Predicting Analog Forecasting Errors using Dynamical Systems

Paul Platzer, Pascal Yiou, Pierre Tandeo, Philippe Naveau, Jean-François Filipot

PREDICTING ANALOG FORECASTING ERRORS USING DYNAMICAL SYSTEMS

Paul Platzer[1,2,3], Pascal Yiou[1], Pierre Tandeo[2], Philippe Naveau[1], Jean-François Filipot[3]

***Abstract*—Nearest neighbor algorithms called Analog Forecasting have been used to produce short-range to long-range forecasts in many atmospheric and oceanic applications. Analog Forecasting is often treated as a purely empirical method, independent from physical equations. In this paper, we investigate Analog Forecasting error from a dynamical systems point of view. Assuming that analogs follow the same dynamics as the system of concern, we evaluate statistical properties of Analog Forecasting errors. We further design dynamics-based systematic error correction methods for standard Analog Forecasting techniques. These procedures are tested on the 3-dimensional Lorenz-63 system.**

I. MOTIVATION

Two states of a system are called ”analogs” when they meet a similarity criterion such as a low Euclidean distance. Analog Forecasting (AF) is based on the assumption that similar states will have similar evolution, and produces forecasts based on the ”successors” in time of the analogs of the current state. AF has been used in a wide variety of applications in atmospheric prediction, see for instance [1], [2] or [3]. Although AF is less precise than forecasting methods based on the numerical resolution of physical equations, AF can provide statistical forecasts at a low computational cost, outperforming persistence and climatological forecasts. Also, increasing observational data and computer memory make analogs a promising forecast tool.

Although the concept of analogs was originally introduced by [4] to gain information on the dynamics of the atmosphere, today AF is mostly used as a purely empirical method and the link between AF and the underlying dynamics of the system is rarely mentioned. However [5] could improve the skills of AF by using dynamics-based weights. [6] used a dynamical systems framework to study analogs, but focusing on recurrence time statistics rather than AF performance.

Corresponding author: P Platzer, paul.platzer@imt-atlantique.fr [1]Laboratoire des Sciences du Climat et de l’Environnement, Saclay, France [2]IMT-Atlantique, Plouzané, France [3]France Énergies Marines, Plouzané, France

Here we propose to improve our understanding of forecasting errors associated with AF by expressing them as a function of the dynamics of the system of concern. We assume, as a first step, that the analogs and the system follow exactly the same dynamics. With calculations similar to [7], we give a simple approximation for the error of standard AF techniques for small lead times. This enables us to evaluate AF root mean square error (RMSE). We further propose two systematic error correction methods for standard AF techniques. We test our methods on the famous Lorenz-63 system [8] in numerical experiments.

This study aims at bridging the gap between purely data-driven AF and purely model-driven methods. We show that understanding the role of the system’s dynamics can not only provide statistical information about AF error, but also help improving AF performances.

II. METHOD

Let $\mathbf{x}_0$ represent the system state at time $t = 0$. We are interested in estimating $\mathbf{x}_t$, for time $t > 0$. AF starts with finding a finite number of analogs of $\mathbf{x}_0$ inside a large database called the catalog. We note $\mathbf{a}_0^k$ the k-th analog of $\mathbf{x}_0$. Then AF considers their successors at time t, where the k-th is noted $\mathbf{a}_t^k$. Then, a common AF technique is the Locally Constant (LC) forecast [9], which takes a weighted average of all successors (with weights ω_k) as a forecast. LC forecast is written:

$$\mathbf{LC}_t = \sum_k \omega_k \mathbf{a}_t^k \, . \tag{1}$$

The Locally Incremental (LI) forecast [9] makes use of the differences between successors and analogs called ”increments”, i.e. $(\mathbf{a}_t^k - \mathbf{a}_0^k)$, rather than successors $\mathbf{a}_t^k$. LI takes the sum of the initial state $\mathbf{x}_0$, and the weighted average of increments $\sum_k \omega_k (\mathbf{a}_t^k - \mathbf{a}_0^k)$. It can also be written:

$$\mathbf{LI}_t = \mathbf{LC}_t - \boldsymbol{\varepsilon}_0, \tag{2}$$

where $\boldsymbol{\varepsilon}_0 = \mathbf{LC}_0 - \mathbf{x}_0$. We suppose that $\mathbf{x}$ follows the dynamical equation $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x})$, as well as all $\mathbf{a}^k$. We write $\mathbf{J}_t = \nabla \mathbf{f}(\mathbf{x}_t)^{\mathrm{T}}$ (T-superscript is the transpose and ∇ is the gradient operator) the Jacobian matrix of $\mathbf{f}$ along the trajectory $\mathbf{x}$ at time t. If we further assume that $\omega_k \|\mathbf{a}_0^k - \mathbf{x}_0\|$ is small enough (up to a given norm) for all k, the error for the LC forecast is:

$$\mathbf{LC}_t - \mathbf{x}_t \approx \left[\mathbf{I} + t\mathbf{J}_0 + \frac{t^2}{2}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0) \right] \varepsilon_0, \quad (3)$$

and the error for the LI forecast is:

$$\mathbf{LI}_t - \mathbf{x}_t \approx \left[t\mathbf{J}_0 + \frac{t^2}{2}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0) \right] \varepsilon_0. \quad (4)$$

where $\mathbf{I}$ is the identity matrix and $\dot{\mathbf{J}}_0 = \frac{\mathrm{d}\mathbf{J}_0}{\mathrm{d}t}$. The derivation of these formulae is similar to the calculations of [7]. From those formulae we can:

1) infer statistical properties of AF errors associated with LC-type and LI-type AF techniques,
2) apply a systematic error correction to improve the skills of those techniques.

From Eqs. (3–4), it follows that LF and LI are unbiased as long as ε_0 is of zero mean. Squared error of LC can be evaluated up to the order t^2 by taking the square of Eq. (3):

$$\|\mathbf{LC}_t - \mathbf{x}_t\|^2 \approx \varepsilon_0^{\mathrm{T}} \left(\mathbf{I} + 2t\mathbf{J}_0 + t^2(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0 + \mathbf{J}_0^{\mathrm{T}}\mathbf{J}_0) \right) \varepsilon_0, \quad (5)$$

and taking the square of Eq. (4) gives the square error of LI up to the order t^3:

$$\|\mathbf{LI}_t - \mathbf{x}_t\|^2 \approx \varepsilon_0^{\mathrm{T}} \left(t^2\mathbf{J}_0^{\mathrm{T}}\mathbf{J}_0 + t^3\mathbf{J}_0^{\mathrm{T}}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0) \right) \varepsilon_0. \quad (6)$$

Eqs. (5–6) involve products of $\mathbf{J}_0$-dependent terms and ε_0-dependent terms. We assume that $\mathbf{J}_0$ and ε_0 are statistically independent. Thus, when we estimate the RMSE of LC and LI, we can calculate separately averages of $\mathbf{J}_0$-dependent terms, and averages of ε_0-dependent terms, and multiply them. The first averages are taken over the attractor's invariant distribution, which is equivalent to taking an average over a very long trajectory. The second averages are taken over the attractor's invariant distribution (because ε_0 depends on $\mathbf{x}_0$) and over possible realizations of the catalog (because ε_0 depends on $\mathbf{LC}_0$). Both could be estimated using the analog database called the "catalog".

Finally, we define the following error-corrected AF techniques:

$$\mathbf{GDC}_t = \mathbf{LI}_t - \left[t < \mathbf{J}_0 > + \frac{t^2}{2} < \mathbf{J}_0^2 > \right] \varepsilon_0 \,, \quad (7)$$

where GDC stands for "global dynamics correction", and

$$\mathbf{LDC}_t = \mathbf{LI}_t - \left[t\mathbf{J}_0 + \frac{t^2}{2}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0) \right] \varepsilon_0 \,, \quad (8)$$

where LDC stands for "local dynamics correction". The means (symbol $<>$) are taken over the invariant distribution of the attractor. GDC only needs average information from the Jacobian matrix, which can be inferred offline. LDC needs local information on the dynamics, which needs to be inferred online. Both global and local information about the Jacobian could be estimated using the catalog.

The quality of the estimation of $\mathbf{J}_0$-dependent quantities from the catalog depends on the dimension of the attractor and the size and quality of the catalog. However, this problematic is not studied further here and in our numerical experiments we use directly analytical expressions of the Jacobian at $\mathbf{x}_0$.

III. EVALUATION

We use the 3-dimensional Lorenz Model [8] in its standard non-dimensional form, with parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$. Time integration is done through a 4th-order Runge-Kutta finite difference numerical scheme with a time step of 0.01 (non-dimensional units). For this model the Jacobian at any point $\mathbf{x}_0$ is:

$$\mathbf{J}_0 = \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho - z_0 & -1 & -x_0 \\ y_0 & x_0 & -\beta \end{pmatrix} , \quad (9)$$

where $\mathbf{x}_0 = (x_0, y_0, z_0)$. We build 50 independent catalogs of 40000 analogs each. For each catalog we draw 400 initial vectors $\mathbf{x}_0$ from the invariant distribution of the attractor. For each of these points we apply AF with our four different methods (LC, LI, GDC and LDC) at 20 different lead times from $t = 0.01$ to $t = 0.2$.

Each analog forecast goes through the following initial steps:

1) select the 40 analogs of $\mathbf{x}_0$ with the lowest Euclidian distance to $\mathbf{x}_0$,
2) if two or more selected analog follow each other in time by one integration time step, make a group of all these analogs and keep only the one with the lowest Euclidian distance to $\mathbf{x}_0$,
3) use Gaussian kernels for the weights $\omega_k \propto \exp\left(-0.5\|\mathbf{x}_0 - \mathbf{a}_0^k\|^2/\lambda^2\right)$ with λ set to the median of $\|\mathbf{x}_0 - \mathbf{a}_0^k\|$ inside the small set of analogs used for the forecast.

For GDC, the averages of $\mathbf{J}_0$-dependent terms are estimated offline, using Eq. (9) and within a random trajectory of 100000 points on the attractor. For LDC, $\mathbf{J}_0$-dependent terms are computed online, using Eq. (9). As mentioned earlier, all those terms could be estimated without knowledge of the model equations, provided that the catalog is large and precise enough.

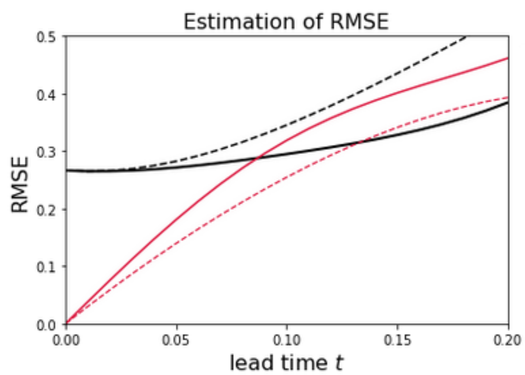

Fig. 1. RMSE associated with locally constant and locally incremental analog forecasting. Black, full: LC, empirical. Black, dashed: LC, predicted using Eq. (5). Red, full: LI, empirical. Red, dashed: LI, predicted using Eq. (6).

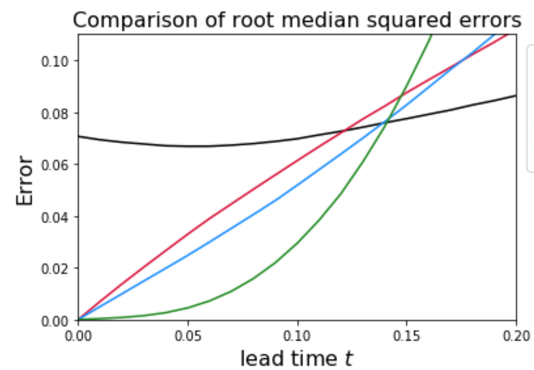

Fig. 2. Empirical root median squared errors associated with different analog forecasting techniques. Black: locally constant (LC). Red: locally incremental (LI). Blue: global dynamics correction (GDC). Green: local dynamics correction (LDC).

Empirical RMSE associated with LC and LI are shown in the full lines (black and red) of figure 1, with estimations from Eqs. (5–6) in dashed lines. The averages of $\mathbf{J}_0$-dependent terms are estimated on a trajectory of 100000 points on the attractor. The averages of ε_0-dependent terms are estimated over the 50*40000 different analog forecasts, which span 50 different catalogs and 40000 different points $\mathbf{x}_0$ inside the attractor. Estimations based on Eqs. (5-6) are perfect for $t \rightarrow 0$, and as t grows several neglected terms influence the validity of our estimations. In particular, we neglected higher-order terms in the t-expansion of Eqs. (5-6) and non-linear terms in ε_0. However, our estimations capture correctly the behavior and order of magnitude of RMSE for all lead times considered here.

Root median square errors associated with our four different forecasting methods are shown in figure 2. The median values of LC (black) and LI (red) square errors are smaller than their mean values, this is due to rare but large values of analog forecasting errors. At small lead times, LDC (green) outperforms GDC (blue), which outperforms LI (red), which outperforms LC (black). However, our different types of error corrections have a negative influence on the performances for large lead times, as the hypotheses that were used to build these error corrections are not fulfilled for large lead times.

As $\mathbf{x}_0$ spans the attractor, the ranges of variations of the x_0, y_0 and z_0 are -19 to 18, -26 to 24, and 3 to 47, respectively. Thus all forecasting methods presented here show good performances with RMSE and root median square errors of only a few percent of the state variables' ranges of variations.

IV. CONCLUSION

AF is an empirical method, but its performances can be interpreted in the framework of dynamical systems, allowing for systematic error correction. We have expressed the leading terms of the errors of common AF techniques for short lead times. We then used those expressions on the Lorenz system as a means of estimating the RMSE and for systematic error correction, yielding positive results for short lead times.

Our work is limited to AF error due to imprecision of the analogs, i.e. non-vanishing ε_0, but we wish to extend this formalism to the case the analogs and the system state follow different dynamics, creating additional error with potential bias. We should then compare the way AF variance is estimated empirically in AF applications (as in [10] and [9]) with our estimations based on the dynamics. Also, we could use more general formulae than Eq. (4) to find bounds for the AF error, possibly independent from local dynamics, to use it as error bound for AF. Finally, this formalism should be extended to the case of partial observations of a high-dimensional dynamical system, which is a more realistic situation. Indeed, in most atmospheric and oceanic applications only a few physical variables (such as temperature, pressure, humidity...) are observed, and observations are limited in time and space due to operational constraints. In such situations, AF is usually combined with Takens' time-lagged embeddings, using analogs on partial observations at current and past times [11].

ACKNOWLEDGMENTS

Funding for the authors was provided by ANR No. 10-IEED-0006-26 and ERC grant No. 338965-A2C2.

REFERENCES

[1] 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. 1974, pp. 3498–3516, 2013.

[2] A. Ayet and P. Tandeo, "Nowcasting solar irradiance using an analog method and geostationary satellite images," *Solar Energy*, vol. 164, no. July 2017, pp. 301–315, 2018.

[3] P. Yiou and C. Déandréis, "Stochastic ensemble climate forecast with an analogue model," *Geoscientific Model Development*, vol. 12, no. 2, pp. 723–734, 2019.

[4] E. N. Lorenz, "Atmospheric Predictability as Revealed by Naturally Occurring Analogues," *Journal of the Atmospheric Sciences*, vol. 26, no. 4, pp. 636–646, 1969.

[5] Z. Zhao and D. Giannakis, "Analog forecasting with dynamics-adapted kernels," *Nonlinearity*, vol. 29, no. 9, pp. 2888–2939, 2016.

[6] C. Nicolis, "Atmospheric Analogs and Recurrence Time Statistics: Toward a Dynamical Formulation," *Journal of the Atmospheric Sciences*, vol. 55, pp. 465–475, 1998.

[7] C. Nicolis, P. Perdigao, and S. Vannitsem, "Dynamics of Prediction Errors under the Combined Effect of Initial Condition and Model Errors," *Journal of the Atmospheric Sciences*, vol. 66, pp. 766–778, 2009.

[8] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.

[9] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, "The analog data assimilation," *Monthly Weather Review*, vol. 145, no. 10, pp. 4093–4107, 2017.

[10] P. Yiou, "AnaWEGE: a weather generator based on analogues of atmospheric circulation," *Geoscientific Model Development*, vol. 7, pp. 531–543, apr 2014.

[11] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980*, pp. 366—-381, Springer, 1981.