



Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review

Sibo Cheng, César Quilodrán-Casas, Said Ouala, Alban Farchi, Che Liu, Pierre Tandeo, Ronan Fablet, Didier Lucor, Bertrand Iooss, Julien Brajard, et al.

► To cite this version:

Sibo Cheng, César Quilodrán-Casas, Said Ouala, Alban Farchi, Che Liu, et al.. Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review. *IEEE/CAA Journal of Automatica Sinica*, inPress, 10 (6), pp.1361-1387. 10.1109/JAS.2023.123537. hal-04039094

HAL Id: hal-04039094

<https://hal.science/hal-04039094>

Submitted on 21 Mar 2023

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

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

Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review

Sibo Cheng, César Quilodrán-Casas, Said Ouala, Alban Farchi, Che Liu, Pierre Tandeo, Ronan Fablet, Didier Lucor, Bertrand Iooss, Julien Brajard, Dunhui Xiao, Tijana Janjic, Weiping Ding, Yike Guo, Alberto Carrassi, Marc Bocquet, Rossella Arcucci*

Abstract—Data Assimilation (DA) and Uncertainty quantification (UQ) are extensively used in analysing and reducing error propagation in high-dimensional spatial-temporal dynamics. Typical applications span from computational fluid dynamics (CFD) to geoscience and climate systems. Recently, much effort has been given in combining DA, UQ and machine learning (ML) techniques. These research efforts seek to address some critical challenges in high-dimensional dynamical systems, including but not limited to dynamical system identification, reduced order surrogate modelling, error covariance specification and model error correction. A large number of developed techniques and methodologies exhibit a broad applicability across numerous domains, resulting in the necessity for a comprehensive guide. This paper provides the first overview of the state-of-the-art researches in this interdisciplinary field, covering a wide range of applications. This review aims at ML scientists who attempt to apply DA and UQ techniques to improve the accuracy and the interpretability of their models, but also at DA and UQ experts who intend to integrate cutting-edge ML approaches to their systems. Therefore, this article has a special focus on how ML methods can overcome the existing limits of DA and UQ, and vice versa. Some exciting perspectives of this rapidly developing research field are also discussed.

Keywords: Machine Learning; Deep learning; Data assimila-

tion; Uncertainty quantification; Reduced-order-modelling.

I. INTRODUCTION

The rapid growth of Machine Learning (ML) has been witnessed in a wide range of research fields, including computer vision [1], natural language processing [2] and AI for science [3]. In particular, literature shows that the application of ML algorithms, from conventional methods to deep neural networks, is present in nearly all aspects of spatio-temporal problems [4], [5], [6]. Adopting ML algorithms yields considerable improvements in forecasting complex high-dimensional dynamics. However, the black-box nature of ML algorithms makes them to exhibit poor interpretability, lack of robustness, weak reliability, and vulnerability to adversarial attacks and noisy systems. On the other hand, Data Assimilation (DA) [7] and Uncertainty quantification (UQ) [8] are reference frameworks that deal with model/data noises and error propagation inside dynamical systems. Compared to ML, they provide interpretable and explicit solutions based on some mathematical assumptions, such as linearity and Gaussianity [7], [9], [10].

As [11] pointed out, substantial mathematical similarities exist between ML and DA, in particular, variational-type assimilation methods [7]. In fact, the latter also relies on gradient descent techniques to minimise a cost function measuring the difference between model outputs and prior estimation/observation. Several works have examined the connections between the acquisition, interpretation, and use of data in ML and DA. Integrations of DA and ML have been introduced in [12], [13], [14]. The link between probabilistic ML approaches and differential equations is highlighted when the frameworks of DA and ML are combined from a Bayesian perspective. This equivalency, which demonstrates the parallels between the two areas, is presented formally in [11], [15]. Here, they show how to approximate Bayesian inverse methods (i.e., Variational data assimilation (VarDA) in DA and back-propagation in ML) can be utilised to combine the four-dimensional VarDA (4D-Var) and Recurrent Neural Network (RNN) fields. In [16], VarDA and ML are considered already incorporated in a Weak Constraint VarDA, offering a somewhat different viewpoint. As demonstrated in [17], [15], [18], these approaches are also particularly well adapted to systems using Gaussian processes. These are data-driven algorithms capable of estimating model statistics and learning nonlinear, space-dependent,

Sibo Cheng, César Quilodrán-Casas, Che Liu, Yike Guo and Rossella Arcucci are with Data Science Institute, Department of computing, Imperial College London, SW7 2AZ London, UK.

César Quilodrán-Casas, Che Liu and Rossella Arcucci are also with Department of Earth Science and Engineering, Imperial College London, SW7 2AZ London, UK.

Yike Guo is also with Department of Computer Science and Engineering, Hong Kong university of science and technology, Hong Kong, 999077, China.

Said Ouala, Pierre Tandeo and Ronan Fablet are with IMT Atlantique, Lab-STICC, UMR CNRS 6285, France and Odyssey, Inria/IMT, France.

Pierre Tandeo is also with RIKEN Center for Computational Science, Kobe, Japan

Alban Farchi and Marc Bocquet are with CERE, École des Ponts and EDF R&D, île-de-France, France.

Didier Lucor is with the Laboratoire Interdisciplinaire des Sciences du Numérique, CNRS, Paris-Saclay university, F-91403, Orsay, France.

Bertrand Iooss is with Electricité de France (EDF), 78401 Chatou, France, Institut de Mathématiques de Toulouse, 31062 Toulouse, France and SINCLAIR AI Lab, Saclay, France.

Julien Brajard is with Sorbonne University, Paris, France and Nansen Environmental and Remote Sensing Center (NERSC), Bergen, Norway.

Dunhui Xiao is with School of Mathematical Sciences, Tongji University, 200092 Shanghai, China.

Tijana Janjic is with Mathematical institute for machine learning and data science, KU Eichstätt-Ingolstadt, Bavaria, Germany.

Weiping Ding is with School of Information Science and Technology, Nantong University, 226019 Nantong, China.

Alberto Carrassi is with Department of Physics and Astronomy “Augusto Righi”, University of Bologna, 40124 Bologna, Italy.

Corresponding author: Rossella Arcucci (r.arcucci@imperial.ac.uk)

cross-correlations in a unified manner. Specifically, for high-dimensional systems, DA is often combined with Reduced-Order Modelling (ROM), such as classical Proper Orthogonal Decomposition (POD) [19] and ML-based autoencoders [20] to reduce the computational cost. Both practical uses of these fusion algorithms, such as air quality forecasting using data-driven artificial intelligence [21], [22] and more theoretical ones, like spatiotemporal oscillations of the Partial Differential Equation (PDE) [23] using numerically computed approximations of Koopman eigenfunctions and eigenvalues, have been presented. Other methods, such as those in [12], [24], which iteratively apply an Ensemble Kalman Filter (EnKF) and a neural network to imitate hidden dynamics and forecast future states, are more akin to the works reported in this paper. A modular approach integrating neural network and DA has been presented in [14] which shows several methods to combine neural network and DA to overcome limitations in applying these fields to real-world data.

ML methods fail under their primary form in providing any guarantees of convergence or quantifying the error/uncertainty associated with their predictions, thus it is critical to provide UQ to ML predictions in order to anticipate and explain model failure to generalise. Model UQ is crucial for instance to help choosing what data to learn from, or exploring an agent's environment efficiently. In reality, data collection can be very expensive and time-consuming in dynamical system; and in extreme cases, only sparse and discrete batches of noisy data can be observed overtime. In these cases, UQ for ML helps in learning from small amounts of labelled data. UQ is also extensively used for quantifying error propagation in dynamical systems [25] through Monte Carlo methods and Polynomial Chaos [26], [27]. Monte Carlo methods are known for their broad applicability, and Polynomial Chaos is proven to have significant advantages in terms of computational efficiency and interpretability [25]. When dealing with noisy dynamical systems, DA and UQ can be naturally combined. For example, [28] made use of polynomial chaos expansion to model and reduce the sampling errors in EnKF. A number of papers [29], [30] applied Monte Carlo methods to estimate the error covariance matrices, which played a pivotal role in DA algorithms.

Growing research efforts were devoted to combining and comparing DA and UQ with ML under different contexts. The number of published articles (including preprints on open-access repositories) from 2012 to 2021 that involved the concept of DA, UQ and ML is illustrated in Figure 1. A sudden increase can be noticed, especially from 2015 when Deep Learning (DL) [31] started to become the reference approach in many research areas. The applications of these methods cover a large range of fields, including climate science, fluid dynamics and image analysis. In this review, the related researches are mainly classified into two categories: *DA using ML techniques* and *ML assisted by DA and UQ*, respectively. The former focuses on ML-based solutions to the long-standing challenges of DA, including the correction of forward model errors and the error covariance specification. The second category gives attention to how DA and UQ can assist ML in predicting high-dimensional dynamical systems.

Specifically we concentrate on the challenge of noisy partial data and the use of real-time observations to progressively adjust ML surrogate models.

Figure 2 illustrates conceptually the related technologies as a function of problem dimension (x-axis) and noise level (y-axis) for their usual use cases. Different challenges presented in this review are also displayed where the colours indicate the technologies involved. It is worth mentioning that the ROM plays a pivotal role in enabling the combination of ML and DA methods, especially in real-world applications, by reducing the computational cost.

This review aims to cover most of the cutting-edge articles in the related research fields. This paper can thus serve as a comprehensive guide for navigating these fast growing techniques and methodologies. We stress that the objective of this work is not to compare the performance of existing methods since they were developed to address different problems. In summary, we made the following contributions in this paper:

- To the best of the authors' knowledge, this is the first review that addresses the combination of ML, DA and UQ for dynamical systems.
- This paper has a special focus on how ML methods can contribute to the key challenges of DA and UQ, and vice versa.
- This review includes a range of main applications in DA, UQ and ML, such as Numerical Weather Prediction (NWP), environmental modelling and Computational Fluid Dynamics (CFD).
- Some promising and insightful research perspectives and challenges are discussed.

The rest of the paper is organized as follows. Section II introduces the background and preliminaries for DA, UQ and ML applied to high-dimensional dynamical systems. In Section III and IV, we describe how the cutting-edge ML techniques can be used to address the key challenges in DA and UQ, and vice versa. Other approaches and perspectives that combine ML with DA or UQ are discussed in Section V. We finish this review with a conclusion in Section VI.

II. BACKGROUND AND PRELIMINARIES

In this section, we briefly summarise the foundation of UQ, DA and ML with a particular attention given to the application on high-dimensional dynamical systems.

A. Uncertainty and error quantification for complex and dynamical systems

In statistical terms, there are different types of uncertainties [32], [33], including

- *Aleatoric uncertainty* which originates in noisy input data (gappy, noisy, discordant or multimodal), where homoscedastic uncertainty denotes the variance that stays constant for all input parameters;
- *Heteroscedastic uncertainty* represents the variance that depends on the input parameters and can potentially be predicted as a model output. In general, regardless of the quality of a model or the amount of training data, this uncertainty is irreducible;

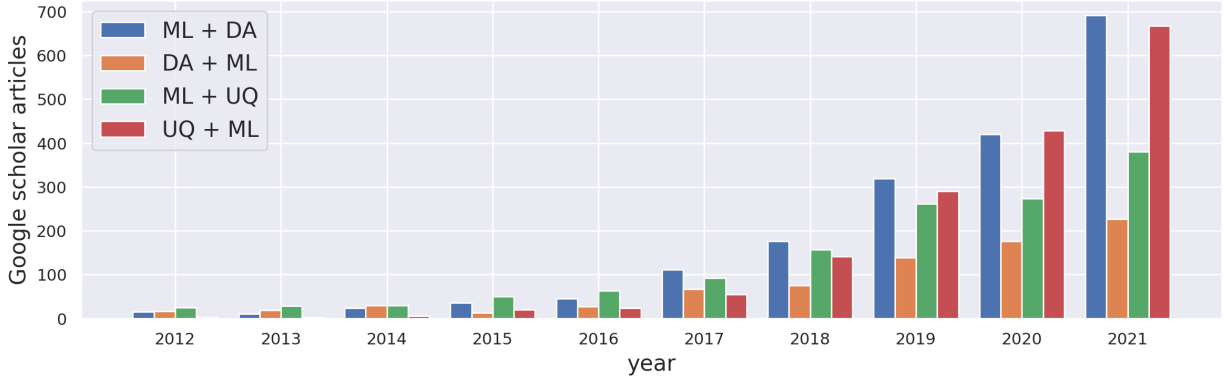


Fig. 1: Number of published articles combining ML, DA and UQ according to Google scholar. 'A + B' denotes the number of articles which include 'A' in the title and 'B' in the text.

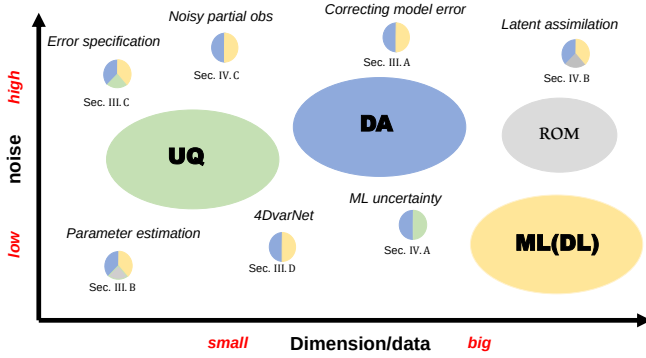


Fig. 2: Combination of ML, DA and UQ methods and challenges versus available data dimension and noise level

- *Uncertainty in model parameters* that best explain the observed data, for instance a large number of models are able to explain a given dataset, in which case we might be uncertain which model parameters to choose to predict with;
- *Structural uncertainty*, i.e., what model structure should we use, how do we specify our model to interpolate and extrapolate well.

The latter two can be grouped under model uncertainty that are epistemic uncertainties. Epistemic uncertainties describe the fidelity of the model in its representation of the data—barring aleatoric uncertainties. Typically in ML, epistemic uncertainties decrease as the training data size increases. Combining aleatoric and epistemic uncertainties provides us with the predictive uncertainty that is the confidence level of the model accounting for noise it can explain and noise it cannot.

Sensitivity analysis (SA) is the primary tool in UQ to analyse error propagation in complex and dynamical systems through understanding and distinguishing the effects of various uncertainties on model output [34], [35], [36], [37]. SA can be used to determine which input variables contribute most to an output behavior, and which inputs are not influential, or to check for certain interaction effects in the model. This can

be extremely useful when the model interpretability is poor or the explicit formulas of the governing equation is out of reach, as in a majority of ML models. The SA process involves calculating and analyzing sensitivity indices of input variables relative to a given quantity of interest in the model output (e.g., its mean, its variance, a particular quantile, its maximum). Considering the variance as an uncertainty metrics, sensitivity indices of each uncertain input variable on the variance of the model output allow for a better understanding of the model behavior, in order to reduce the uncertainties in the output in the most efficient manner. For example, identifying the most influential inputs will reduce their uncertainties, and then the model output uncertainty [38].

In [36], four settings have been defined for the needs of SA in practice. First, the model exploration setting aims at understanding the behavior of the model by investigating the input-output relationship, e.g., via graphical tools. Second, the factors fixing setting aims at reducing the number of uncertain inputs by finding then fixing non-influential inputs. Third, the factors prioritization setting aims at precisely quantifying the effects of the most influential inputs. Last, the input distribution robustness setting aims at analyzing the variations in the quantity of interest with respect to uncertainty in inputs' distributions.

Recently, it has been recognized (see, e.g., [39], [40]) a wide analogy between SA and the topic of interpretability in ML [41]. As in SA, four settings have been defined in [39] for the needs of ML interpretability in practice: visualization of the relation between the predicted output label and the input features, identification of the most important features in ML prediction, important measures of explanatory variables and robustness of the decision boundary. UQ can also be extremely useful in DA models, especially in determining the prior and posterior errors of states and observations, which play an important role in DA [42]. In fact, as shown in the next section, DA algorithms, particularly Kalman-based procedures, are probabilistic approaches where the state of the system is a Gaussian random variable. Thus, the state vector is defined at each time step by a mean vector and a covariance matrix. This UQ of the assimilation results is important because it provides information on : the observation sampling in both space and

time, the uncertainty correlation between variables, and the confidence in the estimate of the state vector.

B. DA for dynamical systems

Data assimilation aims at predicting physical fields and estimating model parameters [7], [43] by aggregating information from different sources. Applying DA to dynamical systems allows continuous corrections to model predictions while accounting for model and observation errors/uncertainties.

A typical time dependent DA framework over a discrete time window $[0, \dots, T]$ involves variables and parameters as listed below.

- states $\mathbf{x}_t$: the target field of estimation, which is assumed unobservable in DA;
- true states $\mathbf{x}_t^{\text{true}}$: theoretical values of $\mathbf{x}_t$;
- background states $\mathbf{x}_t^b$: prior estimations of $\mathbf{x}_t^{\text{true}}$, often obtained via predictive models;
- observations $\mathbf{y}_t$: observable quantities, for example, from sensors or satellites;
- transformation operators $\mathcal{H}_t$: functions that map the state variables to the observations;
- transformation operators $\mathcal{M}_t$: functions that map the dynamical system from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$;
- prior errors $\epsilon_t^x, \epsilon_t^y$: estimation and prediction errors associated to $\mathbf{x}_t^b$ and $\mathbf{y}_t$ respectively;
- error covariances $\mathbf{B}_t, \mathbf{R}_t, \mathbf{Q}_t$: auto-covariance matrices of background, observation and model errors;
- analysis states $\mathbf{x}_t^a$: output of DA models;

where $\mathbf{y}_t$ exists only when the observation at time t is available. In the rest of this paper, we denote $\mathbf{x}_{0:T} = \{\mathbf{x}_0, \dots, \mathbf{x}_T\}$ and $\mathbf{y}_{0:T} = \{\mathbf{y}_0, \dots, \mathbf{y}_T\}$ as a sequence of states and observations, respectively. We denote $\mathbf{H}_t, \mathbf{M}_t$ the linearisation of $\mathcal{H}_t, \mathcal{M}_t$.

Operational DA models are mainly twofold: Kalman filter-based methods from estimation theory and variational DA related to control theory. Both families of DA approaches can be derived from Bayes' theorem [44]. The analysis states $\mathbf{x}_t^a$ obtained from DA could be viewed as a compromise between $\mathbf{x}_t^b$ and $\mathbf{y}_t$, where the weights are determined by $\mathbf{B}_t, \mathbf{Q}_t$ and $\mathbf{R}_t$. In DA, the state vector is Gaussian because the errors terms $\epsilon_t^x, \epsilon_t^y$ in the state- and observation-space model are often assumed to be additive, Gaussian, and centered [7]. The background error represents the error propagation in the dynamical model and the fact that the initial condition is not necessarily well estimated [45]. The observation error represents the mismatch between the observation vector $\mathbf{y}_t$ and the state vector projected in the observation space. This mismatch is mainly due to the error of representativity between the two vectors [46] and instrumental errors. Those background and observation errors are characterised by the covariance matrices $\mathbf{B}_t$ and $\mathbf{R}_t$, respectively. They show error amplitudes, error spatial correlations, and shared errors between variables.

a) *Variational DA*: Following DA's statistical framework [44], we wish to directly apply Bayes' rule [44] over $[0, T]$, with batches of observations $\mathbf{y}_{0:T}$. We can focus on

the estimation of the conditional Probability Density Function (PDF) of $p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T})$. Applying Bayes' rule, we obtain:

$$p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T}) = \frac{p(\mathbf{y}_{0:T}|\mathbf{x}_{0:T})p(\mathbf{x}_{0:T})}{p(\mathbf{y}_{0:T})} \propto p(\mathbf{y}_{0:T}|\mathbf{x}_{0:T})p(\mathbf{x}_{0:T}), \quad (1a)$$

where the evidence $p(\mathbf{y}_{0:T})$ is inessential here. We further assume that the observation errors are Gaussian and uncorrelated in time, with covariance matrices $\mathbf{R}_{0:T}$, so that:

$$p(\mathbf{y}_{0:T}|\mathbf{x}_{0:T}) = \prod_{t=0}^T p(\mathbf{y}_t|\mathbf{x}_t) \quad (2a)$$

$$\propto \exp \left[-\frac{1}{2} \sum_{t=0}^T \|\mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t)\|_{\mathbf{R}_t^{-1}}^2 \right]. \quad (2b)$$

Moreover, the prior PDF $p(\mathbf{x}_{0:T})$ is assumed *Markovian*, i.e. the state $\mathbf{x}_t$ conditional on the previous state $\mathbf{x}_{t-1}$ does not depend on all other previous past states:

$$p(\mathbf{x}_{0:T}) = p(\mathbf{x}_0) \prod_{t=1}^T p(\mathbf{x}_t|\mathbf{x}_{0:t-1}) \quad (3a)$$

$$\stackrel{\text{Markov}}{=} p(\mathbf{x}_0) \prod_{t=1}^T p(\mathbf{x}_t|\mathbf{x}_{t-1}). \quad (3b)$$

We also assume Gaussian statistics for the model error and the initial background which are uncorrelated in time, with zero bias and covariance matrices $\mathbf{Q}_{0:T}$ and $\mathbf{B}_0$, respectively. Therefore, $p(\mathbf{x}_{0:T})$ is proportional to:

$$p(\mathbf{x}_0) \exp \left[-\frac{1}{2} \sum_{t=1}^T \|\mathbf{x}_t - \mathcal{M}_t(\mathbf{x}_{t-1})\|_{\mathbf{Q}_t^{-1}}^2 \right], \quad (4)$$

with

$$p(\mathbf{x}_0) = \exp \left[-\frac{1}{2} \|\mathbf{x}_0 - \mathbf{x}_0^b\|_{\mathbf{B}_0^{-1}}^2 \right]. \quad (5)$$

Now, we can gather the likelihood and prior pieces to obtain the cost function associated to the conditional PDF $p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T})$:

$$\begin{aligned} \mathcal{J}(\mathbf{x}_{0:T}) &= -\ln p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T}) \\ &= \frac{1}{2} \left(\|\mathbf{x}_0 - \mathbf{x}_0^b\|_{\mathbf{B}_0^{-1}}^2 + \sum_{t=0}^T \|\mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t)\|_{\mathbf{R}_t^{-1}}^2 \right. \\ &\quad \left. + \sum_{t=1}^T \|\mathbf{x}_t - \mathcal{M}_t(\mathbf{x}_{t-1})\|_{\mathbf{Q}_t^{-1}}^2 \right), \end{aligned} \quad (6a)$$

up to constants that do not depend on the control variables $\mathbf{x}_{0:T}$. This is the cost/loss function of the *weak-constraint 4D-Var* [47], with which several connections to ML can be made. The cost/loss function can be minimised via gradient-based nonlinear optimisation, hence finding a compromise between the constraints of the observations, of the model, and of the background, whose errors are weighted against their statistics in the loss function. The argument of the minimum is the analysis trajectory. Note that to be applicable to geofluid models, this algorithm must be cycled in time, sequentially, i.e. the dynamical numerical model must be applied to this analysis.

Even though very powerful, these variational DA methods require the tangent linear and adjoint models of $\mathcal{M}_k$ and $\mathcal{H}^k$ [48] which hampered their widespread adoption, but might now be boosted by the developments of differentiable models from ML. More details are given in Section III-D of this paper.

In variational DA, well modelled error covariance matrices, including $\mathbf{B}_t$, $\mathbf{R}_t$ and $\mathbf{Q}_t$, are required since they are crucial to spread information between observed and non-observed variables within the analysis of the scheme. However, due to the high-dimensionality, defining these covariances as a sequence of operators can be intricate and computationally demanding, as further discussed in Section III-C.

b) *Kalman-filter-based DA*: Processing the measurements as they become available is what is done in a *filter*. As opposed to estimating the full PDF $p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T})$ all at once, filters sequentially estimate the marginal PDFs $p(\mathbf{x}_t|\mathbf{y}_{0:t})$, for all $t \in [0, T]$. The process alternates an *analysis step*, based on Bayes' rule [49], where the conditional PDF $p(\mathbf{x}_t|\mathbf{y}_{0:T})$ is updated using the latest observation $\mathbf{y}_t$, with a *forecast step* which propagates this PDF to the next observation batch [50].

This Bayesian approach is very difficult already in problems of moderate model dimension due to the cost of sampling and evolving these PDFs. Similar to the **4DVar!** (**4DVar!**), we assume that the uncertainties about observations, model, and prior are all Gaussian distributed: the PDFs are now defined by only means and covariance. By further assuming that the dynamical and observational models are both linear, with time-in-time and mutually uncorrelated errors, we get the Kalman Filter (KF), which is the exact analytic solution to the Gaussian estimation problem:

$$\text{Forecast Step} \quad \mathbf{x}_t^f = \mathbf{M}_{t-1}\mathbf{x}_{t-1}^a, \quad (7a)$$

$$\mathbf{P}_t^f = \mathbf{M}_{t-1}\mathbf{P}_{t-1}^a\mathbf{M}_{t-1}^T + \mathbf{Q}_t. \quad (7b)$$

$$\text{Analysis step} \quad \mathbf{K}_t = \mathbf{P}_t^f\mathbf{H}_t^T(\mathbf{H}_t^k\mathbf{P}_t^f\mathbf{H}_t^T + \mathbf{R}_t^{-1}), \quad (8a)$$

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^f), \quad (8b)$$

$$\mathbf{P}_t^a = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_t^f. \quad (8c)$$

Equations (7a)–(8c) sequentially estimate the state, $\mathbf{x}_t^f$, $\mathbf{x}_t^a$, and error covariance, $\mathbf{P}_t^f$, $\mathbf{P}_t^a$. The matrix $\mathbf{K}_t$ is the Kalman gain containing the coefficients of the optimal linear combination between the prior mean and the observations. The analysis $\mathbf{x}_t^a$, has minimum error variance and is unbiased. The KF is very powerful and, by solving for mean and covariance, provides a time dependent estimate of the system's state and associated uncertainty. Its biggest limitations are the linear assumptions and the computational cost for storing, evolving and manipulating the matrices.

The extended Kalman filter (Extended Kalman Filter (EKF), [50]) is a first-order expansion of the KF for nonlinear dynamics. It operates a linearisation of the nonlinear model equations around the model's solution. The nonlinear model is used to propagate the state but the tangent linear model for the error covariance evolution. The EKF also assumes Gaussian errors, but under the action of nonlinear dynamics, even an initial Gaussian error may become non Gaussian.

The EKF is therefore a good approximation as long as the observations interval is shorter than the time scale of the error growing modes [51]. Although the EKF has been successful in a number of pioneering applications, including DA for the geosciences, see *e.g.* [52], [53], [54], it is also plagued by the same huge computational requirements as the KF.

A Monte Carlo approach is at the basis of a class of algorithms referred to as EnKFs [55]. The EnKFs use the KF statistical framework and mimics its analysis updates, but the estimation and propagation of the errors is approximated by a finite ensemble of model realisations. While the accuracy of the EnKFs is linked to the size of the computationally affordable ensemble, the EnKFs gave proofs of extraordinary capabilities even in high-dimensional problems ($\mathcal{O}(10^8)$), by using “only” as few as 100 members. The reasons behind this success are multiple and concurrent. The ensemble members are used to approximate a Gaussian distribution, as opposed to the vastly more complex task of estimating a generic, non parametric, PDFs. The latter is attempted by particle filters [56] that are in fact strongly affected by the curse of dimensionality. The application of the EnKFs to chaotic dynamics, such as for geofluids, is challenged by the instabilities and the low predictability. Nevertheless, the EnKFs benefit from the chaotic systems' tendency to confine error growths within a smaller subspace than the full system's dimension. Tracking this relatively low-dimensional unstable subspace with the finite ensemble is easier than affording the error description in the fully dimensional space [57].

Finally and in practice, the success of the EnKFs in high dimensions is related to two ad-hoc fixes: *inflation* and *localisation* [7]. Inflation consists in artificially increasing the ensemble-based error covariance, to combat error underestimation due to under sampling. Even more impactful, localisation acts to boost the ensemble-based error covariance rank and span, by reducing or even eliminating the small long distance correlations that are unavoidably poorly estimated with a small ensemble [58]. A recent surveys on EnKF from the mean field perspective and for both discrete and continuous time can be found in [59] while a review on state-of-the-art ensemble-based approaches in general, including ensemble smoother and ensemble variational methods, is given in [60].

C. ML with UQ

Supervised and self-supervised ML approaches are proven to be very efficient in many real-world applications and have still great potential for improving industrial means of production as well as research and development aspects. However, all these opportunities are still subject to methodological challenges as, among others, the bias-variance trade-off, the balance between the complexity of the underlying model to learn and the available amount of training data, a possible high dimensional input space, the presence of heterogeneous noise in the observations [86]. One of these challenges concerns the UQ associated to ML predictions.

ML techniques can be categorized in different families [87], [88], each of them having specific characteristics, for example, theoretical properties, practical performance in complex

TABLE I: ML approaches considered in this review

Categories	Methods	Application/section	References
<i>Linear and polynomial</i>	linear operator	POD (II-D), KF (II-B), equation identification (IV-C)	[61], [19], [62]
	SINDY	ROM (II-D), equation identification (IV-C)	[63], [64], [65]
	GLA (IV-B)	latent DA (IV-B), parameter estimation (III-C)	[66], [67], [68]
	PCE	UQ (II-A and IV-A)	[27], [28]
<i>Neighborhood</i>	Kriging	equation identification (IV-C)	[69]
	KNN	ROM (II-D)	[67], [70], [71]
<i>Ensemble</i>	RF	ROM (II-D), parameter estimation (III-C)	[67], [70]
	DE	UQ (IV-A)	[72]
<i>Deep learning</i>	CNN	ROM (II-D), error specification (III-C), latent DA (IV-B)	[73], [74], [75], [76]
	RNN	ROM (II-D), error specification (III-C), latent DA (IV-B), equation identification (IV-C)	[77], [73], [78], [79]
	BNN	parameter estimation (III-C), UQ (IV-A)	[80], [81]
	GNN	ROM (II-D)	[82]
	Transformer	ROM (II-D), equation identification (IV-C)	[83], [84], [85]

problems related to the amount of data it requires, efficiency in high dimension, stability, computational complexity and interpretability capabilities. From the simplest to the most complex one, we distinguish the following four grand families:

- Linear and polynomial models (into which the method of polynomial chaos fits [89]). Confidence intervals associated to predictions performed by these models can be obtained, as these models provide an analytical formula for leave-one-out error as well as mathematical properties of their regression coefficients [90];
- Neighborhood models (into which the kriging method fits). The Gaussian process assumption behind the kriging model allows to associate easily computable confidence intervals associated to each prediction [89], [91];
- Ensemble models, especially those based on regression trees (e.g., random forests, gradient boosting). Obtaining confidence intervals for this kind of models is more difficult but two ways have been achieved: quantile regression [92] and a specific subsampling procedure [93];
- Deep learning, also known as Deep Neural Networks (DNN). In this category, different families of UQ methods have been proposed based on Bayesian frameworks, for example, Bayesian Neural Network (BNN) [81] and Monte-Carlo Dropout (MCD) [94]. The latter consists of ensembles of Neural Network (NN) optimization iterates or independently trained NNs (e.g., deep ensembles: DE [72]). Thorough overviews of uncertainty quantification in DL, are provided by very recent review papers, e.g., [95], [8], [96].

Finally, the method of conformal predictions [97] appears to be a valuable way to provide confidence intervals for any type of ML models.

In the particular case of predicting high-dimensional dynamical systems, the state-of-the-art ML methods often consist of DL-based ROM (see Section II-D). These reduced order models lie in a combination of different errors and uncertainties, including observation/data uncertainties, compression errors and predictive errors. The specification of these errors and the correction of the DL-based ROMs are discussed in detail in Section III-C and IV-A, respectively. In this review, we

consider the combination of DA and UQ with a wide range of ML algorithms as shown in Table I. However, the main focus is given to DL approaches since they are state-of-the-art in predicting dynamical systems, especially in high-dimensional spaces.

D. ML for predicting high-dimensional dynamical systems

ML algorithms for predicting high-dimensional dynamics often rely on ROM. More precisely, data are first compressed into a reduced latent space to decrease the computational cost. Predictive models are then used to surrogate the dynamics in the reduced space. Despite its efficiency, such a system will introduce several terms of errors and uncertainties, including compression and prediction errors. UQ and DA can be employed to specify and correct these errors, as discussed in detail in Section IV-A and IV-B. In this section, we review the state-of-the-art ML approaches for both ROM and time-series predictions.

1) *Reduced-order-modelling*: Reducing the dimension of complex dynamical systems has been a long-standing research problem [98], [99]. Projection-based approaches, such as POD and Proper Generalized Decomposition (PGD) have been extensively applied in a large range of engineering problems [100], where the explicit transition from the full physical space to a low dimensional latent space relies on linear projection operators. In the past decade, much attention has been given in enhancing the ROM using ML methods, in particular, DL-based autoencoders [77]. Autoencoder is a specific type of self-supervised neural network that has identical inputs and outputs. A typical autoencoder consists of an encoder E which maps the input variables $\mathbf{x}_t$ to the reduced latent space and a decoder D which reconstructs the full physical field $\mathbf{x}_t^{\text{AE}}$ from the latent representation $\mathbf{z}_t$, that is,

$$\mathbf{z}_t = E(\mathbf{x}_t) \quad \text{and} \quad \mathbf{x}_t^{\text{AE}} = D(\mathbf{z}_t). \quad (9)$$

The encoder E and the decoder D are trained jointly with the objective to minimise the reconstruction loss, for instance,

quantified by the Mean Square Error (MSE),

$$\mathcal{L}^{\text{MSE}} = \mathbb{E}(\|\mathbf{x}_t - D \circ E(\mathbf{x}_t)\|_F^2), \quad (10)$$

where $\mathcal{L}^{\text{MSE}}$ is the loss function. $\|\cdot\|_F$ and $\circ$ denote the Frobenius norm and the composition function, respectively. $\mathbb{E}$ is the expectation operator.

Autoencoders (AEs) show great potential in capturing non-linear patterns compared to projection-based methods such as POD [101], [85], [102]. However, the geometry of the latent space obtained by AEs can be chaotic, and thus, less interpretable [103]. Continuous efforts have been given in combining DL autoencoders with traditional dimension reduction methods. Carlberg et al. [104] used dimensionality reduction methods comprising Principal Component Analysis (PCA) and AEs to recover missing data. The works of [105] and [106] used AE to learn the Koopman invariant subspace for Dynamic Mode Decomposition (DMD). A number of studies have also successfully applied POD-based AEs for urban air pollution [107] and nuclear engineering [108], [68]. These methods benefit from both the accuracy of DL AEs and the interpretability of projection-based approaches. To tackle the issue of chaotic latent space, Variational Autoencoder (VAE) was proposed by [109], where a regularisation term is added in the loss function. The latent variables were constrained by Gaussian distributions through the Kullback–Leibler Divergence (KLD) to ensure the smoothness of the latent space geometry [109]. The explicit latent space could naturally improve the interpretability of AEs. More recently, the work of [110] introduced Vector Quantized Variational Autoencoders (VQ-VAE) which generate a discrete latent space instead of a continuous one like standard VAEs. Some recent researches [111], [85], [112] also attempted to enhance the reconstruction performance by employing an attention-based mechanism [113]. For example, Rui et al [85] presented a nonlinear non-intrusive using an AE and self-attention DL method. Stacked AE is used to perform the nonlinear model reduction and a self-attention mechanism is used to represent the fluid dynamics.

Among different structures of AEs, Convolutional Autoencoder (CAE) (including convolutional VAE), is by far the most widespread architecture. However, it can be cumbersome to apply CAE for unstructured data, for instance, in CFD with irregular meshes. To address this bottleneck, Graph Neural Network (GNN) architectures [82], [114] were proposed and applied with success for modelling liquids and granular materials. Moreover, [82] showed how graph-based ML can also learn adaptive remeshing. There is increased attention on using meshes for learned geometry and shape processing, but despite the widespread use in classical simulators, adaptive mesh representations have yet to see much use in learnable prediction models. To deal with incomplete data, the pioneering work of [115] introduced Masked autoencoders, capable of reconstructing the full field using a limited number of observable patches.

The development of different AEs is illustrated in Figure 3 with a particular focus on the transitions between ‘explicit’ and ‘implicit’ latent spaces.

2) *Predictive models*: Forecasts produced by ROMs cost only a fraction compared to high-dimensional model solution. Non-intrusive ROMs were broadly used in predicting reduced variables. Traditional approaches often relied on, for instance, radial basis functions [116] or shallow machine learning techniques, such as K-Nearest Neighbours (KNN) and Random Forest (RF) [67]. Recently, RNNs have been used to model and predict temporal dependencies between inputs and outputs of ROMs. ROMs and RNNs are used together in previous studies, e.g. [24], [117], [118] where the surrogate forecast systems can easily reproduce subsequent time-steps. Long Short-Term Memory (LSTM) networks, originally described in [119], are used extensively to learn the underlying dynamics in the reduced space [120], [121]. LSTM is a special variant of RNN that is stable and powerful enough to be able to model long-range time dependencies [122] and overcomes the vanishing gradient problem [123]. Some recent works in ROM also focused on the state-of-the-art ML predictive models, namely Transformer [83] and adversarial predictions [124], [22]. Transformers, originally developed in Natural Language Processing (NLP), have been successfully implemented with latent space representations of videos [83], or time-series forecasting [125]. However, the model efficiency and model adaptation are hampered during the implementations of transformers due to the computation and memory complexity of the self-attention modules in the transformer [84]. Some other approaches aim to learn the underlying governing equations with dynamical data as input. For example, Sparse Identification of Nonlinear Dynamics from Data (SINDy) [63], [126] present a procedure that extracts sparse dynamic system models from time series data. SINDy has been successful in generating robust, high quality models for physical systems, even with a ROM obtained via PCA [127], [128], [129] or deep AE [130]. Conversely, the accuracy of predictions can also be increased by including specialised knowledge about the system modelled in the form of loss terms [131], [132], or by physics-informed feature normalisation [133]. In summary, various ML predictive models were paired with ROM to release the computational burden in high-dimensional system modelling. However, when the predicted output is used as an input for the prediction of the subsequent time sequence (known as the ‘rollout’ process), the results can detach quickly from the underlying physical model solution when encountering out-of-distribution data. This detachment is mainly due to the error/uncertainty propagation and accumulation during iterative predictions over rollouts [134].

III. DATA ASSIMILATION USING MACHINE LEARNING TECHNIQUES

In this section, we focus on how ML techniques are used to address the key challenges of DA algorithms, including model error correction (Section III-A), parameter estimation (Section III-B), error covariance specification (Section III-C) and end-to-end learning of DA system (Section III-D).

A. ML to correct model errors in DA

As discussed at the end of Section II-D2, when predicting complex dynamical systems, physics-based or data-driven nu-

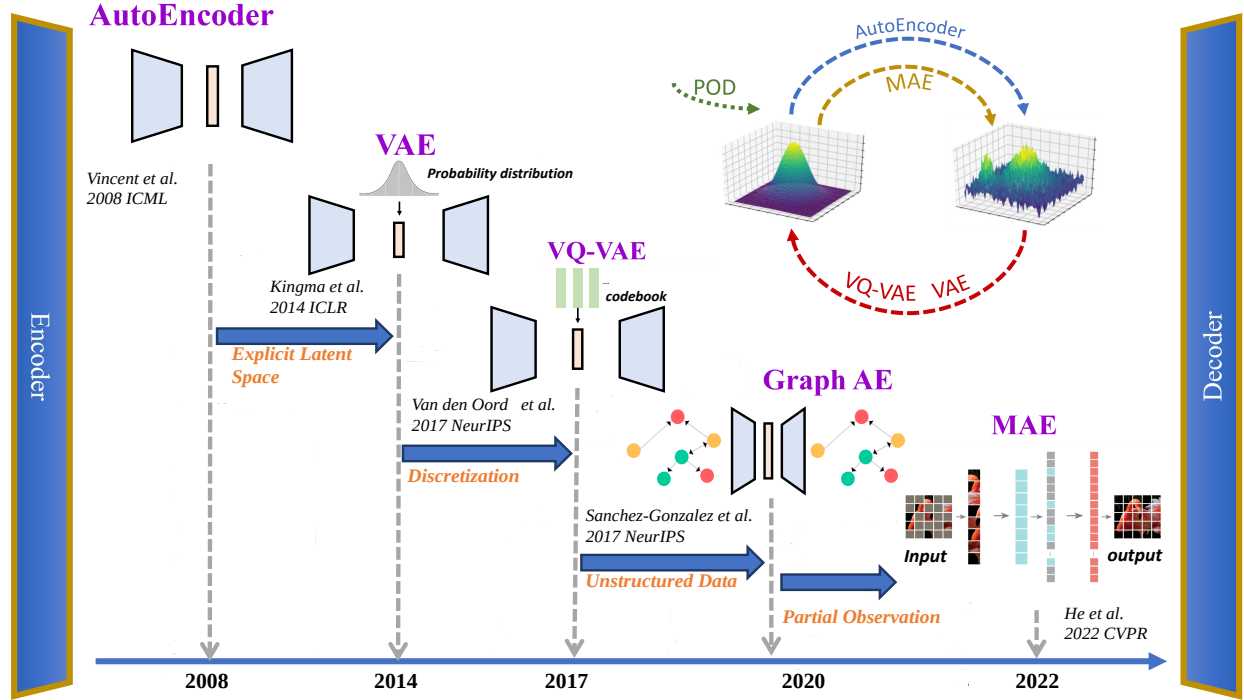


Fig. 3: Progression of ML-based reduced-order-modelling

merical models are inevitably affected by errors. Typical model error correction approaches consist of using a statistical model (typically a neural network) to correct a physical, knowledge-based model. In practice, this implies that we try to build a hybrid physical/statistical model [135], [136], [137], [138], [139], [140], [141], [142], [143]. A physical model is usually defined by a set of differential equations. These equations are discretised and implemented to form the model tendencies. A numerical scheme is then used to integrate the tendencies over a small time interval and several integration steps are composed to iterate from one time to the next. From here, there exist various ways to design the hybrid model, depending on how the statistical correction is introduced [144]. The easiest possibility is to include a single correction per model integration:

$$\mathbf{x}_{t+1} = \mathcal{M}_t(\mathbf{x}_t) + \mathcal{F}_t(\mathbf{x}_t), \quad (11)$$

where $\mathcal{M}_t$ is the resolvent of the physical model from t to $t+1$ and $\mathcal{F}_t$ is the statistical correction, written in an additive form for simplicity (there are of course other possibilities such as multiplicative correction). This is called *resolvent correction* because, in this case, the correction is added to the resolvent. At the other end of the spectrum, the statistical correction can be included in the model tendencies. The *tendencies correction* is potentially more efficient because the errors can be corrected before they manifest (i.e. before they are integrated) and because the statistical correction benefits from the interaction (via the integration scheme) with the physical model. However, a tendency correction is by construction intrusive (even and prominently at the level of physical and statistical models' codes interdependence) and hence more difficult to implement

than a resolvent correction. This approach can be formalised in the general form,

$$\mathbf{x}_{t+1} = \mathcal{F}_t(\mathcal{M}_t(\mathbf{x}_t), \mathbf{x}_t). \quad (12)$$

It enables representation of non-additive error [142] or to increase the model resolution from the original resolution of the physical model to a higher dimension [145]. On the other hand, the correction can no longer be directly related to the analysis increment.

In a DA context, where observations are usually sparse and noisy, the statistical correction can be trained using a series of DA analyses, where the system state has first been estimated from observations [12]. In the case of the resolvent correction, the contribution of the physical and statistical models is independent. Therefore, it is possible to show that the statistical model can predict the analysis increments [143], [146]. This independence is convenient since the analysis increments are usually products of the NWP centres. However, even though analysis increments can be seen as a proxy for model errors, they are usually affected by other sources of errors, e.g., approximations in the observation operator or in the DA method itself [147], [148]. An illustration is provided in Figure 4, where a two-dimensional quasi-geostrophic model is corrected using a neural network. The neural network is trained using the analysis increments over one or two days. The discrepancy between the neural network predictions and the actual model error illustrates the difference between model error and analysis increments.

More generally, this training process can be interpreted as the first step of a coordinate descent [15], where DA steps alternate with ML steps to learn both the system state and

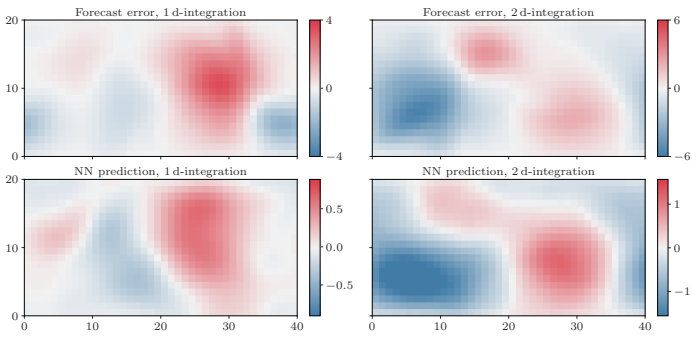


Fig. 4: Correction of a two-dimensional quasi-geostrophic model using a neural network. Top panels: true model error snapshots for a 1-day (left) or 2-day (right) integration. Bottom panels: corresponding neural network predictions, based on the analysis increments. Figure reproduced from [143].

the statistical correction from observations as illustrated in Figure 5. By construction, this is an offline learning strategy, because the ML step does not start until the entire analysis trajectory is available. This means that one can benefit from all the tools developed for deep learning while relying on the DA infrastructure that could be in place and under continuous development over the year (as in, e.g., weather forecast centres). Online learning methods have been recently developed as an alternative [149], [144], [150], [151]. The idea is to use augmented state DA techniques to estimate at the same time the system state and the parameters of the statistical correction. Compared to offline learning, online learning approaches naturally fit a sequential context (where observations become available over time) and allow model error correction once the first observation is available. From a machine learning perspective, this two-step approach can also be put in a fully differentiable framework using auto-differentiable to access the gradient of the DA itself [152].

B. ML and DA for parameter estimations

Parameters in dynamical systems are considered as additional variables, other than state vector, which determine the dynamic characteristics [153]. Online parameter estimation for high-dimensional dynamical systems has been a long-standing challenge [154]. In geoscience, hybrid mechanistic-empirical models are widely used for large-scale problems, for instance, in climate [155] and wildfire [156] forecasting. The parameter estimation of these models often relies on case-by-case tuning [157] or posterior diagnosis/analysis [158] that can be computationally difficult due to the complexity of the predictive model. Much effort has been given in applying DA, especially variational assimilation, for parameter estimation using real-time observations [159], [160]. In particular, an augmented state approach that jointly estimates state and parameters has been used [50]. In augmented state approach, the parameters are updated through cross-correlations with the observed state. For this method to work well, there needs to be a substantial correlation between observed values and parameters as pointed out by [161]. Two main difficulties are associated with the estimation of model parameters [162]: one is the strong nonlinear coupling between the parameter and

model equations and the second is that the parameters are usually between certain value range (for example parameters are positive as a rule). Due to both these properties, a DA method such as the EnKF or variational methods that relies on Gaussian assumptions and uses only the first two statistical moments in the analysis step, needs to be modified in order to be able to deal with probability density functions poorly approximated by the normal distribution. The work of [163], [164], and [165] presented techniques in parameter estimation that have been successfully applied in low-resolution non-chaotic systems. In addition, [166], [167] adopted a new algorithm of [168] to the estimation of cloud microphysical parameters that use higher than second order moments. [169] compares several of the nonlinear DA algorithms for joint state and parameter estimation and shows the benefits of including higher order moments or physical constraints in DA. The work of [170] combines a data-driven simulator for forecasting regional wildfire front position and an EnKF for wind and biomass fuel parameters estimation. In this framework, a surrogate model based on a Polynomial Chaos approximation is iteratively adapted to capture the nonlinearities of the forward model and considerably improves the EnKF efficiency. However, all these sample-based algorithms, require stochastic models for parameters to ensure continuous updates of parameters based on new data [169]. As illustrated in [161], in real world applications, the accuracy of the parameter estimates is quite sensitive to the stochastic model chosen. Thus, significant effort is required to properly tune the stochastic model of the parameters.

On the other hand, since the parameters of numerical models are not observed, ML is not an obvious method of choice for the estimation of the parameters. However, in a hybrid setting with DA, ML has shown to bring several benefits to this problem as well. [171] proposed the use of RNN to replace the standard compartmental model in epidemic modelling for COVID-19. Thanks to its efficiency, this approach can incorporate new data to adjust model parameters via DA in real-time. [172] applied similar ideas to analyse cryptocurrency markets. [173] implemented deep residual neural networks to surrogate the assimilation process thus enhancing model forecasts. The proposed approach managed to handle both parameter and state estimation with sparse and noisy observations [173]. ML can also be used to estimate parameters [80] and their uncertainties as an alternative to the augmented state approach. Although purely offline training produces a good, averaged value of the parameters, we are often in a situation where parameter values might differ due to various reasons like season, or weather situation in NWP systems. In this case, combining ML with DA is quite beneficial, since at the times observations are assimilated, the online improvements to ML model can be made. [80] goes beyond deterministic point predictions and learns probabilistic neural networks: a deep ensemble of point estimate neural network and BNN. After training, these two types of neural networks are incorporated within a DA system, where DA is used for state estimation and ML for parameter estimation. BNNs are additionally trained online during the DA cycle using a realistic number of forecast/analysis ensemble members allowing further improve-

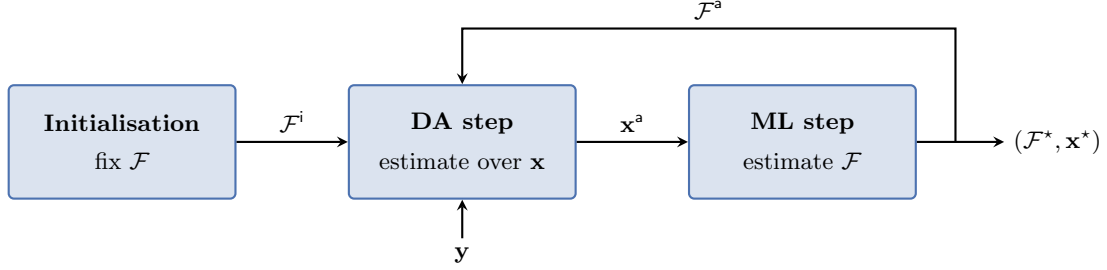


Fig. 5: Schematic illustration of the coordinate descent minimisation. DA steps are used to estimate the system state from sparse and noisy observations, and ML steps are used to estimate the parameters of the statistical correction based on the estimated states. The process can be iterated for increased accuracy in the estimation. Here, $\mathcal{F}^i, \mathcal{F}^a, \mathcal{F}^*$ denote "initial-", "analysis-" and "optimal correction" respectively.

ments to a ML model. By including the parameter estimates obtained from the BNNs in the DAcycle, the results show reduced state errors and increased ensemble spread compared to the case without parameter estimation and with unknown parameters. However, even though the BNNs can accurately estimate the model parameters and their uncertainties, the high computational cost poses an obstacle. Therefore other methods as random forests are currently considered for parameter estimation problems [174].

Building efficient forward surrogate models is an alternative solution to reduce the computational burden of inverse modellings [175], including parameter estimation [176]. The very recent work of [67] proposed the use of Generalised Latent Assimilation (GLA) (see [66] or our Section IV-B) to integrate real-time observations for model parameter adjustment that can yield more accurate future predictions. Learning from a high-fidelity mechanistic model, the authors first constructed a ROM- and ML-based surrogate for predicting wildfire dynamics. The model coefficients related to the fire spread rate are then consistently updated using real-time satellite observations. The same technology has been applied to nuclear reactor physics [68] with spatially-sparse observations. By construction, this approach can incorporate observations with flexible length time windows where pure ML can get into difficulties with unfixed input dimension.

C. Error specification in DA: traditional and ML methods

Typical DA methods are tuned by some statistical parameters, especially those associated with the error terms, i.e. ϵ_t^x the model error, and ϵ_t^y the observation error. Assuming that those two error terms are Gaussians, the corresponding model and observation error covariance matrices are noted $\mathbf{Q}_t$ and $\mathbf{R}_t$ (see Section II-B). They play a crucial role on the quality of the DA reconstruction by controlling the weight given to the forecast and the observations in the DA algorithms. This is illustrated in Figure 2 of [177].

Several methods have been proposed in the DA literature to jointly estimate $\mathbf{Q}_t$ (or alternatively the background covariance $\mathbf{B}_t$) and $\mathbf{R}_t$, they are summarised in [177]. All the methods are based on the innovation difference, noted $\mathbf{d}_t^{o-b}$, between the observation $\mathbf{y}_t$ and the background projected in the observation space $\mathcal{H}_t(\mathbf{x}_t^b)$, i.e.,

$$\mathbf{d}_t^{o-b} = \mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t^b). \quad (13)$$

When considering Gaussian errors ϵ_t^x and ϵ_t^y , the innovation $\mathbf{d}_t^{o-b}$ is also Gaussian, and its covariance matrix is $\mathbf{H}_t \mathbf{B}_t \mathbf{H}_t^\top + \mathbf{R}_t$. In order to jointly estimate $\mathbf{Q}_t$ (or $\mathbf{B}_t$) and $\mathbf{R}_t$, innovation alone is not enough, and the authors proposed to examine other innovation statistics (e.g., observation minus analysis, noted $\mathbf{d}_t^{o-a}$) in the observation space: this is called the Desroziers method [178]. Alternatively, several works among [179], based on Mehra theory [180], had a look at the lag-innovation statistics, the difference between two consecutive innovations. Both Desroziers and Mehra methods use two different innovations to retrieve the two unknown covariances. As an example, the Desroziers innovation statistics should verify the following equations:

$$\mathbf{R}_t = \mathbb{E} \left[\mathbf{d}_t^{o-a} (\mathbf{d}_t^{o-b})^\top \right], \quad (14)$$

$$\mathbf{H}_t \mathbf{B}_t \mathbf{H}_t^\top + \mathbf{R}_t = \mathbb{E} \left[\mathbf{d}_t^{o-b} (\mathbf{d}_t^{o-b})^\top \right]. \quad (15)$$

Many works based on the principle of maximum likelihood approaches also use the innovation to find the most likely covariances $\mathbf{Q}_t$ and $\mathbf{R}_t$ [181]. In DA, those likelihood-based approaches use either the Bayesian or the frequentist framework. In the Bayesian framework, prior distributions are proposed for the shape parameters of the two covariances (typically, noise levels and spatial correlation lengths). Then, two-stage procedures are used to estimate the state of the system and the posterior distribution of those shape parameters [182]. In the frequentist framework, covariance matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$ (or parametric versions of them) are tuned to maximise the total likelihood of the state-space representation. This tuning is often achieved by implementing Expectation-Maximisation (EM) algorithms [183]. The latter consists of a two-stage procedure, where the state of the system is estimated, and then covariance parameters are updated on an iterative basis [184], [185], [186], [187], [188], [189].

Classical error covariance estimation algorithms (e.g., [190], [178]) often rely on posterior analysis, and require iterative applications of DA. This can be computationally difficult for high-dimensional systems. Furthermore, careful attention must be paid when making the initial guess of error covariances, which may crucially impact the algorithm performance [191]. Continuous effort sought to enhance error covariance modelling, in particular, with ML techniques. Covariance Estimation and Learning through Likelihood Optimization (CELLO)

was proposed by Vega-Brown et al. [192] to provide a fast prediction of the observation error covariance $\mathbf{R}_t$ based on a Bayesian non-parametric learning methods. CELLO achieved similar results compared to empirically estimated covariances using manually annotating sensor regimes. The authors have also shown that the learned covariances can provide substantial enhancement to state estimation accuracy during online filtering. A Convolutional Neural Network (CNN)-based approach, named Deep Inference for Covariance (DICE), has been developed to learn the measurement error distribution in a supervised manner [74]. Relying on the Gaussian assumption, KLD was employed as the loss function to train $\mathbf{R}_t$. Thanks to the capacity of CNNs in capturing local spatial patterns, DICE considerably outperformed CELLO on both simulated and real data.

In fact, $\mathbf{R}_t$ is often considered time-invariant in a wide range of DA applications [46]. Therefore, observed values $\mathbf{y}_t$ at different time steps are jointly considered to predict $\mathbf{R}_t$. Following this idea, the very recent work of [78] proposed a RNN-based framework for observation matrix specification. More precisely, synthetically generated observation sequences are considered as RNN inputs while the corresponding $\mathbf{R}_t$ is the output target during the training process. In particular, LSTM was used to build the model because of its strength in dealing with long-term time dependencies. The workflow of offline model training and online prediction is illustrated in Figure 6. As an important advantage, this approach managed to handle both non-parametric and parametric (e.g., with a pre-selected covariance kernel) covariance modellings. The comparison of different error covariance specification approaches is given in Table II, where the computational efficiency refers to low online computational cost and the temporal dependency signifies if the method can use time-varying data to estimate error covariances.

TABLE II: Comparison of covariance specification approaches

Methods	Training free	good interpretability	computational efficiency	temporal dependency
EM [186]	✓	✓	✗	✓
Desroziers [178]	✓	✓	✗	✗
CELLO [192]	✗	✗	✓	✗
DICE [74]	✗	✗	✓	✗
LSTM-based [78]	✗	✗	✓	✓

D. End-to-end learning of DA systems

Instead of using ML techniques to address difficulties in DA algorithms (e.g., model error correction, parameter estimation and error covariance specification), some recent works focused on building end-to-end learning schemes for the whole DA system. End-to-end learning [31] naturally arises as an appealing feature of deep learning schemes to address a given inverse problem from raw input data through the combination of elementary neural blocks. The key property here is the differentiability of the elementary blocks which leads to the differentiability of the end-to-end architecture. As such, one can train the latter at once using a supervised or partially-supervised learning strategy. This end-to-end learning strategy has been at the core of the breakthroughs of deep learning in many application fields, including signal processing and computer vision [31]. Over the last decade, the elementary blocks or layers available to design neural architectures have also greatly expanded from initial dense, convolution, pooling, activation and recurrent layers [31] to more complex blocks, including among others attention blocks [113], multi-scale neural architectures [193], finite-difference and spectral solvers [194], neural optimizer [195], and physics-informed neural networks [196].

This diversity of neural components provides the basis to address DA through an end-to-end learning framework. This may simply consist in training a state-of-the-art neural architecture to map observation data to the targeted state sequence or model parameters [197], [198]. Here, we focus on neural approaches which take a closer look at DA schemes to design DA-inspired neural schemes. Broadly speaking, as sketched in Fig.7, this applies both to sequential DA schemes and non-sequential variational DA ones:

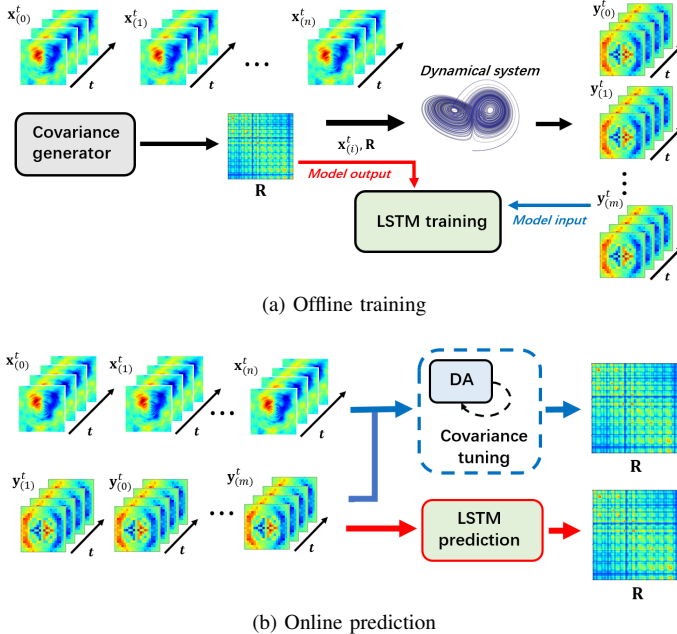


Fig. 6: Workflow of training and applying RNN for observation covariance estimation

- Sequential-DA-inspired neural schemes: in sequential DA, two main operators naturally arise, a forecasting operator to forecast or sample the state at the next time step given the current state and an analysis operator to update the state given new observation data (see Section II-B for details). From a deep learning point of view, both operators naturally relate to RNN. As such, one may explore state-of-the-art RNN such as LSTM and Gated Recurrent Unit (GRU) [31] as in [199], [79]. When introducing a latent representation $\mathbf{z}_t$ (for example, obtained from ROM as shown in Section II-D1) for the physical state $\mathbf{x}_t$ or its PDFs, this leads to a trainable

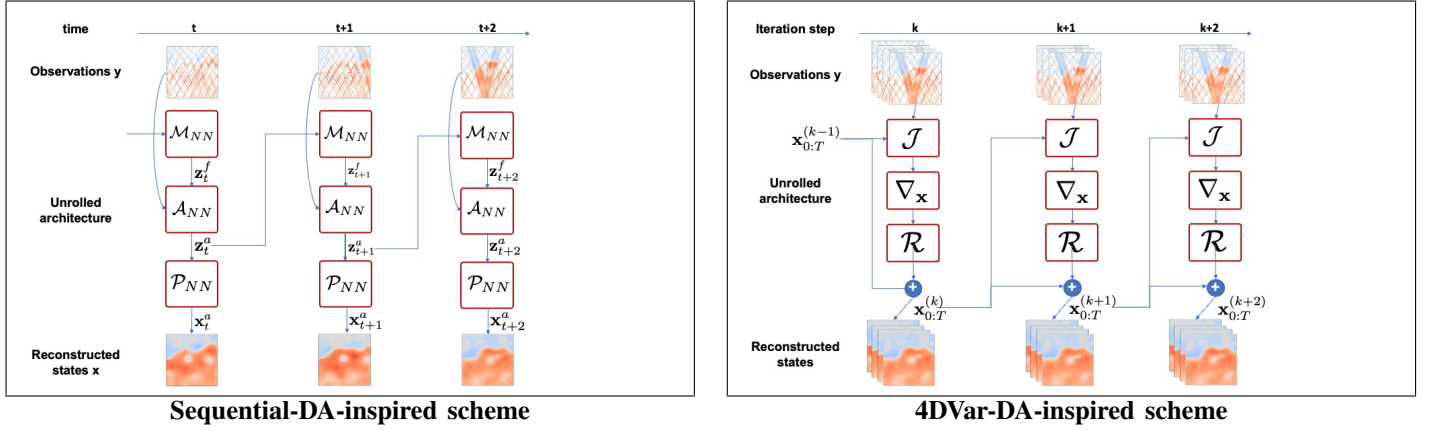


Fig. 7: **Sketch of end-to-end neural architectures for DA:** some architectures mimic the forecasting and analysis steps of sequential DA at each time step (see Eq.16) (left panel) whereas others implement iterative gradient descents for a variational DA criterion (right panel).

recursion of the following form at time step t :

$$\begin{cases} \mathbf{z}_t^f, \mathbf{h}_t &= \mathcal{M}_{NN}(\mathbf{z}_{t-1}^a, \mathbf{h}_{t-1}) \\ \mathbf{z}_t^a &= \mathcal{A}_{NN}(\mathbf{z}_t^f, \mathbf{y}_t) \\ \mathbf{x}_t^f &= \mathcal{P}_{NN}(\mathbf{z}_t^a) \end{cases} \quad (16)$$

where $\mathcal{M}_{NN}$, $\mathcal{A}_{NN}$ and $\mathcal{P}_{NN}$ are neural networks and $\mathbf{h}_t$ is the internal state of recurrent networks, if any. One may also explore physically-constrained parameterisations, typically neural Ordinary Differential Equation (ODE)/PDE schemes for the forecasting operator if the underlying physics [200], [201], [194] are known and/or explicit Kalman recursion rule under additional linear-Gaussian hypothesis for the posterior and the observation operator [202], [79]. Regarding the learning step, these approaches may adapt classic stochastic optimisation algorithms [31] with randomised re-initialisation steps of the internal states of the recurrent blocks [199];

- 4DVar-DA-inspired neural schemes: from a neural perspective, variational DA combines a variational cost (see Section II-B) and a gradient-based optimizer using an adjoint method [48]. Assuming that both the observation and dynamical operators are implemented as neural operators, the automatic differentiation embedded in deep learning schemes makes it convenient to apply a gradient descent with respect to state sequence and/or model parameters, with no need to explicitly code the adjoint operators. Besides, trainable optimizers have also emerged as computationally-efficient solvers for minimisation problems [195], [203]. The combination of these two elements as proposed in 4DVarNet approach [204] arises as an appealing solution to learn unknown terms in the variational formulation jointly to computationally-efficient solvers for the DA problem. It relies on the weak-constrained 4DVar cost function $\mathcal{J}$ over a time window $[0, T]$ with state variables $\mathbf{x}_{0:T}$ and observations $\mathbf{y}_{0:T}$ as defined in Equation 6a in Section II-B. The associated trainable solver exploits the following update rule from some initial condition $\mathbf{x}_{0:T}^{(0)}$:

$$\mathbf{x}_{0:T}^{(k+1)} = \mathbf{x}_{0:T}^{(k)} + \mathcal{R} \left[\nabla_{\mathbf{x}} \mathcal{J} \left(\mathbf{x}_{0:T}^{(k)}, \mathbf{y}_{0:T} \right) \right] \quad (17)$$

where k stands for the iteration index, $\mathcal{R}$ a recurrent neural network, typically a LSTM, and $\nabla_{\mathbf{x}} \mathcal{J}$ for the automatic differentiation of the cost function with respect to state $\mathbf{x}_{0:T}$. We want to reiterate that the definition of the cost function $\mathcal{J}$ depends on the initial background $\mathbf{x}_0^b$ and the transformation operators $\mathcal{H}_{0:T}$ (see Equation 6a). Beyond these physics-informed parameterisations of the variational cost, especially using neural differential schemes, one may also explore non-sequential representations of the dynamics [204].

From a theoretical point of view, the learning stage for these neural schemes relates to bi-level optimisation problems [205], for instance for 4DVar-inspired schemes given by:

$$\begin{aligned} \arg \min_{\mathcal{H}_{0:T}, \mathbf{x}_0^b} \mathcal{L} \left(\{(\mathbf{x}_{0:T}^{\text{true}})_n, (\widehat{\mathbf{x}}_{0:T})_n\}_n \right), \text{ such that,} \\ \forall n, (\widehat{\mathbf{x}}_{0:T})_n = \arg \min_{\mathbf{x}_{0:T}} \mathcal{J}(\mathbf{x}_{0:T}, (\mathbf{y}_{0:T})_n) \end{aligned} \quad (18)$$

$\{(\mathbf{x}_{0:T}^{\text{true}})_n, (\widehat{\mathbf{x}}_{0:T})_n\}_n$ is the training dataset (n is the sample index) and $\mathcal{L}$ the considered training loss. Here $\mathbf{x}_{0:T}^{\text{true}}$ denotes the theoretical value of the states which are supposed to be known during the training process. Optimal interpolation [206] solves such a bi-level formulation for a minimum-variance criterion and a linear-Gaussian state-space. End-to-end DA schemes then open new research avenues to explore optimal DA schemes and reduce estimation biases for nonlinear and/or non Gaussian systems as illustrated for partially-observed nonlinear dynamics by [204]. This also applies to the shift from general-purpose DA pipelines to application-centric ones optimized for specific observing systems, states and/or diagnosis variables. Beyond applications on toy examples, recent demonstrations for the reconstruction of sea surface dynamics from satellite-derived observations [207] support the relevance of these schemes to advance the state-of-the-art for real DA problems. A key challenge is their application to complex spatial-temporal DA problems currently solved by operational DA systems in climate simulation, operational oceanography and weather forecast. In such contexts, we may emphasise the great flexibility in terms of state definition and model parameterisation opened by the end-to-end learning framework,

including for instance augmented state [208], multimodal formulation [207] and uncertainty representation [209].

IV. MACHINE LEARNING ASSISTED BY DATA ASSIMILATION AND UNCERTAINTY QUANTIFICATION

In this section, we discuss how DA and UQ techniques can be used to enhance ML models in dynamical systems regarding both prediction accuracy and interpretability. This consists of uncertainty analysis for ML approaches (Section IV-A), latent DA methods for correcting ML surrogate models (Section IV-B), identification of governing equations using DA (Section IV-C) and forecasting partially observed dynamical systems (Section IV-D).

A. Uncertainty analysis for ML approaches

Different families of UQ methods in ML, especially DL, have been proposed based on Bayesian frameworks, for example, BNN [81] and MCD [94]. The latter consists of ensembles of NN optimization iterates or independently trained NNs (e.g., deep ensembles: DE [72]).

In both BNN and Deep Ensembles (DE) methods, epistemic uncertainty is often estimated by looking at an ensemble of trained models where the sampling approach from the set of possible models varies from method to method. Similar to the process of EnKF (see Section II-B) in DA, the spread of predictions obtained from different models is then used as an estimate of epistemic uncertainty.

In particular, BNN offers a probabilistic interpretation of DL models by inferring distributions over the models' weights. They place a prior distribution over NN weights, which induces a distribution over a parametric set of functions. BNNs thus offer robustness to over-fitting, uncertainty estimates, and can learn from small datasets [81]. The Bayesian framework quickly explained here after plays an important role in the foundation of BNN and some of the proposed UQ methods.

Given a set of paired noisy observations $\mathcal{S}_o = \{\alpha_i, \beta_i\}_{i=1}^N$ and a set $\mathcal{A}$ of user assumptions and preferences (e.g., NN architecture or likelihood function), the goal is to construct a conditional distribution $p(\tilde{\mathbf{x}}|\mathbf{x}, \mathcal{S}_o, \mathcal{A})$ of the quantity of interest $\tilde{\mathbf{x}}$ given an input vector $\mathbf{x}$. For each input, it is assumed that the response contains both a deterministic as well as some additive aleatoric noise. We aim at identifying the parameters ω of the NN mapping function $\mathcal{M}_\omega$ that fits the function inputs $\mathbf{x}$ and outputs $\tilde{\mathbf{x}}$ with maximum likelihood,

$$\tilde{\mathbf{x}} = \mathcal{M}_\omega(\mathbf{x}) + \epsilon(\mathbf{x}), \quad (19)$$

where $\epsilon(\mathbf{x})$ represents the sum of the aleatoric noises. In the case of a dynamical system, Equation 19 can be written as

$$\mathbf{x}_{t+1} = \mathcal{M}_t(\mathbf{x}_t) + \epsilon_t^x, \quad (20)$$

following the notation of Section II-B. We then assume a likelihood function $p(\tilde{\mathbf{x}}|\mathbf{x}, \omega)$ with parameters to be inferred from the available data. In order to do so, we construct a model function $\tilde{\mathbf{x}}_\omega(\mathbf{x})$ of the parameters that captures the deterministic part of the response and assume a model the

distribution for the noise (e.g. multivariate factorized Gaussian likelihood function [210]). To obtain the posterior distribution for any new input $\mathbf{x}$, we must marginalise over the model parameters,

$$p(\tilde{\mathbf{x}}|\mathbf{x}, \mathcal{S}_o) = \mathbb{E}_{\omega|\mathcal{S}_o} [p(\tilde{\mathbf{x}}|\mathbf{x}, \omega)], \quad (21)$$

where $p(\omega|\mathcal{S}_o)$ is obtained from Bayes' formula and therefore requires the likelihood of the data to be evaluated. Obtaining the posterior exactly is computationally and analytically intractable. Indeed, characterizing uncertainty over NN parameters is challenging due to the high-dimensionality and potential complex dependencies of the weights. Moreover the influence of the prior distribution is difficult to be understood. To address this obstacle, approximate inference methods [211], [8] aim to approximate the posterior by another distribution and/or obtaining samples from the posterior. M is denoted as the sampling size. All methods obtain a set of parameters samples $\{\tilde{\omega}_j\}_{j=1}^M$ that maybe used to approximate the integration via Monte Carlo (MC) estimation as,

$$p(\tilde{\mathbf{x}}|\mathbf{x}, \mathcal{S}_o) \approx \sum_{j=1}^M \tilde{\mathbf{x}}_{\tilde{\omega}_j}(\mathbf{x})/M, \quad (22)$$

which provides an approximate distribution of the total predictive uncertainty. Under Gaussian assumption of the model likelihood, there exists simple close forms for the mean and standard deviation of the posterior prediction. In this case, the approximate total uncertainty standard deviation is a combination of the aleatoric and the epistemic parts of the total uncertainty, respectively.

However, in practice $p(\omega|\mathcal{S}_o)$ may be approximated by variational parameters, i.e., by a parametrized function $q_\theta(\omega)$. The aim is to approximate a distribution that is close to the posterior distribution obtained by the model. As such, the KLD between those two distributions (i.e., $KL(q_\theta(\omega)||p(\omega|\mathcal{S}_o))$) may be minimized with regard to θ . The KLD minimisation is also equivalent to maximising the evidence lower bound (ELBO) [212] with respect to the variational parameters. This procedure is also known as the variational inference (VI) [32]. VI is a technique which replaces the Bayesian modelling marginalisation with optimisation (i.e., replace the calculation of integrals with that of derivatives) which can considerably reduce the computational cost.

The posterior inference may also be approximated by various Monte Carlo (MC) based methods such as Markov-chain Monte Carlo (MCMC) techniques, drawing NN parameters samples from the posterior by using a Markov chain with the distribution of the parameters given the data as its invariant(i.e., stationary distribution). In particular, Hamiltonian Monte Carlo (HMC) is a reference sampling algorithm but is extremely computationally demanding [213], [214]. Stochastic gradient MCMC approaches based on Langevin and Hamiltonian dynamics have been proposed to alleviate the computational burden of MCMC algorithms thanks to stochastic approximation to the gradients [215], [216], [217]. It is important to emphasize that these approaches do not rely on any prior assumptions about the form of the posterior distribution.

For Deep Ensembles (DE) [72], the concept is very simple and one needs only to retrain the same network many times with different weights initialisations. The inherent randomness then provides different samples of the trained network parameters, meaning identifying multiple minimums of the NN parameter loss landscape (i.e., different Maximum a Posteriori (MAP) estimates). If one optimizes the networks with a MSE loss, this provides only a measure of epistemic uncertainty. If one optimizes the data log likelihood, it estimate both aleatoric and epistemic uncertainties. Variants such as – Snapshots Ensembles (SEn) obtain the set of multiple minimums without incurring any additional cost as compared to standard training, thanks to letting the algorithms to converge to M different local optimums during a single optimization trajectory; or – Stochastic Weight averaging-Gaussian (SWAG) extends SEn by also fitting a Gaussian distribution to the aforementioned local optimums [218].

Recently, much attention has been given in addressing the ML explainability using UQ techniques [219], [220], [221]. The latter can contribute directly to the counterfactual explanations [222], for instance, under which condition the decision has been made and with what degree of freedom [219]. Furthermore, UQ can provide information about model noises, which is crucial for DA algorithms when being applied to dynamical systems (see Section III-C for details).

B. ML and DA with ROM

In this section we present the algorithms and applications which combine DA and ROM (especially ML-based ones) to get benefits from both technologies.

On the one hand, due to the high-dimensionality and the complexity of the transformation function, implementing DA for high-dimensional systems can be computationally challenging. Classical solutions consist of dimensionality reduction via projection-based ROM, such as POD (see Section II-D1). DA is then performed in the reduced space. The optimal choice of the reduced space dimension has also been extensively investigated [223], [224], [225], [226]. Many recent research efforts sought to address this challenge of efficiency by performing DA with ML-based AEs (see Section II-D1). Such algorithms, known as Latent Assimilation (LA), can benefit from the efficiency of ML and the accuracy of DA. More precisely, [227] proposes to learn assimilated results using RNN in a reduced space to enhance future predictions. Similar ideas can be found in [12] which introduces an iterative DA-ML scheme is introduced. However, when applying this algorithm, retraining of the neural networks is required when new observations become available. In past two years, online LA has raised significant research attention. For sparse and unstructured data, domain decomposition techniques [228] can also be used to reduce the problem dimension, for example, via community detection through a connection graph [229], [230].

On the other hand, as discussed in Section II-D1 and II-D2, despite its great efficiency, ML surrogate models can introduce prediction errors in a cumulative manner because of the iterative forecasts [231]. DA with ROM can address this problem

by updating the surrogate prediction consistently using real-time observations collected from local sensors or satellites. As shown in Figure 8, this is an online iterative process that can be used to update the starting point of the next time-level forecast in the latent space, thus improving the accuracy of long-term predictions. Considerable research efforts also sought to adjust the latent error covariance which crucially impacts the assimilation performance (see [177] and Section III-C). Ensemble LA was introduced in [76] and [232] to estimate the background matrix, while [73] employed posterior covariance tuning in the latent space. The proposed online LA methods are mainly split into two groups:

- LA° [233], [234] where observations in the full physical space are used to correct/adjust the reduced-order models;
- LA° [75], [235], [73], [232] where the state variables and the observations are compressed into a same latent space.

The latter can perform more efficient assimilation especially for dense observation mappings, as demonstrated in wildfire forecast [73], air pollution estimation [75] and fluid mechanics [232]. However, encoding state and observation variables into a same reduced space is challenging, especially with highly nonlinear state-observation transformation mappings, which exists in a majority of real-world DA problems. Therefore, separate AE are often required for the states and the observations, leading to heterogeneous latent spaces.

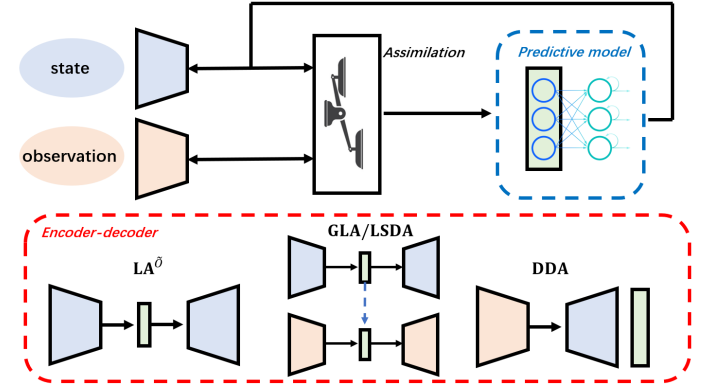


Fig. 8: Workflow of online LA with surrogate modelling and different encoding-decoding strategies

TABLE III: Comparison of Latent Assimilation approaches

Methods	Reduced state	Reduced observation	nonlinear mapping	Non-explicit mapping
RODDA [227]	✓	✗	✗	✗
LA [233], [234]	✓	✗	✗	✗
LA+ [75], [73], [232]	✓	✓	✗	✗
GLA [66]	✓	✓	✓	✗
LSDA [236]	✓	✓	✓	✗
DDA [237], [238]	✗	✗	✓	✓

To tackle this bottleneck, GLA and Latent Space data assimilation (LSDA) were proposed in the very recent works of [66] and [236] which make use of local surrogate functions (i.e., polynomial functions [66] and Multi layer perceptron (MLP) [236]) to connect multiple latent spaces. Afterward, variational DA can then be performed by solving a local

optimisation problem using smooth surrogate functions as shown in Figure 8. However, the computation of the local surrogate functions around the predicted latent variables must be performed online, resulting in relatively high computational cost. More importantly, considerable uncertainties can be introduced when mapping the two latent spaces, especially when the choice of the approximation range is inappropriate. Recent research of [237] addresses the difficulty of complex state-observation mapping by proposing a new DA scheme, named Deep data assimilation (DDA), which trained jointly an observation-domain encoder and a state-domain decoder. A similar idea can be found in the work of [238]. Applying such method, the observation data can be directly transferred to the state space. The characteristics of different LA approaches are summarised in Table III.

C. ML for dynamical systems assisted by DA

Except for ML surrogate models that learn directly from the state variables (see Section II-D2 and IV-B), there are several examples of data-driven models derived from observations that show forecasting abilities [239], [240]. Those models can take various forms (e.g. neural networks) but all assume the observations to be perfect and complete, i.e. very little noise and a spatio-temporal complete coverage of the processes of interest. Nevertheless, these conditions are almost never met in reality and dynamical systems of natural processes are usually observed noisily and sparsely. DA techniques, on the other hand, can provide a direct methodological formulation that supports the inference of dynamical systems. This formulation can be obtained from a collection of observations that can be irregular both in space and time, noisy, and may also miss some of the degrees of freedom that constitute the underlying dynamics. Following the notation defined in Section II-B, let us consider the following state space model:

$$\begin{cases} \mathbf{x}_t = \mathcal{M}_t(\mathbf{x}_{t-1}, \epsilon_{t-1}^x) \\ \mathbf{y}_t = \mathcal{H}_t(\mathbf{x}_t, \epsilon_t^y) \end{cases} \quad (23)$$

The prior errors ϵ_t^x and ϵ_t^y are considered as random processes accounting for the uncertainties in the dynamical and observation models.

In an identification scenario, neither the dynamical model $\mathcal{M}_t$ nor the state variable $\mathbf{x}_t$ is known. Instead, we are only provided with observations $\mathbf{y}_t$ that are related in some way to the hidden states through the observation operator $\mathcal{H}_t$. From this point of view, state-of-the-art identification approaches can be discussed based on the nature of the elements of the state space model (Equation (23)).

1) Noise-free, direct measurements of the state variables:

When provided with direct measurements of $\mathbf{x}_t$ and assuming that the model and observation noise ϵ_t^x and ϵ_t^y are zero, the problem may be regarded as the identification of the most appropriate basis function that explains the temporal variability of the observations.

One may distinguish data-driven approaches into two prominent families. The first (traditional) category involves an expansion of $\mathcal{M}_t$ as a combination of nonlinear basis, where polynomial representations are typical examples [241]. The

combination of such representations with sparse optimisation techniques, as shown in SINDy framework (see Section II-D2), recently opened new research avenues in the context of deriving interpretable dynamical models (see [63], [242] and Section II-D2). The SINDy methodology has the advantage of interpretability and fewer parameters compared to other ML models, which significantly reduces the chances of overfitting. These models were successfully applied to a variety of canonical problems in fluid dynamics [63], [243], electrohydrodynamic [244] and magnetohydrodynamics [245]. The main drawback of these approaches remains is that they rely on estimates of the derivatives of the time series. This reliance hampers the direct application to real problems where data can be noisy and irregular. DA allows to bypass some of the issues in SINDy by providing estimates of state space variables. The work of [64] applied the bootstrap technique in an ensemble-SINDY modelling to address the challenge of noises and uncertainties in observation data.

A second category adopts a ML point of view and states the identification issue as a regression problem between consecutive observations. Beyond non-parametric regression models based on analog forecasting [246], recent state-of-the-art research investigates several methodologies based on different ML tools. For instance, reservoir computing approaches [247], [248] were shown to be well-suited for learning dynamical systems from data [249], [250]. Furthermore, the link between residual neural networks (ResNets) and ODE motivated a large body of work in deriving differential equations that are parameterised by neural networks [251], [252], [208]. These new techniques show great flexibility and can be applied to a variety of problems. They can also build from the extensive advances in neural networks and deep learning to tackle challenging issues such as discontinuities in the observations [253]. These methods, however, may suffer from generalisation issues, which motivate the use of various regularisation techniques, based on prior knowledge of the dynamics, to promote generalisation and interpretability of these data-driven models [196], [254], [255], [256], [257], [258], [259], [260], [261], [262].

2) *Noisy observations of the state variables:* When the observation operator $\mathcal{H}_t$ relates to all the states $\mathbf{x}_t$ of the system through an irregular space-time sampling and the noise processes ϵ_t^x and ϵ_t^y are not zero, the derivation of governing equations typically passes through an inversion step. This inversion means that one should estimate the state variables $\mathbf{x}_t$ from the observations in order to perform the identification. To address this challenge we proposed to leverage on DA in a similar fashion to what is described in Section III-A for the estimate of the model error and the construction of a hybrid physics+ML model. Here however, we assume that the model $\mathcal{M}_t$ (see, for example, Equation (6a)) is fully unknown. The lack of a physical model, renders the first cycle of DA/ML very critical: as no model is known at the first cycle, the analysis provided by the DA can be very far from the underlying “truth” and many optimisation cycles may be required, or in the worst case the procedure could fail to converge. To mitigate this, in some cases, a purely data-driven interpolation (e.g., Kriging) can be performed in place of the first DA step to produce

the first estimate of $\mathbf{x}_{0:T}$. Another option is to emulate the dynamical system $\mathcal{M}_t$ using analog forecasting methods, and plug it into an ensemble DA technique [263], [264]. In absence of any original physical-based model, the data-driven model can only reconstruct dynamics on variables of the system that are observed, even though the problem can be circumvented by using Takens's delay embedding theorem [265], [49] as it is detailed in Section IV-D. If an original model is available, it has been shown that it was possible to correct the dynamics, including non-observed variables [142]. At the end of the DA-ML cycles, the analysis can be used as an initial condition and the data-driven model as a forecast model.

The above approach has shown to improve the forecast skill of small-dimensional dynamical systems [12] and to be equivalent to an expectation maximisation method [15]. In [12], it is also shown that the data-driven model trained on noisy and sparse data has a skill of the same order of magnitude as a data-driven model trained on complete and noiseless observations. In the cases where the dynamical system is partially known, fewer DA/ML cycles are necessary [143], [140]. Sensitivity studies using UQ (see Section IV-A) showed that this approach was not very sensitive to the density of observations, up to a certain point. This is likely to be case specific to vary depending on the application. It has also been shown that the approach is correcting the effect of noisy observation, but the final result is still very sensitive to the noise in the data, as confirmed by other studies [266]. Implicitly, all studies on a data-driven models trained on reanalysis [267], [268] are doing one cycle of this method: first DA to produce a reanalysis and then ML to train a data-driven model. One obvious limitation of this approach is the computing cost. Several iterations with successive application of the DA and the training of the data-driven model are necessary. Therefore, the question of finding the compromise between the improvement of the model and the corresponding cost is crucial. Another open question about the use of data-driven or hybrid models in forecast experiments is whether the improvement brought in forecast skill can help understand the deficiency of the current physical-based model. In that perspective, explainable Artificial Intelligence (AI) tools [269], [270], [271] are needed to build operational and trust-worthy systems.

D. ML with DA for partially observed dynamical systems

In practice, high-dimensional dynamical systems are often only partially observable [272], [273]. Let us consider the same problem as Section IV-C with observations only related to a subset of the state vector $\mathbf{x}_t$. Therefore, the derivation of meaningful (Markovian) governing equations in the observation space is (as long as the governing equation of $\mathbf{x}_t$ cannot be decoupled) not possible. This issue was discussed in depth by [274] in the context of closure modelling of a known mechanistic or empirical model. In their work, [274] constructed a mathematical framework that unifies many of the common approaches for blending mechanistic and ML models. The authors studied both discrete and continuous time models and discussed ML based closure models that

can be both memoryless (Markovian) and memory-dependent. Their representations were also combined with DA methods to mitigate noise.

When there is no prior knowledge about the dynamics, a popular path is based on the phase-space reconstruction methodology. In this framework, we seek at projecting the observations into a higher dimensional space that forms an embedding of the hidden state space variables $\mathbf{x}_t$. The temporal evolution of the variables of the embedding is then deterministic and can, in theory, be used to define a model. The most employed embedding methodology in signal processing is the celebrated Takens delay embedding theorem [265]. It shows that by considering delayed observations, one can unfold a phase space that can be topologically similar to the one of the unseen state variables. Several identification techniques have been used on such representations, including polynomial representations [275], recurrent neural networks [276], support vector regression [277], non-parametric models [278] and reservoir computing [279]. Delay embedding representations were also combined with DA frameworks in [49] to infer dynamical models from noisy and partial observations.

Interestingly, the idea of using delay embeddings of the observations can also be found at the heart of recent advances in the inference of latent spaces in state space models based on deep learning architectures [280], [281], [252], [282]. In such methodologies, latent variables are inferred from a posterior distribution given a sequence of observations. This posterior distribution is parameterised by a neural network and optimized using the evidence lower bound. These frameworks have the advantage of bypassing classical assumptions used in traditional DA algorithms such as the Gaussianity of the noise, but suffer, similarly to all models defined based on delay representations, from the problem of correctly parameterising the embedding parameters.

The parameterisation of a delay embedding [283], [284] is a complex task (especially when given high dimensional observations) and the model-making is highly sensitive to this parameterisation. To address these limitations, the Neural Embedding of Dynamical Systems framework (NbedDyn) [285], [286] proposed to solve the embedding problem jointly with the optimisation of a dynamical model. Specifically, NbedDyn defines a new latent state $\mathbf{z}_t$ as follows:

$$\mathbf{z}_t^T = [\mathcal{R}(\mathbf{x}_t)^T, \mathbf{y}_t^T] \quad (24)$$

with $\mathbf{y}_t$ the unobserved component of latent state $\mathbf{z}_t$ and $\mathcal{R}$ an invertible operator used to reduce the dimensionality of the observations. The augmented latent space evolves in time according to the following state space model:

$$\begin{cases} \mathbf{z}_t = \mathcal{M}_{\theta,t}(\mathbf{z}_{t-1}, \epsilon_{t-1}^z) \\ \mathbf{y}_t = \mathcal{R}^{-1}(\mathbf{G}\mathbf{z}_t) + \epsilon_t^y \end{cases} \quad (25)$$

where $\mathcal{M}_{\theta,t}$ is the approximate dynamical operator of $\mathcal{M}_t$ and $\mathbf{G}$ is a projection matrix that satisfies $\mathcal{R}(\mathbf{x}_t) = \mathbf{G}\mathbf{z}_t$. The optimisation of the parameters of the model $\mathcal{M}_{\theta,t}$, as well as the reconstruction of $\mathbf{y}_t$, are carried out jointly using DA. Several optimisation strategies can be defined, depending on the form of the dynamical and observation model as well

as the uncertainties ϵ_t^z and ϵ_t^y . For instance, when considering noise-free observations, a 4D-var formulation was used in [285], [286] to derive nonlinear dynamical models from partial observations of the state space. In related works, a KF-based identification was proposed for linear dynamical and observation models with Gaussian uncertainties [287], [288].

In practice, when using phase space reconstruction techniques, one should not forget about the assumptions that this theory is built on. For any embedding to work, we are assuming that the dynamical model in Equation (23) exists and can be represented by an ordinary differential equation [289]. For several realistic applications, this ODE may not exist or can have an extremely large dimension. In geosciences, for instance, the dimension of a state space variable can reach $O(10^9)$ [7]. In these situations, reconstructing such high-dimensional phase space becomes significantly more challenging. Reducing the dimension of the problem as demonstrated in [285], [286] can help making this problem tractable but may lead to closure issues. In practice, the model returned by any embedding technique can be complemented by an appropriate closure. The form of this closure term can be deterministic using, for example, the framework of [274] or stochastic through an appropriate calibration of a noise forcing. Dynamical system identification approaches introduced in Section IV-C and IV-D are summarised in Table IV, where the partial observation stands for partial observations of the state space and the computational efficiency refers to low online computational cost.

V. OTHER APPROACHES, CHALLENGES & PERSPECTIVES

In this section, we briefly introduced some other approaches and future works combining ML with DA/UQ that have not been discussed in detail.

a) Forecasting Multi-scale dynamical systems: The question of defining multi-scale representations of dynamical systems is an established practice in theory-guided modelling and empirical representations of dynamical systems. In chemical dynamics for instance, several types of reactions are described by stiff differential equations [290]. In finances, stochastic representations are a common tool for representing the impact of unresolved parameters on large-scale quantities such as stock prices [291]. In geosciences, The continuity of phenomena across spatiotemporal scales motivated a large body of work for the derivation of multi-scale models that can be both deterministic [292] and stochastic [293], [294], [295]. From a ML of view, defining multi-scale dynamical systems from data has been investigated mainly in prototypical scenarios. The work of [296] developed sampling strategies for the definition of multi-scale models using state space observations and the definition of closure models for approximating the impact of small-scale variables in the resolution of PDE models is by today standards a common practice [297], [298], [299], [300]. Rethinking such ML solutions in terms of real observations may require considering carefully designed, DA schemes, in order to deal with noise and irregularities in the observations.

b) Mode-switching dynamics: Beyond multi-scale variability, real-world dynamical systems are constantly prone to

switching between different dynamical modes. Making data-driven representations aware of such issues may help us understand and predict these tipping phenomena in complex dynamical systems from data. From this viewpoint, recent state-of-the-art works started investigating the possibility of finding canonical bifurcations of dynamical systems from toy examples [301]. Generalising such approaches to real data is a challenging question that requires finding these critical transitions and accounting for their dynamics and aftereffects in real-time with DA.

c) Learning state-observation mapping in data assimilation: In operational DA, the transformation operator $\mathcal{H}_t$ which maps the state variables to the observations y_t can be complex and highly nonlinear [302], [303], leading to difficulties in minimising the cost function (see Equation (6a)). Furthermore, as pointed by [304], [237] and our Section IV-D, the observations are often incomplete and only relate to a subset of state variables. Recently, much effort [237] has been given to compute machine learned transformation operators that can decrease the computational burden, and address the missing information in the observation field. [305] applied fully connected neural networks to surrogate the mapping from brightness temperature to microwave radiometer observations. The learned mapping function was then used as the transformation operator in an EnKF for DA. Similar ideas can be found in [306], [238], [66], [236]. As mentioned in Section IV-B, [66], [236] compute the surrogate operator in some reduced latent space can further enhance the computational efficiency. On the other hand, [307] aimed to learn directly the inverse (i.e., observation-to-state) transformation operator to speed-up the convergence of DA algorithms. However, since the inverse mapping in DA is often not well-defined, observations are still required during the assimilation process. The idea of learning state-observation mapping is naturally related to some cutting-edge ML concepts, such as transfer learning [308] and domain adaption [309]. It opens promising avenues in solving multi-domain and multi-physics problems.

VI. CONCLUSION

The combination of ML with DA and UQ techniques advanced the state-of-the-art of data-driven modelling in various fields and applications. In this overview, we presented an (as much as possible to our knowledge) exhaustive description and discussion of state-of-the-art approaches that involve DA (or UQ) and ML. In particular, we insist on that these hybrid models provide strengths in interpretability and noise reduction. The development trends and future challenges of this fast-growing field are also investigated. Significant space for further breakthrough advances still exists, especially in applying these approaches in operational contexts. From a methodological perspective, future research efforts could concentrate on, in particular, the integration of ML and DA in high dimensional, multimodal and multi-scale systems, such as NWP and ocean dynamics [310]. We hope that this review paper will benefit scientists working in this vibrant area by providing guidance on the use of DA and UQ in ML and vice versa, as well as to prompt further developments.

TABLE IV: Comparison of governing equations identification approaches

Methods	good interpretability	noisy & irregular observation	partial observation	computational efficiency
polynomial [241]	✓	✗	✗	✓
SINDY [63], [242]	✓	✗	✗	✓
ResNet-based [251], [252], [208]	✗	✗	✗	✓
DA-ML cycle [12]	✗	✓	✗	✗
delay embedding [280], [281], [252], [282]	✗	✗	✓	✗

ACKNOWLEDGEMENT

SC acknowledges the support of Leverhulme Centre for Wildfires, Environment and Society through the Leverhulme Trust, grant no. RC-2018-023. SC, CQ and RA acknowledge the support of the PREMIERE project (grant no. EP/T000414/1). CQ and RA acknowledge the EPSRC grant EP/T003189/1 Health assessment across biological length scales for personal pollution exposure and its mitigation (IN-HALE), and the EPSRC grant EP/V040235/1 New Generation Modelling Suite for the Survivability of Wave Energy Convertors in Marine Environments (WavE-Suite). DX would like to acknowledge the support of EPSRC grant: PURIFY (EP/V000756/1) and the Fundamental Research Funds for the Central Universities. MB, JB, AC and AF acknowledge the support of the project SASIP (grant no. 353) funded by Schmidt Futures – a philanthropic initiative that seeks to improve societal outcomes through the development of emerging science and technologies. TJ would like to thank DFG for her Heisenberg Programm Award (JA 1077/4-1). WD would like to acknowledge the National Natural Science Foundation of China under Grant 61976120 and the Natural Science Key Foundation of Jiangsu Education Department under Grant 21KJA510004.

ACRONYMS

NN Neural Network
UQ Uncertainty quantification
ML Machine Learning
LA Latent Assimilation
DA Data Assimilation
AE Autoencoder
VAE Variational Autoencoder
CAE Convolutional Autoencoder
VAE Variational Autoencoder
GNN Graph Neural Network
BNN Bayesian Neural Network
BLUE Best Linear Unbiased Estimator
RNN Recurrent Neural Network
CNN Convolutional Neural Network
LSTM Long Short-Term Memory
POD Proper Orthogonal Decomposition
PCA Principal Component Analysis
MAE Masked Autoencoders
NLP Natural Language Processing
PGD Proper Generalized Decomposition
ROM Reduced-Order Modelling
DMD Dynamic Mode Decomposition
CFD Computational Fluid Dynamics
PDF Probability Density Function
NWP Numerical Weather Prediction
MSE Mean Square Error
MAE Mean Absolute Error
ODE Ordinary Differential Equation
PDE Partial Differential Equation
AI Artificial Intelligence

DL Deep Learning
KNN K-Nearest Neighbours
RF Random Forest
KF Kalman Filter
EKF Extended Kalman Filter
GLA Generalised Latent Assimilation
3Dvar Three-dimensional variational data assimilation
4Dvar Four-dimensional variational data assimilation
MLP Multi layer perceptron
DDA Deep data assimilation
LSDA Latent Space data assimilation
CELLO Covariance Estimation and Learning through Likelihood Optimization
DICE Deep Inference for Covariance
KLD Kullback–Leibler Divergence
VarDA Variational data assimilation
EnKF Ensemble Kalman Filter
KF Kalman Filter
GRU Gated Recurrent Unit
SINDy Sparse Identification of Nonlinear Dynamics from Data
MCD Monte-Carlo Dropout
DE Deep Ensembles
BNN Bayesian Neural Network
MAP Maximum a Posteriori

REFERENCES

- [1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep learning for computer vision: A brief review,” *Computational intelligence and neuroscience*, vol. 2018, 2018.
- [2] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent trends in deep learning based natural language processing,” *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.
- [3] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, “Physics-informed machine learning,” *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.
- [4] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge *et al.*, “Skillful precipitation nowcasting using deep generative models of radar,” *Nature*, vol. 597, no. 7878, pp. 672–677, 2021.
- [5] J. S. Drams, “70 years of machine learning in geoscience in review,” *Advances in geophysics*, vol. 61, pp. 1–55, 2020.
- [6] S. L. Brunton and J. N. Kutz, *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.
- [7] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, “Data assimilation in the geosciences: An overview of methods, issues, and perspectives,” *Wiley Interdisciplinary Reviews: Climate Change*, vol. 9, no. 5, p. e535, 2018.
- [8] M. Abdar, F. Pourpanah, S. Hussain, D. Rezadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” *Information Fusion*, vol. 76, pp. 243–297, 2021.
- [9] R. Smith, *Uncertainty quantification*. SIAM, 2014.
- [10] T. Sullivan, *Introduction to uncertainty quantification*. Springer, 2015.
- [11] A. J. Geer, “Learning earth system models from observations: machine learning or data assimilation?” *ECMWF, Tech. Rep.* 863, 05 2020.
- [12] 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,” *Journal of Computational Science*, vol. 44, p. 101171, 2020.

- [13] R. Arcucci, J. Zhu, S. Hu, and Y.-K. Guo, "Deep data assimilation: integrating deep learning with data assimilation," *Applied Sciences*, vol. 11, no. 3, p. 1114, 2021.
- [14] C. Buizza, C. Q. Casas, P. Nadler, J. Mack, S. Marrone, Z. Titus, C. Le Cornec, E. Heylen, T. Dur, L. B. Ruiz *et al.*, "Data learning: Integrating data assimilation and machine learning," *Journal of Computational Science*, vol. 58, p. 101525, 2022.
- [15] M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino, "Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization," *Foundations of Data Science*, vol. 2, no. 1, p. 55, 2020.
- [16] M. Bonavita, S. Massart, P. Laloyaux, and M. Chrust, "Data assimilation and machine learning science at ecmwf," 2014.
- [17] M. Raissi, P. Perdikaris, and G. Karniadakis, "Inferring solutions of differential equations using noisy multi-fidelity data," *Journal of Computational Physics*, vol. 335, 07 2016.
- [18] P. Perdikaris, M. Raissi, A. Damianou, N. Lawrence, and G. Karniadakis, "Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, vol. 473, p. 20160751, 02 2017.
- [19] G. Berkooz, P. Holmes, and J. L. Lumley, "The proper orthogonal decomposition in the analysis of turbulent flows," *Annual review of fluid mechanics*, vol. 25, no. 1, pp. 539–575, 1993.
- [20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [21] H. Lin, J. Jin, and H. Herik, "Air quality forecast through integrated data assimilation and machine learning," in *International Conference on Agents and Artificial Intelligence*, 02 2019, pp. 787–793.
- [22] C. Quilodr n-Casas, R. Arcucci, L. Mottet, Y. Guo, and C. Pain, "Adversarial autoencoders and adversarial lstm for improved forecasts of urban air pollution simulations," *arXiv preprint arXiv:2104.06297*, 2021.
- [23] M. Williams, C. Rowley, I. Mezic, and I. Kevrekidis, "Data fusion via intrinsic dynamic variables: An application of data-driven koopman spectral analysis," *EPL (Europhysics Letters)*, vol. 109, 11 2014.
- [24] C. Quilodr n Casas, "Fast ocean data assimilation and forecasting using a neural-network reduced-space regional ocean model of the north Brazil current," Ph.D. dissertation, Imperial College London, 2018.
- [25] I. Mezi  and T. Runolfsson, "Uncertainty propagation in dynamical systems," *Automatica*, vol. 44, no. 12, pp. 3003–3013, 2008.
- [26] R. Ghanem and J. Red-Horse, "Propagation of probabilistic uncertainty in complex physical systems using a stochastic finite element approach," *Physica D: Nonlinear Phenomena*, vol. 133, no. 1-4, pp. 137–144, 1999.
- [27] D. Lucor, C.-H. Su, and G. E. Karniadakis, "Generalized polynomial chaos and random oscillators," *International Journal for Numerical Methods in Engineering*, vol. 60, no. 3, pp. 571–596, 2004.
- [28] J. Li and D. Xiu, "A generalized polynomial chaos based ensemble kalman filter with high accuracy," *Journal of computational physics*, vol. 228, no. 15, pp. 5454–5469, 2009.
- [29] A. Borovikov, M. Rienecker, C. L. Keppenne, and G. C. Johnson, "Multivariate error covariance estimates by monte carlo simulation for assimilation studies in the pacific ocean," *Monthly weather review*, vol. 133, no. 8, pp. 2310–2334, 2005.
- [30] N. Boussez, D. Henze, A. Perkins, K. Bowman, M. Lee, J. Liu, F. Deng, and D. Jones, "Improved analysis-error covariance matrix for high-dimensional variational inversions: Application to source estimation using a 3d atmospheric transport model," *Quarterly Journal of the Royal Meteorological Society*, vol. 141, no. 690, pp. 1906–1921, 2015.
- [31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [32] Y. Gal, "Uncertainty in Deep Learning," Phd Thesis, University of Cambridge, England, 2016.
- [33] E. H llermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Mach Learn*, vol. 110, no. 3, pp. 457–506, Mar. 2021.
- [34] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity analysis in practice: A guide to assessing scientific models*. Wiley, 2004.
- [35] R. Ghanem, D. Higdon, and H. Owahdi, Eds., *Springer Handbook on Uncertainty Quantification*. Springer, 2017.
- [36] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur, *Basics and Trends in Sensitivity Analysis. Theory and Practice in R*. SIAM, 2021.
- [37] F. A. Rihan, "Sensitivity analysis for dynamic systems with time-lags," *Journal of Computational and Applied Mathematics*, vol. 151, no. 2, pp. 445–462, 2003.
- [38] D. G. Cacuci, M. Ionescu-Bujor, and I. M. Navon, *Sensitivity and uncertainty analysis, volume II: applications to large-scale systems*. CRC press, 2005.
- [39] B. Iooss, R. Kennet, and P. Secchi, "Different views of interpretability," in *Interpretability for Industry 4.0: Statistical and Machine Learning Approaches*, A. Lepore, B. Palumbo, and J.-M. Poggi, Eds. Springer, 2022.
- [40] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J.-M. Loub s, "Quantile-constrained wasserstein projections for robust interpretability analyses of numerical and machine learning models," *Preprint, arXiv:2209.11539*, 2022.
- [41] C. Molnar, *Interpretable machine learning: A guide for making black-box models explainable (2nd ed.)*. github, 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- [42] M. D'Elia and A. Veneziani, "Uncertainty quantification for data assimilation in a steady incompressible navier-stokes problem," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 47, no. 4, pp. 1037–1057, 2013.
- [43] B. Wang, X. Zou, and J. Zhu, "Data assimilation and its applications," *Proceedings of the National Academy of Sciences*, vol. 97, no. 21, pp. 11 143–11 144, 2000.
- [44] A. C. Lorenc, "Analysis methods for numerical weather prediction," *Q. J. R. Meteorol. Soc.*, vol. 112, pp. 1177–1194, 1986.
- [45] R. N. Bannister, "A review of forecast error covariance statistics in atmospheric variational data assimilation. ii: Modelling the forecast error covariance statistics," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 134, no. 637, pp. 1971–1996, 2008.
- [46] T. Janji , N. Bormann, M. Bocquet, J. Carton, S. Cohn, S. L. Dance, S. Losa, N. K. Nichols, R. Potthast, J. A. Waller *et al.*, "On the representation error in data assimilation," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 713, pp. 1257–1278, 2018.
- [47] Y. Tr molet, "Accounting for an imperfect model in 4D-Var," *Q. J. R. Meteorol. Soc.*, vol. 132, pp. 2483–2504, 2006.
- [48] L. Hasco t, "Adjoint by automatic differentiation," in *Advanced data assimilation for geosciences*,  . Blayo, M. Bocquet, E. Cosme, and L. F. Cugliandolo, Eds. Les Houches school of physics: Oxford University Press, 2014, pp. 349–369.
- [49] G. A. Gottwald and S. Reich, "Combining machine learning and data assimilation to forecast dynamical systems from noisy partial observations," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 10, p. 101103, 2021.
- [50] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.
- [51] R. N. Miller, M. Ghil, and F. Gauthiez, "Advanced data assimilation in strongly nonlinear dynamical systems," *Journal of Atmospheric Sciences*, vol. 51, no. 8, pp. 1037–1056, 1994.
- [52] M. Ghil and P. Malanotte-Rizzoli, "Data assimilation in meteorology and oceanography," in *Advances in geophysics*. Elsevier, 1991, vol. 33, pp. 141–266.
- [53] P. de Rosnay, G. Balsamo, C. Albergel, J. Mu oz-Sabater, and L. Isaksen, "Initialisation of land surface variables for numerical weather prediction," *Surveys in Geophysics*, vol. 35, no. 3, pp. 607–621, 2014.
- [54] G. Evensen, "Using the extended kalman filter with a multilayer quasi-geostrophic ocean model," *Journal of Geophysical Research: Oceans*, vol. 97, no. C11, pp. 17 905–17 924, 1992.
- [55] G. Evensen *et al.*, *Data assimilation: the ensemble Kalman filter*. Springer, 2009, vol. 2.
- [56] P. J. Van Leeuwen, H. R. K nsch, L. Nerger, R. Potthast, and S. Reich, "Particle filters for high-dimensional geoscience applications: A review," *Quarterly Journal of the Royal Meteorological Society*, vol. 145, no. 723, pp. 2335–2365, 2019.
- [57] A. Carrassi, M. Bocquet, J. Demaeyer, C. Grudzien, P. Raanes, and S. Vannitsem, "Data assimilation for chaotic dynamics," in *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV)*, S. K. Park and L. Xu, Eds. Cham: Springer International Publishing, 2022, pp. 1–42.
- [58] M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*, ser. Fundamentals of Algorithms. SIAM, Philadelphia, 2016.
- [59] E. Calvello, S. Reich, and A. M. Stuart, "Ensemble kalman methods: A mean field perspective," *arXiv preprint arXiv:2209.11371*, 2022.

- [60] G. Evensen, F. C. Vossepoel, and P. J. van Leeuwen, "Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem," 2022.
- [61] R. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82(Series D), pp. 35–45, 1960.
- [62] S. Ouala, R. Fablet, L. Drumetz, B. Chapron, A. Pascual, F. Collard, and L. Gaultier, "End-to-end kalman filter for the reconstruction of sea surface dynamics from satellite data," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 7414–7417.
- [63] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.
- [64] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton, "Ensemble-sindy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control," *Proceedings of the Royal Society A*, vol. 478, no. 2260, p. 20210904, 2022.
- [65] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Sparse identification of nonlinear dynamics with control (sindyc)," *IFAC-PapersOnLine*, vol. 49, no. 18, pp. 710–715, 2016.
- [66] S. Cheng, J. Chen, C. Anastasiou, P. Angeli, O. K. Matar, Y.-K. Guo, C. C. Pain, and R. Arcucci, "Generalised latent assimilation in heterogeneous reduced spaces with machine learning surrogate models," *Journal of Scientific Computing*, vol. 94, no. 1, pp. 1–37, 2023.
- [67] S. Cheng, Y. Jin, S. P. Harrison, C. Quilodrán-Casas, I. C. Prentice, Y.-K. Guo, and R. Arcucci, "Parameter flexible wildfire prediction using machine learning techniques: Forward and inverse modelling," *Remote Sensing*, vol. 14, no. 13, p. 3228, 2022.
- [68] H. Gong, S. Cheng, Z. Chen, Q. Li, C. Quilodrán-Casas, D. Xiao, and R. Arcucci, "An efficient digital twin based on machine learning svd autoencoder and generalised latent assimilation for nuclear reactor physics," *Annals of Nuclear Energy*, vol. 179, p. 109431, 2022.
- [69] A. D. Carnerero, D. R. Ramirez, and T. Alamo, "State-space kriging: A data-driven method to forecast nonlinear dynamical systems," *IEEE Control Systems Letters*, vol. 6, pp. 2258–2263, 2022.
- [70] Z. Bai and L. Peng, "Non-intrusive nonlinear model reduction via machine learning approximations to low-dimensional operators," *Advanced Modeling and Simulation in Engineering Sciences*, vol. 8, no. 1, pp. 1–24, 2021.
- [71] H. Gong, S. Cheng, Z. Chen, and Q. Li, "Data-enabled physics-informed machine learning for reduced-order modeling digital twin: application to nuclear reactor physics," *Nuclear Science and Engineering*, vol. 196, no. 6, pp. 668–693, 2022.
- [72] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.
- [73] S. Cheng, I. C. Prentice, Y. Huang, Y. Jin, Y.-K. Guo, and R. Arcucci, "Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting," *Journal of Computational Physics*, p. 111302, 2022.
- [74] K. Liu, K. Ok, W. Vega-Brown, and N. Roy, "Deep inference for covariance estimation: Learning gaussian noise models for state estimation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1436–1443.
- [75] M. Amendola, R. Arcucci, L. Mottet, C. Q. Casas, S. Fan, C. Pain, P. Linden, and Y.-K. Guo, "Data assimilation in the latent space of a neural network," 2020.
- [76] Y. Zhuang, S. Cheng, N. Kovalchuk, M. Simmons, O. K. Matar, Y.-K. Guo, and R. Arcucci, "Ensemble latent assimilation with deep learning surrogate model: application to drop interaction in a microfluidics device," *Lab on a Chip*, vol. 22, no. 17, pp. 3187–3202, 2022.
- [77] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.
- [78] S. Cheng and M. Qiu, "Observation error covariance specification in dynamical systems for data assimilation using recurrent neural networks," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13 149–13 167, 2022.
- [79] G. Revach, N. Shlezinger, X. Ni, A. Escoriza, R. van Sloun, and Y. Eldar, "KalmanNet: Neural Network Aided Kalman Filtering for Partially Known Dynamics," *IEEE Trans. on Sig. Proc.*, vol. 70, pp. 1532–1547, 2022.
- [80] S. Legler and T. Janjic, "Combining data assimilation and machine learning to estimate parameters of a convective-scale model," *Q J R Meteorol Soc*, vol. 148, pp. 860–874, 2022.
- [81] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on bayesian neural networks—a tutorial for deep learning users," *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29–48, 2022.
- [82] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P. W. Battaglia, "Learning mesh-based simulation with graph networks," *arXiv preprint arXiv:2010.03409*, 2020.
- [83] R. Rakhimov, D. Volkonskiy, A. Artemov, D. Zorin, and E. Burnaev, "Latent video transformer," *arXiv preprint arXiv:2006.10704*, 2020.
- [84] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, 2022.
- [85] R. Fu, D. Xiao, I. Navon, and C. Wang, "A data driven reduced order model of fluid flow by auto-encoder and self-attention deep learning methods," *arXiv preprint arXiv:2109.02126*, 2021.
- [86] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, 2nd ed. Springer, 2009.
- [87] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.
- [88] DEEL Certification Workgroup, "Machine Learning in Certified Systems," DEpendable & Explainable Learning (DEEL), IRT Saint Exupéry, Tech. Rep., 2020, report No. S079L03T00-005.
- [89] L. Le Gratiet, S. Marelli, and B. Sudret, "Metamodel-based sensitivity analysis: Polynomial chaos expansions and Gaussian processes," in *Springer Handbook on Uncertainty Quantification*, R. Ghanem, D. Higdon, and H. Owahdi, Eds. Springer, 2017, pp. 1289–1325.
- [90] G. Blatman, *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. PhD Thesis of Blaise Pascal - Clermont II University, 2009.
- [91] C. Demay, B. Iooss, L. L. Gratiet, and A. Marrel, "Model selection for Gaussian Process regression: an application with highlights on the model variance validation," *Quality and Reliability Engineering International Journal*, 2021.
- [92] N. Meinshausen, "Forest Garotte," *Electronic Journal of Statistics*, vol. 3, p. 1288–1304, 2009.
- [93] L. M. ad G. Hooker, "Quantifying uncertainty in random forests via confidence intervals and hypothesis tests," *Journal of Machine Learning Research*, vol. 17, pp. 1–41, 2016.
- [94] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.
- [95] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher et al., "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.
- [96] A. F. Psaros, X. Meng, Z. Zou, L. Guo, and G. E. Karniadakis, "Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons," *arXiv preprint arXiv:2201.07766*, 2022.
- [97] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.
- [98] G. Rega and H. Troger, "Dimension reduction of dynamical systems: methods, models, applications," *Nonlinear Dynamics*, vol. 41, no. 1, pp. 1–15, 2005.
- [99] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, "Data-driven model reduction and transfer operator approximation," *Journal of Nonlinear Science*, vol. 28, no. 3, pp. 985–1010, 2018.
- [100] D. J. Lucia, P. S. Beran, and W. A. Silva, "Reduced-order modeling: new approaches for computational physics," *Progress in aerospace sciences*, vol. 40, no. 1-2, pp. 51–117, 2004.
- [101] K. Lee and K. T. Carlberg, "Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders," *Journal of Computational Physics*, vol. 404, p. 108973, 2020.
- [102] D. Wu, Y. He, X. Luo, and M. Zhou, "A latent factor analysis-based approach to online sparse streaming feature selection," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 11, pp. 6744–6758, 2021.
- [103] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," in *Medical Imaging with Deep Learning*. PMLR, 2021, pp. 728–744.
- [104] K. T. Carlberg, A. Jameson, M. J. Kochenderfer, J. Morton, L. Peng, and F. D. Witherden, "Recovering missing cfd data for high-order discretizations using deep neural networks and dynamics learning," *Journal of Computational Physics*, vol. 395, pp. 105–124, 2019.
- [105] S. E. Otto and C. W. Rowley, "Linearly recurrent autoencoder networks for learning dynamics," *SIAM Journal on Applied Dynamical Systems*, vol. 18, no. 1, pp. 558–593, 2019.

- [106] N. Takeishi, Y. Kawahara, and T. Yairi, "Learning koopman invariant subspaces for dynamic mode decomposition," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [107] C. Quilodrán Casas, R. Arcucci, and Y. Guo, "Urban air pollution forecasts generated from latent space representation," in *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020, p. 26.
- [108] T. R. Phillips, C. E. Heaney, P. N. Smith, and C. C. Pain, "An autoencoder-based reduced-order model for eigenvalue problems with application to neutron diffusion," *International Journal for Numerical Methods in Engineering*, vol. 122, no. 15, pp. 3780–3811, 2021.
- [109] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [110] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.
- [111] A. Polyak and L. Wolf, "Attention-based wavenet autoencoder for universal voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6800–6804.
- [112] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2138–2148, 2019.
- [113] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [114] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. Battaglia, "Learning to simulate complex physics with graph networks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8459–8468.
- [115] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.
- [116] D. Xiao, P. Yang, F. Fang, J. Xiang, C. C. Pain, and I. M. Navon, "Non-intrusive reduced order modelling of fluid-structure interactions," *Computer Methods in Applied Mechanics and Engineering*, vol. 303, pp. 35–54, 2016.
- [117] C. Quilodrán-Casas, R. Arcucci, C. Pain, and Y. Guo, "Adversarially trained LSTMs on reduced order models of urban air pollution simulations," *arXiv preprint arXiv:2101.01568*, 2021.
- [118] S. B. Reddy, A. R. Magee, R. K. Jaiman, J. Liu, W. Xu, A. Choudhary, and A. Hussain, "Reduced order model for unsteady fluid flows via recurrent neural networks," in *International Conference on Offshore Mechanics and Arctic Engineering*, vol. 58776. American Society of Mechanical Engineers, 2019, p. V002T08A007.
- [119] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [120] T. Nakamura, K. Fukami, K. Hasegawa, Y. Nabae, and K. Fukagata, "Convolutional neural network and long short-term memory based reduced order surrogate for minimal turbulent channel flow," *Physics of Fluids*, vol. 33, no. 2, p. 025116, 2021.
- [121] A. T. Mohan and D. V. Gaitonde, "A deep learning based approach to reduced order modeling for turbulent flow control using lstm neural networks," *arXiv preprint arXiv:1804.09269*, 2018.
- [122] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.
- [123] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.
- [124] M. Cheng, F. Fang, C. C. Pain, and I. Navon, "Data-driven modelling of nonlinear spatio-temporal fluid flows using a deep convolutional generative adversarial network," *Computer Methods in Applied Mechanics and Engineering*, vol. 365, p. 113000, 2020.
- [125] J. Tong, L. Xie, W. Yang, and K. Zhang, "Probabilistic decomposition transformer for time series forecasting," *arXiv preprint arXiv:2210.17393*, 2022.
- [126] E. Kaiser, J. N. Kutz, and S. L. Brunton, "Sparse identification of nonlinear dynamics for model predictive control in the low-data limit," *Proceedings of the Royal Society A*, vol. 474, no. 2219, p. 20180335, 2018.
- [127] A. A. Kaptanoglu, B. M. de Silva, U. Fasel, K. Kaheman, A. J. Goldschmidt, J. Callahan, C. B. Delahunt, Z. G. Nicolaou, K. Champion, J.-C. Loiseau, J. N. Kutz, and S. L. Brunton, "Pysindy: A comprehensive python package for robust sparse system identification," *Journal of Open Source Software*, vol. 7, no. 69, p. 3994, 2021.
- [128] C. Paglia, A. Stiehl, and C. Uhl, "Identification of low-dimensional nonlinear dynamics from high-dimensional simulated and real-world data," in *CONTROLO 2022*, L. Brito Palma, R. Neves-Silva, and L. Gomes, Eds. Cham: Springer International Publishing, 2022, pp. 205–213.
- [129] Y. Cai, X. Wang, G. Joos, and I. Kamwa, "An online data-driven method to locate forced oscillation sources from power plants based on sparse identification of nonlinear dynamics (sindy)," *IEEE Transactions on Power Systems*, 2022.
- [130] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, "Data-driven discovery of coordinates and governing equations," *Proceedings of the National Academy of Sciences*, vol. 116, no. 45, pp. 22445–22451, 2019.
- [131] B. Kim, V. C. Azevedo, N. Thuerey, T. Kim, M. Gross, and B. Solenthaler, "Deep fluids: A generative network for parameterized fluid simulations," in *Computer Graphics Forum*. Wiley Online Library, 2019, pp. 59–70.
- [132] S. Wiewel, M. Becher, and N. Thuerey, "Latent space physics: Towards learning the temporal evolution of fluid flow," in *Computer graphics forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 71–82.
- [133] N. Thuerey, K. Weißenow, L. Prantl, and X. Hu, "Deep learning methods for reynolds-averaged navier-stokes simulations of airfoil flows," *AIAA Journal*, vol. 58, no. 1, pp. 25–36, 2020.
- [134] A. Sanchez-Gonzalez and K. Stachenfeld, "Learning general-purpose cnn-based simulators for astrophysical turbulence," in *SimDL workshop at ICLR2021*, 2021.
- [135] S. Rasp, M. S. Pritchard, and P. Gentile, "Deep learning to represent subgrid processes in climate models," *Proceedings of the National Academy of Sciences*, vol. 115, no. 39, pp. 9684–9689, 2018.
- [136] T. Bolton and L. Zanna, "Applications of deep learning to ocean data inference and subgrid parameterization," *Journal of Advances in Modeling Earth Systems*, vol. 11, no. 1, pp. 376–399, 2019.
- [137] X. Jia, J. Willard, A. Karpate, J. Read, J. Zwart, M. Steinbach, and V. Kumar, "Physics guided rnns for modeling dynamical systems: A case study in simulating lake temperature profiles," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2019, pp. 558–566.
- [138] P. A. G. Watson, "Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction," *Journal of Advances in Modeling Earth Systems*, vol. 11, no. 5, pp. 1402–1417, 2019.
- [139] M. Bonavita and P. Laloyaux, "Machine learning for model error inference and correction," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 12, 2020.
- [140] J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino, "Combining data assimilation and machine learning to infer unresolved scale parametrization," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200086, 2021.
- [141] D. J. Gagne, H. M. Christensen, A. C. Subramanian, and A. H. Monahan, "Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz '96 model," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 3, 2020.
- [142] A. Wikner, J. Pathak, B. R. Hunt, I. Szunyogh, M. Girvan, and E. Ott, "Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 5, p. 053114, 2021.
- [143] A. Farchi, P. Laloyaux, M. Bonavita, and M. Bocquet, "Using machine learning to correct model error in data assimilation and forecast applications," *Quarterly Journal of the Royal Meteorological Society*, vol. 147, no. 739, pp. 3067–3084, 2021.
- [144] A. Farchi, M. Bocquet, P. Laloyaux, M. Bonavita, and Q. Malartic, "A comparison of combined data assimilation and machine learning methods for offline and online model error correction," *Journal of Computational Science*, vol. 55, p. 101468, 2021.
- [145] S. Barthélémy, J. Brajard, L. Bertino, and F. Counillon, "Super-resolution data assimilation," *Ocean Dynamics*, vol. 72, no. 8, pp. 661–678, 2022.
- [146] M. A. Sacco, J. J. Ruiz, M. Pulido, and P. Tandeo, "Evaluation of machine learning techniques for forecast uncertainty quantification,"

- Quarterly Journal of the Royal Meteorological Society*, vol. 148, no. 749, pp. 3470–3490, 2022.
- [147] D. P. Dee, “Bias and data assimilation,” *Quarterly Journal of the Royal Meteorological Society*, vol. 131, no. 613, pp. 3323–3343, 2005.
- [148] A. Carrassi and S. Vannitsem, “Treatment of the error due to unresolved scales in sequential data assimilation,” *International Journal of Bifurcation and Chaos*, vol. 21, no. 12, pp. 3619–3626, 2011.
- [149] M. Bocquet, A. Farchi, and Q. Malartic, “Online learning of both state and dynamics using ensemble kalman filters,” *Foundations of Data Science*, vol. 3, no. 3, pp. 305–330, 2021.
- [150] Q. Malartic, A. Farchi, and M. Bocquet, “State, global, and local parameter estimation using local ensemble kalman filters: Applications to online machine learning of chaotic dynamics,” *Quarterly Journal of the Royal Meteorological Society*, p. qj.4297, 2022.
- [151] A. Farchi, M. Chrut, M. Bocquet, P. Laloyaux, and M. Bonavita, “Online model error correction with neural networks in the incremental 4d-var framework,” 2022.
- [152] Y. Chen, D. Sanz-Alonso, and R. Willett, “Autodifferentiable ensemble kalman filters,” *SIAM Journal on Mathematics of Data Science*, vol. 4, no. 2, pp. 801–833, 2022.
- [153] J. Shogren, *Encyclopedia of energy, natural resource, and environmental economics*. Newnes, 2013.
- [154] H. Modares, A. Alf, and M.-M. Fateh, “Parameter identification of chaotic dynamic systems through an improved particle swarm optimization,” *Expert Systems with Applications*, vol. 37, no. 5, pp. 3714–3720, 2010.
- [155] C. W. Brown, R. R. Hood, W. Long, J. Jacobs, D. Ramers, C. Wazniak, J. Wiggert, R. Wood, and J. Xu, “Ecological forecasting in chesapeake bay: using a mechanistic–empirical modeling approach,” *Journal of Marine Systems*, vol. 125, pp. 113–125, 2013.
- [156] F. A. Albini, *Estimating wildfire behavior and effects*. Department of Agriculture, Forest Service, Intermountain Forest and Range ..., 1976, vol. 30.
- [157] C. Lautenberger, “Wildland fire modeling with an eulerian level set method and automated calibration,” *Fire Safety Journal*, vol. 62, pp. 289–298, 2013.
- [158] A. Alessandri, P. Bagnnerini, M. Gaggero, and L. Mantelli, “Parameter estimation of fire propagation models using level set methods,” *Applied Mathematical Modelling*, vol. 92, pp. 731–747, 2021.
- [159] P. J. Smith, S. L. Dance, M. J. Baines, N. K. Nichols, and T. R. Scott, “Variational data assimilation for parameter estimation: application to a simple morphodynamic model,” *Ocean Dynamics*, vol. 59, no. 5, pp. 697–708, 2009.
- [160] N. Wanders, M. F. Bierkens, S. M. de Jong, A. de Roo, and D. Karssen-berg, “The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models,” *Water resources research*, vol. 50, no. 8, pp. 6874–6891, 2014.
- [161] Y. Ruckstuhl and T. Janjić, “Combined state-parameter estimation with the letkf for convective-scale weather forecasting,” *Monthly Weather Review*, vol. 148, no. 4, pp. 1607–1628, 2020.
- [162] G. Kivman, “Sequential parameter estimation for stochastic systems,” *Nonlinear Processes in Geophysics*, vol. 10, no. 3, pp. 253–259, 2003.
- [163] J. Annan, J. Hargreaves, N. Edwards, and R. Marsh, “Parameter estimation in an intermediate complexity Earth system model using an ensemble Kalman filter,” *Ocean Modelling*, vol. 8, pp. 135–154, 2005.
- [164] A. A. Emerick and A. C. Reynolds, “History matching time-lapse seismic data using the ensemble kalman filter with multiple data assimilations,” *Computational Geosciences*, vol. 16, no. 3, pp. 639–659, 2012.
- [165] J. J. Ruiz, M. Pulido, and T. Miyoshi, “Estimating model parameters with ensemble-based data assimilation: A review,” *Journal of the Meteorological Society of Japan. Ser. II*, vol. 91, no. 2, pp. 79–99, 2013.
- [166] D. J. Posselt and C. H. Bishop, “Nonlinear parameter estimation: Comparison of an ensemble kalman smoother with a markov chain monte carlo algorithm,” *Monthly weather review*, vol. 140, no. 6, pp. 1957–1974, 2012.
- [167] D. J. Posselt, D. Hodyss, and C. H. Bishop, “Errors in ensemble kalman smoother estimates of cloud microphysical parameters,” *Monthly Weather Review*, vol. 142, no. 4, pp. 1631–1654, 2014.
- [168] D. Hodyss, “Ensemble state estimation for nonlinear systems using polynomial expansions in the innovation,” *Monthly Weather Review*, vol. 139, no. 11, pp. 3571–3588, 2011.
- [169] Y. M. Ruckstuhl and T. Janjić, “Parameter and state estimation with ensemble kalman filter based algorithms for convective-scale applications,” *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 712, pp. 826–841, 2018.
- [170] M. C. Rochoux, S. Ricci, D. Lucor, B. Cuenot, and A. Trouvé, “Towards predictive data-driven simulations of wildfire spread—part i: Reduced-cost ensemble kalman filter based on a polynomial chaos surrogate model for parameter estimation,” *Natural Hazards and Earth System Sciences*, vol. 14, no. 11, pp. 2951–2973, 2014.
- [171] P. Nadler, R. Arcucci, and Y. Guo, “A neural sir model for global forecasting,” in *Machine Learning for Health*. PMLR, 2020, pp. 254–266.
- [172] P. Nadler, R. Arcucci, and Y.-K. Guo, “Data assimilation for parameter estimation in economic modelling,” in *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2019, pp. 649–656.
- [173] X. Li, C. Xiao, A. Cheng, and H. Lin, “Joint estimation of parameter and state with hybrid data assimilation and machine learning,” *Preprint*, 2022.
- [174] L. A. Ferrat, M. Goodfellow, and J. R. Terry, “Classifying dynamic transitions in high dimensional neural mass models: A random forest approach,” *PLoS computational biology*, vol. 14, no. 3, p. e1006009, 2018.
- [175] M. Frangos, Y. Marzouk, K. Willcox, and B. van Bloemen Waanders, “Surrogate and reduced-order modeling: a comparison of approaches for large-scale statistical inverse problems,” *Large-Scale Inverse Problems and Quantification of Uncertainty*, pp. 123–149, 2010.
- [176] L. Cai, L. Ren, Y. Wang, W. Xie, G. Zhu, and H. Gao, “Surrogate models based on machine learning methods for parameter estimation of left ventricular myocardium,” *Royal Society open science*, vol. 8, no. 1, p. 201121, 2021.
- [177] P. Tandeo, P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, and Y. Zhen, “A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation,” *Monthly Weather Review*, vol. 148, no. 10, pp. 3973–3994, 2020.
- [178] G. Desroziers, L. Berre, B. Chapnik, and P. Poli, “Diagnosis of observation, background and analysis-error statistics in observation space,” *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 131, no. 613, pp. 3385–3396, 2005.
- [179] T. Berry and T. Sauer, “Adaptive ensemble kalman filtering of nonlinear systems,” *Tellus A: Dynamic Meteorology and Oceanography*, vol. 65, no. 1, p. 20331, 2013.
- [180] R. Mehra, “On the identification of variances and adaptive kalman filtering,” *IEEE Transactions on automatic control*, vol. 15, no. 2, pp. 175–184, 1970.
- [181] D. P. Dee, “On-line estimation of error covariance parameters for atmospheric data assimilation,” *Monthly weather review*, vol. 123, no. 4, pp. 1128–1145, 1995.
- [182] J. R. Stroud, M. Katzfuss, and C. K. Wikle, “A bayesian adaptive ensemble kalman filter for sequential state and parameter estimation,” *Monthly weather review*, vol. 146, no. 1, pp. 373–386, 2018.
- [183] R. H. Shumway and D. S. Stoffer, “An approach to time series smoothing and forecasting using the em algorithm,” *Journal of time series analysis*, vol. 3, no. 4, pp. 253–264, 1982.
- [184] P. Tandeo, M. Pulido, and F. Lott, “Offline parameter estimation using enfk and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization,” *Quarterly Journal of the Royal Meteorological Society*, vol. 141, no. 687, pp. 383–395, 2015.
- [185] M. Pulido, G. Scheffler, J. J. Ruiz, M. M. Lucini, and P. Tandeo, “Estimation of the functional form of subgrid-scale parametrizations using ensemble-based data assimilation: a simple model experiment,” *Quarterly Journal of the Royal Meteorological Society*, vol. 142, no. 701, pp. 2974–2984, 2016.
- [186] D. Dreano, P. Tandeo, M. Pulido, B. Ait-El-Fquih, T. Chonavel, and I. Hoteit, “Estimating model-error covariances in nonlinear state-space models using kalman smoothing and the expectation–maximization algorithm,” *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 705, pp. 1877–1885, 2017.
- [187] M. Pulido, P. Tandeo, M. Bocquet, A. Carrassi, and M. Lucini, “Stochastic parameterization identification using ensemble kalman filtering combined with maximum likelihood methods,” *Tellus A: Dynamic Meteorology and Oceanography*, vol. 70, no. 1, pp. 1–17, 2018.
- [188] T. J. Cocucci, M. Pulido, M. Lucini, and P. Tandeo, “Model error covariance estimation in particle and ensemble kalman filters using an online expectation–maximization algorithm,” *Quarterly Journal of the Royal Meteorological Society*, vol. 147, no. 734, pp. 526–543, 2021.

- [189] T. T. Chau, P. Ailliot, V. Monbet, and P. Tandeo, "Comparison of simulation-based algorithms for parameter estimation and state reconstruction in nonlinear state-space models," *Discrete and Continuous Dynamical Systems-Series S*, pp. 1–24, 2022.
- [190] G. Desroziers and S. Ivanov, "Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation," *Quarterly Journal of the Royal Meteorological Society*, vol. 127, no. 574, pp. 1433 – 1452, 2001.
- [191] R. Ménard, "Error covariance estimation methods based on analysis residuals: Theoretical foundation and convergence properties derived from simplified observation networks," *Quarterly Journal of the Royal Meteorological Society*, vol. 142, no. 694, pp. 257–273, 2016.
- [192] W. Vega-Brown, A. Bachrach, A. Bry, J. Kelly, and N. Roy, "Cello: A fast algorithm for covariance estimation," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 3160–3167.
- [193] O. Cicek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Proc. MICCAI*, 2016, pp. 424–432.
- [194] Z. Long, Y. Lu, and B. Dong, "PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network," *Journal of Computational Physics*, vol. 399, p. 108925, 2019.
- [195] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," *Advances in neural information processing systems*, vol. 29, 2016.
- [196] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.
- [197] A. Barth, A. Alvera-Azcárate, M. Licer, and J.-M. Beckers, "DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations," *Geosci. Mod. Dev.*, vol. 13, no. 3, pp. 1609–1622, Mar. 2020.
- [198] G. E. Manucharyan, L. Siegelman, and P. Klein, "A Deep Learning Approach to Spatiotemporal Sea Surface Height Interpolation and Estimation of Deep Currents in Geostrophic Ocean Turbulence," *JAMES*, vol. 13, no. 1, p. e2019MS001965, 2021.
- [199] P. Boudier, A. Fillion, S. Gratton, and S. Gürol, "DAN – An optimal Data Assimilation framework based on machine learning Recurrent Networks," *arXiv:2010.09694*, Oct. 2020.
- [200] T. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural Ordinary Differential Equations," in *Proc. NIPS'2018*, 2018, pp. 6571–6583.
- [201] M. Raissi, "Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations," *JMLR*, vol. 19, pp. 1–24, 2018.
- [202] S. Ouala, R. Fablet, C. Herzet, B. Chapron, A. Pascual, F. Collard, and L. Gaultier, "Neural-Network-based Kalman Filters for the Spatio-Temporal Interpolation of Satellite-derived Sea Surface Temperature," *Remote Sensing*, 2018.
- [203] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv:2004.05439*, 2020.
- [204] R. Fablet, B. Chapron, L. Drumetz, E. Memin, O. Pannekoucke, and F. Rousseau, "Learning Variational Data Assimilation Models and Solvers," *JAMES*, vol. 13, no. e2021MS002572, 2021.
- [205] R. Liu, L. Ma, X. Yuan, S. Zeng, and J. Zhang, "Bilevel Integrative Optimization for Ill-posed Inverse Problems," *arXiv:1907.03083*, 2019.
- [206] N. Cressie and C. Wikle, *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2015.
- [207] R. Fablet, Q. Febvre, and B. Chapron, "Multimodal 4DVarNets for the reconstruction of sea surface dynamics from SST-SSH synergies," *arXiv:2207.01372*, 2022.
- [208] S. Ouala, L. Debreu, A. Pascual, B. Chapron, F. Collard, L. Gaultier, and R. Fablet, "Learning runge-kutta integration schemes for ode simulation and identification," *arXiv preprint arXiv:2105.04999*, 2021.
- [209] N. Lafon, R. Fablet, and P. Naveau, "Uncertainty quantification when learning dynamical models and solvers with variational methods," *arXiv*, 2022.
- [210] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, 2006.
- [211] J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez, "Quality of uncertainty quantification for bayesian neural network inference," *arXiv preprint arXiv:1906.09686*, 2019.
- [212] Y. Huang, Y. Zhang, and J. A. Chambers, "A novel kullback-leibler divergence minimization-based adaptive student's t-filter," *IEEE Transactions on signal Processing*, vol. 67, no. 20, pp. 5417–5432, 2019.
- [213] R. Neal, "MCMC Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, Eds. Chapman and Hall/CRC, May 2011, vol. 20116022, series Title: Chapman & Hall/CRC Handbooks of Modern Statistical Methods.
- [214] A. D. Cobb and B. Jalaian, "Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 675–685.
- [215] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 681–688.
- [216] C. Nemeth and P. Fearnhead, "Stochastic gradient markov chain monte carlo," *Journal of the American Statistical Association*, vol. 116, no. 533, pp. 433–450, 2021.
- [217] B. Staber and S. da Veiga, "Quantitative performance evaluation of Bayesian neural networks," 2022, arXiv:2206.06779. [Online]. Available: <http://arxiv.org/abs/2206.06779>
- [218] W. Maddox, T. Garipov, P. Izmailov, D. Vetrov, and A. G. Wilson, "A Simple Baseline for Bayesian Uncertainty in Deep Learning," *arXiv:1902.02476 [cs, stat]*, Dec. 2019, arXiv: 1902.02476. [Online]. Available: <http://arxiv.org/abs/1902.02476>
- [219] D. Seuß, "Bridging the gap between explainable ai and uncertainty quantification to enhance trustability," *arXiv preprint arXiv:2105.11828*, 2021.
- [220] G. Agarwal, L. Hay, I. Iashvili, B. Mannix, C. McLean, M. Morris, S. Rappoccio, and U. Schubert, "Explainable ai for ml jet taggers using expert variables and layerwise relevance propagation," *Journal of High Energy Physics*, vol. 2021, no. 5, pp. 1–36, 2021.
- [221] X. Zhang, F. T. Chan, and S. Mahadevan, "Explainable machine learning in image classification models: An uncertainty quantification perspective," *Knowledge-Based Systems*, vol. 243, p. 108418, 2022.
- [222] K. Sokol and P. A. Flach, "Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety," *SafeAI@ AAAI*, 2019.
- [223] R. Arcucci, L. Mottet, C. Pain, and Y.-K. Guo, "Optimal reduced space for variational data assimilation," *Journal of Computational Physics*, vol. 379, pp. 51–69, 2019.
- [224] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, "Data assimilation in reduced modeling," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 5, no. 1, pp. 1–29, 2017.
- [225] S. Cheng, D. Lucor, and J.-P. Argaud, "Observation data compression for variational assimilation of dynamical systems," *Journal of Computational Science*, vol. 53, p. 101405, 2021.
- [226] D. Xiao, J. Du, F. Fang, C. Pain, and J. Li, "Parameterised non-intrusive reduced order methods for ensemble kalman filter data assimilation," *Computers & Fluids*, vol. 177, pp. 69–77, 2018.
- [227] C. Q. Casas, R. Arcucci, P. Wu, C. Pain, and Y.-K. Guo, "A reduced order deep data assimilation model," *Physica D: Nonlinear Phenomena*, vol. 412, p. 132615, 2020.
- [228] D. Xiao, F. Fang, C. E. Heaney, I. Navon, and C. Pain, "A domain decomposition method for the non-intrusive reduced order modelling of fluid flow," *Computer Methods in Applied Mechanics and Engineering*, vol. 354, pp. 307–330, 2019.
- [229] S. Cheng, J.-P. Argaud, B. Iooss, A. Ponçot, and D. Lucor, "A graph clustering approach to localization for adaptive covariance tuning in data assimilation based on state-observation mapping," *Mathematical Geosciences*, vol. 53, no. 8, pp. 1751–1780, 2021.
- [230] X. Luo, Z. Liu, M. Shang, J. Lou, and M. Zhou, "Highly-accurate community detection via pointwise mutual information-incorporated symmetric non-negative matrix factorization," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 1, pp. 463–476, 2020.
- [231] P. R. Vlachas, W. Byeon, Z. Y. Wan, T. P. Sapsis, and P. Koumoutsakos, "Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 474, no. 2213, p. 20170844, 2018.
- [232] C. Liu, R. Fu, D. Xiao, R. Stefanescu, P. Sharma, C. Zhu, S. Sun, and C. Wang, "Enkf data-driven reduced order assimilation system," *Engineering Analysis with Boundary Elements*, vol. 139, pp. 46–55, 2022.
- [233] M. Peyron, A. Fillion, S. Gürol, V. Marchais, S. Gratton, P. Boudier, and G. Goret, "Latent space data assimilation by using deep learning," *Quarterly Journal of the Royal Meteorological Society*, vol. 147, no. 740, pp. 3759–3777, 2021.
- [234] R. Maulik, V. Rao, J. Wang, G. Mengaldo, E. Constantinescu, B. Lusch, P. Balaprakash, I. Foster, and R. Kotamarthi, "Efficient

- high-dimensional variational data assimilation with machine-learned reduced-order models,” *arXiv preprint arXiv:2112.07856*, 2021.
- [235] J. Mack, R. Arcucci, M. Molina-Solana, and Y.-K. Guo, “Attention-based convolutional autoencoders for 3d-variational data assimilation,” *Computer Methods in Applied Mechanics and Engineering*, vol. 372, p. 113291, 2020.
- [236] S. Mohd Razak, A. Jahandideh, U. Djuraev, and B. Jafarpour, “Deep learning for latent space data assimilation in subsurface flow systems,” *SPE Journal*, pp. 1–21, 2022.
- [237] Y. Wang, X. Shi, L. Lei, and J. C.-H. Fung, “Deep learning augmented data assimilation: Reconstructing missing information with convolutional autoencoders,” *Monthly Weather Review*, 2022.
- [238] A. Storto, G. De Magistris, S. Falchetti, and P. Oddo, “A neural network-based observation operator for coupled ocean–acoustic variational data assimilation,” *Monthly Weather Review*, vol. 149, no. 6, pp. 1967–1985, 2021.
- [239] L. Espeholt, S. Agrawal, C. Sønderby, M. Kumar, J. Heek, C. Bromberg, C. Gaze, R. Carver, M. Andrychowicz, J. Hickey *et al.*, “Deep learning for twelve hour precipitation forecasts,” *Nature communications*, vol. 13, no. 1, pp. 1–10, 2022.
- [240] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, “Deep learning for multi-year ENSO forecasts,” *Nature*, vol. 573, no. 7775, pp. 568–572, 2019.
- [241] J. Paduart, L. Lauwers, J. Swevers, K. Smolders, J. Schoukens, and R. Pintelon, “Identification of nonlinear systems using polynomial nonlinear state space models,” *Automatica*, vol. 46, no. 4, pp. 647–656, 2010.
- [242] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, “Data-driven discovery of partial differential equations,” *Science advances*, vol. 3, no. 4, p. e1602614, 2017.
- [243] J.-C. Loiseau, B. R. Noack, and S. L. Brunton, “Sparse reduced-order modelling: sensor-based dynamics to full-state estimation,” *Journal of Fluid Mechanics*, vol. 844, pp. 459–490, 2018.
- [244] Y. Guan, S. L. Brunton, and I. Novosselov, “Sparse nonlinear models of chaotic electroconvection,” *Royal Society Open Science*, vol. 8, no. 8, p. 202367, 2021.
- [245] A. A. Kaptanoglu, K. D. Morgan, C. J. Hansen, and S. L. Brunton, “Physics-constrained, low-dimensional models for magnetohydrodynamics: First-principles and data-driven approaches,” *Physical Review E*, vol. 104, no. 1, p. 015206, 2021.
- [246] E. N. Lorenz, “Predictability: A problem partly solved,” in *Proc. Seminar on predictability*, vol. 1, no. 1, 1996.
- [247] W. Maass, T. Natschlager, and H. Markram, “Real-time computing without stable states: A new framework for neural computation based on perturbations,” *Neural computation*, vol. 14, no. 11, pp. 2531–2560, 2002.
- [248] H. Jaeger and H. Haas, “Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication,” *science*, vol. 304, no. 5667, pp. 78–80, 2004.
- [249] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, “Using machine learning to replicate chaotic attractors and calculate lyapunov exponents from data,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 12, p. 121102, 2017.
- [250] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, “Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach,” *Physical review letters*, vol. 120, no. 2, p. 024102, 2018.
- [251] R. Fablet, S. Ouala, and C. Herzet, “Bilinear residual Neural Network for the identification and forecasting of dynamical systems,” *SciRate*, dec 2017. [Online]. Available: <https://scirate.com/arxiv/1712.07003>
- [252] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, “Neural ordinary differential equations,” in *Advances in Neural Information Processing Systems*, 2018, pp. 6571–6583.
- [253] C. Herrera, F. Krach, and J. Teichmann, “Neural jump ordinary differential equations: Consistent continuous-time prediction and filtering,” *arXiv preprint arXiv:2006.04727*, 2020.
- [254] R. Wang, R. Walters, and R. Yu, “Incorporating symmetry into deep dynamics models for improved generalization,” *arXiv preprint arXiv:2002.03061*, 2020.
- [255] H. Ma, X. Hu, Y. Zhang, N. Thuerey, and O. J. Haidn, “A combined data-driven and physics-driven method for steady heat conduction prediction using deep convolutional neural networks,” 2020.
- [256] R. Wang, K. Kashinath, M. Mustafa, A. Albert, and R. Yu, “Towards physics-informed deep learning for turbulent flow prediction,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1457–1466.
- [257] X. Jin, S. Cai, H. Li, and G. E. Karniadakis, “Nsfnets (navier-stokes flow nets): Physics-informed neural networks for the incompressible navier-stokes equations,” *Journal of Computational Physics*, vol. 426, p. 109951, 2021.
- [258] S. Cai, Z. Wang, S. Wang, P. Perdikaris, and G. Karniadakis, “Physics-informed neural networks (PINNs) for heat transfer problems,” *Journal of Heat Transfer*, 2021.
- [259] K. Kashinath, M. Mustafa, A. Albert, J. Wu, C. Jiang, S. Esmaeilzadeh, K. Aizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh *et al.*, “Physics-informed machine learning: case studies for weather and climate modelling,” *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200093, 2021.
- [260] M. Mahmoudabadbozchelou, M. Caggioni, S. Shahsavari, W. H. Hartt, G. E. Karniadakis, and S. Jamali, “Data-driven physics-informed constitutive metamodeling of complex fluids: A multifidelity neural network (mfnn) framework,” *Journal of Rheology*, vol. 65, no. 2, pp. 179–198, 2021.
- [261] E. Kharazmi, Z. Zhang, and G. E. Karniadakis, “hp-vpinns: Variational physics-informed neural networks with domain decomposition,” *Computer Methods in Applied Mechanics and Engineering*, vol. 374, p. 113547, 2021.
- [262] D. Lucor, A. Agrawal, and A. Sergeant, “Simple computational strategies for more effective physics-informed neural networks modeling of turbulent natural convection,” *Journal of Computational Physics*, vol. 456, p. 111022, 2022.
- [263] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, “The analog data assimilation,” *Monthly Weather Review*, vol. 145, no. 10, pp. 4093–4107, 2017.
- [264] Y. Zhen, P. Tandeo, S. Leroux, S. Metref, T. Penduff, and J. Le Sommer, “An adaptive optimal interpolation based on analog forecasting: application to ssh in the gulf of mexico,” *Journal of Atmospheric and Oceanic Technology*, vol. 37, no. 9, pp. 1697–1711, 2020.
- [265] F. Takens, “Detecting strange attractors in turbulence,” in *Dynamical Systems and Turbulence, Warwick 1980*, D. Rand and L.-S. Young, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981, pp. 366–381.
- [266] M. Sangiorgio, F. Dercole, and G. Guariso, “Forecasting of noisy chaotic systems with deep neural networks,” *Chaos, Solitons & Fractals*, vol. 153, p. 111570, 2021.
- [267] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, “Weatherbench: a benchmark data set for data-driven weather forecasting,” *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, p. e2020MS002203, 2020.
- [268] T. Kurth, S. Subramanian, P. Harrington, J. Pathak, M. Mardani, D. Hall, A. Miele, K. Kashinath, and A. Anandkumar, “Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators,” *arXiv preprint arXiv:2208.05419*, 2022.
- [269] A. McGovern, R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, “Making the black box more transparent: Understanding the physical implications of machine learning,” *Bulletin of the American Meteorological Society*, vol. 100, no. 11, pp. 2175–2199, 2019.
- [270] B. A. Toms, E. A. Barnes, and I. Ebert-Uphoff, “Physically interpretable neural networks for the geosciences: Applications to earth system variability,” *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 9, p. e2019MS002002, 2020.
- [271] C. Irrgang, N. Boers, M. Sonnewald, E. A. Barnes, C. Kadow, J. Staneva, and J. Saynisch-Wagner, “Towards neural earth system modelling by integrating artificial intelligence in earth system science,” *Nature Machine Intelligence*, vol. 3, no. 8, pp. 667–674, 2021.
- [272] I. Ayed, E. de Bézenac, A. Pajot, J. Brajard, and P. Gallinari, “Learning dynamical systems from partial observations,” *arXiv preprint arXiv:1902.11136*, 2019.
- [273] B. Boots and G. Gordon, “An online spectral learning algorithm for partially observable nonlinear dynamical systems,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 293–300.
- [274] M. Levine and A. Stuart, “A framework for machine learning of model error in dynamical systems,” *Communications of the American Mathematical Society*, vol. 2, no. 07, pp. 283–344, 2022.
- [275] H. D. I. Abarbanel, *Modeling Chaos*. New York, NY: Springer New York, 1996, pp. 95–114.
- [276] J. Frank, S. Mannor, and D. Precup, “Activity and gait recognition with time-delay embeddings,” in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, ser. AAAI’10. AAAI Press, 2010, pp. 1581–1586.
- [277] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, “Support vector regression with chaos-based firefly algorithm for stock

- market price forecasting,” *Applied Soft Computing*, vol. 13, no. 2, pp. 947–958, 2013.
- [278] T. Berry and J. Harlim, “Forecasting turbulent modes with nonparametric diffusion models: Learning from noisy data,” *Physica D: Nonlinear Phenomena*, vol. 320, pp. 57–76, 2016.
- [279] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. Barbosa, “Next generation reservoir computing,” *Nature communications*, vol. 12, no. 1, pp. 1–8, 2021.
- [280] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]*, december 2015, arXiv: 1512.03385. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [281] R. Krishnan, U. Shalit, and D. Sontag, “Structured inference networks for nonlinear state space models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [282] D. Nguyen, S. Ouala, L. Drumetz, and R. Fablet, “Em-like learning chaotic dynamics from noisy and partial observations,” *arXiv preprint arXiv:1903.10335*, 2019.
- [283] H. D. I. Abarbanel, *Choosing Time Delays*. New York, NY: Springer New York, 1996, pp. 25–37.
- [284] H. D. Abarbanel and H. D. Abarbanel, “Choosing the dimension of reconstructed phase space,” *Analysis of observed chaotic data*, pp. 39–67, 1996.
- [285] S. Ouala, D. Nguyen, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier, and R. Fablet, “Learning latent dynamics for partially observed chaotic systems,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 10, p. 103121, 2020.
- [286] S. Ouala, S. L. Brunton, A. Pascual, B. Chapron, F. Collard, L. Gaultier, and R. Fablet, “Bounded nonlinear forecasts of partially observed geophysical systems with physics-constrained deep learning,” *arXiv preprint arXiv:2202.05750*, 2022.
- [287] S. Ouala, P. Tandeo, B. Chapron, F. Collard, and R. Fablet, “End-to-end kalman filter in a high dimensional linear embedding of the observations,” in *Stochastic Transport in Upper Ocean Dynamics*, B. Chapron, D. Crisan, D. Holm, E. Mémin, and A. Radomska, Eds. Cham: Springer International Publishing, 2023, pp. 211–221.
- [288] P. Tandeo, P. Ailliot, and F. Sévellec, “Data-driven reconstruction of partially observed dynamical systems,” *EGU sphere*, vol. 2022, pp. 1–11, 2022. [Online]. Available: <https://egusphere.copernicus.org/preprints/egusphere-2022-1316/>
- [289] T. Sauer, J. A. Yorke, and M. Casdagli, “Embedology,” *Journal of statistical Physics*, vol. 65, no. 3, pp. 579–616, 1991.
- [290] R. M. Noyes and R. J. Field, “Oscillatory chemical reactions,” *Annual Review of Physical Chemistry*, vol. 25, no. 1, pp. 95–119, 1974.
- [291] K. P. Sharp, “Stochastic differential equations in finance,” *Applied mathematics and Computation*, vol. 37, no. 2, pp. 131–148, 1990.
- [292] U. Piomelli, “Large-eddy simulation: achievements and challenges,” *Progress in aerospace sciences*, vol. 35, no. 4, pp. 335–362, 1999.
- [293] D. D. Holm, “Variational principles for stochastic fluid dynamics,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 471, no. 2176, p. 20140963, 2015.
- [294] E. Mémin, “Fluid flow dynamics under location uncertainty,” *Geophysical & Astrophysical Fluid Dynamics*, vol. 108, no. 2, pp. 119–146, 2014.
- [295] B. Chapron, P. Dérian, E. Mémin, and V. Resseguier, “Large-scale flows under location uncertainty: a consistent stochastic framework,” *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 710, pp. 251–260, 2018.
- [296] K. P. Champion, S. L. Brunton, and J. N. Kutz, “Discovery of nonlinear multiscale systems: Sampling strategies and embeddings,” *SIAM Journal on Applied Dynamical Systems*, vol. 18, no. 1, pp. 312–333, 2019.
- [297] H. Frezat, G. Balarac, J. Le Sommer, R. Fablet, and R. Lguensat, “Physical invariance in neural networks for subgrid-scale scalar flux modeling,” *Physical Review Fluids*, vol. 6, no. 2, p. 024607, 2021.
- [298] H. Frezat, J. L. Sommer, R. Fablet, G. Balarac, and R. Lguensat, “A posteriori learning for quasi-geostrophic turbulence parametrization,” *arXiv preprint arXiv:2204.03911*, 2022.
- [299] R. Vinuesa and S. L. Brunton, “Enhancing computational fluid dynamics with machine learning,” *Nature Computational Science*, vol. 2, no. 6, pp. 358–366, 2022.
- [300] B. Ouyang, L.-T. Zhu, Y.-H. Su, and Z.-H. Luo, “A hybrid mesoscale closure combining cfd and deep learning for coarse-grid prediction of gas-particle flow dynamics,” *Chemical Engineering Science*, vol. 248, p. 117268, 2022.
- [301] T. M. Bury, R. Sujith, I. Pavithran, M. Scheffer, T. M. Lenton, M. Anand, and C. T. Bauch, “Deep learning for early warning signals of tipping points,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 39, p. e2106140118, 2021.
- [302] P. J. Van Leeuwen, “Nonlinear data assimilation in geosciences: an extremely efficient particle filter,” *Quarterly Journal of the Royal Meteorological Society*, vol. 136, no. 653, pp. 1991–1999, 2010.
- [303] R. H. Reichle, “Data assimilation methods in the earth sciences,” *Advances in water resources*, vol. 31, no. 11, pp. 1411–1418, 2008.
- [304] X. Zou, I. Navon, and F. Le Dimet, “Incomplete observations and control of gravity waves in variational data assimilation,” *Tellus A: Dynamic meteorology and oceanography*, vol. 44, no. 4, pp. 273–296, 1992.
- [305] J. Liang, K. Terasaki, and T. Miyoshi, “A machine learning approach to the observation operator for satellite radiance data assimilation,” 2023.
- [306] S. Jing, “Data assimilation with a machine learned observation operator and application to the assimilation of satellite data for sea ice models,” *University of North Carolina at Chapel Hill*, 2019.
- [307] 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*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 3449–3458.
- [308] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, “A comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.
- [309] W. M. Kouw and M. Loog, “A review of domain adaptation without target labels,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 766–785, 2019.
- [310] R. Schneider, M. Bonavita, A. Geer, R. Arcucci, P. Dueben, C. Vitolo, B. Le Saux, B. Demir, and P.-P. Mathieu, “Esa-ecmwf report on recent progress and research directions in machine learning for earth system observation and prediction,” *npj Climate and Atmospheric Science*, vol. 5, no. 1, p. 51, 2022.

AUTHORS BIBLIOGRAPHY

Sibo Cheng (sibo.cheng@imperial.ac.uk) completed his Ph.D at Université Paris-Saclay, France, in 2020. He is currently a research associate at the Data science institute, Department of Computing of Imperial College London. His research mainly focuses on physics-related machine learning, data assimilation and reduced order modelling for high-dimensional dynamical systems with a wide range of applications including wildfire prediction, computational fluid dynamics, microfluidic drop modelling and Medical AI

César Quilodrán-Casas (c.quilodran@imperial.ac.uk) ' areas of expertise are data-driven methods and machine learning. He was awarded his PhD from Imperial College London on data-driven oceanography in the Space and Atmospheric Physics Group. As a Research Associate at the Data Science Institute, Imperial College London he has worked in diverse multidisciplinary teams and applied his ML expertise on adversarial and generative networks to different applications such as urban air pollution, microfluidics, and wave energy.

Said Ouala (said.ouala@imt-atlantique.fr) received the M.S. degree in AI and signal processing from the Sorbonne University, Paris, France, and the Ph.D. degree in AI for geophysical dynamics from IMT-Atlantique, Brest, France. He is currently a postdoctoral research assistant in the Stochastic Transport in Upper Ocean Dynamics (STUOD) program. His research interests are at the intersection of dynamical systems and artificial intelligence, with applications in geosciences.

Alban Farchi (alban.farchi@enpc.fr) recieved his PhD in physics and environmental sciences (Uni. Paris-Est, 2019). He is a recently hired permanent researcher at CERE. He works in the field of data assimilation for the geosciences with application to atmospheric chemistry. Since 2020, he has been working with ECMWF on the use of machine learning techniques to correct model error in data assimilation and forecast applications.

Che Liu (che.liu21@imperial.ac.uk) is currently pursuing the Ph.D. degree with the School of Earth Science and Engineering, Imperial College London, London, UK. He is also affiliated with Data Science Institute, Imperial College London, London, UK. His research is at the interface between deep learning and multimode medical data processing, such as biomedical signal processing, medical image segmentation, clinical text processing, and image-text fusion.

Pierre Tandeo (pierre.tandeo@imt-atlantique.fr) was born in France in 1983. He received the M.S. degree in applied statistics from Agrocampus

Ouest, Rennes, France, and the Ph.D. degree from the Oceanography from Space Laboratory at IFREMER, Brest, France, in 2010. Then, he spent two years as a Postdoctoral Researcher with the Atmospheric Science Research Group, University of Corrientes, Argentina, and three years at Télécom Bretagne, Brest, France. Since 2015, he is an associate professor at IMT Atlantique, Brest, France, and a researcher at Lab-STICC, CNRS, France. Since 2019, he is an associate researcher at the Data Assimilation Research Team, RIKEN Center for Computational Science, Kobe, Japan. His main research interests are focused on IA, data assimilation, and inverse problems for geophysics.

Ronan Fablet (ronan.fablet@imt-atlantique.fr) currently holds a full Professor position at IMT Atlantique and is a research scientist at Lab-STICC in INRIA team Odyssey. With engineer and PhD degrees in Applied Math and Signal Processing, he has a significant experience in interdisciplinary research at the interface between data science and ocean science, especially space oceanography and marine ecology. He currently leads Research Chair OceaniX on Physics-informed AI for Ocean Monitoring and Surveillance. He co-authored more than 200 articles and communications in peer-reviewed conferences and journals. He is also a member of the scientific committees of French and European programs and institutes for his dual ocean-AI expertise. His current research interest includes deep learning for data assimilation, ocean modeling and ocean observation.

Didier Lucor (didier.lucor@lisn.upsaclay.fr) is a CNRS research director and deputy director of the Interdisciplinary Laboratory of Numerical Sciences (LISN), part of Paris-Saclay campus in Orsay France. He received his PhD in Applied Mathematics in 2004 from Brown University, USA and he was a postdoctoral fellow in the department of Ocean Engineering at MIT, USA. He is the coordinator of the France West Ercoftac pilot centre. His research interests relate to stochastic modelling, computational mechanics and probabilistic scientific computing, with emphasis on physics-informed machine learning, uncertainty quantification and data assimilation and applications ranging from turbulence, environmental flows, energy systems and biomechanics.

Bertrand Iooss (bertrand.iooss@edf.fr) obtained a Ph.D thesis in Geostatistics at the Paris School of Mines (1998) and an habilitation thesis in Statistics at Toulouse University (2009). He works as a senior researcher in industrial statistics at Electricité de France (EDF), having managed (2015-2021) a research project named “Uncertainty quantification and machine learning” for electricity production needs, in particular for the nuclear industry. His main research works involve the design, analysis, modeling, uncertainty quantification and validation of computer experiments, related to nuclear safety and environmental problems. He is currently interested in the topics of interpretability and validation of machine learning techniques.

Julien Brajard (julien.brajard@nersc.no) is an associate professor at Sorbonne University (Paris, France) since 2009 and a researcher at NERSC (Nansen Environmental and Remote Sensing Center) since 2018. He works at proposing new methodologies at the crossroads between data assimilation and machine learning. His objective is to improve the forecast of key climate variables by correcting model errors using data, in particular from remote sensing.

Dunhui Xiao (xiaodunhui@tongji.edu.cn) is a professor at Tongji University (Shanghai, China). He obtained his PhD from Imperial College London where he did his Post-doc. His research interests include numerical modelling with a focus on non-intrusive reduced-order modelling of Navier-Stokes equations, fluid-structure interactions, and multiphase flows in porous media. He is also interested in data-driven modelling, data science, physical data combined machine learning and optimisation. He is the PI of a number of grants. He sits on the editorial boards for a number of journals and he is the reviewer for many journals and the EPSRC.

Tijana Janjic (tijana.pfander@physik.uni-muenchen.de) is Heisenberg Professor of Data Assimilation at the Mathematical Institute for Machine Learning and Data Science, KU Eichstaett-Ingolstadt, Germany. She received her PhD in Applied Mathematics from University of Maryland and her BSc in Mathematics from the University of Belgrade. To date, her international and interdisciplinary career led to significant contributions to both theory and real-life applications in the field of data assimilation. Her scientific papers span topics in convective scale data assimilation, numerical methods, uncertainty quantification, large-scale linear and nonlinear constrained optimization and data science. She has been serving as associate editor for Journal of Advances in Modeling Earth Systems and QJRM. She is also a recipient of prestigious Heisenberg funding of German Science Foundation

(DFG).

Weiping Ding (ding.wp@ntu.edu.cn) is a Full Professor with the School of Information Science and Technology, Nantong University, Nantong, China. His main research directions involve deep neural networks, multimodal machine learning, and medical images analysis. He received the Ph.D. degree in Computer Science, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. In 2016, He was a Visiting Scholar at National University of Singapore, Singapore. From 2017 to 2018, he was a Visiting Professor at University of Technology Sydney, Australia. He has published over 200 articles, including over 80 IEEE journal papers. His fifteen authored/co-authored papers have been selected as ESI Highly Cited Papers. He serves as an Associate Editor/Editorial Board member of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Fuzzy Systems, IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on Intelligent Transportation Systems, Information Fusion, Information Sciences, Neurocomputing, Applied Soft Computing. He is the Leading Guest Editor of Special Issues in several prestigious journals, including IEEE Transactions on Evolutionary Computation, IEEE Transactions on Fuzzy Systems, and Information Fusion.

Yike Guo (yikeguo@ust.hk) is the Provost of the Hong Kong University of Science and Technology science December 2022. He is also a Professor of Computing Science in the Department of Computing at Imperial College London. He is the founding Director of the Data Science Institute at Imperial College. His main research interests lie in the field of machine learning and large-scale data analysis management. He is a Fellow of the Royal Academy of Engineering (FREng), Member of Academia Europaea (MAE), Fellow of British Computer Society and a Trustee of The Royal Institution of Great Britain. He received his PhD in Computational Logic from Imperial College in 1993. Professor Guo has published over 250 articles, papers and reports. Projects he has contributed to have been internationally recognised.

Alberto Carrassi (alberto.carrassi@unibo.it) is trained as a physicist, with a PhD thesis on dynamical systems. He has been always working at the crossroad between applied mathematics and climate. Research breakthroughs include contribution to the development of algorithms known as the assimilation in the unstable subspace unifying dynamical systems and data assimilation or of a novel paradigm for causal inference using data assimilation for detection and attribution of climate change. In recent years, he contributed to pioneer powerful combined methods merging data assimilation and machine learning, making possible using machine learning with the very sparse and noisy climatic dataset. In 2019 he became Professor at the University of Reading (UK) until November 2021 when he became Professor at the University of Bologna (IT).

Marc Bocquet (marc.bocquet@enpc.fr) is Professor at École des Ponts (France) and deputy director of CEREAs laboratory. He works on the methods of data assimilation, machine learning, inverse problems and environmental statistics, with applications to dynamical systems, atmospheric chemistry and transport, and meteorology. He is a fellow of ECMWF, and editor for the QJRM, Foundation of Data Science, and Frontiers in Applied Mathematics and Statistics. He has published over 115 peer-reviewed papers.

Rossella Arcucci (r.arcucci@imperial.ac.uk) is an Assistant Professor (Lecturer) in Data Science and Machine Learning at Imperial College London (ICL). She is an elected member of the World Meteorological Organization and the elected speaker of the AI Network of Excellence at ICL where she represents more than 270 academics working on AI. She has been with the Data Science Institute at ICL since 2017 where she has created, and she leads the Data Assimilation and Machine Learning (Data Learning) group. Her work involves developing AI models for climate, health and environmental impact and she also collaborates with the Leonardo Centre on Business for Society at Imperial College Business School. Ph.D. in Computational and Computer Science in February 2012 and she received the acknowledgement of Marie Skłodowska-Curie fellow from European Commission Research Executive Agency in 2017. She is co-investigator of several grants and projects.