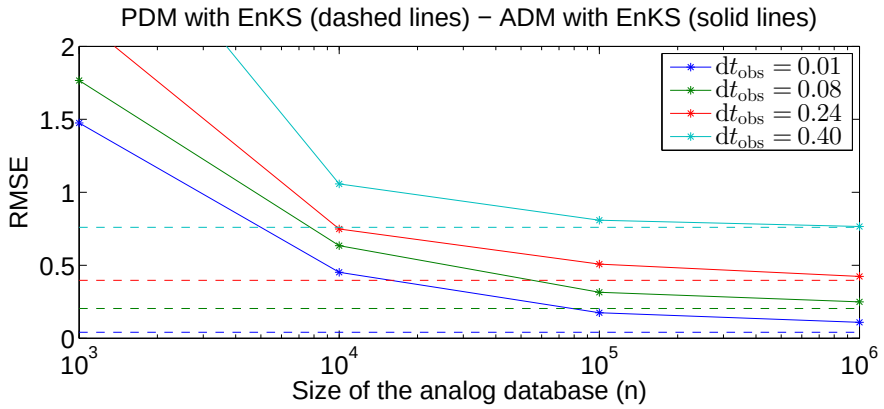


Fig. 3 Root Mean Square Error (RMSE) for the three variables of the Lorenz-63 model as a function of the size of the catalog (n) and the time sampling rate between consecutive observations ($\mathrm{d}t_{\mathrm{obs}}$). Dashed and solid lines refer respectively to the reanalysis (smoothed estimates) for the classical EnKS using PDM and the proposed EnKS using ADM (see Fig. 1 for the difference between the two approaches).

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 2 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 [17], 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

1. Evensen G. (2007) Data assimilation. *Springer*.

2. Friedman JH, Bentley JL, Finkel RA. (1977) An algorithm for finding best matches in logarithmic expected time *ACM Transactions on Mathematical Software (TOMS)* 3(3):209-226.
3. 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. *Quarterly Journal of the Royal Meteorological Society* 126(570):2991-3012.
4. Bertino L, Evensen G, Wackernagel H. (2003) Sequential data assimilation techniques in oceanography. *International Statistical Review* 71(2):223-241.
5. Delle Monache L, Eckel FA, Rife DL, Nagarajan B, Searight K. (2013) Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review* 141(10):3498-3516.
6. Lorenz EN. (1963) Deterministic nonperiodic flow. *Journal of the atmospheric sciences* 20(2):130-141.
7. Monbet V, Ailliot P, Marteau P-F. (2008) L1-convergence of smoothing densities in non-parametric state space models. *Statistical Inference for Stochastic Processes* 11(3):311-325.
8. Van Leeuwen PJ. (1999) Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society* 136(653):1991-1999.
9. Anderson JL, Anderson SL. (1999) A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review* 127(12):2741-2758.
10. Dormand JR, Prince PJ. (1980) A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics* 6(1):19-26.
11. Van den Dool H. (2006) Empirical methods in short-term climate prediction. *Oxford University Press*.
12. Lorenz EN. (1969) Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences* 26(4):636-646.
13. Timbal B, Dufour A, McAvaney B. (2003) An estimate of future climate change for western France using a statistical downscaling technique. *Climate Dynamics* 20(7-8):807-823.
14. Schenk F, Zorita E. (2012) Reconstruction of high resolution atmospheric fields for Northern Europe using analog-upscaling. *Climate of the Past Discussions* 8(2):819-868.
15. Yiou P, Salameh T, Drobinski P, Menut L, Vautard R, Vrac M. (2013) Ensemble reconstruction of the atmospheric column from surface pressure using analogues. *Climate dynamics* 41(5-6):1333-1344.
16. Miller RN, Ghil M, Gauthiez F. (1994) Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of the Atmospheric Sciences* 51(8):1037-1056.
17. Sprott JC. (2003) Chaos and time-series analysis. *Oxford: Oxford University Press*.
18. Ruiz J, Saulo C, Nogués-Paegle J. (2010) WRF model sensitivity to choice of parameterization over South America: validation against surface variables. *Monthly Weather Review* 138(8):3342-3355.
19. Lott F, Miller MJ. (1997) A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quarterly Journal of the Royal Meteorological Society* 123(537):101-127.
20. 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. *Monthly Weather Review* 136(1): 317-334.
21. Pham DT. (2001) Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review* 129(5): 1194-1207.