Uncertainty Quantification When Learning Dynamical Models and Solvers With Variational Methods

Nicolas Lafon, Ronan Fablet, Philippe Naveau

Uncertainty quantification when learning dynamical models and solvers with variational methods

N. Lafon[1], R. Fablet[2]and P. Naveau[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement, EstimR, IPSL-CNRS, CEA Saclay, Gif-sur-Yvette, France
[2]IMT Atlantique, UMR CNRS Lab-STICC, Brest, France

Key Points:

- We propose an approach to jointly solve data assimilation and uncertainty quantification problems using a variational Bayes formulation
- Our end-to-end neural architecture implements a learnable gradient-based solver for a ELBO criterion (Evidence Lower Bound)
- Studies conducted on Lorenz 63 dynamics and on river discharge from the Danube river network highlight the potential of our approach

Corresponding author: Nicolas Lafon, `nicolas.lafon@lsce.ipsl.fr`

Abstract

In geosciences, data assimilation (DA) addresses the reconstruction of a hidden dynamical process given some observation data. DA is at the core of operational systems such as weather forecasting, operational oceanography and climate studies. Beyond the reconstruction of the mean or most likely state, precise inference of the state posterior distribution remains a key challenge, i.e. quantify uncertainties as well as to inform intrinsical stochastic variabilities. Indeed, DA schemes, such as variational DA and Kalman methods, can have difficulty in dealing with complex non-linear processes. A growing literature investigates the cross-fertilization of DA and machine learning. This study proposes an end-to-end Neural Network (NN) scheme based on a Variational Bayes (VB) inference formulation. It combines an ELBO (Evidence Lower BOund) variational cost to a trainable gradient-based solver to infer the state posterior pdf given observation data. The inference of the posterior and the trainable solver are learnt jointly. We demonstrate the relevance of the proposed scheme for a Gaussian parameterization of the posterior and different case-study experiments. It includes a benchmark w.r.t. state-of-the-art schemes. We discuss further the generalization of the proposed approach to non-Gaussian posterior.

Plain Language Summary

The spatio-temporal reconstruction of a dynamical process from some observational data is at the core of a wide range of applications in geosciences. This is particularly true for weather forecasting, operational oceanography and climate studies. However, the reconstruction of a given dynamic and the prediction of future states must take into account the uncertainties that affect the system. Thus, the available observational measurements are only provided with a limited accuracy. Besides, the encoded physical equations that model the evolution of the system do not capture the full complexity of the real system. Finally, the numerical approximation generates a non-negligible error. For these reasons, it seems relevant to calculate a probability distribution of the state system rather than the most probable state. Using recent advances in machine learning techniques for inverse problems, we propose an algorithm that jointly learns a parametric distribution of the state, the dynamics governing the evolution of the parameters, and a solver. Experiments conducted on synthetic reference datasets, as well as on datasets describing environmental systems, validate our approach.

1 Introduction

The reconstruction and forecasting of dynamical systems from available observations are key challenges in Earth science (see, e.g. Welch et al., 1995). These tasks have been classically addressed by DA approaches, especially variational DA and ensemble Kalman schemes (see, e.g. Evensen et al., 2009). In this general context, the quantification of estimation uncertainties as well as the inference of the intrinsical variabilities of the processes in play are crucial. It is especially important when dealing with misrepresented physical processes (Machenhauer & Kirchner, 2000) and unresolved small-scale processes (Hamill & Whitaker, 2005) with respect to the space-time observational resolution.

In a variational setting, assumptions (Le Dimet & Talagrand, 1986) moved from explicit model formulations to weaker ones (see, e.g. Trémolet, 2007). This allows to take into account the model error. However, standard variational methods do not always allow to estimate uncertainties of the predicted state-space. Concerning ensemble methods, estimating model error also became crucial and efforts were made to adapt ensemble methods to non-deterministic model (Sasaki, 1970). This led to the development of the iterative ensemble Kalman filter in presence of additive noise by Sakov et al. (2018). Unlike variational methods, ensemble methods provide a Gaussian estimate of the pos-

terior distribution of the state-space through a covariance matrix updated at each step (see, e.g. Evensen, 2003a; Evensen & Van Leeuwen, 2000).

Recently, a rich literature has emerged to apply machine learning (ML) paradigms to address DA issues. ML schemes are particularly efficient to solve complex and high-dimensional optimization problems. Its numerous successful applications in various fields are proof of this. Applications include image classification (Le, 2013; Krizhevsky et al., 2012), natural language processing (Otter et al., 2020), language translation (Sutskever et al., 2014) computational physics (Raissi et al., 2017; Mohan et al., 2020)...

Regarding DA, ML-based algorithms offer new means to learn the governing equations of the dynamics (Fablet et al., 2018; Long et al., 2018) and the associated flow operator (Fablet et al., 2021; Bocquet et al., 2020; Scher & Messori, 2019), or model correction terms (Long et al., 2018; Farchi et al., 2021), directly from the data. These algorithms can be used in a plug-and-play manner in state-of-the-art DA schemes. When considering variational DA, trainable emulators of the adjoint operator of the dynamics (Nonnenmacher & Greenberg, 2021) or directly of the gradient-based DA solver (Fablet et al., 2021) emerged as appealing solutions. Similarly, recent studies have explored learning-based Kalman techniques (de Bézenac et al., 2020). The latter is particularly relevant to address uncertainty quantification. The underlying assumption of the existence of the linear-Gaussian latent space may however restrict their application to real-world case-studies. Generative adversarial networks also naturally arose as appealing ML tools to develop new ensemble DA schemes (Silva et al., 2021).

In this paper, we further explore how to bridge learning-based schemes and DA to infer the posterior distribution of the state of a dynamical system given a set of observations. Following a VB inference formulation, we develop an end-to-end neural architecture to retrieve a parametric approximation of the posterior. Our neural scheme derives from an underlying ELBO cost and exploits a trainable surrogate representation of the dynamics and a trainable gradient-based solver. It can be regarded as an extension of Fablet et al. (2021) to a state-space associated with the parameters of the posterior. To the best of our knowledge, this is the first study which combines a trainable solver for variational DA along with a VB formulation. We demonstrate the relevance of the proposed scheme for different case-studies using a Gaussian approximation for the posterior pdf. We further discuss the generalization of the proposed approach to non-Gaussian posterior and related works.

This paper is structured as follows. Section 2 introduces DA and uncertainty quantification. Section 3 presents the proposed approach, based on ELBO maximization, and the associated end-to-end neural framework. Numerical experiments on Lorenz 63 dynamics and discharges on Danube river network are reported in Section 4. Finally, concluding remarks are provided in Section 5.

2 Problem statement

DA relies on state-space formulation for some time-dependent state x and associated time-dependent observations y. Within a time-continuous setting, it leads to (see (Trémolet, 2007)):

$$\begin{aligned} \frac{\partial x}{\partial t}(t) &= \mathcal{M}(x(t)) + \eta(t) \\ y(t) &= \mathcal{H}(x(t)) + \epsilon(t), \quad t \in \Omega \end{aligned} \tag{1}$$

with $\mathcal{M}$ the dynamical model and $\mathcal{H}$ the observation operator. Variable η and ϵ represent respectively the model error and the observational error. Assuming a zero-mean random process η, the weak-constraint variational DA formulation (Sasaki, 1970) states the reconstruction or forecasting of x given y as the following minimization issue:

$$U_\phi(x, y) = ||\mathcal{H}(x) - y||_R^2 + ||x - \phi(x)||_Q^2, \tag{2}$$

where ϕ is the flow operator associated with dynamical operator $\mathcal{M}$. ϕ is also referred to a time-stepping operator:

$$\phi(x)(t) = x(t-\Delta) + \int_{t-\Delta}^{t} \mathcal{M}(x(u))du. \tag{3}$$

On the right side of Equation 2, the first term represents the data fidelity term with respect to the observation, whereas the second one penalizes the discrepancy between the state and the underlying dynamics. The considered norms are Mahalanobis norm with respect to covariance matrices R and Q. R is the observational error matrix while Q is the model error matrix. The estimation of these matrices is of paramount importance (see, e.g. Tandeo et al., 2018; Trémolet, 2007). Especially, in Kalman methods (Evensen & Van Leeuwen, 2000), the error matrices drove the inference of the posterior. Particle filters (Gordon et al., 1993; Van Leeuwen, 2009), lag-innovation (Belanger, 1974), and Bayesian inference-based methods such as Stroud et al. (2018); Tandeo et al. (2015) addressed the estimation of these matrices. Particle filters suffer high-dimensional problems (Snyder et al., 2008)). Other approaches assume additive time-independent noise processes. This may restrict their applicability when considering time-varying and state-dependent noise processes.

We aim at addressing these shortcuts thanks to the modeling versatility of the deep learning schemes. As shown in Fablet et al. (2021), one may learn jointly dynamical prior ϕ (Equation 3) and a gradient-based solver for the minimization of cost U_ϕ (Equation 2). This joint learning feature lets us introduce an augmented state-space formulation to directly account for a parametric approximation of the posterior of the state $p(x|y)$ rather than state x. As detailed hereafter, we exploit a ELBO criterion (see, e.g. Hoffman et al., 2013) to extend Equation 2 to this augmented state-space formulation.

3 Proposed approach

3.1 VB formulation

For a state-space formulation such as Equation 1, VB inference (Kingma & Welling, 2013) relies on the approximation of the true posterior $p(x|y)$ by a parametric target distribution $q_\theta(x)$. θ refers to the parameters of this approximation. The ELBO provides a lower-bound to the likelihood of the observations y:

$$\log p(y) \geq \mathbf{E}_{x\sim q_\theta} \log\left(\frac{p(x,y)}{q_\theta(x)}\right),$$

with equality whenever $q_\theta(x) = p(x|y)$. We can equivalently rewrite this inequation :

$$\log p(y) \geq \mathbf{E}_{x\sim q_\theta} \log\left(p(y|x)\right) - D_{KL}(q_\theta||p(x)), \tag{4}$$

where D_{KL} denotes the Kullback-Leibler divergence and measures how two distributions differ from each other. The Kullback-Leibler divergence between two distributions is given, for two pdf p and q, by the following expression :

$$\mathbf{E}_{x\sim q} \log\left(\frac{q(x)}{p(x)}\right).$$

The ELBO can then lead to a computationally-tractable maximization of a lower-bound of the likelihood $p(y)$ (Hoffman et al., 2013).

Let us further assume a Gaussian approximation for target distribution q_θ and a Gaussian additive noise model for observation likelihood $p(y|x)$, that is to say $q_\theta \sim \mathcal{N}(\mu, \Sigma)$ and $p(y|x) \sim \mathcal{N}(x, R)$. For a linear observation operator $\mathcal{H}$, we then derive:

$$\mathbf{E}_{x\sim q_\theta} \log\left(p(y|x)\right) = -\frac{1}{2}(Tr(R^{-1}\Sigma) + ||\mathcal{H}(\mu) - y||_R^2),$$

up to a function of R. Under the assumption that norm of the posterior covariance is significantly smaller than that of the observation covariance, this term reduces to $-\frac{1}{2}||\mathcal{H}(\mu)-y||^2_R$.

Concerning the Kullback-Leibler divergence in ELBO expression of Equation 4, if we further assume that the prior satistifies $p(x) \sim \mathcal{N}(m, S)$, then we can derive the following analytical expression:

$$-D_{KL}(q_\theta||p(x)) = -\frac{1}{2}\left(Tr(S^{-1}\Sigma) + ||\mu - m||^2_S + \log\left(\frac{|S|}{|\Sigma|}\right)\right). \tag{5}$$

If we no longer assume a specific form for the prior $p(x)$, the expression $-D_{KL}(q_\theta||p(x))$ is a non-positive function of approximate posterior parameters θ. Let us call g this non-negative function. To match the generic formulation of the prior term in Equation 2, we consider the following form for $g(\theta)$:

$$g(\theta) = -||\phi(\theta) - \theta||^2.$$

This form is also widely used in machine learning regularizing techniques experimented by Ryu et al. (2019); Venkatakrishnan et al. (2013) and referred to as plug-and-play methods for inverse problems. Besides, we may note that Equation 5 may be rewritten in this form. As ϕ is usually unknown, we should rely on an estimator $\tilde{\phi}$ of ϕ to compute g. Overall, from the ELBO formulation, we infer the parameters $\theta = (\mu, \mathbf{\Sigma})$ of a Gaussian approximation of the true posterior $p(x|y)$ according to the minimization of a variational cost given by

$$U_{\tilde{\phi}}(\theta, y) = \lambda||\mathcal{H}(\mu) - y||^2 + ||\theta - \tilde{\phi}(\theta)||^2. \tag{6}$$

For the sake of simplicity, we have further assume a scalar covariance matrix R with parameter λ. This variational formulation is similar to that considered in Fablet et al. (2021). These authors provided a NN approach to jointly learn the operator $\tilde{\phi}$ and a gradient-based solver of variational cost Equation 6 to infer the approximate posterior $q_\theta(x)$ from available observations y.

3.2 Proposed neural architecture

We introduce the proposed end-to-end neural architecture based on variational cost defined in Equation 6. This neural architecture combines two main components: a neural parameterization for operator $\tilde{\phi}$, and a trainable gradient-based solver.

Fablet et al. (2021) noticed that a learned $\tilde{\phi}$ leads to better results than imposing the dynamics. Consequently, we used a constrained convolutional NN representation of $\tilde{\phi}$. Recall from Equation 6 that the minimum of the stochastic variational cost w.r.t $\tilde{\phi}$ is reached whenever $\tilde{\phi}$ is equal to the identity. As long as we want to find dynamical trends in the evolution of parameters, we constrain the convolutional NN to differ from the identity. We impose the constraint in the architecture of the convolutional NN itself. As in Fablet et al. (2021), we use Gibbs Energy NN (Perez et al., 1998) with two scale representations.

Once we design a neural formulation for the dynamical operator $\tilde{\phi}$, the minimisation of stochastic variational cost Equation 6 is performed by means of a neural solver. We use a ResNet architecture with Long Short-Term Memory blocks (Schmidhuber et al., 1997)). Each block is fed on one side with the increment between the estimated parameters at the entry of the block and the input parameters $\theta^{(0)}$, and on the other side by the gradient of the variational cost with respect to θ applied on the current estimated parameters. The number of iterations has been tuned during experiments and optimal values are comprised between 5 and 10 iterations. Figure 1 shows the working principle of the end-to-end architecture. The proposed neural architecture was implemented using pytorch framework and the Adam optimizer.

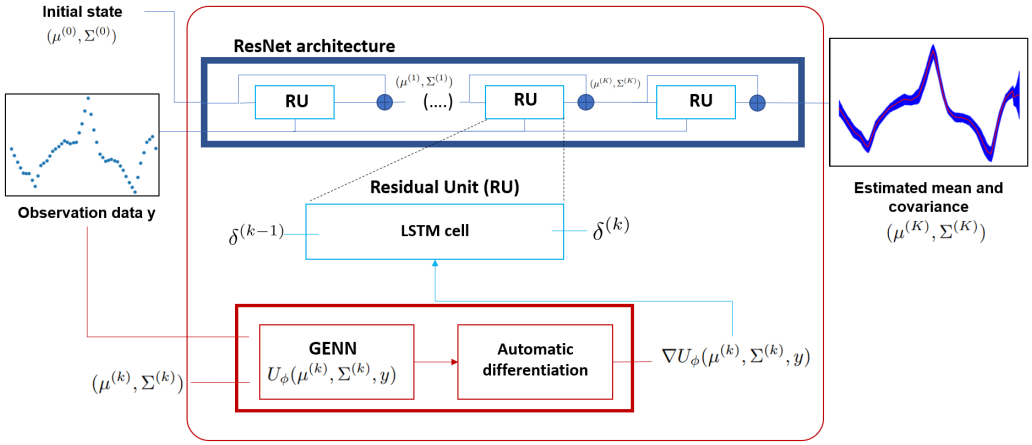

Figure 1. Proposed end-to-end architecture. Illustration comes from L63 experiment. Given a partial observation piece of data y and an initial pdf state (μ_0, Σ_0), the proposed network calculates the optimized parameters (μ_K, Σ_K) after K steps in the solver. On the right handside, red curve contains μ and the blue envelope is a rescaled visualisation of Σ. $\delta^{(k)}$ is the difference between the parameters at iteration step (k) and at iteration step $(k \quad - \quad 1)$. LSTM and GENN stands respectively for Long-Short Term Memory and Gibbs Energy NN.

3.3 Learning setting

A cost function to measure the proximity between q_θ and $x|y$ needs to be chosen. As we are predicting probability distribution, the logarithmic score is a proper scoring rule (see Dawid & Musio, 2014). Considering a supervised setting with access to true state during training, we only have a single realization of $x|y$ for each y. We cannot calculate the proper score $\mathbf{E}_{z\sim x|y}(-\log(q_\theta(z)))$ because there is no exact ground truth for the distribution $x|y$. An approximation is required.

Let us consider a training dataset which comprises observation series $\{\mathbf{y}_1, ..., \mathbf{y}_N\}$, and true states $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, with each $\mathbf{y}_i \in \mathbf{R}^{d_y} \times \mathbf{R}^{N_t}$ and $\mathbf{x}_i \in \mathbf{R}^{d_x} \times \mathbf{R}^{N_t}$. d_y is the spatial dimension of the observation domain, d_x is the spatial dimension on which we wish to reconstruct the posterior. N_t is the length of the time window. Let note Γ the neural solver and Φ the NN dynamical operator. The output of the system for an input parameter $\theta^{(0)}$ and an observation series $\mathbf{y}$ will be noted as $\Theta_{\Phi,\Gamma}(\theta^{(0)}, \mathbf{y})$. For a dataset of size N, we have N outputs $\Theta_{\Phi,\Gamma}(\theta_i^{(0)}, \mathbf{y}_i) = \{\mu_k^i, \Sigma_k^i,\ k \in [0, N_t]\}$. Score S is set as a log-likelihood criterion, given by the following :

$$S(\Theta_{\Phi,\Gamma}(\theta_i^{(0)}, \mathbf{y}_i), \mathbf{x}_i) = -\sum_{k=0}^{N_t} \log(p_{\{\mu_k,\Sigma_k\}}((\mathbf{x}_i)_k). \tag{7}$$

$p_{\{\mu_k,\Sigma_k\}}$ is the pdf of a gaussian law of parameters $\{\mu_k, \Sigma_k\}$. The overall partially supervised criterion is :

$$\mathcal{N} = \frac{1}{N}\sum_{i=0}^{N} S(\Theta_{\Phi,\Gamma}(\theta_i^{(0)}, \mathbf{y}_i), \mathbf{x}_i). \tag{8}$$

4 Numerical experiments

To assess the relevance of the proposed approach, we consider two case-studies: namely, the Lorenz 63 dynamics and an application to a real dataset corresponding to the monitoring of Danube river discharges. In the following, our approach will be referred to as 4DvarnetSto. The baseline approach is the Ensemble Kalman Filter and will be abbreviated as EnKF.

4.1 L63 dynamics

4.1.1 Standard L63 dynamics

The Lorenz dynamics is a system made of the following ordinary differential equations (Lorenz, 1963)):

$$\begin{aligned} \frac{dx_1}{dt} &= \sigma(x_2 - x_1) \\ \frac{dx_2}{dt} &= \rho x_1 - x_2 - x_1 x_3 \\ \frac{dx_3}{dt} &= x_1 x_2 - \beta x_3. \end{aligned} \tag{9}$$

We use the following parametrization : $\sigma = 8$, $\rho = 28$, and $\beta = \frac{8}{3}$. In this setup, the Lorenz 63 system has a chaotic solution. An RK4 (Butcher, 1996) integration scheme with 0.01 time step enables us to simulate the time series. Figure 2 (a) is a trajectory of this dynamics for 200 time steps.

4.1.2 Stochastic L63 dynamics

In order to introduce model noise in L63 dynamics, we use the stochastic framework designed by Chapron et al. (2018). It intends to mimic stochastic behaviour in large-scale geophysical flow dynamics. The ordinary differential equation (Equation 9) becomes a stochastic differential equation :

$$\begin{aligned} dX_1 &= \left(\sigma(X_2 - X_1) - \frac{4}{2\Gamma}X_1\right)dt \\ dX_2 &= \left(\rho X_1 - X_2 - X_1 X_3 - \frac{4}{2\Gamma}\right)dt + \frac{\rho - X_3}{\Gamma^{\frac{1}{2}}}dB_t \\ dX_3 &= \left(X_1 X_2 - \beta X_3 - \frac{8}{2\Gamma}X_3\right)dt + \frac{X_2}{\Gamma^{\frac{1}{2}}}dB_t. \end{aligned} \tag{10}$$

dB_t is a white noise, formally the derivative of a standard Brownian motion. Γ is the new parameter of our model which is fixed to 2 in our experiments. Note that if $\Gamma \longrightarrow \infty$, we recover the original model. Figure 2 (b) is a 3D plot for a time series of this stochastic Lorenz 63 version. Adding the model noise strongly deteriorate the smoothness and the convergence to standard Lorenz attractor.

4.1.3 Training setting and results

For both dynamics, we consider time series of 200 time steps. Training set contains 10000 time series, validation and test set 2000 each. Observations of the real state are made available solely for the first variable of the system, every 8 timesteps. We train our NN in two stages. First, we constraint the covariance matrix to be diagonal and we find a first optimum. In a second step, we start a new learning session to find a non-diagonal covariance matrix initialized by the previous diagonal matrix.

We compare our method with the EnKF of Evensen (2003b). In our experiment, the EnKF has 1000 ensemble members and the initial state is chosen as if the first time

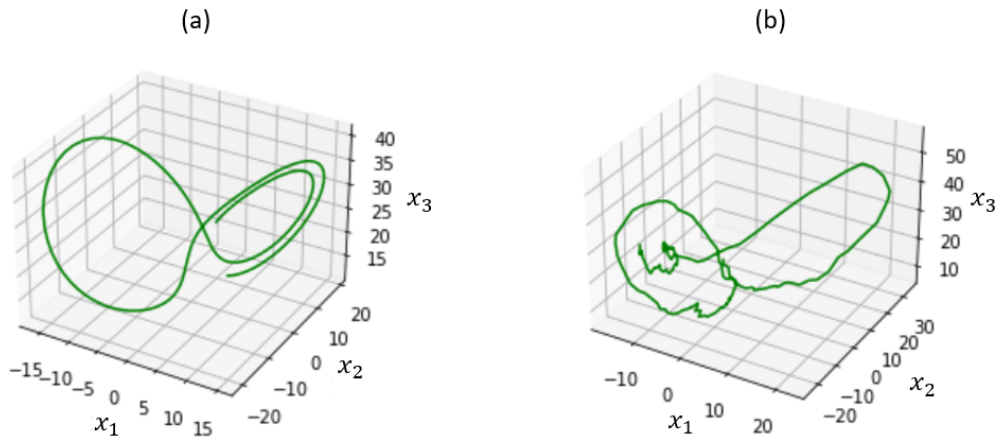

Figure 2. Evolution of Lorenz dynamics for (a) standard model (see Equation 9) and (b) stochastic model of Chapron et al. (2018) (Equation 10) for 200 time steps of 0.01 length each. The value on each axis has been standardized and normalized according to scalar mean and standard deviation calculated on the training set.

step of the time window was entirely observed. Notice that the inference with the ensemble method is done by filtering and not by smoothing (see, e.g. Evensen & Van Leeuwen, 2000) which would have led to better performance. The objective here is not so much to compare the reconstruction error of our method and the best possible ensemble method but rather to demonstrate that our approach is better suited for entropy-based minimization criterion. Table 1 compiles the important results for the appropriate scores. If the first variable is observed for both our approach and EnKF, the 4DVarnetSto outperforms the EnKF in each score for both dynamics. By adding observed variables in EnKF experiment, the R-score and P-score decrease. For the standard dynamics, the P-score for the EnKF with all variables observed becomes lower than its value in the 4DvarnetSto experiment, but the R-score stay above. To minimize the P-score with the EnKF, we inflated the covariance matrix by 1.15. Covariance matrix inflation is a common techniques in ensemble methods to avoid filter divergence (see, e.g. Anderson & Anderson, 1999). Notice that without covariance inflation, the R-score is much lower and reaches 0.36 when all variables are observed. For the stochastic dynamics, the results of our approach with one observed variable are comparable to the EnKF with x_1 and x_2 observed. In this case, the model noise is sufficient to avoid the use of an inflation factor in the EnKF experiments.

Figure 3 compares our estimated states (orange) and the associated 95% confidence interval with the real states (blue) defined by Equation 9 in the context of standard dynamics. Figure 4 presents the same elements for the stochastic dynamics defined by Equation 10. Both figures represent time series for which the attractor changes its wing. The change of wing is realized when the variables x_1 and x_2 simultaneously go from a maximum to a minimum or vice versa. In Figure 3, the mean state estimated by our approach (top three graphs) and the actual state of the system are almost merged. Moreover, the area representing the uncertainty is also very thin but widens for a given variable when an extremum is reached. The uncertainty is slightly larger for the unobserved variables x_2 and x_3 than for the observed variable x_1. Comparatively, the EnKF with only x_1 observed shows an estimated state further from the true state and a higher uncertainty (middle three graphs). This is particularly noticeable during the transition from one wing to

Model	Model noise	R-score	P-score
4DvarnetSto	No	0.45	-4.60
with x_1 observed	Yes	3.51	-1.42
EnKF	No	1.40	-0.48
with x_1 observed	Yes	23.8	4.9
EnKF	No	1.70	-3.60
with x_1 and x_2 observed	Yes	4.44	-1.85
EnKF	No	0.84	-5.08
with all variables observed	Yes	2.60	-2.47

Table 1. Scores of 4DVarnetSto and EnKF for L63 simulations for both dynamics. Model noise sets to "No" indicates standard dynamics (see Equation 9), "Yes" implies stochastic one (see Equation 10). Only the first variable is observed when performing 4DvarnetSto. In EnKF experiments, from one to all variables are considered as observed. For the standard dynamics, the covariance matrix is inflated by a factor of 1.15. Two benchmark score are evaluated : the mean square error of the reconstruction of the true state (R-score), and the mean of the negative log-likelihood of the predicted parametric distribution applied in true state (P-score, see Equation 8).

the other (between $t = 50$ and $t = 125$). The uncertainty is then very high because the sequential approach of the EnKF does not allow to predict in advance which wing the trajectory is heading for. When observing the 3 variables for the EnKF (bottom three graphs), the estimated state and uncertainty are improved. Except for the first few time steps corresponding to a calibration phase of the EnKF and the change of wing, the estimated and actual state almost coincide.

4.2 Danube river network for discharge measurements

The upper Danube basin is an European river network which covers a large part of Austria, Switzerland and of the south of Germany. Figure 5 shows the topography of the Danube basin as well as the locations of the 31 stations at which daily measurements of river discharge are available. Stations considered as observed or unobserved in our experiment are colored differently. The daily measurements series have lengths from 51 to 110 years. We restrict ourselves to the period 1960-2010 for which all stations have available measurements. This dataset have been widely studied in the community of multivariate extremes (see for example Mhalla et al. (2020); Asadi et al. (2015)).

This experiment with a real dataset aims to meet several objectives. Learning an unknown dynamics and associated uncertainties is challenging. The data-driven models that can be learned lacks important variables (precipitation, snow melt) to be highly reliable, and consequently encompass high error model. Thus, the ability of our approach to adapt to a high level of model error is studied. Finally, we assess how informative the estimated fluctuation of variances is.

The main difference between this experiment and those performed for Lorenz dynamics is that our dataset does not include true state but solely observations. Consequently, we cannot exactly use the supervised criterion of Equation 8. Assuming that the observational error is negligible compared to the model error, we replace Equation 7 by

$$S(\Theta_{\Phi,\Gamma}(\theta_i^{(0)}, \mathbf{y}_i, M_i), \mathbf{y}_i) = -\sum_{k=0}^{N_t} \log(p_{\{\mu_k,\Sigma_k\}}((\mathbf{y}_i)_k). \tag{11}$$

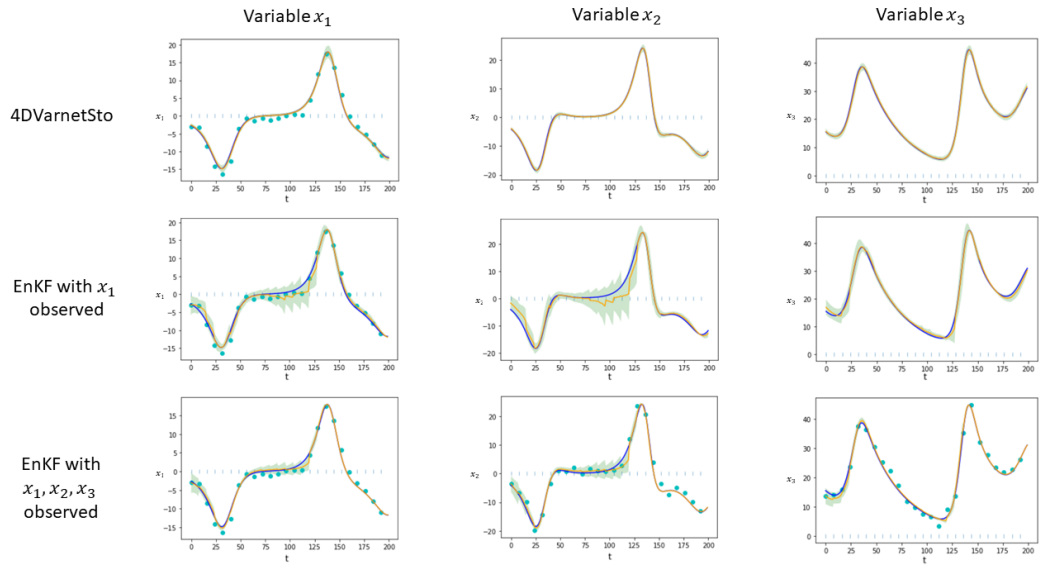

Figure 3. Experiments with standard Lorenz dynamics (Equation 9). For a set of observations (cyan dots) on given timesteps (light blue dashes on the time axis), the true state (blue curve) and estimated state (orange curve) are plotted for our approach and EnKF with one or all variables observed. The estimated 95% confidence intervals are represented by the green area.

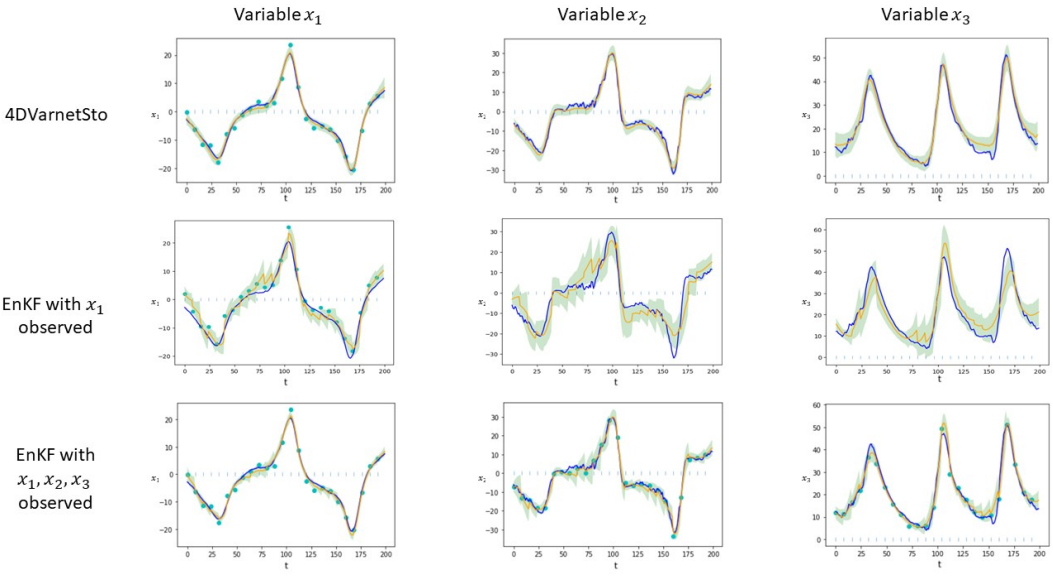

Figure 4. Experiments with the stochastic Lorenz dynamics of Chapron et al. (2018) (Equation 10). See Figure 3 for details.

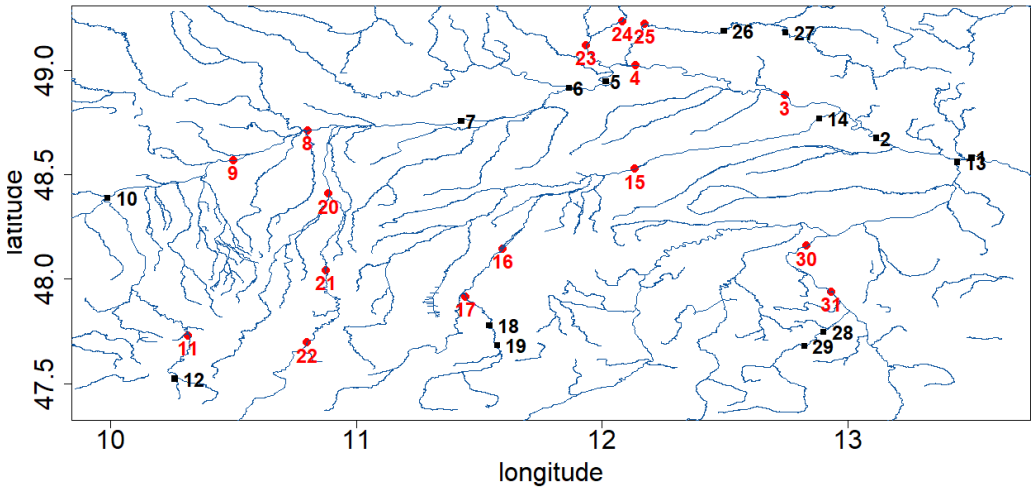

Figure 5. Topographic map of the upper Danube basin with the 31 gauging stations. A dataset of 50 years of daily measurements is considered (from 1960 to 2010). In training setting, we assume that some stations are observed (red dots) and the other are unobserved (black squares). We further assume that the observed stations have available observations only once every four days.

Using this framework, half of the stations are considered to be observed every four days (see red locations in Figure 5). We consider time series of 48 consecutive days. For each time series, our goal is to estimate the mean and covariance of the approximate posterior distribution of flow on each day of the time series and at each station, wherever observations are missing. The training dataset comprises 9999 time series of 48 days, validation and test set 1749 each. To construct these datasets, we divided the 51 years of daily measurement into 550-day blocks. In each block, the first 350 days create 303 time series for the training dataset. The 200 remaining days are divided in two and create 53 time series for validation set and as many for the test set. Figures 6 & 7 show the estimated mean state, the confidence interval and the observations for a summer and winter month, respectively. The stations are identical from one figure to another. Seasonality plays an important role in discharge analysis, and here, we focus on the summer and winter seasons. In summer, flows are lower than in winter and subject to important variations in absolute value. This is linked essentially to snow or ice melts at altitude, as well as to episodes of heavy precipitation. For similar reasons, different station elevations, and thus different positions along the river system, were chosen. Stations upstream of the river system have lower flows than those downstream. Flows at upstream stations vary greatly depending on local weather and climate events.

The relative variance estimated by our approach is larger in Figure 6 than in Figure 7. This finding is consistent with the initial considerations about variances in summer and winter. The estimated variance is also more constant in summer than in winter. One can assume that the model error is such that it becomes difficult to detect patterns that would reduce the uncertainty. In winter, on the other hand, the estimated confidence interval varies significantly, and seems to widen at the peaks reached by the flow. We notice that our predictions are sometimes biased for a large number of consecutive time steps. This is particularly true in Figure 7 where a negative bias between the observations and the predicted mean exist, especially for unobserved stations. The presence of available observations drastically reduces the bias.

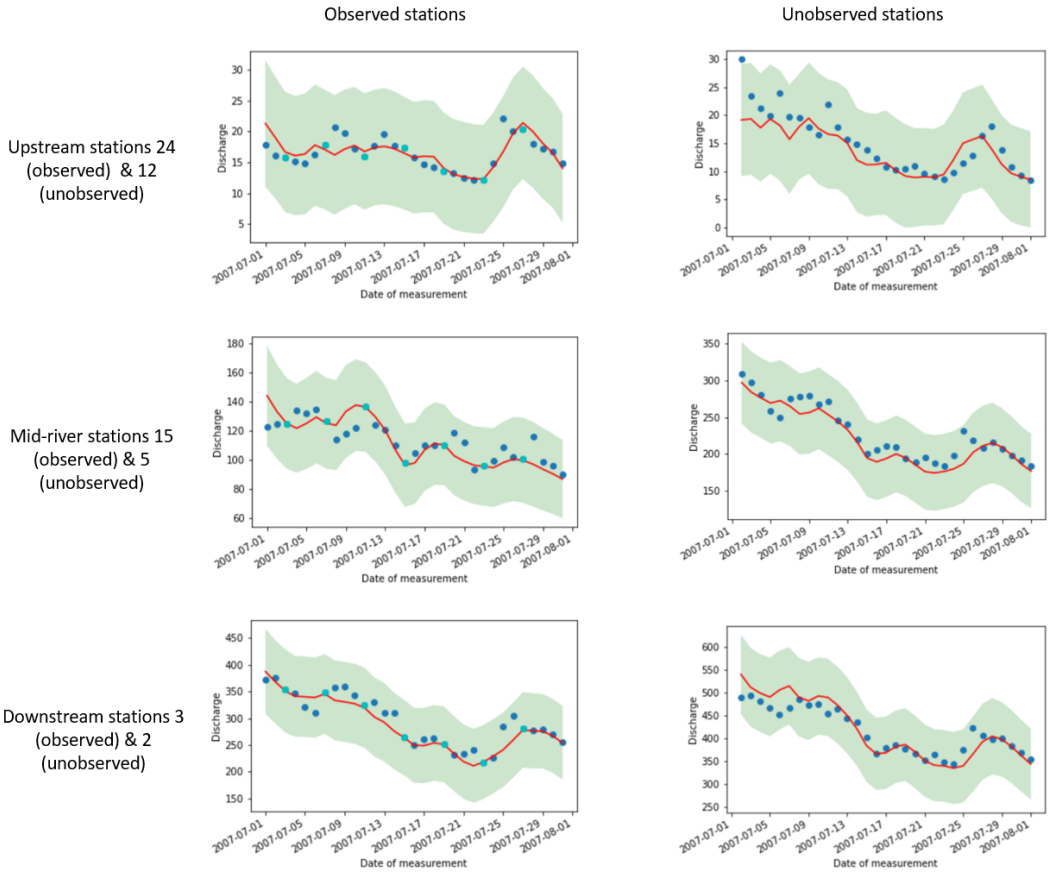

Figure 6. For a summer month (July 2007), we show the estimated discharge (red curve), the 95% confidence interval (green area) estimated by our method for observed and unobserved stations at different elevations. The daily measurements are also represented according to whether they are available (light blue dots) or unavailable(deeper blue) as inputs. The discharges are expressed in m^3/s.

In order to estimate the quality of the variance predictions of our approach, we first calculated the entropy defined by the score defined in Equation 11, with μ_k and Σ_k obtained by our approach, and averaged it over test dataset. Then, we created a comparative approach. To do so, we replaced our obtained Σ_k with a constant covariance matrix Σ_{diag}. This covariance is diagonal and each diagonal coefficient is the variance between our mean estimation μ_k and true observations $(y_i)_k$ for a given station. The entropy computed with our approach is 0.068 against 1.06 with the naive approach. The variations of variances given by our approach allow a significant improvement of the entropy criterion.

5 Conclusion

Based on previous works which introduced an end-to-end learning framework for variational assimilation problems, we extend this approach to stochastic prediction given an ad hoc parametric family, namely the gaussian. Using a stochastic variational cost derived from an ELBO maximization w.r.t a target gaussian distribution, we have been able to find a gaussian approximation of the pdf of the posterior. The learning framework comprises a neural-network representation of the dynamics of the parameters and

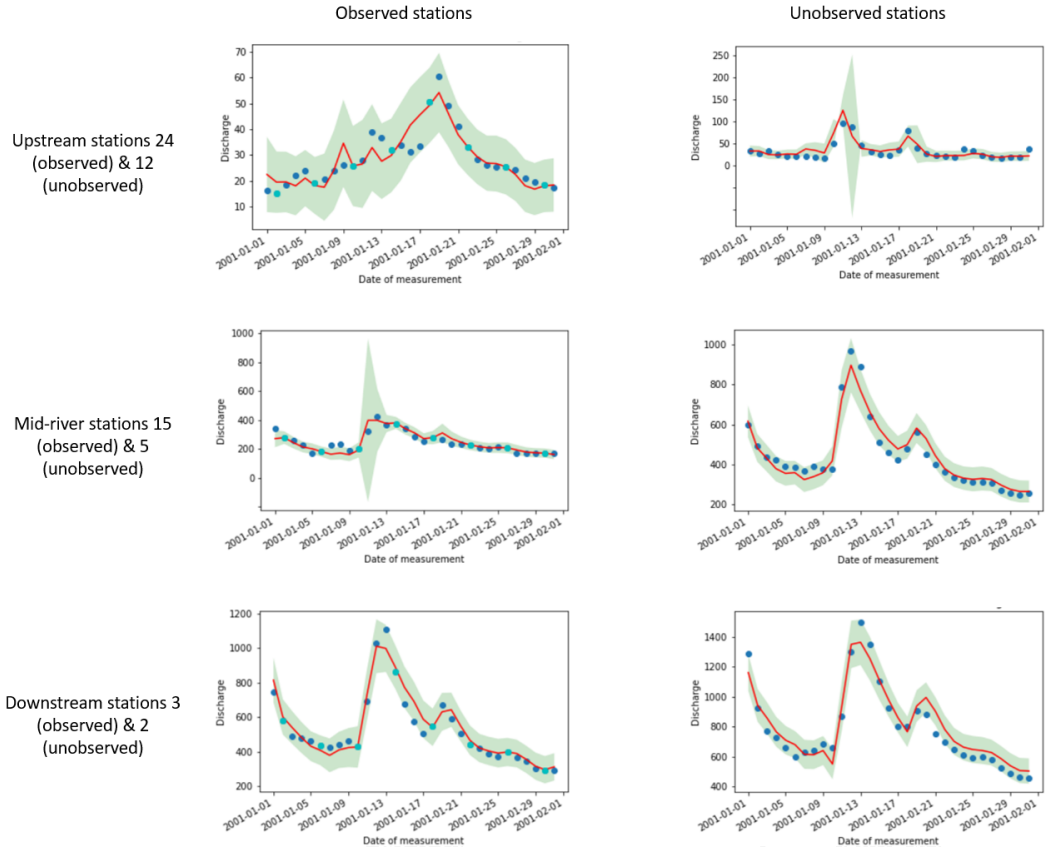

Figure 7. Winter month (January 2000) (see Figure 6 for details).

a neural solver for the considered stochastic variational cost. Both solver and parameters are learnt jointly in a context of entropy optimization. This joint learning process offers new perspectives for VB-based cost minimization in DA problems.

Lorenz 63 dynamics and discharges on Danube river networks have been studied. Concerning Lorenz dynamics, our approach captures well the dynamics and the uncertainty. When adding state-dependent model noise, we have been able to retrieve complex type of uncertainty structure. The experiments on the Danube river system provide a setting where the dynamics are unknown, and the data to estimate them incomplete. In this context, our approach allows us to calculate a consistent estimate of the flow, the associated dynamics and the uncertainties.

Our findings also underlines that beyond state-of-the-art results obtained for mean squared error of reconstruction, our approach is well-suited for entropy criterion. This is a real improvement over reference ensemble methods which suffer from limitations and require careful adaptation to obtain good performance for an entropy criterion.

Future works will focus on improving the accuracy of the upper quantile of the predicted distribution. In fact, for both experiments, our predicted confidence intervals for high quantile are narrower than the empirical ones. A parameterization of the posterior by heavy tail distribution (see e.g. Resnick, 2007) could be an improvement track. Moreover, as discharges are positive values, a Gaussian parametrization is not ideal to infer uncertainties. More broadly, symmetrical distribution cannot consistently large uncertainty in this problem as it could cover negative flow value. Extending our approach to non-symmetrical distribution would be of interest.

6 Open Research

We provide our associated code available at `https://github.com/Nicolasecl16/These/tree/main/4Dvarnetstochastic` which comprises the generation of the synthetic datasets. Danube river network dataset is make available by the Bavarian Environmental Agency at `http://www.gkd.bayern.de`.

Acknowledgments
The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-Melody (ANR-19-CE46-0011). Part of this work was supported by 80 PRIME CNRS-INSU, ANR-20-CE40-0025-01 (T-REX project), and the European H2020 XAIDA (Grant agreement ID: 101003469).

References

Anderson, J. L., & Anderson, S. L. (1999). A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly weather review*, *127*(12), 2741–2758.

Asadi, P., Davison, A. C., & Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, *9*(4), 2023–2050.

Belanger, P. R. (1974). Estimation of noise covariance matrices for a linear time-varying stochastic process. *Automatica*, *10*(3), 267–275.

Bocquet, M., Brajard, J., Carrassi, A., & Bertino, L. (2020). Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *arXiv preprint arXiv:2001.06270*.

Butcher, J. C. (1996). A history of runge-kutta methods. *Applied numerical mathematics*, *20*(3), 247–260.

Chapron, B., Dérian, P., Mémin, E., & Resseguier, V. (2018). Large-scale flows under location uncertainty: a consistent stochastic framework. *Quarterly Journal of the Royal Meteorological Society*, *144*(710), 251–260.

Dawid, A. P., & Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, *72*(2), 169–183.

de Bézenac, E., Rangapuram, S. S., Benidis, K., Bohlke-Schneider, M., Kurle, R., Stella, L., ... Januschowski, T. (2020). Normalizing kalman filters for multivariate time series analysis. *Advances in Neural Information Processing Systems*, *33*, 2995–3007.

Evensen, G. (2003a). The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, *53*(4), 343–367.

Evensen, G. (2003b). The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, *53*(4), 343–367.

Evensen, G., et al. (2009). *Data assimilation: the ensemble kalman filter* (Vol. 2). Springer.

Evensen, G., & Van Leeuwen, P. J. (2000). An ensemble kalman smoother for nonlinear dynamics. *Monthly Weather Review*, *128*(6), 1852–1867.

Fablet, R., Chapron, B., Drumetz, L., Mémin, E., Pannekoucke, O., & Rousseau, F. (2021). Learning variational data assimilation models and solvers. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002572.

Fablet, R., Ouala, S., & Herzet, C. (2018). Bilinear residual neural network for the identification and forecasting of geophysical dynamics. In *2018 26th european signal processing conference (eusipco)* (pp. 1477–1481).

Farchi, A., Laloyaux, P., Bonavita, M., & Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, *147*(739), 3067–3084.

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *Iee proceedings f-radar*

and signal processing (Vol. 140, pp. 107–113).

Hamill, T. M., & Whitaker, J. S. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Monthly weather review*, *133*(11), 3132–3147.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8595–8598).

Le Dimet, F.-X., & Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A: Dynamic Meteorology and Oceanography*, *38*(2), 97–110.

Long, Z., Lu, Y., Ma, X., & Dong, B. (2018). Pde-net: Learning pdes from data. In *International conference on machine learning* (pp. 3208–3216).

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, *20*(2), 130–141.

Machenhauer, B., & Kirchner, I. (2000). Diagnosis of systematic initial tendency errors in the echam agcm using slow normal mode data assimilation of ecmwf reanalysis data. *CLIVAR Exchanges*, *5*(4), 9–10.

Mhalla, L., Chavez-Demoulin, V., & Dupuis, D. J. (2020). Causal mechanism of extreme river discharges in the upper danube basin network. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(4), 741–764.

Mohan, A. T., Lubbers, N., Livescu, D., & Chertkov, M. (2020). Embedding hard physical constraints in convolutional neural networks for 3d turbulence. In *Iclr 2020 workshop on integration of deep neural models and differential equations.*

Nonnenmacher, M., & Greenberg, D. S. (2021). Deep emulators for differentiation, forecasting, and parametrization in earth science simulators. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002554.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, *32*(2), 604–624.

Perez, P., et al. (1998). *Markov random fields and images.* IRISA.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*.

Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling.* Springer Science & Business Media.

Ryu, E., Liu, J., Wang, S., Chen, X., Wang, Z., & Yin, W. (2019). Plug-and-play methods provably converge with properly trained denoisers. In *International conference on machine learning* (pp. 5546–5557).

Sakov, P., Haussaire, J.-M., & Bocquet, M. (2018). An iterative ensemble kalman filter in the presence of additive model error. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1297–1309.

Sasaki, Y. (1970). Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, *98*(12), 875–883.

Scher, S., & Messori, G. (2019). Generalization properties of feed-forward neural networks trained on lorenz systems. *Nonlinear processes in geophysics*, *26*(4), 381–399.

Schmidhuber, J., Hochreiter, S., et al. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735–1780.

Silva, V. L., Heaney, C. E., & Pain, C. C. (2021). Gan for time series prediction, data assimilation and uncertainty quantification. *arXiv preprint arXiv:2105.13859*.

Snyder, C., Bengtsson, T., Bickel, P., & Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, *136*(12), 4629–4640.

Stroud, J. R., Katzfuss, M., & Wikle, C. K. (2018). A bayesian adaptive ensemble kalman filter for sequential state and parameter estimation. *Monthly weather review*, *146*(1), 373–386.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., & Zhen, Y. (2018). Joint estimation of model and observation error covariance matrices in data assimilation: a review.

Tandeo, P., Pulido, M., & Lott, F. (2015). Offline parameter estimation using enkf and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quarterly journal of the royal meteorological society*, *141*(687), 383–395.

Trémolet, Y. (2007). Model-error estimation in 4d-var. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *133*(626), 1267–1280.

Van Leeuwen, P. J. (2009). Particle filtering in geophysical systems. *Monthly Weather Review*, *137*(12), 4089–4114.

Venkatakrishnan, S. V., Bouman, C. A., & Wohlberg, B. (2013). Plug-and-play priors for model based reconstruction. In *2013 ieee global conference on signal and information processing* (pp. 945–948).

Welch, G., Bishop, G., et al. (1995). An introduction to the kalman filter.