The Analog Ensemble Kalman Filter and Smoother

Pierre Tandeo, Pierre Ailliot, Ronan Fablet, Juan Ruiz, François Rousseau, Bertrand Chapron

4th International Workshop on Climate Informatics
September 25-26, 2014
Hosted by the National Center for Atmospheric Research in Boulder, CO

THE ANALOG ENSEMBLE KALMAN FILTER AND SMOOTHER

Pierre Tandeo[1], Pierre Ailliot[2], Ronan Fablet[1], Juan Ruiz[3], Bertrand Chapron[4]

***Abstract*—In classical data assimilation using sequential Monte Carlo methods, a physical model is run at each time steps to simulate members corresponding to different forecast scenarios. In this paper, we propose to use statistical analogs provided by observational or model-simulated data to emulate the dynamical model and generate relevant forecast members. This new methodology is called AnEnKF/AnEnFS for Analog Ensemble Kalman Filter and Smoother. We test our methodology using the Lorenz-63 model. The simulations indicate that, for a rich analog database, the assimilation results with the AnEnKF/AnEnFS are comparable to those obtained using the Lorenz dynamical equations into a classical Ensemble Kalman Filter/Smoother.**

I. MOTIVATION

Data assimilation methods combines information of a physical dynamical model and observations (see e.g., [1] and reference therein). Nowadays, due to their flexibility, sequential Monte Carlo filters are widely used in geoscience ([2]). In these methods, several members are generated and compared to the observations at each time step. This generally leads to intensive computation in practical applications since the physical model need to be run with different initial conditions at each time step in order to generate the members. This number of members must be high enough to explore the state space of the physical model.

The amount of observational and model-simulated data has grown very quickly in the last decades. These datasets may now provide enough information to build realistic statistical emulators of the dynamics of the geophysical variables and generate members at a lower computational cost compared to running a physical dynamical model. In this study, we propose to use the analog (or nearest neighbors) method to generate the members ([3]) and the classical Ensemble Kalman recursions to combine these members with the observations (see [1] for more details). The feasibility of our method is illustrated on the classical Lorenz-63 model ([4]).

Corresponding author: P. Tandeo, Telecom Bretagne, Technopôle Brest Iroise - CS 83818, 29238 BREST Cedex, France, pierre.tandeo@telecom-bretagne.eu [1]CNRS UMR 6285 LabSTICC - Pôle CID, France [2]Laboratoire de Mathématiques de Bretagne Atlantique, UMR 6205, Université de Brest, France [3]National Scientific and Technical Research Council, Buenos Aires, Argentina [4]Oceanography from Space Laboratory, IFREMER, Plouzané, France

II. METHOD

Sequential data assimilation techniques are generally formulated using a nonlinear state space model (see e.g. [2])

$$\mathbf{x}_t = \mathcal{M}(\mathbf{x}_{t-1}, \boldsymbol{\eta}_t) \quad (1)$$

$$\mathbf{y}_t = \mathcal{H}(\mathbf{x}_t, \boldsymbol{\epsilon}_t) \quad . \quad (2)$$

The dynamical (or "process") model given in Eq. (1) describes the evolution of the "true" geophysical process $\mathbf{x}_t$ and includes a random perturbation $\boldsymbol{\eta}_t$ which accounts for the various sources of uncertainties (e.g. boundary conditions, forcing terms, physical parameterization, etc...). The observation (or "data") model given in Eq. (2) links the observation $\mathbf{y}_t$ with the true state at the same time t. It also includes a random noise $\boldsymbol{\epsilon}_t$ which models observation error, change of support and so on.

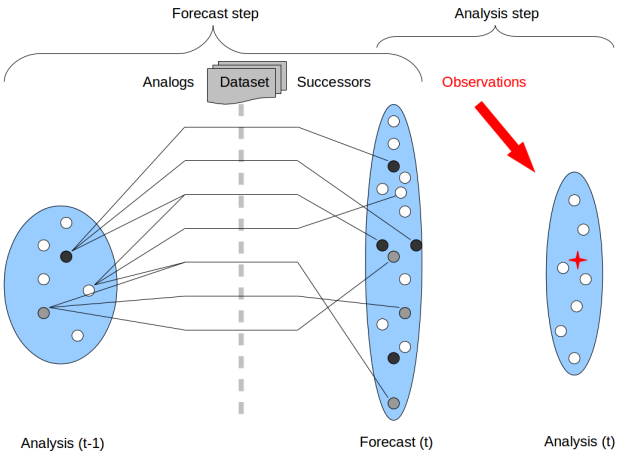

Fig. 1. Scheme of the Analog Ensemble Kalman Filter.

The main originality of the methodology proposed in this paper consists in using statistical analogs to approximate the dynamical model given in Eq. (1). More

4th International Workshop on Climate Informatics
September 25-26, 2014
Hosted by the National Center for Atmospheric Research in Boulder, CO

precisely, as illustrated in Fig. 1 ("Forecast step"), we assume that a catalog describing the time evolution of the state $\mathbf{x}_t$ is available. This catalog is used to build an emulator of the dynamical model $\mathcal{M}$ and the associated error $\boldsymbol{\eta}$ which can be run much faster than the physical model. In practice, if $\mathbf{x}_{t-1}$ denotes the state at time $t-1$, then the analogs are the points in the catalog which are close to $\mathbf{x}_{t-1}$ and the successors of these analogs may be used to generate possible forecast states of the geophysical process at time t. Here, to select the best analogs, we use a classical machine learning algorithm: the k-Weighted Nearest Neighbors (k-WNN).

Then, as described in Fig. 1 ("Analysis step"), this estimate can then be plugged in a standard algorithm (e.g. Kalman recursions or particle filters) to estimate the filtering or smoothing probabilities for the state space model (1-2). The convergence of these estimated filtering and smoothing probabilities to the true ones, when the size of the catalog tends to infinity, is discussed in [5]. In the next section, we perform a simulation study to assess the behavior of the method on a classical toy example which has been extensively used in the literature on data assimilation.

III. NUMERICAL EVALUATION

Here, we generate three different datasets (true state, noisy observations and analog database) using the exact Lorenz-63 differential equations with the classical parameters $\rho = 28$, $\sigma = 10$, $\beta = 8/3$ and the delta time dt $= 0.01$. From a random initial condition and after 500 time steps, the trajectory converges to the attractor and we start to generate the data. 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 a Gaussian white noise with variance 2. To generate the analog database, we use another random initial condition and after 500 time steps, we start to store the consecutive states vectors $\mathbf{x}(t)$ in the analog database.

In this paper, we assimilate the noisy observations $\mathbf{y}(t)$ with (i) the classical EnKF/EnKS using the Pure Dynamical Model (PDM) corresponding to the exact Lorenz-63 differential equations and (ii) the AnEnKF/AnEnKS using the Analog Dynamical Model (ADM) evaluated at each iteration using the k-WNN algorithm. We perform different simulations varying (j) the time step between two consecutive observations $\text{dt}_{\text{obs}} = \{0.01, 0.08, 0.24, 0.40\}$ and (jj) the size of the analog database $n = \{10^3, 10^4, 10^5, 10^6\}$. For each experiment, we compute the Root Mean Square Error (RMSE) between the true state and the different reanalysis obtained by the smoothing probabilities using $N = 100$ members.

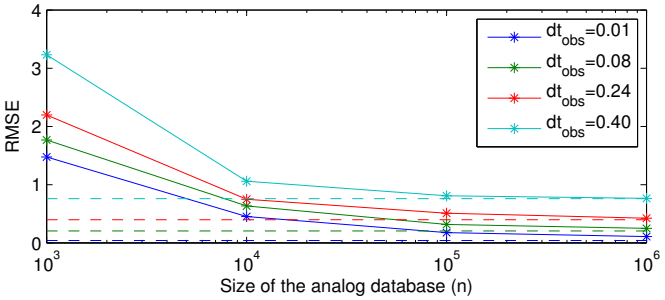

Fig. 2. Root Mean Square Error for the three variables of the Lorenz-63 model as a function of the size of the training database n and the observation time sampling dt_{obs}. Full and straight lines represent respectively the reanalysis results for the EnKS using PDM and the AnEnKS using ADM.

Experiment results are given in Fig. 2. As benchmark curves, in dashed lines, we plot the results of the classical EnKS using the PDM. In full lines, we can see the rapid decrease of the error when the size of the analog database n increases (x-axis in log scale). It also shows that the difference of RMSE between the two kinds of reanalysis (PDM and ADM) decreases when the time step (and thus the forecast error) between two consecutive observations dt_{obs} increases (colours in legend).

IV. CONCLUSIONS AND PERSPECTIVES

In this paper, we show that the statistical combination of Monte Carlo members and k-WNN procedures is able to model the nonlinearities of the chaotic Lorenz-63 model. In future works, we plan to apply this AnEnKF/AnEnKS methodology to archives of remote sensing data and model-simulated data for the interpolation of geophysical parameters at the surface of the ocean. We also plan to use the analogs together with more flexible particle filters and smoothers.

REFERENCES

[1] G. Evensen, *Data assimilation*. Springer, 2007.
[2] L. Bertino, G. Evensen, and H. Wackernagel, "Sequential data assimilation techniques in oceanography," *International Statistical Review*, vol. 71, no. 2, pp. 223–241, 2003.
[3] L. Delle Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, "Probabilistic weather prediction with an analog ensemble," *Monthly Weather Review*, vol. 141, no. 10, pp. 3498–3516, 2013.
[4] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.
[5] V. Monbet, P. Ailliot, and P.-F. Marteau, "L1-convergence of smoothing densities in non-parametric state space models," *Statistical Inference for Stochastic Processes*, vol. 11, no. 3, pp. 311–325, 2008.