

Data-driven reconstruction of partially observed dynamical systems

Pierre Tandeo, Pierre Ailliot, Florian Sévellec

▶ **To cite this version:**

Pierre Tandeo, Pierre Ailliot, Florian Sévellec. Data-driven reconstruction of partially observed dynamical systems. Nonlinear Processes in Geophysics, 2023, 30 (2), pp.129 - 137. 10.5194/npg-30-129-2023 . hal-04131312v2

HAL Id: hal-04131312

https://imt-atlantique.hal.science/hal-04131312v2

Submitted on 16 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nonlin. Processes Geophys., 30, 129–137, 2023
https://doi.org/10.5194/npg-30-129-2023
© Author(s) 2023. This work is distributed under
the Creative Commons Attribution 4.0 License.

Nonlinear Processes in Geophysics
Open Access
EGU

Data-driven reconstruction of partially observed dynamical systems

Pierre Tandeo[1,2,3], **Pierre Ailliot**[4], **and Florian Sévellec**[5,2]

[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238, Brest, France
[2]Odyssey, Inria/IMT/CNRS, Rennes, France
[3]RIKEN Center for Computational Science, Kobe, 650-0047, Japan
[4]Laboratoire de Mathematiques de Bretagne Atlantique, Univ Brest, UMR CNRS 6205, Brest, France
[5]Laboratoire d'Océanographie Physique et Spatiale, Univ Brest CNRS IRD Ifremer, Brest, France

Correspondence: Pierre Tandeo (pierre.tandeo@imt-atlantique.fr)

Received: 22 November 2022 – Discussion started: 29 November 2022
Revised: 14 April 2023 – Accepted: 2 May 2023 – Published: 9 June 2023

Abstract. The state of the atmosphere, or of the ocean, cannot be exhaustively observed. Crucial parts might remain out of reach of proper monitoring. Also, defining the exact set of equations driving the atmosphere and ocean is virtually impossible because of their complexity. The goal of this paper is to obtain predictions of a partially observed dynamical system without knowing the model equations. In this data-driven context, the article focuses on the Lorenz-63 system, where only the second and third components are observed and access to the equations is not allowed. To account for those strong constraints, a combination of machine learning and data assimilation techniques is proposed. The key aspects are the following: the introduction of latent variables, a linear approximation of the dynamics and a database that is updated iteratively, maximizing the likelihood. We find that the latent variables inferred by the procedure are related to the successive derivatives of the observed components of the dynamical system. The method is also able to reconstruct accurately the local dynamics of the partially observed system. Overall, the proposed methodology is simple, is easy to code and gives promising results, even in the case of small numbers of observations.

1 Introduction

In geophysics, even if one has perfect knowledge of the studied dynamical system, it remains difficult to predict because of the existence of nonlinear processes (Lorenz, 1963). Beyond this important difficulty, achieving this perfect knowledge of the system is often impossible. Consequently, the governing differential equations are often not known in full because of their complexity, in particular regarding scale interactions (e.g., turbulent closures are often assumed rather than "known" per se). On top of these two major difficulties, the state of the system is not and cannot be exhaustively observed. Potentially crucial components are and might remain partly or fully out of reach of proper monitoring (e.g., deep ocean or small-scale features). Predicting a partially observed and partially known system is therefore a key issue in current geophysics and in particular for ocean, climate and atmospheric sciences.

A typical example of such a framework is the use of climate indices (e.g., global mean temperature, Niño 3.4 index, North Atlantic Oscillation index) and the study of their links and their dynamics. In this context, the direct relationship between those indices is unknown, even if their more indirect and complex relations exist, through full knowledge of the climate dynamics. Also, it is highly possible that climate indices are dependent on components of the climate that are not currently considered key indices and so are not fully monitored. However, these key indices could be sufficient to describe the most important aspect of climate, leading to accurate and reliable predictions and enabling cost-effective adaptation and mitigation.

Hence, an alternative to physics-based models is to use available observations of the system and statistical approaches to discover equations and then make predictions. This has been introduced in several papers using combinations and polynomials of observed variables as well as sparse regressions or model selection strategies (Brunton et al., 2016; Rudy et al., 2017; Mangiarotti and Huc, 2019). Those methods have then been extended to the case of noisy and irregular observation sampling, using a Bayesian framework as in data assimilation (Bocquet et al., 2019; North et al., 2022). Alternatively, some authors used data assimilation and local linear regressions based on analogs (Tandeo et al., 2015; Lguensat et al., 2017) or iterative data assimilation coupled with neural networks (Brajard et al., 2020; Fablet et al., 2021; Brajard et al., 2021) to make data-driven predictions without discovering equations.

However, many approaches cited above assume that the full state of the system is observed, which is a strong assumption. Indeed, in a lot of applications in geophysics, important components of the system are never or only partially observed, such as the deep ocean (see, e.g., Jayne et al., 2017), and data-driven methods fail to make good predictions. To deal with those strong constraints, i.e., when the model is unknown and when some components of the system are never observed, combination of data assimilation and machine learning shows potential (see, e.g., Wikner et al., 2021). Additionally, an option is to use time-delay embedding of the available components of the system (Takens, 1981; Brunton et al., 2017), whereas another option is to find latent representations of the dynamical system (see, e.g., Talmon et al., 2015; Ouala et al., 2020). In this study, we will show that there are strong relationships between those two approaches.

Here, we propose a simple algorithm using linear and Gaussian assumptions based on a state-space formulation. This classic Bayesian framework, used in data assimilation, is able to deal with a dynamical model (physics- or data-driven) and observations (partial and noisy). Three main ideas are used: (i) augmented state formulation (Kitagawa, 1998), (ii) global linear approximation of the dynamical system (Korda and Mezić, 2018) and (iii) estimation of the dynamical parameters using an iterative algorithm combined with Kalman recursions (Shumway and Stoffer, 1982). The current paper is thus an extension of Shumway and Stoffer (1982) to never-observed components of a dynamical system, using a state-augmentation strategy. The proposed framework is probabilistic, where the state of the system is approximated using a Gaussian distribution (with a mean vector and a covariance matrix). The algorithm is iterative, where a catalog is updated at each iteration and used to learn a linear dynamical model. The final estimate of this catalog corresponds to a new system of variables, including latent ones.

The proposed methodology is based on an important assumption: the surrogate model is linear. Although it can be considered a disadvantage compared to nonlinear models, this linear assumption also has interesting properties. Indeed, nonlinear models combined with state augmentation are a very broad family of models and may lead to identifiability issues. Using linear dynamics already leads to a very flexible family of models since the latent variable may describe nonlinearities and include, for example, any transformation of the observed or non-observed components of a dynamical model. Furthermore, it allows rigorous estimation of the parameters using well-established statistical algorithms which can be run at a low computational cost. The proposed methodology is evaluated on a low-dimensional and weakly nonlinear chaotic model. As this paper is a proof of concept, a linear surrogate model is certainly well suited for this situation.

The paper is organized as follows. Firstly, the methodology is explained in Sect. 2. Secondly, Sect. 3 describes the experiment using the Lorenz-63 system. Thirdly, the results are reported in Sect. 4. The conclusions and perspectives are given in Sect. 5.

2 Methods

The methodology proposed in this paper is borrowed from data assimilation, machine learning and dynamical systems. It is summarized in Fig. 1 and explained below.

In data assimilation, the goal is to estimate, from partial and noisy observations $\boldsymbol{y}$, the full state of a system $\boldsymbol{x}$. When the dynamical model used to propagate $\boldsymbol{x}$ in time is available (i.e., when model equations are given), classic data assimilation techniques are used to retrieve unobserved components of the system. For instance, in the Lorenz-63 system (Lorenz, 1963), if only two variables (x_2 and x_3 in the example defined below) are observed, knowing the Lorenz equations (system of three ordinary differential equations), it is possible to retrieve the unobserved one (x_1 in our example below). However, this estimation requires good estimates of model and observation error statistics (see, e.g., Dreano et al., 2017; Pulido et al., 2018).

Now, if the model equations are not known and observations of the system are available over a sufficient period of time, it is possible to use data-driven methods to mathematically approximate the system dynamics. In this paper, a linear approximation is used to model the relationship of the state vector $\boldsymbol{x}$ between two time steps. It is parameterized with the matrix $\mathbf{M}$, whose dimension is equal to the square of the state space. Moreover, a linear observation operator is introduced to relate the partial observations $\boldsymbol{y}$ and the state $\boldsymbol{x}$. It is written using a matrix $\mathbf{H}$, with its dimension equal to the observation-space times the state-space dimensions. Nonlinear and adaptive operators and noisy observations could be taken into account but, for the sake of simplicity, only the linear and non-noisy case is considered in this paper.

Mathematically, matrices ($\mathbf{M}$, $\mathbf{H}$) and vectors ($\boldsymbol{x}$, $\boldsymbol{y}$) are linked using a Gaussian and linear state-space model such

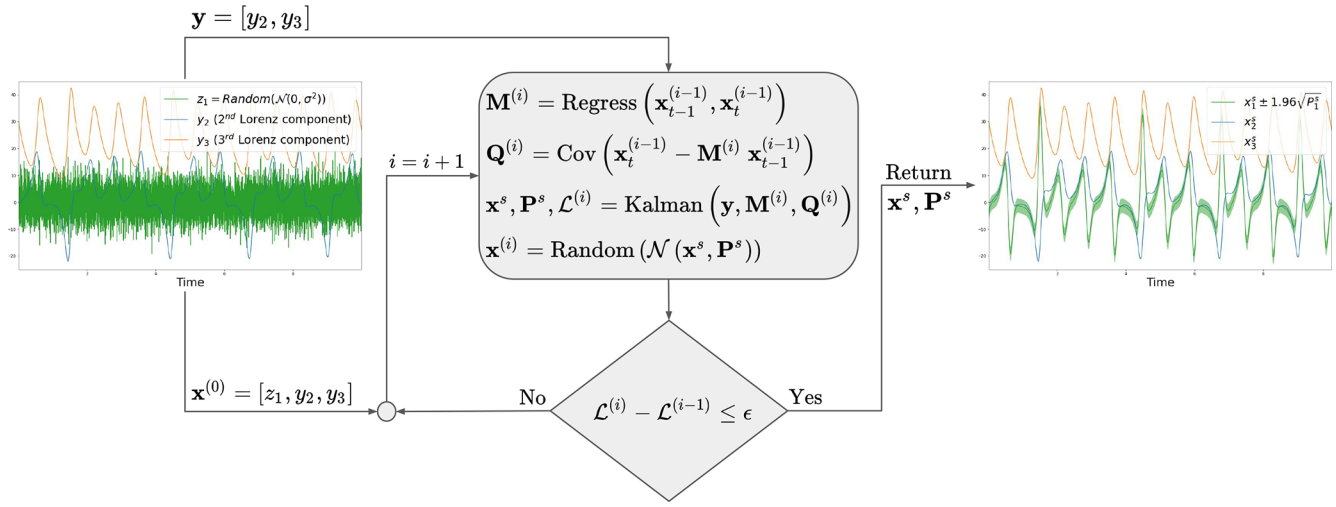


Figure 1. Schematic of the proposed methodology, illustrated using the Lorenz-63 system. The algorithm is initialized with a Gaussian random noise for the hidden component (i.e., z_1) and with partial observations of the system (i.e., y_2 and y_3). Then, an iterative procedure is applied with a linear regression, a covariance computation, the Kalman recursions and a random sampling. This algorithm iteratively maximizes the likelihood of the observations denoted $\mathcal{L}$. After convergence of the algorithm, a hidden component z_1 is stabilized and represented by a Gaussian distribution represented by the mean x_1^{s} and variance P_1^{s}.

that

$$\boldsymbol{x}_t = \mathbf{M}\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_t, \tag{1a}$$

$$\boldsymbol{y}_t = \mathbf{H}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t, \tag{1b}$$

where t is the time index and $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are unbiased Gaussian vectors, representing the model and observation errors, respectively. Their error covariance matrices are denoted $\mathbf{Q}$ and $\mathbf{R}$, respectively. Those matrices indirectly control the respective weight given to the model and to the observations. It constitutes an important tuning part of the state-space models (see Tandeo et al., 2020, for a more in-depth discussion).

In such a data-driven problem where only a part of the system is observed, a first natural step is to consider that the state $\boldsymbol{x}$ is directly related to the observations $\boldsymbol{y}$. For instance, in the example of the Lorenz-63 system introduced previously, observations correspond to the second and third components of the system (i.e., x_2 and x_3, formally defined later).

In this paper, we propose introducing a hidden vector denoted $\boldsymbol{z}$, corresponding to one or more hidden components that are not observed. For this purpose, the state is augmented using this hidden component $\boldsymbol{z}$, the observation vector $\boldsymbol{y}$ does not change, and the operator $\mathbf{H}$ is a truncated identity matrix. The use of augmented state space is classic in data assimilation and mostly refers to the estimation of unknown parameters of the dynamical model (see Ruiz et al., 2013, for further details).

The hidden vector $\boldsymbol{z}$ is now accounted in the linear model $\mathbf{M}$ given in Eq. (1a) whose dimension has increased. The hidden components are completely unknown and thus randomly initialized using Gaussian white noises and are parameterized by σ^2, their level of variance. The next step is to infer $\boldsymbol{z}$ using a statistical estimation method. Starting from the random initialization, an iterative procedure is proposed based on the maximization of the likelihood.

The proposed approach is based on a linear and Gaussian state-space model given in Eq. (1) and thus uses the classic Kalman filter and smoother equations. The Kalman filter (forward in time) is used to get the information of the likelihood, whereas the Kalman smoother (forward and backward in time) is used to get the best estimate of the state. The proposed approach is inspired by the expectation-maximization algorithm (denoted EM; see Shumway and Stoffer, 1982) and is able to iteratively estimate the matrices $\mathbf{M}$ and $\mathbf{Q}$. In this paper, $\mathbf{R}$ is assumed to be known and negligible. The criterion used to update those matrices is based on the innovations defined by the difference between the observations $\boldsymbol{y}$ and the forecast of the model $\mathbf{M}$, denoted $\boldsymbol{x}^{\mathrm{f}}$. The likelihood of the innovations, denoted $\mathcal{L}$, is computed using T time steps such that

$$\mathcal{L} \triangleq p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T | \boldsymbol{x}_1^{\mathrm{f}}, \ldots, \mathbf{x}_T^{\mathrm{f}}\right) \propto \prod_{t=1}^{T} \exp\left(-\left(\boldsymbol{y}_t - \mathbf{H}\mathbf{x}_t^{\mathrm{f}}\right)^{\top} \boldsymbol{\Sigma}_t^{-1} \left(\boldsymbol{y}_t - \mathbf{H}\mathbf{x}_t^{\mathrm{f}}\right)\right), \tag{2}$$

where $\boldsymbol{\Sigma}_t = \mathbf{H}\mathbf{P}_t^{\mathrm{f}}\mathbf{H}^{\top} + \mathbf{R}$, with $\mathbf{P}_t^{\mathrm{f}} = \mathbf{M}\mathbf{P}_{t-1}^{a}\mathbf{M}^{\top} + \mathbf{Q}$ and $\mathbf{P}_{t-1}^{a}$ corresponding to the state covariance estimated by the Kalman filter at time $t-1$. The innovation likelihood given in Eq. (2) is interesting because it corresponds to the squared distance between the observations and the forecast normalized by their uncertainties, represented by the covariance $\boldsymbol{\Sigma}_t$.

At each iteration of the augmented Kalman procedure, the estimate of the matrix $\mathbf{M}$ is given by the least-square estima-

tor, using a linear regression such that

$$\mathbf{M}^{(i)} = \sum_{t=2}^{T} \frac{\left(\boldsymbol{x}_{t-1}^{(i-1)}(\boldsymbol{x}_{t-1}^{(i-1)})^{\top}\right)^{-1} \boldsymbol{x}_{t}^{(i-1)}(\boldsymbol{x}_{t-1}^{(i-1)})^{\top}}{T-1}, \quad (3)$$

where $\boldsymbol{x}^{(i-1)}$ corresponds to the output catalog of the previous iteration (the result of a Kalman smoothing and a Gaussian sampling, explained in more detail below). Following Eq. (1a), the covariance $\mathbf{Q}$ is estimated empirically using the estimate of $\mathbf{M}$ given in Eq. (3), such that

$$\mathbf{Q}^{(i)} = \sum_{t=2}^{T} \frac{\left(\boldsymbol{x}_{t}^{(i-1)} - \mathbf{M}^{(i)}\boldsymbol{x}_{t-1}^{(i-1)}\right)\left(\boldsymbol{x}_{t}^{(i-1)} - \mathbf{M}^{(i)}\boldsymbol{x}_{t-1}^{(i-1)}\right)^{\top}}{T-1}. \quad (4)$$

Then, a Kalman smoother is applied using the $\mathbf{M}^{(i)}$ and $\mathbf{Q}^{(i)}$ matrices estimated in Eqs. (3) and (4). At each time t, it results in a Gaussian mean vector $\boldsymbol{x}_t^{\mathrm{s}}$ and a covariance matrix $\mathbf{P}_t^{\mathrm{s}}$. As input of the next iteration of the algorithm, the catalog $\mathbf{x}^{(i)}$ is updated using a Gaussian random sampling using $\mathbf{x}_t^{\mathrm{s}}$ and $\mathbf{P}_t^{\mathrm{s}}$ at each time t. This random sampling is used to exploit the linear correlations between the components of the state vector that appear in the nondiagonal terms of $\mathbf{P}^{\mathrm{s}}$. The random sampling is also used to avoid being trapped in a local maximum, as in stochastic EM procedures (Delyon et al., 1999).

The likelihood calculated at each iteration of the procedure increases until convergence. The algorithm is stopped when the likelihood difference between two iterations becomes small. The solutions of the proposed method are the last Gaussian mean vectors $\boldsymbol{x}_t^{\mathrm{s}}$ and covariance matrices $\mathbf{P}_t^{\mathrm{s}}$ calculated at each time t. The component corresponding to the latent component $\boldsymbol{z}$ is finally retrieved with information on its uncertainty.

3 Experiment and evaluation metrics

The methodology is tested on the Lorenz-63 system (Lorenz, 1963). This three-dimensional dynamical system models the evolution of the convection (x_1) as a function of horizontal (x_2) and vertical temperature gradients (x_3). The evolution of the system is governed by three ordinary differential equations, i.e.,

$$\dot{x_1} = 10(x_2 - x_1), \quad (5a)$$

$$\dot{x_2} = x_1(28 - x_3) - x_2, \quad (5b)$$

$$\dot{x_3} = x_1 x_2 - \frac{8}{3}x_3. \quad (5c)$$

Runge–Kutta 4-5 is used to integrate the Lorenz-63 equations to generate x_1, x_2 and x_3. In this paper, it is assumed that x_1 is never observed: only x_2 and x_3 are observed on 10 model time units of the Lorenz-63 system every $\mathrm{d}t = 0.001$ time steps (Fig. 2a). The observation vector is thus $\boldsymbol{y} = [y_2, y_3]$. In what follows, only those data are available, not the set of Eq. (5).

The methodology is applied to the Lorenz-63 system, adding sequentially a new hidden component in the state of the system as follows. At the beginning, the state is augmented such that $\boldsymbol{x} = [x_2, x_3, z_1]$, where z_1 is randomly initialized with a white noise, with variance $\sigma^2 = 5$. The observations are stored in the vector $\boldsymbol{y} = [y_2, y_3]$. The observation operator is thus the 2×3 matrix $\mathbf{H} = [1, 0, 0|0, 1, 0]$. After 30 iterations of the algorithm presented in Sect. 2, the hidden component z_1 has converged. After that, a new white noise z_2 is used to augment the state such that $\boldsymbol{x} = [x_2, x_3, z_1, z_2]$, the vector $\boldsymbol{y} = [y_2, y_3]$ remains the same, and the iterative algorithm is applied until stabilization of z_2. As long as the stabilized likelihood continues to increase with the addition of a hidden component, this state-augmentation procedure is repeated.

Note that several hidden components can be added all at once, with a similar performance to the sequential procedure described above (results not shown). In this all-at-once case, the interpretation of the retrieved components is not as informative, and thus we decided to retain the sequential case. Note also that the methodology has been tested with larger $\mathrm{d}t$ (i.e., 0.01 and 0.1). The conclusion is that, by increasing the time delay between observations, it significantly increases the number of latent variables (results not shown). Finally, the assimilation window length corresponds to 10^4 time steps. By reducing this length (e.g., to 10^3, 10^2 or 10^1), the conclusions remain the same as for $\mathrm{d}t = 0.001$.

4 Results

Using the experiment presented in Sect. 3, three hidden components z_1, z_2 and z_3 were sequentially added. They are reported in Fig. 2 with the true Lorenz components x_1, x_2 and x_3. Although they do not fit the hidden variable x_1 of the Lorenz system, the first two hidden components z_1 and z_2 show time variations. By contrast, z_3 remains close to 0, with a large confidence interval. This suggests that our method has identified that two hidden variables are enough to retrieve the dynamics of the two observed variables. This result is consistent with the effective dimension of the Lorenz-63 system, which is between two and three. Here, as the estimated dynamical model $\mathbf{M}$ is a linear approximation, the dimension of the augmented state and the observed components is higher than the effective one.

This is confirmed by the evaluation of the likelihood of the observations y_2 and y_3 with different linear models, obtained with or without the use of hidden components $\boldsymbol{z}$ (Fig. 3). This likelihood is useful for diagnosing the optimal number of dimensions needed to emulate the dynamics of the observed components. As the proposed method is stochastic, 50 independent realizations of the likelihood are shown for each experiment. The 50 realizations vary from the random values

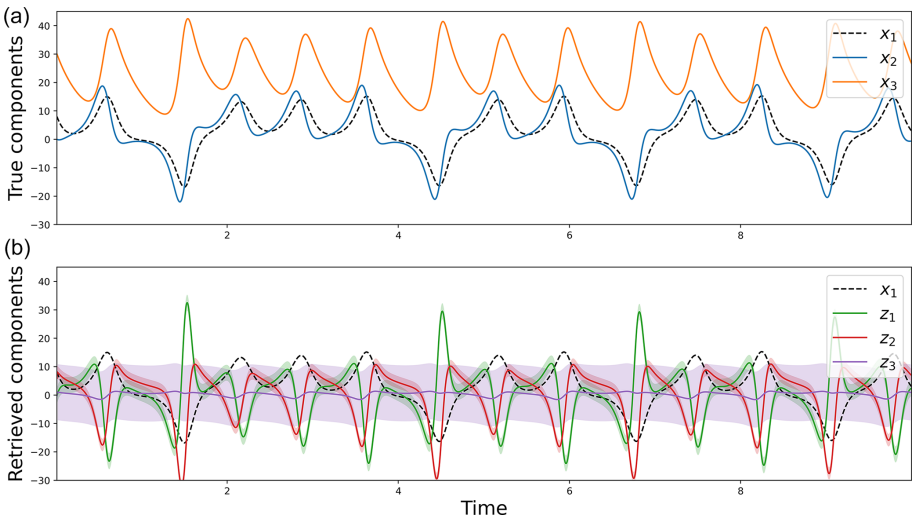


Figure 2. True components of the Lorenz-63 model **(a)** and hidden components estimated using the iterative and augmented Kalman procedure **(b)**. The shaded colors correspond to the 95 % Gaussian confidence intervals.

given to the added hidden variable at the beginning of the iterative procedure. In the naive case where the state of the system is $[x_2, x_3]$ (black dashed line), the likelihood is small. Then, adding successively z_1 (green lines) and z_2 (red lines), after 30 iterations of the proposed algorithm, the likelihood significantly increases. Finally, due to a significant increase in the forecast covariance $\mathbf{P}^{\mathrm{f}}$ in Eq. (2), the inclusion of z_3 reduces the likelihood (purple lines). This suggests that a third variable is not needed and is even detrimental to the skill of the reconstruction. Those results indicate that the best linear model for predicting the variations of the observations y_2 and y_3 is the one using two hidden components. Thus, for the rest of the paper, the focus is placed on the model with the following augmented state: $\boldsymbol{x} = [x_2, x_3, z_1, z_2]$.

The question is now the following: what is the significance of those hidden components z_1 and z_2 estimated using the proposed methodology? Are they correlated with the unobserved component x_1 or with the observed ones x_2 and x_3? Are they somehow proxies of the unobserved component? Using symbolic regression (i.e., using basic mathematical transformations of x_2 and x_3 as regressors to explain z_1 and z_2), it has been found that the hidden components $\boldsymbol{z}$ correspond to linear combinations of the derivatives of the observations such that

$$z_1 = a_2\dot{x_2} + a_3\dot{x_3}, \tag{6a}$$

$$z_2 = b_1\dot{z_1} + b_2\dot{x_2} + b_3\dot{x_3}. \tag{6b}$$

When developing Eq. (6b) using Eq. (6a), the second hidden component is written as $z_2 = b_2\dot{x_2} + b_3\dot{x_3} + b_1a_2\ddot{x_2} + b_1a_3\ddot{x_3}$. It shows that z_1 uses the first derivative of x_2 and x_3, whereas z_2 uses the second derivatives. This result makes the link with the Taylor and Takens theorem, which shows that an unobserved component (i.e., x_1) can be replaced by the observed components (i.e., x_2 and x_3) at different time lags. Note that, due to the stochastic behavior of the algorithm, the a and b coefficients are not fixed, and several combinations of them can reach the same performance in terms of likelihood. This is illustrated in Fig. 3a, with 50 independent realizations of the proposed algorithm. When considering only z_1 (green lines), the algorithm converges to various solutions but is mainly restricted around two solutions (corresponding to a minimum and a maximum of likelihood). As shown in Fig. 3b, the minimum likelihood corresponds to $a_3 = 0$ and the maximum likelihood corresponds to $a_2 = 0$. Thus, the likelihood of $z_1 = a_3\dot{x_3}$ is higher than $z_1 = a_2\dot{x_2}$. This suggests that $\dot{x_3}$ is more important than $\dot{x_2}$ in explaining the variations of the Lorenz system (this is consistent with the investigation of Sévellec and Fedorov, 2014, in a modified version of the Lorenz-63 model). Interestingly, the scatter plot between a_2 and a_3 shows a circular relationship. This is also the case for b_2 and b_3 (results not shown). Then, in Fig. 3a, when considering z_1 and z_2 (red lines), the 50 independent realizations reach the same likelihood after 30 iterations. This means that if $a_3 = 0$ when considering only z_1, then $b_3 \neq 0$ when introducing z_2. In terms of forecast performance, this is similar to $a_2 = 0$ and $b_2 \neq 0$, because the likelihoods converge to the same value (red lines after 30 iterations).

To compare the performance of the naive linear model $\mathbf{M}$ with $[x_2, x_3]$ and the ones with $[x_2, x_3, z_1]$ or $[x_2, x_3, z_1, z_2]$, their forecasts are evaluated. After applying the proposed algorithm, the $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{Q}}$ estimated matrices are used to derive probabilistic forecast, starting from the last available observation $\boldsymbol{y}_t$, using

$$E[\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t] = \widehat{\mathbf{M}}E[\boldsymbol{x}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t], \tag{7a}$$

$$\mathrm{Cov}[\boldsymbol{x}_{t+1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t] = \widehat{\mathbf{M}}\mathrm{Cov}[\boldsymbol{x}_t|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t]\widehat{\mathbf{M}}^T + \widehat{\mathbf{Q}}, \tag{7b}$$

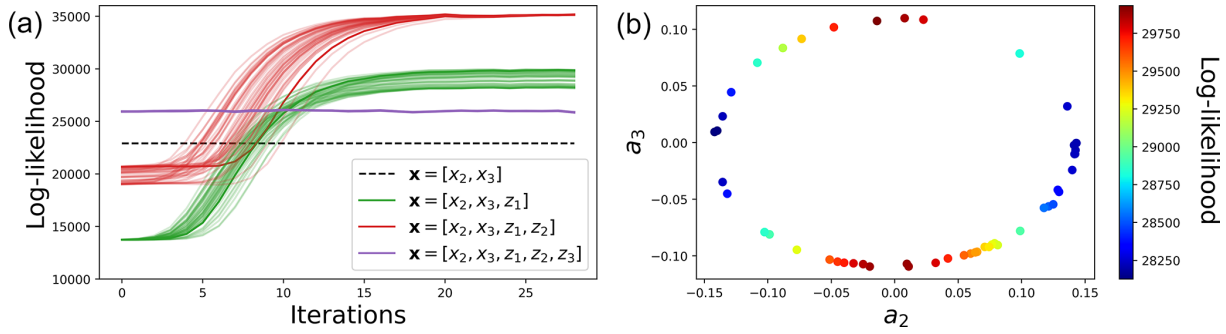


Figure 3. Likelihoods as a function of the iteration of the augmented Kalman procedure **(a)** and estimation of the a_2 and a_3 parameters **(b)**. Different dynamical models are considered, from none to three hidden components in $\boldsymbol{z}$, whereas only x_2 and x_3 are observed in the Lorenz-63 model. The likelihoods of 50 independent realizations of the iterative and augmented Kalman procedure are shown.

with E and Cov the expectation and the covariance, respectively. To test the predictability of the different linear models (i.e., with or without hidden components $\boldsymbol{z}$), a test set has been created, starting from the end of the sequence of observations $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ used in the assimilation window. This test set also corresponds to 10^4 time steps with $\mathrm{d}t = 0.001$. It is used to compute two metrics, the root mean square error (RMSE) and the coverage probability at 50 %. The RMSE is used to evaluate the precision of the forecasts, comparing the true x_2 and x_3 components to the estimated ones, whereas the coverage probability is used to evaluate the reliability of the prediction, evaluating the proportion of true trajectories falling within the 50 % prediction interval of x_2 and x_3. Examples of predictions are given in Fig. 4. It shows bad linear predictions of the model with only $[x_2, x_3]$ (dashed black lines). As the $\mathbf{M}$ operator is not time-dependent, the predictions are quite similar, close to the persistence. Then, adding one (green) or two (red) hidden components in the $\mathbf{M}$ operators creates some nonlinearities in the forecasts.

In Fig. 5, the predictions are evaluated over the whole test dataset for different lead times. By introducing hidden components, the RMSE decreases for both x_2 and x_3 components (panels a and b). For instance, for a lead time of 0.05, when considering two hidden components, the RMSE is halved when it is compared to the naive linear model without hidden components. The coverage probability metric is also largely improved (panels c and d). Indeed, the results with two hidden components are close to 50 %, the optimal value.

To evaluate where the linear model with $[x_2, x_3, z_1, z_2]$ performs better than the one with $[x_2, x_3]$, the Euclidean distances between the forecasts (for a lead time of 0.1) and the truth are computed. Those errors are evaluated at each time step of the test dataset, in the (x_2, x_3) space. Based on those errors, Fig. 6 shows the relative improvement between the model without and the model with hidden components. When the two models have similar performance, values are close to 0 (white), and when the model including z_1 and z_2 is better, values are close to 1 (red). Figure 6 clearly shows that error reduction is not homogeneous in the attractor. The improvement is moderate on the outside of the wings of the attractor but important in the wing transition. This suggests that the introduction of the hidden components z_1 and z_2 makes it possible to provide information on the position in the attractor and thus to make better predictions, especially in bifurcation regions.

5 Conclusions

In this article, the goal is to retrieve hidden components of a dynamical system that is partially observed. The proposed methodology is purely data-driven, not physics-driven (i.e., without the use of any equations of the dynamical model). It is based on the combination of data assimilation and machine learning techniques. Three main ideas are used in the methodology: an augmented state strategy, a linear approximation of a dynamical system and an iterative procedure. The methodology is easy to implement using simple strategies and well-established algorithms: Kalman filter and smoother, linear regression using least squares, an iterative procedure inspired by the EM recursions and Gaussian random sampling for the stochastic aspect.

The methodology is tested on the Lorenz-63 system, where only two components of the system are observed in a short period of time. Several hidden components are introduced sequentially in the system. Although the hidden components are initialized randomly, only a few iterations of the proposed algorithm are necessary to retrieve relevant information. The recovered components are expressed with Gaussian distributions. The new components correspond to linear combinations of successive derivatives of the observed variables. This result is consistent with the theorems of Taylor and Takens, which show that time-delay embedding is useful for improving the forecasts of the system. In our case, this is evaluated using the likelihood, a metric that evaluates the innovation (i.e., the difference between Gaussian forecasts and Gaussian observations).

Using our methodology, we do not retrieve the true missing Lorenz component and need two hidden variables to rep-

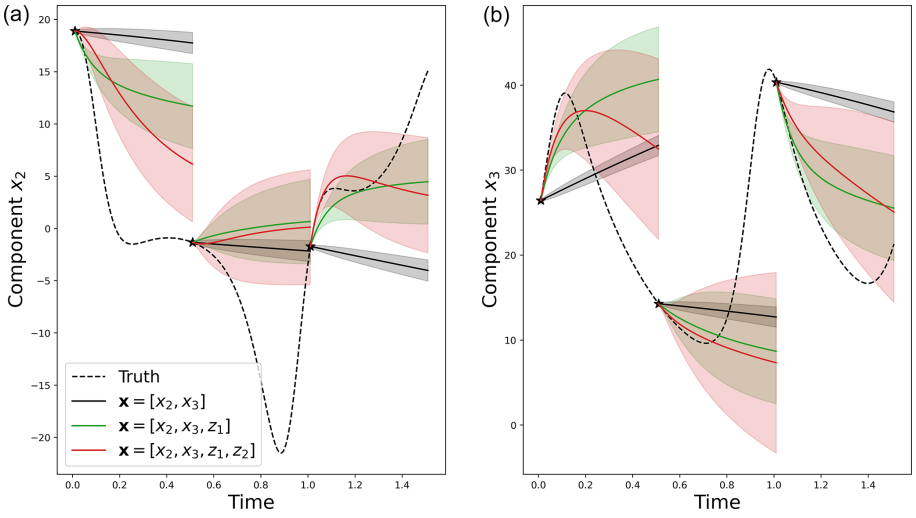


Figure 4. Example of three statistical forecasts of x_2 **(a)** and x_3 **(b)** with their 50 % prediction interval using three different linear operators with no hidden component (dashed black), one hidden component (green) and two hidden components (red). These predictions are obtained using sequential statistical forecasts, as explained in Eq. (7), on an independent test dataset.

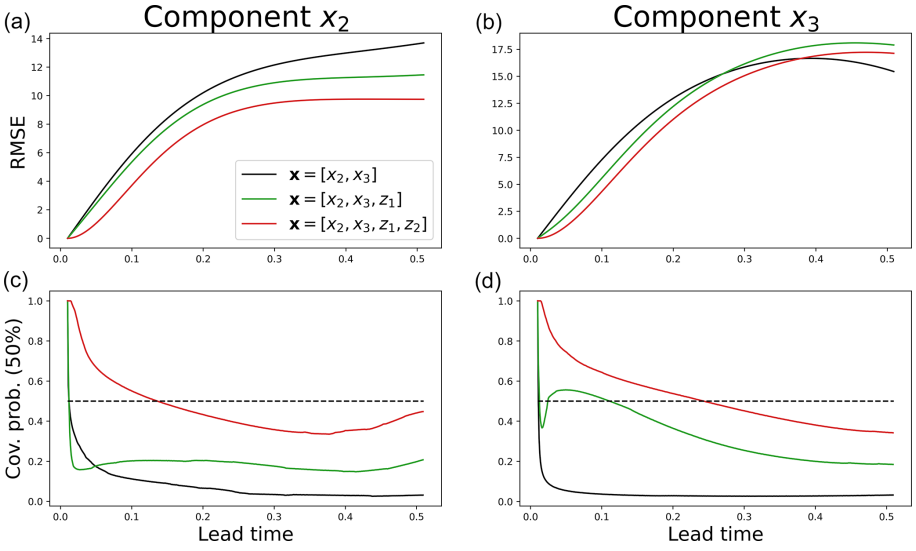


Figure 5. Root mean square error **(a, b)** and 50 % coverage probability **(c, d)** as a function of the lead time (x axis) for the reconstruction of the components x_2 **(a, c)** and x_3 **(b, d)**. These metrics are evaluated on an independent test dataset.

resent a single missing one. The reason for this mismatch is two-fold and is mainly the linear approximation of the dynamical system, which implies that (1) the true missing component, which does not have to be linear combinations of the observed variables, is impossible to retrieve in our framework and (2) two variables, using combinations of the time derivatives of the observed variables, are needed to accurately represent the complexity of the dynamics. However, it is important to note that, even if two variables are needed to replace a single one, the dynamical evolution of the system is relatively well captured, for short lead times, with our methodology. This correct representation of the evolution might ultimately be the most important (e.g., for accurate and reliable forecasting).

The proposed methodology uses a strong assumption: the linear approximation of the dynamical system is global (i.e., fixed for the whole observation period). A perspective is to use adaptive approximations of the model using local linear regressions. This strategy is computationally more expensive because a linear regression is adjusted at each time step but shows some improvements in chaotic systems (see Platzer et al., 2021a, b). In this context of an adaptive linear dynamical model, the proposed methodology could be easily plugged into an ensemble Kalman procedure based on

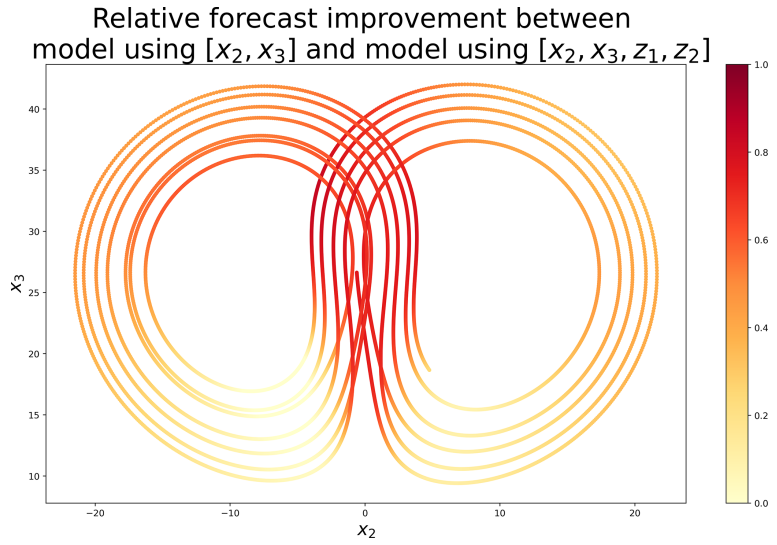


Figure 6. Relative forecast improvement measured as 1 minus the ratio between two Euclidean distances: the one calculated with model $[x_2, x_3, z_1, z_2]$ (at the numerator) and the one calculated with model $[x_2, x_3]$ (at the denominator). The Euclidean distances are calculated in the (x_2, x_3) space and correspond to the error between the forecasts (for a lead time of 0.1) and the truth, evaluated on an independent test dataset.

analog forecasts (Lguensat et al., 2017). In future works, we plan to compare the global and local linear approaches (i.e., a fixed or adaptive linear surrogate model). We also plan to compare them to nonlinear surrogate models, based on neural network architectures with latent information encoded in an augmented space or in hidden layers (e.g., long short-term memory – LSTM).

In this paper, we have demonstrated the feasibility of the method on an idealized and comprehensive problem using the Lorenz-63 system. In the future, we plan to apply the methodology to more challenging problems, like the Lorenz-96 system or a quasi-geostrophic model. For application to real data, we plan to use a database of observed climate indices and try to find latent variables that help to make data-driven predictions.

Code and data availability. The Python code is available at https://github.com/ptandeo/Kalman under the GNU license and the data are generated using the Lorenz-63 system (Lorenz, 1963).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/npg-30-129-2023-supplement.

Author contributions. PT wrote the article. PT and PA developed the algorithm. FS and PA helped with the redaction of the paper.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Nonlinear Processes in Geophysics*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. This paper is the result of a project proposed in a course on "Data Assimilation" in the masters program "Ocean Data Science" at Univ Brest, ENSTA Bretagne, and IMT Atlantique, France. The authors would like to thank the students for their participation in the project: Nils Niebaum, Zackary Vanche, Benoit Presse, Dimitri Vlahopoulos, Yanis Grit and Joséphine Schmutz. The authors would like to thank Noémie Le Carrer for her proofreading of the paper and Paul Platzer, Said Ouala, Lucas Drumetz, Juan Ruiz, Manuel Pulido and Takemasa Miyoshi for their valuable comments.

Financial support. This work was supported by ISblue project, Interdisciplinary graduate school for the blue planet (ANR-17-EURE-0015) and co-funded by a grant from the French government under the program "Investissements d'Avenir" embedded in France 2030. This work was also supported by LEFE program (LEFE IMAGO projects ARVOR).

Review statement. This paper was edited by Natale Alberto Carrassi and reviewed by two anonymous referees.

References

Bocquet, M., Brajard, J., Carrassi, A., and Bertino, L.: Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models, Nonlin. Processes Geophys., 26, 143–162, https://doi.org/10.5194/npg-26-143-2019, 2019.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model, Journal of Computational Science, 44, 101171, https://doi.org/10.1016/j.jocs.2020.101171, 2020.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to infer unresolved scale parametrization, Philos. T. Roy. Soc. A, 379, 2194, https://doi.org/10.1098/rsta.2020.0086, 2021.

Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems, P. Natl. Acad. Sci. USA, 113, 3932–3937, 2016.

Brunton, S. L., Brunton, B. W., Proctor, J. L., Kaiser, E., and Kutz, J. N.: Chaos as an intermittently forced linear system, Nat. Commun., 8, 19, https://doi.org/10.1038/s41467-017-00030-8, 2017.

Delyon, B., Lavielle, M., and Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm, Ann. Stat., 27, 94–128, 1999.

Dreano, D., Tandeo, P., Pulido, M., Ait-El-Fquih, B., Chonavel, T., and Hoteit, I.: Estimating model-error covariances in nonlinear state-space models using Kalman smoothing and the expectation–maximization algorithm, Q. J. Roy. Meteor. Soc., 143, 1877–1885, 2017.

Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., and Rousseau, F.: Learning variational data assimilation models and solvers, J. Adv. Model. Earth Sy., 13, e2021MS002572, https://doi.org/10.1029/2021MS002572, 2021.

Jayne, S. R., Roemmich, D., Zilberman, N., Riser, S. C., Johnson, K. S., Johnson, G. C., and Piotrowicz, S. R.: The Argo program: present and future, Oceanography, 30, 18–28, 2017.

Kitagawa, G.: A self-organizing state-space model, J. Am. Stat. Assoc., 93, 1203–1215, 1998.

Korda, M. and Mezić, I.: Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control, Automatica, 93, 149–160, 2018.

Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., and Fablet, R.: The analog data assimilation, Mon. Weather Rev., 145, 4093–4107, 2017.

Lorenz, E. N.: Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963 (data available at: https://github.com/ptandeo/Kalman, last access: 26 May 2023).

Mangiarotti, S. and Huc, M.: Can the original equations of a dynamical system be retrieved from observational time series?, Chaos: An Interdisciplinary Journal of Nonlinear Science, 29, 023133, https://doi.org/10.1063/1.5081448, 2019.

North, J. S., Wikle, C. K., and Schliep, E. M.: A Bayesian Approach for Data-Driven Dynamic Equation Discovery, J. Agr. Biol. Envir. S., 27, 728–747, https://doi.org/10.1007/s13253-022-00514-1, 2022.

Ouala, S., Nguyen, D., Drumetz, L., Chapron, B., Pascual, A., Collard, F., Gaultier, L., and Fablet, R.: Learning latent dynamics for partially observed chaotic systems, Chaos: An Interdisciplinary Journal of Nonlinear Science, 30, 103121, https://doi.org/10.1063/5.0019309, 2020.

Platzer, P., Yiou, P., Naveau, P., Filipot, J.-F., Thiébaut, M., and Tandeo, P.: Probability distributions for analog-to-target distances, J. Atmos. Sci., 78, 3317–3335, 2021a.

Platzer, P., Yiou, P., Naveau, P., Tandeo, P., Filipot, J.-F., Ailliot, P., and Zhen, Y.: Using local dynamics to explain analog forecasting of chaotic systems, J. Atmos. Sci., 78, 2117–2133, 2021b.

Pulido, M., Tandeo, P., Bocquet, M., Carrassi, A., and Lucini, M.: Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods, Tellus A, 70, 1442099, https://doi.org/10.1080/16000870.2018.1442099, 2018.

Rudy, S. H., Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Data-driven discovery of partial differential equations, Science Advances, 3, e1602614, https://doi.org/10.1126/sciadv.1602614, 2017.

Ruiz, J. J., Pulido, M., and Miyoshi, T.: Estimating model parameters with ensemble-based data assimilation: A review, J. Meteorol. Soc. Jpn., 91, 79–99, 2013.

Sévellec, F. and Fedorov, A. V.: Millennial variability in an idealized ocean model: predicting the AMOC regime shifts, J. Climate, 27, 3551–3564, 2014.

Shumway, R. H. and Stoffer, D. S.: An approach to time series smoothing and forecasting using the EM algorithm, J. Time Ser. Anal., 3, 253–264, 1982.

Takens, F.: Detecting strange attractors in turbulence, in: Dynamical systems and turbulence, Warwick 1980, Springer, 366–381, ISBN 978-3-540-38945-3, https://doi.org/10.1007/BFb0091924, 1981.

Talmon, R., Mallat, S., Zaveri, H., and Coifman, R. R.: Manifold learning for latent variable inference in dynamical systems, IEEE T. Signal Proces., 63, 3843–3856, 2015.

Tandeo, P., Ailliot, P., Ruiz, J., Hannart, A., Chapron, B., Cuzol, A., Monbet, V., Easton, R., and Fablet, R.: Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system, in: Machine learning and data mining approaches to climate science, Springer, 3–12, ISBN 978-3-319-17220-0, https://doi.org/10.1007/978-3-319-17220-0_1, 2015.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., and Zhen, Y.: A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation, Mon. Weather Rev., 148, 3973–3994, 2020.

Wikner, A., Pathak, J., Hunt, B. R., Szunyogh, I., Girvan, M., and Ott, E.: Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components, Chaos: An Interdisciplinary Journal of Nonlinear Science, 31, 053114, https://doi.org/10.1063/5.0048050, 2021.