Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble

Bocquet, M. ORCID: https://orcid.org/0000-0003-2675-0347, Farchi, A. ORCID: https://orcid.org/0000-0002-4162-8289, Finn, T. S. ORCID: https://orcid.org/0000-0001-9585-8349, Durand, C. ORCID: https://orcid.org/0000-0001-5588-549X, Cheng, S. ORCID: https://orcid.org/0000-0002-8707-2589, Chen, Y. ORCID: https://orcid.org/0000-0002-2319-6937, Pasmans, I. ORCID: https://orcid.org/0000-0001-5076-5421 and Carrassi, A. ORCID: https://orcid.org/0000-0003-0722-5600 (2024) Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble. Chaos: An Interdisciplinary Journal of Nonlinear Science, 34 (9). 091104. ISSN 1089-7682 doi: 10.1063/5.0230837 Available at https://centaur.reading.ac.uk/118943/

Chaos
An Interdisciplinary Journal of Nonlinear Science

RESEARCH ARTICLE | SEPTEMBER 30 2024

Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble

Marc Bocquet ; Alban Farchi ; Tobias S. Finn ; Charlotte Durand ; Sibo Cheng ; Yumeng Chen ; Ivo Pasmans ; Alberto Carrassi

Chaos 34, 091104 (2024)
https://doi.org/10.1063/5.0230837

Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble

Cite as: Chaos 34, 091104 (2024); doi: 10.1063/5.0230837
Submitted: 26 July 2024 · Accepted: 8 September 2024 ·
Published Online: 30 September 2024

Marc Bocquet,[1,a] Alban Farchi,[1,b] Tobias S. Finn,[1] Charlotte Durand,[1] Sibo Cheng,[1] Yumeng Chen,[2] Ivo Pasmans,[2] and Alberto Carrassi[3]

AFFILIATIONS

[1]CEREA, École des Ponts and EDF R&D, Île-de-France, France
[2]Department of Meteorology and National Centre for Earth Observation, University of Reading, Earley Gate, PO Box 243, Reading RG6 6BB, United Kingdom
[3]Department of Physics and Astronomy, University of Bologna, Viale Carlo Berti Pichat, 6/2, Bologna 40127, Italy

[a]**Author to whom correspondence should be addressed:** marc.bocquet@enpc.fr. **URL:** https://cerea.enpc.fr/HomePages/bocquet/
[b]**Now at:** The European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany.

ABSTRACT

We investigate the ability to discover data assimilation (DA) schemes meant for chaotic dynamics with deep learning. The focus is on learning the analysis step of sequential DA, from state trajectories and their observations, using a simple residual convolutional neural network, while assuming the dynamics to be known. Experiments are performed with the Lorenz 96 dynamics, which display spatiotemporal chaos and for which solid benchmarks for DA performance exist. The accuracy of the states obtained from the learned analysis approaches that of the best possibly tuned ensemble Kalman filter and is far better than that of variational DA alternatives. Critically, this can be achieved while propagating even just a single state in the forecast step. We investigate the reason for achieving ensemble filtering accuracy without an ensemble. We diagnose that the analysis scheme actually identifies key dynamical perturbations, mildly aligned with the unstable subspace, from the forecast state alone, without any ensemble-based covariances representation. This reveals that the analysis scheme has learned some multiplicative ergodic theorem associated to the DA process seen as a non-autonomous random dynamical system.

Data assimilation (DA) estimates the state of dynamical systems from sparse and noisy observations and is used worldwide in numerical weather prediction centers. Accurate DA demands the representation of the time-dependent errors in this state estimate, usually achieved through the propagation of an ensemble of states. Using deep learning, we discover the update step of DA applied to chaotic dynamics. We show that a simple convolutional neural network (CNN) can learn DA, reaching an accuracy as good as that of ensemble-based DA. Crucially, the CNN can achieve this best accuracy with single state forecasts. This is explained by the CNN's ability to identify local space patterns from this one state, which are used to assess the errors in the analysis. This suggests building a new class of efficient deep learning-based ensemble-free DA algorithms.

I. INTRODUCTION

A. Context and problem

In a simplified but quintessential framework, the goal of data assimilation (DA) and, in particular, filtering algorithms is to accurately estimate states $\mathbf{x}_k^{\mathrm{t}} \in \mathbb{R}^{N_{\mathrm{x}}}$, where "t" stands for truth, at equally spaced times τ_k for $k = 0, \ldots, K$ along a trajectory of a dynamical system. Hence, they are related by

$$\mathbf{x}_{k+1}^{\mathrm{t}} = \mathscr{M}\left(\mathbf{x}_k^{t}\right), \tag{1a}$$

where $\mathscr{M}$ is the resolvent over $\tau_{k+1} - \tau_k$ of known autonomous, i.e., time-independent, dynamics. Such goal is achieved from the knowledge of the dynamics $\mathscr{M}$ and of observation vectors $\mathbf{y}_k \in \mathbb{R}^{N_{\mathrm{y},k}}$ obtained from the non-accessible states $\mathbf{x}_k^{\mathrm{t}}$ via observation operators

$\mathscr{H}_k$ and perturbed by a white-in-time Gaussian noise $\boldsymbol{\varepsilon}_k$ of mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_k$,

$$\mathbf{y}_k = \mathscr{H}_k(\mathbf{x}^{\mathrm{t}}_k) + \boldsymbol{\varepsilon}_k, \quad \boldsymbol{\varepsilon}_k \sim N(\mathbf{0}, \mathbf{R}_k). \tag{1b}$$

Applied to chaotic hence dynamically unstable dynamics, sequential (in time) algorithms must be used.[1,2] They alternate an *analysis* step that, from the newly acquired observation vector $\mathbf{y}_k$ in Eq. (1b) and the current estimate of the state $\mathbf{x}^{\mathrm{f}}_k$, provides an updated optimal estimate of the state $\mathbf{x}^{\mathrm{a}}_k$ called the analysis. The subsequent state estimate $\mathbf{x}^{\mathrm{f}}_{k+1}$ stems from the *forecast* step that relies on Eq. (1a). The estimates in both steps can either be deterministic or probabilistic, often leveraging an ensemble in the latter case. Such sequential DA is widely used in numerical weather prediction (NWP), and in many areas of climate sciences,[2] as a suite of both research and operational tools.

Classical DA methods are classified into (i) variational methods, such as 3D-Var and 4D-Var, (ii) ensemble-based statistical methods, such as the ensemble Kalman filter (EnKF), and (iii) ensemble variational methods that inherit the assets of the two previous categories.[1] On the one hand, variational methods account for the nonlinearity of models (dynamical model and observation operators), leveraging nonlinear optimization techniques. Ensemble-based methods, on the other hand, can capture the *errors of the day*, i.e., time-dependent error statistics, via an ensemble meant to diagnose sample error statistics. Those are key properties that drive the performance of these DA methods in mildly nonlinear chaotic models. For low-order, chaotic dynamics such as the celebrated Lorenz 96 (L96) model,[3] the EnKF significantly outperforms 3D-Var or a moderately long window 4D-Var in terms of accuracy owing to its dynamical representation of the errors. This has been emphasized and illustrated in twin experiments.[4] In fact, current implementations of 4D-Var in NWP centers incorporate a forecast ensemble so as to capture the errors-of-the-day.[5,6] However, in high-dimensional models, these ensemble-based error statistics must necessarily be regularized using techniques known as localization and possibly inflation.[7] With a focus on the time-dependent error statistics of sequential DA, it has been conjectured[8,9] then proven[10–12] that for linear dynamics and when localization is unnecessary, the forecast and analysis error covariance matrices of the EnKF are confined to the unstable-neutral subspace, denoted as $\mathscr{U}$ from now on, of the dynamics. This subspace is spanned by the covariant Lyapunov vectors associated to non-negative Lyapunov exponents.[13] It is precisely when the ensemble size is smaller than the dimension of this subspace that localization is required to avoid divergence of the EnKF. Deviating from linear dynamics turns those exact results into approximations, for which these findings were nonetheless numerically confirmed.[14–16]

This paper focuses on methodological DA and on what can be discovered from deep learning (DL) techniques to improve state-of-the-art DA schemes such as those mentioned above. Hence, we hereby give a brief account on the recent introduction of DL techniques for DA applied to chaotic dynamics.[17]

It was first proposed to learn DA analysis through DL from the data produced by existing DA schemes.[18,19] One can alternatively replace the solver of a 4D-Var over a long DA window by a DL operator that would learn the outcome of the 4D-Var cost function minimization.[20–24] However, the latter approaches do not consider cycling sequential DA, the focus of the present paper. A systematic, formal Bayesian view on the use of DL in the critical components of sequential DA has been proposed[25] and called *data assimilation network* (DAN). In the present paper, a simplified variant of this DAN concept is used. As far as ensemble and Kalman-related DA methods are concerned, it has been proposed to learn their Kalman gain[26,27] or parameters thereof possibly relying on an auto-differentiable implementation of the (En)KF.[28–30] As a step further, it was also proposed to learn the full analysis operator using (self-)supervision.[25,31] Finally, bypassing the need for dynamical models and DA schemes altogether, DL-based *end-to-end* methods aim at estimating states of the system from the observations only,[32,33] yet so far with a focus on the feasibility of such endeavor.

B. Objectives

In this paper, the forecast model in Eq. (1a) is assumed to be known so as to avoid intricate interactions when learning the DA operators and the dynamics simultaneously.

Our objective is to learn the analysis operator of a sequential DA scheme meant for chaotic dynamics from a long trajectory of the dynamical system and the associated set of noisy, possibly sparse observations. Hence, it stands out from past studies that exploited DL to learn the dynamics, possibly using DA.[34–39] The resulting DL-based analysis operator will be referred to as $a_{\boldsymbol{\theta}}$, while the full resulting DA scheme will be called DAN.

We will first explain how to learn such analysis operator from DL. It will then be shown numerically that $a_{\boldsymbol{\theta}}$ can surprisingly perform as accurately as a well-optimized EnKF, even *without using an ensemble*. This strongly contrasts with common beliefs in methodological DA. To interpret this result, we will show using innovative concepts based on a Taylor expansion of the learned $a_{\boldsymbol{\theta}}$, that $a_{\boldsymbol{\theta}}$ directly discovers and utilizes a fine knowledge of the dynamics, as opposed to agnostic classical DA. The nature of these dynamical structures learned through $a_{\boldsymbol{\theta}}$ will then be discussed and interpreted.

II. EXPERIMENTAL SETUP

With the goal to learn an analysis operator $a_{\boldsymbol{\theta}}$ as a key step of a filtering DA scheme for chaotic dynamics, we build a twin experiment within the framework offered by Eq. (1).

A. Analysis operator and its neural network representation

Let us define a filtering DA scheme, based on an analysis and forecast ensemble. The ith members of the analysis and forecast ensembles at time τ_k are noted $\mathbf{x}^{\mathrm{a},i}_k$ and $\mathbf{x}^{\mathrm{f},i}_k$, respectively. By denoting $\mathscr{S}_{\mathrm{e}} = 1, \ldots, N_{\mathrm{e}}$, the corresponding analysis and forecast ensembles are $\mathbf{E}^{\mathrm{a}}_k = \{\mathbf{x}^{\mathrm{a},i}_k\}_{i\in\mathscr{S}_{\mathrm{e}}} \subset \mathscr{E}^{\mathrm{e}}$ and $\mathbf{E}^{\mathrm{f}}_k = \{\mathbf{x}^{\mathrm{f},i}_k\}_{i\in\mathscr{S}_{\mathrm{e}}} \subset \mathscr{E}^{\mathrm{e}}$, respectively, where $\mathscr{E}^{\mathrm{e}} = \mathbb{R}^{N_{\mathrm{e}}\times N_{\mathrm{x}}}$. The initial ensemble $\mathbf{E}^{\mathrm{f}}_0$ is obtained from perturbing a random state on the attractor of the dynamics.

The analysis step of the DA scheme is given by the (incremental) analysis operator $a_{\boldsymbol{\theta}}$, which depends on a set of neural network

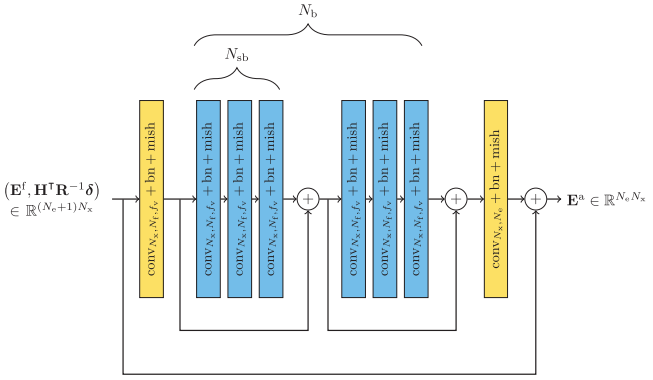

FIG. 1. Architecture of the residual convolutional network, where $N_b = 2$ and $N_{sb} = 3$. $\mathrm{conv}_{N_1,N_2,f}$ is a generic one-dimensional convolutional layer of dimension N_1, with N_2 filters of kernel size f. See text for more details.

weights and biases, a vector $\boldsymbol{\theta}$,

$$\mathbf{E}^a_k = \mathbf{E}^f_k + a_{\boldsymbol{\theta}}\left(\mathbf{E}^f_k, \mathbf{H}^\intercal_k \mathbf{R}^{-1}_k \boldsymbol{\delta}_k\right), \tag{2a}$$

where $\boldsymbol{\delta}_k$, the innovation at time τ_k, is defined by

$$\boldsymbol{\delta}_k \triangleq \mathbf{y}_k - \mathscr{H}_k\left(\bar{\mathbf{x}}^f_k\right), \quad \bar{\mathbf{x}}^f_k \triangleq \frac{1}{N_e}\sum_{i\in\mathscr{S}_e} \mathbf{x}^{f,i}_k. \tag{2b}$$

$\mathbf{H}_k$ is the tangent linear operator of $\mathscr{H}_k$ but any arbitrary injective operator from $\mathbb{R}^{N_{y,k}}$ to $\mathbb{R}^{N_x}$ could be chosen instead. The DA forecast step propagates the analysis ensemble, member-wise:

$$\mathbf{E}^f_{k+1} = \mathscr{M}\left(\mathbf{E}^a_k\right). \tag{3}$$

We choose $a_{\boldsymbol{\theta}}$ to have a simple residual convolutional neural network (CNN) architecture. A schematic of the CNN architecture is displayed in Fig. 1. It begins with an initial convolution that takes $N_e + 1$ channels as inputs and, with N_f filters, outputs N_f channels. This initial layer is followed by N_b residual blocks. Each one of these blocks is a succession of N_{sb} sub-blocks. Each subblock is made of (i) a convolutional layer with N_f channels as inputs, which has N_f filters and a kernel size f_v for each of its filter, (ii) a batch normalization layer, and (iii) an activation function chosen to be mish.[40] The CNN ends with a final convolutional layer that takes N_f channels as inputs and, with N_e filters, outputs N_e channels. The kernel size of the initial and final channels is f_v. Hence, the internal state of the CNN consists of N_f copies of the latent space that we simply choose to be isomorphic to the state space $\mathbb{R}^{N_x}$. Furthermore, the encoder and decoder from state space to latent space and back are chosen to coincide with the identity. We have also tested a depth-separable architecture for this residual CNN, with a number of parameters roughly divided by 3, yielding an accuracy almost as good but longer training times. Note that the fundamental results reported in this paper are agnostic to the details of the architecture: this CNN is a mere functional tool to learn an optimal $a_{\boldsymbol{\theta}}$.

B. Training scheme

Toward efficiently learning an optimal $a_{\boldsymbol{\theta}}$, we consider N_r such DA runs, based on as many independent concurrent trajectories of the dynamics and as many sequences of observation vectors. Hence, the DA runs are specified by $\mathbf{E}^{a,r}_k = \left\{\mathbf{x}^{a,i,r}_k\right\}_{i\in\mathscr{S}_e}$ and $\mathbf{E}^{f,r}_k$ for $r = 1, \ldots, N_r$, and being iterates through N_c cycles, they depend on $\boldsymbol{\theta}$ except for the set of initial conditions $\mathbf{E}^{f,r}_0$. In order to learn an optimal $a_{\boldsymbol{\theta}}$, a loss function is defined,

$$\mathscr{L}(\boldsymbol{\theta}) = \sum_{r=1}^{N_r}\sum_{k=1}^{N_c} \left\|\mathbf{x}^{t,r}_k - \bar{\mathbf{x}}^{a,r}_k(\boldsymbol{\theta})\right\|^2, \quad \bar{\mathbf{x}}^{a,r}_k \triangleq \frac{1}{N_e}\sum_{i\in\mathscr{S}_e} \mathbf{x}^{a,i,r}_k, \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm. Its formulation is based on supervised learning, although self-supervised learning[31,34,35] could have been used instead; it is nonetheless more challenging and rather unrelated to the goals of this paper. This loss matches the analysis ensemble mean trajectory with the true trajectory. The Adam stochastic gradient descent optimization technique[41] is used to minimize it. To avoid the risks of exploding gradients, the inefficiency of vanishing gradients, and huge memory requirements, when computing gradients of Eq. (4), the truncated backpropagation through time technique[42,43] is used; it splits the trajectories in the dataset into chunks of N_{iter} cycles.

It must be pointed out that a successful sequential DA process, when seen as a dynamical system, is stable.[44,45] Hence, after a rough starting phase in the training, the learned $a_{\boldsymbol{\theta}}$ should yield a numerically stable prediction-assimilation dynamical system. In particular, this is likely to avoid exploding gradients. By contrast, the task of learning a DL emulator of the dynamics over many consecutive time steps often fails because of the unstable nature of the chaotic dynamics.

The N_r trajectories are dispatched into a training and validation dataset with a 90%–10% ratio. Overfitting is prevented by an early stopping of the minimization based on the validation score, tantamount to regularization.[46] Moreover, the testing dataset stems from an independently generated trajectory, long enough to yield converged statistics. In the test stage, the DAN scheme is used within a twin experiment using the trajectories and resulting observations of the testing dataset. Its performance is assessed from a single scalar score using the time-averaged root mean square error (aRMSE) of the analysis against the truth,

$$\mathrm{aRMSE} = \frac{1}{K\sqrt{N_x}}\sum_{k=1}^{K} \left\|\mathbf{x}^t_k - \bar{\mathbf{x}}^a_k\right\|, \tag{5}$$

which, in a cycled DA context, is a reliable indicator of the overall performance of the scheme, whatever its purpose.

III. NUMERICAL RESULTS

The $a_{\boldsymbol{\theta}}$ operator is trained on the L96 model,[3] and the results will be interpreted and discussed in the context of this model. L96 is a one-dimensional model defined over a periodic band of latitude of the Earth atmosphere. Its ordinary differential equations read

$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = (x_{n+1} - x_{n-2})x_{n-1} - x_n + F, \tag{6}$$

with $x_{N_x} = x_0$, $x_{-1} = x_{N_x-1}$, $x_{-2} = x_{N_x-2}$, $F = 8$, and $N_x = 40$ in the basic configuration. The model has a Lyapunov time of 0.60. It has 13 positive exponents, and being continuous-in-time and

autonomous, it has one zero Lyapunov exponent. Hence, the dimension of its unstable-neutral subspace $\mathscr{U}$ is $N_u = 14$.

A. Hyperparameters sensitivity analysis

We first carry out a large set of trainings to assess the sensitivity of a_θ's performance to its hyperparameters. We choose $N_{iter} = 16$, without any significant gain beyond this value, while the numerical cost increases due to a deeper backpropagation. We first assume the model to be fully observed with $\mathscr{H}_k = \mathbf{I}_x$, the identity matrix in $\mathbb{R}^{N_x}$, and the observations to be affected by a white-in-time unbiased Gaussian noise of covariance matrix $\mathbf{R}_k = \mathbf{I}_x$ for all time steps. This configuration is the most widely used to benchmark new DA schemes with L96. The ensemble size N_e and the number of filter N_f were selected in a set ranging between 1 and 40. The number N_b of residual blocks in the CNN and number of subblocks N_{sb} in each residual block were both chosen in the set $[\![1, 6]\!]$. Because L96 has short-range correlations in space, we choose a kernel size of $f_v = 5$, even though the CNN receptive field is much larger.

B. First results and robustness

The training dataset size per epoch scales linearly with N_r, which is chosen to be 2^{18} and further discussed in the supplementary material. The subsequent test DA runs with the trained a_θ are actually all stable in time, yielding an aRMSE significantly below 1, as expected if DA has any skill over the mere observations. Unsurprisingly, we found that the larger the hyperparameters N_f, N_b, and N_{sb}, the smaller the test aRMSEs of the resulting DANs, but that $N_f = 40$, $N_b = 5$, and $N_{sb} = 5$ offer a good compromise for accuracy vs training cost and CNN size. This will be the reference configuration, which has about 2×10^5 trainable parameters.

One obvious essential drawback of learning a_θ is its *non-universality*. Specifically, a_θ depends on the observation setup used in the training dataset. This is a critical research path for end-to-end DA. Although not the aim of the present paper, we nonetheless checked the performance of the trained a_θ, with $\mathscr{H}_k = \mathbf{I}_x$ and $\mathbf{R}_k = \sigma_y^2 \mathbf{I}_x$ with $\sigma_y = 1$, in test DA runs with similar observations but generated with σ_y taking value in between 0.1 and 3. Yet, in all test runs, DAN remains robust with slightly degraded aRMSEs for $\sigma_y < 1$ but aRMSEs at least as good for $\sigma_y > 1$, compared to a well-tuned EnKF. Well-tuned EnKF always refers here to an EnKF with an ensemble large enough such that localization is unnecessary and relying on the EnKF-N[47,48] to optimally counteract residual sampling errors such that inflation is unnecessary.

Testing non-trivial $\mathscr{H}_k$, we also learned a single a_θ from observation networks whose density N_y/N_x is randomly and uniformly chosen in the interval $[0, 1]$ at each τ_k and $\sigma_y = 1$. This DAN was then tested on several DA runs, each one with a constant in time observation density N_y/N_x taking value in the interval $[0.2, 1]$. In this configuration, a_θ performs almost as well or better than well-tuned EnKFs for $0.35 < N_y/N_x < 0.65$ and is suboptimal (compared to the EnKF) but still stable outside of this range. These results already pleasantly suggest that these DL-based DA schemes may remain valid well beyond the specifications of observation operators from which a_θ was learned. Plots of these experiments and further discussion are provided in the supplementary material.

C. One state forecast

Using the reference configuration but with an ensemble size N_e taking values in the set $[\![1, 40]\!]$, test aRMSEs fluctuate in between 0.19 and 0.20. By contrast, a sizable ensemble is, as we recalled in Sec. I, one of the key reason for the success of the EnKF. For comparison, we checked that 3D-Var yields an aRMSE of 0.40, that the best linear filter (i.e., a trained a_θ without activation function) yields an aRMSE of 0.384, and that well-tuned EnKFs with $N_e = 20$ and $N_e = 40$ yield aRMSEs of 0.191 and 0.179, respectively. Note that the reference a_θ but with $N_f = 100$ yields an aRMSE of 0.185, closer to the best EnKF with $N_e = N_x = 40$, showing that further improvements are possible even though not the focus of this paper. These key aRMSE scores are arranged in a table in the supplementary material.

However, the pivotal remark is that a single state forecast, $N_e = 1$ in a_θ, is as efficient as using a large ensemble. Furthermore, the need for localization and inflation is completely obviated. We have checked that this is obtained concurrently to a feature collapse in a_θ,[49] i.e., all channels' last layer feature maps converge to the same state. It is likely that a better local minimum of the loss could be obtained with complex encoder and decoder[50] and infusing diversity in the CNN through Monte Carlo dropouts,[31] so as to obtain an a_θ leveraging the ensemble. Nonetheless, the local minimum reached in our trainings yield an accuracy with $N_e = 1$ worthy of a well-tuned EnKF. That is why we shall concentrate in the following on interpreting this astonishing result for which we shall use, especially in Sec. IV, dynamical systems theory.

Therefore, the analysis operator is hereafter learned in the reference configuration but with $N_e = 1$.

IV. INTERPRETATION

In this section, we focus on the remarkable finding that a learned DA method with a single state $N_e = 1$ forecast achieves performance on par with a well-tuned EnKF. We wish to understand the reason for this performance by investigating what a_θ learns. To that end, an innovative expansion of a_θ in terms of more familiar DA operators is carried out.

A. Operator expansion of a_θ

Toward this goal, we look for a classical Kalman update[51,52] that would be a good match to a_θ seen as a mathematical map, at least for small analysis increments. The first diagnostic is the mean anomaly generated by a_θ, i.e., how much $a_\theta(\mathbf{x}, \mathbf{0})$ deviates from $\mathbf{0}$ on average. It should be small since a vanishing innovation $\boldsymbol{\delta}_k$ should not yield any state update. Hence, we define the time-dependent normalized scalar anomalies

$$b_k = \frac{1}{\sqrt{N_x}} \left\| a_\theta(\mathbf{x}_k, 0) \right\|, \tag{7}$$

along with the associated mean bias b and the standard deviation s of b_k in time.

Next, expanding with respect to the innovation, the following functional form for a_θ is assumed:

$$a_\theta\left(\mathbf{x}, \mathbf{H}^\intercal \mathbf{R}^{-1} \boldsymbol{\delta}\right) \approx \mathbf{K}(\mathbf{x}) \cdot \boldsymbol{\delta}, \tag{8}$$

owing to the fact that no state update is needed when the innovation vanishes and only keeping the leading order term in $\boldsymbol{\delta}$. This is an

Ansatz of $a_{\boldsymbol{\theta}}$ where $\mathbf{K}(\mathbf{x}) \in \mathbb{R}^{N_\mathrm{x} \times N_\mathrm{y}}$ is meant to stand as a Kalman gain surrogate. By contrast, with the propagation of a single state, classical sequential DA methods would typically resemble 3D-Var, and the gain would not depend on the forecast state (the first input variable of $a_{\boldsymbol{\theta}}$).

Interestingly, we also learned a simplified $\hat{a}_{\boldsymbol{\theta}}$ replacing Eq. (2a) with $\mathbf{E}^\mathrm{a}_k = \mathbf{E}^\mathrm{f}_k + \hat{a}_{\boldsymbol{\theta}}\left(\mathbf{H}_k^\intercal \mathbf{R}_k^{-1} \boldsymbol{\delta}_k\right)$, whereby losing $a_{\boldsymbol{\theta}}$'s ability to extract information from $\mathbf{E}^\mathrm{f}_k$, similarly to 3D-Var. This yields an aRMSE of 0.382 in test DA runs, unsurprisingly close to 0.40 of our 3D-Var. Hence, learning an optimal constant-in-time $\mathbf{K}$ of an (En)KF,[30] a configuration subsumed by this specific $\hat{a}_{\boldsymbol{\theta}}$, is significantly suboptimal in this context.

B. Identifying the operators in this expansion

Once $a_{\boldsymbol{\theta}}$ has been obtained from training and considering a fixed forecast state $\mathbf{x}$ at a given time step, a large set of innovations $\{\boldsymbol{\delta}_j\}_{j=1,\dots,N_\mathrm{p}}$ are sampled from the observation error statistics: $\boldsymbol{\delta}_j \sim N(\mathbf{0}, \mathbf{R})$. This yields a set of corresponding incremental updates $\left\{\mathbf{a}_j = a_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{H}^\intercal \mathbf{R}^{-1} \boldsymbol{\delta}_j)\right\}_{j=1,\dots,N_\mathrm{p}}$. Since Eq. (8) is only an approximation, $\mathbf{K}(\mathbf{x})$ is estimated with the least squares problem

$$\mathscr{L}_\mathbf{x}(\mathbf{K}) = \sum_{j=1}^{N_\mathrm{p}} \left\| \mathbf{a}_j - \bar{\mathbf{a}} - \mathbf{K}(\mathbf{x}) \cdot \left(\boldsymbol{\delta}_j - \bar{\boldsymbol{\delta}}\right) \right\|^2, \tag{9}$$

where $\bar{\mathbf{a}} = N_\mathrm{p}^{-1} \sum_{j=1}^{N_\mathrm{p}} \mathbf{a}_j$ and $\bar{\boldsymbol{\delta}} = N_\mathrm{p}^{-1} \sum_{j=1}^{N_\mathrm{p}} \boldsymbol{\delta}_j$.

Next, assuming $\mathbf{R}$ is known, we would like to estimate the analysis error covariance matrix $\mathbf{P}^\mathrm{a}$ associated to $a_{\boldsymbol{\theta}}$ in the Kalman gain expansion. It depends on $\mathbf{x}_k$ and, hence, on τ_k. Within the *best linear unbiased estimator* framework, $\mathbf{K}$ is related to $\mathbf{P}^\mathrm{a}$ through $\mathbf{K} = \mathbf{P}^\mathrm{a} \mathbf{H}^\intercal \mathbf{R}^{-1}$ so that from Eq. (8),

$$a_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{H}^\intercal \mathbf{R}^{-1} \boldsymbol{\delta}) \approx \mathbf{P}^\mathrm{a} \mathbf{H}^\intercal \mathbf{R}^{-1} \boldsymbol{\delta}, \tag{10}$$

which suggests that an expansion in the second variable $\boldsymbol{\zeta} \in \mathbb{R}^{N_\mathrm{x}}$ of $a_{\boldsymbol{\theta}}$ yields

$$a_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\zeta}) \approx \mathbf{P}^\mathrm{a}(\mathbf{x}) \cdot \boldsymbol{\zeta}. \tag{11}$$

Hence, $\mathbf{P}^\mathrm{a}$ can be estimated using Eq. (11) either from a least squares loss similar to Eq. (11) or from the Jacobian of $a_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\zeta}$ leveraging auto-differentiable DL libraries.

C. What is learned?—Supporting numerical results

At each τ_k, i.e., over many $\mathbf{x}_k$ on the forecast model's attractor, it is possible to estimate $\mathbf{K}(\mathbf{x}_k)$ and $\mathbf{P}^\mathrm{a}(\mathbf{x}_k)$ from the expansion of $a_{\boldsymbol{\theta}}$. For the sake of simplicity, $\mathscr{H}_k = \mathbf{I}_\mathrm{x}$ and $\mathbf{R}_k = \mathbf{I}_\mathrm{x}$, in which case $\mathbf{P}^\mathrm{a}(\mathbf{x}_k) = \mathbf{K}(\mathbf{x}_k)$.

The analysis mean bias b and its standard deviation s are first computed over a long L96 $a_{\boldsymbol{\theta}}$-based DA run. We obtain $b \simeq 5 \times 10^{-3}$ and $s \simeq 10^{-3}$, which are indeed very small compared to the typical aRMSE of an either DAN or EnKF run, i.e., 0.20, meaning that the bias of $a_{\boldsymbol{\theta}}$ relative to typical updates is roughly 2.5%.

The surrogate $\mathbf{P}^\mathrm{a}$, denoted $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ and estimated from Eq. (11), is compared to that of a concurrent well-tuned EnKF with $N_\mathrm{e} = 40$, whose analysis error covariance matrix is $\mathbf{P}^\mathrm{a}_\mathrm{EnKF}$. $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ is compared to $\mathbf{P}^\mathrm{a}_\mathrm{EnKF}$ using a normalized Bures–Wasserstein distance,[53]

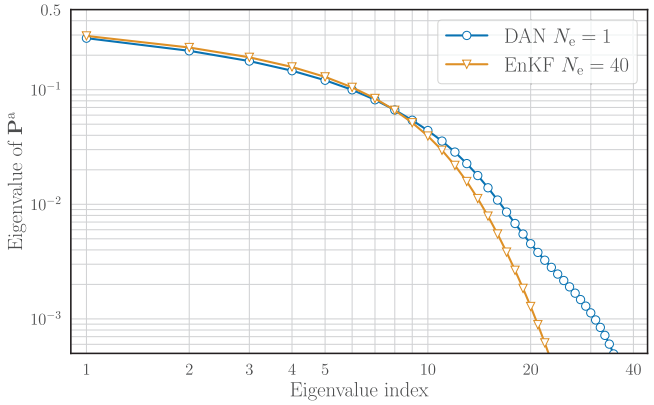

FIG. 2. Time-averaged eigenspectra of $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ and $\mathbf{P}^\mathrm{a}_\mathrm{EnKF}$.

$$d_\mathrm{BW}(\mathbf{U}, \mathbf{V}) = \frac{1}{N_\mathrm{x}} \left[\mathrm{Tr} \left\{ \mathbf{U} + \mathbf{V} - 2\left(\mathbf{V}^{\frac{1}{2}} \mathbf{U} \mathbf{V}^{\frac{1}{2}}\right)^{\frac{1}{2}} \right\} \right]^{\frac{1}{2}}, \tag{12}$$

where $\mathbf{U}$ and $\mathbf{V}$ are two semi-definite symmetric matrices. This metric is expected to smoothly account for the unmatched principal axes of $\mathbf{U}$ and $\mathbf{V}$, but also their associated variances (eigenspectra). The time-averaged d_BW distance between $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ and $\mathbf{P}^\mathrm{a}_\mathrm{EnKF}$ is 0.013, whereas it is 0.048 between $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ and $(0.40)^2 \mathbf{I}_\mathrm{x}$, which approximates $\mathbf{P}^\mathrm{a}$ of a well-tuned 3D-Var. The time-averaged eigenspectra of $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ and $\mathbf{P}^\mathrm{a}_\mathrm{EnKF}$ are plotted in Fig. 2. They are remarkably close to each other for the first ten modes. Beyond these modes, the $a_{\boldsymbol{\theta}}$ operator is likely to selectively apply some (multiplicative) inflation, as one would expect from such stable DA runs.

We further compute the principal angles[14] of the vector subspaces generated by the $N_\mathrm{u} = 14$ dominant eigenvalues of $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ and $\mathbf{P}^\mathrm{a}_\mathrm{EnKF}$. They are reported in Fig. 3. Recall that $N_\mathrm{u} = 14$ is the dimension of the L96 $\mathscr{U}$. The principal angles are intrinsic to the relative position of these subspaces; they do not depend on any coordinate system used to parameterize them. This indicates how close the most unstable directions of $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ and $\mathbf{P}^\mathrm{a}_\mathrm{EnKF}$ are in state space. From Fig. 3, we observe that the simplex formed by the EnKF is on average the most aligned with $\mathscr{U}$.[14] The subspace spanned by the dominant axes of $\mathbf{P}^\mathrm{a}_\mathrm{DAN}$ is also well aligned with $\mathscr{U}$, yet progressively diverges when incorporating less unstable directions. For comparison, the principal angles of $\mathscr{U}$ with an isotropically randomly sampled $N_\mathrm{u} = 14$-dimensional subspace are also plotted in Fig. 3.

D. Main interpretation

These numerical results indicate that $a_{\boldsymbol{\theta}}$ defined through Eq. (2a) depends on the innovation but also on the single forecast state when $N_\mathrm{e} = 1$. This does not hold for the EnKF incremental update that only indirectly depends on the forecast state via the ensemble-based forecast error covariances. Hence, without the need for an ensemble, $a_{\boldsymbol{\theta}}$ extracts from the forecast state critical pieces of

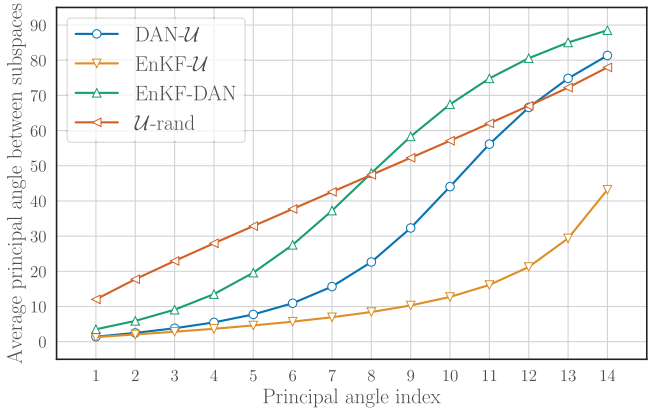

FIG. 3. Time-averaged principal angles (in degrees) formed by the subspaces spanned by the $N_u = 14$ dominant directions of $\mathbf{P}^a_{DAN}$ vs $\mathscr{U}$, $\mathbf{P}^a_{EnKF}$ vs $\mathscr{U}$, $\mathbf{P}^a_{EnKF}$ vs $\mathbf{P}^a_{DAN}$, and $\mathscr{U}$ vs a randomly sampled $N_u = 14$-dimensional subspace.

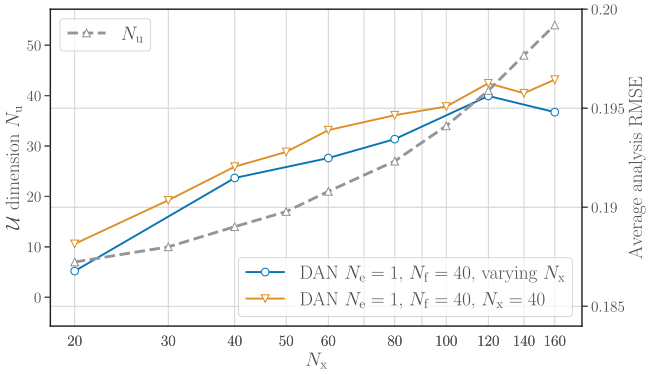

FIG. 4. Test aRMSEs (blue and yellow full lines) of a_θ operators learned from either L96 models with varying N_x or the $N_x = 40$ L96 model but applied to varying N_x L96 models. The dimension N_u of $\mathscr{U}$ (gray dashed line) is much steeper compared to the slowly increasing aRMSE curves.

information on the unstable directions, as shown by the principal angles experiment.

Furthermore, a_θ manages to accurately assess the uncertainty attached to these unstable directions as demonstrated by the spectra of $\mathbf{P}^a_{DAN}$. Overall, $\mathbf{P}^a_{DAN}$ with $N_e = 1$ is on average very close to $\mathbf{P}^a_{EnKF}$ with $N_e = 40$, for the dominant axes, and it applies some inflation onto the less unstable modes as seen by comparing their spectra.[47] We conclude that a_θ directly learns about the dynamics features, as opposed to the regression-based, purely statistical, update in the EnKF.

Essentially, for a_θ, critical pieces of information of the forecast error covariances of the DA run are encoded, and thus exploitable, in the forecast state alone. From the multiplicative ergodic theorem,[54] we know that, in autonomous ergodic dynamical systems such as $\mathscr{M}$, there exists a mapping between each of the system's states and the corresponding Lyapunov covariant vectors. Furthermore, if the DA run (the forecast and analysis cycle) is considered as an ergodic dynamical system of its own,[44] the same theorem guarantees the existence of a mapping between the forecast state and the analysis error covariance matrix that a_θ guesses. The DA process is not autonomous because it indirectly depends on the truth trajectory, the observation noise, and the observation operators; but a generalized variant of the multiplicative ergodic theorem for non-autonomous random dynamics should be applicable.[55–58] Hence, we conjecture that a_θ must learn such mapping, together with how to process this information and combine it with the innovation.

E. Locality and scalability

Next, we have trained a_θ on the L96 model using the reference configuration with $N_e = 1$ but with a changing state space dimension N_x in between 20 and 160. The aRMSEs of well-tuned EnKFs for the changing N_x and picking $N_e = N_x$ have been computed for comparison. The test DAN aRMSEs show no significant dependence on N_x and are all within 5% of the EnKFs. Hence, because the performance of a_θ with an unchanged architecture and the same number of parameters is barely affected by increasing N_x, we conjecture that the learned analysis extracts *local* pieces of information from the forecast state.

If true, the a_θ operator learned for DA on an $N_x = 40$ L96 model could be applied directly to an L96 DA run with a different N_x. Recall that the L96 states exhibit local highs and lows of Rossby-like waves, whose number scales linearly with N_x. Thus, as long as the spatial extent of those waves is captured by the receptive field of the CNN, the same layers of a_θ with the same weights and biases might be able to handle L96 states of distinct dimensionality.

To test this hypothesis, we use the same a_θ operator (same weights and biases) learned as before with $N_e = 1$ and $N_x = 40$ but apply it now to L96 models with N_x ranging from 20 to 160. The corresponding aRMSEs are reported in Fig. 4, which shows that these aRMSEs are roughly the same for all N_x (between 0.188 and 0.197). This demonstrates that this *transdimensional transfer* works surprisingly well.

This strongly supports the fact that a_θ extracts local information from the forecast state (of various dimensions in this experiment), relying on its convolution layers. It is, therefore, able to capture where, in phase-space, the error mass is concentrated. We hypothesize that these localized error structures are related to the *localization* of the dominant covariant Lyapunov vectors.[59–61] A proper mathematical definition of such spatial localization can be found in these references.

V. CONCLUSIONS

Using the L96 chaotic model, we have demonstrated that a learned DL-based analysis a_θ, key part of a sequential DA (often referred to as a filtering scheme), can be almost as accurate as the best possibly tuned EnKF, the benchmark for ensemble filtering methods in this model. More importantly, this learned DA scheme does not require any ensemble and can equally well rely on a single state forecast. Therefore, a_θ appears to be able to retrieve local patterns, representative of unstable and uncertain modes, from the

forecast state alone. We believe that this is fundamentally made possible by some multiplicative ergodic theorem applied to sequential DA seen as a non-autonomous random dynamical system driven by time-dependent true dynamics and observation operators and white-in-time observation errors.

To make sure our conclusions were not entirely bound to the L96 model, we carried out a large number of similar experiments on the well-known chaotic Kuramoto–Sivashinski model.[62,63] They all confirm and support these conclusions.

What is achieved by a_θ resonates with the *parametric EnKF*,[64,65] which encodes the errors of the day in a couple of dynamical ancillary fields, preventing the use of an ensemble. Amazingly, our learned a_θ is even more radical and extracts that information from the state itself.

Taking a step back, we learned from DL that an accurate and efficient DA analysis operator could capture the dynamical error without an ensemble, leveraging model-specific information. This promotes a rethinking of the popular sequential DA schemes for chaotic dynamics.

SUPPLEMENTARY MATERIAL

See the supplementary material for further details on the evaluation of the a_θ-based DA method, as reported in Sec. III. Specifically, a table of the key aRMSE scores mentioned in Sec. III is provided, as well as plots of the aRMSE curves related to the sensitivity experiments mentioned in Sec. III.

ACKNOWLEDGMENTS

The authors are grateful to the Editor Jürgen Kurths and an anonymous Reviewer for their comments and suggestions on the original version of the manuscript. The authors acknowledge the support of the project SASIP (Grant No. G-24-66154) funded by Schmidt Sciences—a philanthropic initiative that seeks to improve societal outcomes through the development of emerging science and technologies. CEREA is a member of Institut Pierre-Simon Laplace (IPSL).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Marc Bocquet: Conceptualization (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (lead). **Alban Farchi:** Conceptualization (supporting); Investigation (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Tobias S. Finn:** Investigation (supporting); Software (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Charlotte Durand:** Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Sibo Cheng:** Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Yumeng Chen:** Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Ivo Pasmans:** Investigation (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting). **Alberto Carrassi:** Methodology (supporting); Validation (supporting); Writing – original draft (supporting); Writing – review & editing (supporting).

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1] M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*, Fundamentals of Algorithms (SIAM, Philadelphia, 2016), p. 324.
[2] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, "Data assimilation in the geosciences: An overview on methods, issues, and perspectives," WIREs Clim. Change **9**, e535 (2018).
[3] E. N. Lorenz and K. A. Emanuel, "Optimal sites for supplementary weather observations: Simulation with a small model," J. Atmos. Sci. **55**, 399–414 (1998).
[4] M. Bocquet and P. Sakov, "Joint state and parameter estimation with an iterative ensemble Kalman smoother," Nonlinear Process. Geophys. **20**, 803–818 (2013).
[5] L. Raynaud, L. Berre, and G. Desroziers, "Accounting for model error in the Météo-France ensemble data assimilation system," Q. J. R. Meteorol. Soc. **138**, 249–262 (2012).
[6] M. Bonavita, L. Isaksen, and E. Hólm, "On the use of EDA background error variances in the ECMWF 4D-Var," Q. J. R. Meteorol. Soc. **138**, 1540–1559 (2012).
[7] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*, 2nd ed. (Springer-Verlag, Berlin, 2009), p. 307.
[8] A. Carrassi, S. Vannitsem, D. Zupanski, and M. Zupanski, "The maximum likelihood ensemble filter performances in chaotic systems," Tellus A **61**, 587–600 (2008).
[9] L. Palatella, A. Carrassi, and A. Trevisan, "Lyapunov vectors and assimilation in the unstable subspace: Theory and applications," J. Phys. A: Math. Theor. **46**, 254020 (2013).
[10] K. S. Gurumoorthy, C. Grudzien, A. Apte, A. Carrassi, and C. K. R. T. Jones, "Rank deficiency of Kalman error covariance matrices in linear time-varying system with deterministic evolution," SIAM J. Control Optim. **55**, 741–759 (2017).
[11] M. Bocquet, K. S. Gurumoorthy, A. Apte, A. Carrassi, C. Grudzien, and C. K. R. T. Jones, "Degenerate Kalman filter error covariances and their convergence onto the unstable subspace," SIAM/ASA J. Uncertainty Quantif. **5**, 304–333 (2017).
[12] D. Crisan and M. Ghil, "Asymptotic behavior of the forecast–assimilation process with unstable dynamics," Chaos **33**, 023139 (2023).
[13] B. Legras and R. Vautard, "A guide to lyapunov vectors," in *ECMWF Workshop on Predictability* (ECMWF, Reading, 1996), pp. 135–146.
[14] M. Bocquet and A. Carrassi, "Four-dimensional ensemble variational data assimilation and the unstable subspace," Tellus A **69**, 1304504 (2017).
[15] Y. Chen, A. Carrassi, and V. Lucarini, "Inferring the instability of a dynamical system from the skill of data assimilation exercises," Nonlinear Process. Geophys. **28**, 633–649 (2021).
[16] A. Carrassi, M. Bocquet, J. Demaeyer, C. Gruzien, P. N. Raanes, and S. Vannitsem, "Data assimilation for chaotic dynamics," in *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, edited by S. K. Park and L. Xu (Springer International Publishing, Cham, 2022), pp. 1–42.
[17] S. Cheng, C. Quilodran-Casas, S. Ouala, A. Farchi, C. Liu, P. Tandeo, R. Fablet, D. Lucor, B. Iooss, J. Brajard, D. Xiao, T. Janjic, W. Ding, Y. Guo, A. Carrassi,

M. Bocquet, and R. Arcucci, "Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review," IEEE/CAA J. Autom. Sin. **10**, 1361–1387 (2023).
[18]T. P. Härter and H. F. de Campos Velho, "Data assimilation procedure by recurrent neural network," Eng. Appl. Comput. Fluid Mech. **6**, 224–233 (2012).
[19]R. S. Cintra and H. F. de Campos Velho, "Data assimilation by artificial neural networks for an atmospheric general circulation model," in *Advanced Applications for Artificial Neural Networks*, edited by A. ElShahat (IntechOpen, 2018), Chap. 17, pp. 265–286.
[20]R. Fablet, B. Chapron, L. Drumetz, E. Mémin, O. Pannekoucke, and F. Rousseau, "Learning variational data assimilation models and solvers," J. Adv. Model. Earth Syst. **13**, e2021MS002572 (2021).
[21]T. Frerix, D. Kochkov, J. Smith, D. Cremers, M. Brenner, and S. Hoyer, "Variational data assimilation with a learned inverse observation operator," in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021), pp. 3449–3458.
[22]N. Lafon, R. Fablet, and P. Naveau, "Uncertainty quantification when learning dynamical models and solvers with variational methods," J. Adv. Model. Earth Syst. **15**, e2022MS003446 (2023).
[23]A. Filoche, J. Brajard, A. Charantonis, and D. Béréziat, "Learning 4DVAR inversion directly from observations," in *Computational Science—ICCS 2023*, edited by J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. Sloot (Springer Nature Switzerland, Cham, 2023), pp. 414–421.
[24]J. D. Keller and R. Potthast, "AI-based data assimilation: Learning the functional of analysis estimation," arXiv:2406.00390 [physics.ao-ph] (2024).
[25]P. Boudier, A. Fillion, S. Gratton, S. Gürol, and S. Zhang, "Data assimilation networks," J. Adv. Model. Earth Syst. **15**, e2022MS003353 (2023).
[26]H. Hoang, P. De Mey, and O. Talagrand, "A simple adaptive algorithm of stochastic approximation type for system parameter and state estimation," in *Proceedings of 1994 33rd IEEE Conference on Decision and Control* (IEEE, 1994), Vol. 1, pp. 747–752.
[27]S. Hoang, R. Baraille, O. Talagrand, X. Carton, and P. De Mey, "Adaptive filtering: Application to satellite data assimilation in oceanography," Dyn. Atmos. Ocean **27**, 257–281 (1998).
[28]T. Haarnoja, A. Ajay, S. Levine, and P. Abbeel, "Backprop KF: Learning discriminative deterministic state estimators," in *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), Vol. 29.
[29]Y. Chen, D. Sanz-Alonso, and R. Willett, "Autodifferentiable ensemble kalman filters," SIAM J. Math. Data Sci. **4**, 801–833 (2022).
[30]E. Luk, E. Bach, R. Baptista, and A. Stuart, "Learning optimal filters using variational inference," arXiv:2406.18066 [cs.LG] (2024).
[31]M. McCabe and J. Brown, "Learning to assimilate in chaotic dynamical systems," in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 12237–12250.
[32]A. McNally, C. Lessig, P. Lean, E. Boucher, M. Alexe, E. Pinnington, M. Chantry, S. Lang, C. Burrows, M. Chrust, F. Pinault, E. Villeneuve, N. Bormann, and S. Healy, "Data driven weather forecasts trained and initialised directly from observations," arXiv:2407.15586[physics.ao-ph] (2024).
[33]A. Vaughan, S. Markou, W. Tebbutt, J. Requeima, W. P. Bruinsma, T. R. Andersson, M. Herzog, N. D. Lane, M. Chantry, J. S. Hosking, and R. E. Turner, "Aardvark weather: End-to-end data-driven weather forecasting," arXiv:2404.00411[physics.ao-ph] (2024).
[34]M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino, "Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models," Nonlinear Process. Geophys. **26**, 143–162 (2019).
[35]J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino, "Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model," J. Comput. Sci. **44**, 101171 (2020).
[36]M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino, "Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization," Found. Data Sci. **2**, 55–80 (2020).
[37]J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino, "Combining data assimilation and machine learning to infer unresolved scale parametrisation," Philos. Trans. R. Soc. A **379**, 20200086 (2021).
[38]Q. Liu, Y. Xu, J. Kurths, and X. Liu, "Complex nonlinear dynamics and vibration suppression of conceptual airfoil models: A state-of-the-art overview," Chaos **32**, 062101 (2022).
[39]X. Wang, J. Feng, Y. Xu, and J. Kurths, "Deep learning-based state prediction of the lorenz system with control parameters," Chaos **34**, 033108 (2024).
[40]D. Misra, "Mish: A self regularized non-monotonic neural activation function," arXiv:1908.08681 (2019).
[41]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, edited by Y. Bengio and Y. LeCun (San Diego, CA, USA, 2015).
[42]H. Tang and J. Glass, "On training recurrent networks with truncated backpropagation through time in speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)* (IEEE, 2018), pp. 48–55.
[43]C. Aicher, N. J. Foti, and E. B. Fox, "Adaptively truncating backpropagation through time to control gradient bias," in *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, Proceedings of Machine Learning Research Vol. 115, edited by R. P. Adams and V. Gogate (PMLR, 2020), pp. 799–808.
[44]A. Carrassi, M. Ghil, A. Trevisan, and F. Uboldi, "Data assimilation as a nonlinear dynamical systems problem: Stability and convergence of the prediction-assimilation system," Chaos **18**, 023112 (2008).
[45]A. Carrassi, A. Trevisan, L. Descamps, O. Talagrand, and F. Uboldi, "Controlling instabilities along a 3DVar analysis cycle by assimilating in the unstable subspace: A comparison with the EnKF," Nonlinear Process. Geophys. **15**, 503–521 (2008).
[46]I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, Cambridge, Massachusetts, 2016), p. 775.
[47]M. Bocquet, P. N. Raanes, and A. Hannart, "Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation," Nonlinear Process. Geophys. **22**, 645–662 (2015).
[48]P. N. Raanes, M. Bocquet, and A. Carrassi, "Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures," Q. J. R. Meteorol. Soc. **145**, 53–75 (2019).
[49]J. van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," arXiv:2102.11409[cs.LG] (2022).
[50]M. Peyron, A. Fillion, S. Gürol, V. Marchais, S. Gratton, P. Boudier, and G. Goret, "Latent space data assimilation by using deep learning," Q. J. R. Meteorol. Soc. **147**, 3759–3777 (2021).
[51]R. E. Kalman, "A new approach to linear filtering and prediction problems," J. Basic Eng. **82**, 35–45 (1960).
[52]M. Ghil and P. Malanotte-Rizzoli, "Data assimilation in meteorology and oceanography," Adv. Geophys. **33**, 141–266 (1991).
[53]R. Bhatia, T. Jain, and Y. Lim, "On the Bures-Wasserstein distance between positive definite matrices," Expo. Math. **37**, 165–191 (2019).
[54]V. I. Oseledec, "A multiplicative ergodic theorem. ljapunov characteristic numbers for dynamical systems," Trans. Moscow Math. Soc. **19**, 197–231 (1968).
[55]L. Arnold, *Random Dynamical Systems* (Springer, Berlin, Heidelberg, 1998), p. 586.
[56]M. D. Chekroun, E. Simonnet, and M. Ghil, "Stochastic climate dynamics: Random attractors and time-dependent invariant measures," Physica D **240**, 1685–1700 (2011).
[57]F. Flandoli and E. Tonello, "An introduction to random dynamical systems for climate" (2021).
[58]M. Ghil and D. Sciamarella, "Review article: Dynamical systems, algebraic topology and the climate sciences," Nonlinear Process. Geophys. **30**, 399–434 (2023).
[59]D. Pazó, M. A. Rodríguez, and J. M. López, "Spatio-temporal evolution of perturbations in ensembles initialized by bred, Lyapunov and singular vectors," Tellus A **62**, 10–23 (2010).
[60]S. Vannitsem and V. Lucarini, "Statistical and dynamical properties of covariant Lyapunov vectors in a coupled atmosphere-ocean model-multiscale effects,

geometric degeneracy, and error dynamics," J. Phys. A: Math. Theor. **49**, 224001 (2016).
[61]B. Giggins and G. A. Gottwald, "Stochastically perturbed bred vectors in multi-scale systems," Q. J. R. Meteorol. Soc. **145**, 642–658 (2019).
[62]Y. Kuramoto and T. Tsuzuki, "Persistent propagation of concentration waves in dissipative media far from thermal equilibrium," Progr. Theor. Phys. **55**, 356–369 (1976).
[63]G. I. Sivashinsky, "Nonlinear analysis of hydrodynamic instability in laminar flames-I. Derivation of basic equations," Acta Astronaut. **4**, 1177–1206 (1977).
[64]O. Pannekoucke, S. Ricci, S. Barthelemy, R. Ménard, and O. Thual, "Parametric Kalman filter for chemical transport model," Tellus A **68**, 31457 (2016).
[65]O. Pannekoucke, M. Bocquet, and R. Ménard, "Parametric covariance dynamics for the nonlinear diffusive Burgers equation," Nonlinear Process. Geophys. **25**, 481–495 (2018).