Neural Incremental Data Assimilation

Matthieu Blanke [1] Ronan Fablet [2] Marc Lelarge [1]

Abstract

Data assimilation is a central problem in many geophysical applications, such as weather forecasting. It aims to estimate the state of a potentially large system, such as the atmosphere, from sparse observations, supplemented by prior physical knowledge. The size of the systems involved and the complexity of the underlying physical equations make it a challenging task from a computational point of view. Neural networks represent a promising method of emulating the physics at low cost, and therefore have the potential to considerably improve and accelerate data assimilation. In this work, we introduce a deep learning approach where the physical system is modeled as a sequence of coarse-to-fine Gaussian prior distributions parametrized by a neural network. This allows us to define an assimilation operator, which is trained in an end-to-end fashion to minimize the reconstruction error on a dataset with different observation processes. We illustrate our approach on chaotic dynamical physical systems with sparse observations, and compare it to traditional variational data assimilation methods.

1. Introduction

Artificial intelligence is transforming many fields, and has a growing number of applications in industry. In the sciences, it has the potential to considerably accelerate the scientific process. Geophysics and weather forecasting are areas where deep learning is particularly active, with recent months seeing an explosion in the number of large neural models for the weather forecasting problem (Pathak et al., 2022; Lam et al., 2022; Hoyer et al., 2023), building on reanalysis datasets such as ERA5 (Muñoz-Sabater et al., 2021) for training. In this work, we focus on the data assimilation problem that underpins weather forecasting: tomorrow's weather forecast is based on today's weather conditions, which are not directly measured, but are estimated from few observations. Data assimilation is the inverse problem of estimating the geophysical state of the globe on the basis of these sparse observations and of prior knowledge of the physics. The estimated state then serves as the starting point for forecasting. While deep learning models are revolutionizing the forecasting problem, they have yet to be applied operationally to data assimilation.

The application of neural networks to inverse problems is an active area of research. The general idea consists in training a neural network to reconstruct a signal, using for training examples a dataset of simulated physical states serving as ground truth. For the data assimilation problem, several approaches have been proposed to incorporate a deep learning in the loop. (Arcucci et al., 2021) propose a sequential scheme where a neural network is trained at regular time steps to combine data assimilation and the forecasting model. Recently, the success of diffusion models for imaging (Ho et al., 2020) has led to the development of so-called "plug and play" methods, where the neural network is trained to learn a prior (Laumont et al., 2022). Once trained, the neural prior can be used to solve a large number of inverse problems. In this line of work, (Rozet & Louppe, 2023) proposed a data assimilation method based on a diffusion model. Another type of approaches called "end-to-end" aim at directly training a neural network to minimize the reconstruction error. They have the benefit of training the network directly on the task of interest, but the versatility of the trained model with respect to the different observational processes is challenging. An end-to-end neural reconstruction algorithm is proposed in (Fablet et al., 2021), and aims at learning the prior distribution of the signal by defining the reconstruction as a maximum a posterior estimate, leading to a bi-level optimization problem. However, the complex prior induced by the neural network may hamper the convergence of this estimate, as it relies on non-convex optimization. Instead, we explore a model where the prior has a sufficiently simple structure to guarantee a convex posterior distribution.

Contributions In this work, we present a neural method for data assimilation. We introduce a data assimilation operator parametrized by a neural Gaussian prior, that is designed to locally improve the likelihood of an estimate. Our model is trained to minimize the reconstruction error in an end-to-end fashion. We show how this operator may be

[1] Inria Paris, DI ENS, PSL Research University [2] IMT Atlantique, Brest, France. Correspondence to: Matthieu Blanke <matthieu.blanke@inria.fr>.

Accepted at at ICML 2024 AI for Science workshop, Vienna, Austria.

iterated to reconstruct complex signals. The effectiveness of our method is demonstrated on simulated nonlinear physical systems. We also show how our method may be used to enhance traditional data assimilation methods.

2. The data assimilation inverse problem

The aim of data assimilation is to reconstruct a state $x \in \mathbb{R}^d$ from partial noisy measurements $y \in \mathbb{R}^m$ of that state (Bouttier & Courtier, 2002; Bocquet et al., 2014). For meteorological applications, for instance, the state x represents the physical quantities on a grid representing the globe, and the observations y are partial measurements, from different sources: in situ measurements, weather balloons, satellites, *etc*. These measurements may be very sparse, with an observation rate m/d that may be of the order 1%, so we cannot generally hope to recover the state as a function of the observations alone. Indeed, for a given observation vector y, a large number of states are compatible, making data assimilation an inverse problem. To reconstruct the state, we need to supplement the partial observations with another source of prior information on the state, which comes from our physical or statistical knowledge of the problem.

The data assimilation problem is then as follows. Given partial observations y and prior information on the state, the aim is to estimate the most probable underlying state x. The Bayesian probabilistic framework lends itself well to the mathematical formalization of the problem : the theoretical information about the state physics is captured by a prior distribution $x \sim p(x)$, and the noisy, partial observations of x can be modeled as $y|x \sim h(x) + \xi$, with h the observation process, and an unbiased additive noise that is typically assumed to be Gaussian $\xi \sim \mathcal{N}(0, R)$ and independent of x. Then, data assimilation can be seen as the estimation of the state maximizing the state posterior distribution $p(x|y) = p(x)p(y|x)/p(y)$. Under the assumption of Gaussian observational noise, this can be formulated as the following minimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad U(x) + \frac{1}{2}\|h(x) - y\|^2_{R^{-1}}, \qquad (2.1)$$

with $U(x) = -\log p(x)$, and where we have adopted the notation $\|z\|_C = \sqrt{z^\top C z}$ for a positive definite matrix C. We assume for simplicity that the observation function h is known, although it may be only partially known in some cases, such as remote sensing (Liang, 2005) or medical imaging (Rangayyan & Krishnan, 2024).

Problem size For weather prediction, the state x represents the geophysical variables on a large spatial grid. It is hence a signal of very high dimension with typically $d \sim 10^6$ or even $d \sim 10^9$. The size of the data assimilation problem makes the computations and memory costs very heavy, severely limiting the computational budget of any numerical method. In the development of new learning-based methods, it is essential to keep this computational constraint in mind if we hope to scale up to real-size systems.

2.1. Least-squares Gaussian interpolation

The first approach considered for data assimilation is naturally that of a linear-quadratic model. Assuming a Gaussian a priori on the state $x \sim \mathcal{N}(\mu, P)$ and a linear observation function $h(x) = Hx$, with $H \in \mathbb{R}^{m \times d}$, the variational Bayesian formulation for data assimilation (2.1) becomes a quadratic least-squares problem:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2}\|x - \mu\|^2_{P^{-1}} + \frac{1}{2}\|Hx - y\|^2_{R^{-1}} \qquad (2.2)$$

whose maximum a posteriori solution takes the form

$$x_{\text{MAP}}(y; \mu, P) := \mu + K(y - H\mu), \qquad (2.3)$$

with the H-dependent Kalman gain

$$K = PH^\top(HPH^\top + R)^{-1} \in \mathbb{R}^{m \times d}. \qquad (2.4)$$

In the remainder of this work, the dependence with respect to H is implicitly assumed in all quantities that depend on the observation vector y.

For meteorological applications, the state x that is optimized for is a snapshot of the set of geophysical variables at a given time, when the observations have been collected. The background term μ is the forecast of this state from the past observations.

Computational cost For large-scale applications, solving (2.2) by computing the closed-form expression (2.3) yields a $\mathcal{O}(m^3 + dm)$ complexity in general, as it involves solving a $m \times m$ linear system and computing a matrix-vector products of size $d \times m$. In operational geophysical applications, this cost may be a bottleneck as d and m may reach prohibitively large values. To avoid such costs, (2.2) is solved by such as conjugate gradient (Fletcher & Reeves, 1964). In the data assimilation community, this variational approach for the estimation of a large-scale geophysical spatial state is called 3D-Var (Courtier et al., 1998).

2.2. Spatio-temporal data assimilation

So far, the prior knowledge of the state has taken the form of a Gaussian distribution, which can capture the proximity of the searched state to an estimate, and the correlations of one state variable to another. Least squares interpolation then searches for the state most faithful to the data, within a fluctuation zone around the estimate. Although simple and analytically solvable, this approach does not use signal physics equations as prior information.

In the 1990s, the quality of data assimilation analyses improved significantly by incorporating a physical model to the reconstruction prior, leading to the state-of-the-art variational assimilation algorithm 4D-Var (Le Dimet & Talagrand, 1986). This algorithm is a generalization of 3D-Var to time-distributed observations, where the estimated signal x is a temporal sequence of the spatial geophysical state on a time window, *i.e.* a trajectory, rather than one single snapshot. The temporal dimension allows formulating the system's dynamical equations as a constraint for the signal. The reconstruction algorithm is applied sequentially on a sliding time window, in combination with a forecasting model, to produce regularly updated estimates of the meteorological variables. Alongside 4D-Var, other algorithms exist for data assimilation of dynamical systems, including sequential methods such as the celebrated Kalman filter, and its extensions to nonlinear models (Jazwinski, 2007), and to ensembling (Evensen, 2003). In this work, we focus on the so-called weak-constraint 4D-Var algorithm (Trémolet, 2007; Fisher et al., 2012), which we briefly explain next. Weak-constraint 4D-Var has the advantage being naturally related to the Bayesian formulation (2.1), and is used in operational systems.

For simplicity, we abstract from the time dimension in our mathematical formalism, and still denote the spatio-temporal signal as $x \in \mathbb{R}^d$. The knowledge of a physical dynamical model materializes as knowledge of a prior distribution $U(x)$ in (2.1), which can be computed and differentiated through with respect to x. In geophysics, this model is typically a fluid dynamics simulator, and its gradients are computed using the adjoint method (Talagrand & Courtier, 1987). Hence, the resulting $U(x)$ is more complex and more informative than a Gaussian prior, but comes with heavy computational costs. In the remained of this work, we assume that the observational processes are linear: $h(x) = Hx$. In practice, h is nonlinear and is sequentially approximated by its linear approximation. We argue that linearizing the physical model is computationally far more expensive than linearizing the observational process, and hence that considering only linear observations does not severely restrict the problem generality.

The weak-constraint 4D-Var algorithm aims at minimizing (2.1) by a Gauss-Newton descent algorithm (Gauss, 1877), with line-serach correction (Nocedal & Wright, 1999). More precisely, a sequence of estimates $\{z_k, 1 \le k \le \ell\}$ approximating the reconstruction signal is iteratively computed. At each iteration k, the objective function is approximated by its quadratic expansion in the vicinity of $z_k \in \mathbb{R}^d$. Specifically, the prior term is approximated as

$$\begin{aligned} U(x) \simeq U(z) &+ \nabla U(z)^\top (x - z) \\ &+ \frac{1}{2}(x - z)^\top \nabla^2 U(z)(x - z). \end{aligned} \tag{2.5}$$

We may express expansion (2.5) as a Gaussian log-likelihood:

$$U(x) \simeq \frac{1}{2}\|x - \mu(z)\|^2_{P(z)^{-1}}, \tag{2.6}$$

with

$$P(z) \simeq \nabla^2 U(z)^{-1}, \tag{2.7a}$$

$$\mu(z) = z - P(z)^{-1}\nabla U(z), \tag{2.7b}$$

the approximation above referring to the gradient-Hessian approximation.

Weak-constraint 4D-Var is described in Algorithm 1. We see that the sequence of estimates (z_k) is iterated with a recursion of the form

$$\begin{aligned} x_k &= A(z_k, y) \\ z_{k+1} &= z_k + \alpha_k (x_k - z_k), \end{aligned} \tag{2.8}$$

where assimilation operator A improves the current estimate z using the observations and the local approximation of the model, by performing a local optimal interpolation:

$$A(z, y) = x_{\mathrm{MAP}}(y; \mu(z), P(z)). \tag{2.9}$$

Limitations The 4D-Var algorithm represents the state of the art for data assimilation in geophysics, and is deployed in operational meteorological centers. Its main limitation is the high computational cost of simulating and differentiating through the physical model. In Algorithm 1, each computation of P_k and μ_k comes with a large cost in addition to the cost of computing (2.9), hence limiting the method's accuracy. Note that this method may also be viewed as an application of the iterative Kalman smoother (Bell & Cathey, 1993; Ménard & Daley, 1996; Fisher et al., 2005; Mandel et al., 2013). As is well known, an additional limitation of this method is that the non-convexity of U may lead to a complex minimization landscape, making the descent algorithm likely to be stuck in local minima (Gratton et al., 2007; Mandel et al., 2013). In the next section, we propose to overcome these limitations by learning operator A from data.

3. Neural data assimilation

Deep neural networks hold great promise for solving inverse problems (Bai et al., 2020), as they can help recover the corrupted signal by using the large amount statistical information acquired on a training dataset. For the data assimilation problem in meteorology or oceanography, the ground truth signals x are not available as the geophysical systems are not observed. However, a promising research direction consists in training a deep neural network to learn a prior on high-resolution simulations, or on the reanalysis datasets

Algorithm 1 Incremental weak-constraint 4D-Var

```
input observation vector $y \in \mathbb{R}^m$, observation matrix $H$, iteration number $\ell$, initial estimate $z_0$, tangent linear physical model $\mu$, $P$
output state estimation $z_\ell$
initialize $z_0 := x_0$
for $0 \le k \le \ell - 1$ do
    compute $P_k := P(z_k), \quad \mu_k = \mu(z_k)$
    estimate $x_k = \mathrm{MAP}(y; \mu_k, P_k)$
    compute line search parameter $\alpha_k$
    update $z_{k+1} = z_k + \alpha_k(x_k - z_k)$
end for
```

Algorithm 2 Incremental neural data assimilation

```
input observation vector $y \in \mathbb{R}^m$, observation matrix $H$, iteration number $\ell$, initial estimate $z_0$, neural models $\mu$, $P$, trained parameter $\theta$
output state estimation $z_\ell$
initialize $z_0 := x_0$
for $0 \le k \le \ell - 1$ do
    compute $P_k := P(z_k; \theta, s_k), \quad \mu_k = \mu(z_k; \theta, s_k)$
    estimate $x_k = \mathrm{MAP}(y; \mu_k, P_k)$
    compute temperature parameter $s_k$
    update $z_{k+1} = z_k + s_k(x_k - z_0)$
end for
```

such as ERA5, like neural weather models (Ben Bouallègue et al., 2024).

Deep learning approaches to inverse problems may be separated in two categories (Mukherjee et al., 2021). A first category of algorithms aims at learning a prior $U(x)$ from a training dataset, using a neural network, independently of the inverse problem. Once trained, the learned prior can be adapted to a reconstruction algorithm to reconstruct the signal. These algorithms are often called "plug-and-play", as the trained neural prior can be used for any downstream inverse problem. In a second category of algorithms, referred to as "end-to-end" learning algorithms, the neural network is explicitly trained to solve the inverse problem. In this case, the training consists of minimizing the neural network's reconstruction error, based on a dataset of state and observations pairs $(x^{(i)}, y^{(i)})$.

One challenge in training end-to-end algorithms is the multiplicity of possible observation processes: the trained neural network must be compatible with all possible (x, y), and hence with varying observation processes H, with different dimensions m for the observations. It should therefore model only the prior distribution $U(x)$, and not depend directly on the observation process H.

3.1. Neural assimilation operator

We adopt an end-to-end learning approach, and we aim at learning a neural assimilation algorithm by minimizing a reconstruction error. We observe that, unlike other inverse problems such as image inpainting, data assimilation often starts with a first physically plausible estimate z of the unknown state. Therefore, rather than learning to interpolate the observations from scratch, we train a neural network to improve the state estimate given z. Drawing inspiration from the 4D-Var algorithm, we learn an assimilation operator $A(z, y; \theta)$, where θ denotes the parameter vector of a neural network. As in (2.6), we model the local prior distribution conditioned on z as a Gaussian prior

$$x|z \sim \mathcal{N}(\mu(z;\theta), P(z;\theta)), \tag{3.1}$$

where $\mu(z;\theta)$ and $P(z;\theta)$ are trainable neural networks. Given this Gaussian prior, the observations are incorporated by solving the least-squares interpolation (2.2):

$$A(z, y; \theta) = x_{\mathrm{MAP}}(y; \mu(z;\theta), P(z;\theta)). \tag{3.2}$$

Versatility As we pointed out, the trained neural network should be compatible with arbitrary observation processes. By formulating it as the solution of a y-dependent interpolation problem, our assimilation operator (3.2) is defined for any observation process (H, y), although the underlying neural networks models only the prior distribution. In particular, the neural networks involved depend neither on y, nor on H: the neural assimilation operator (3.2) combines the observations with a neural prior (3.1) of the state through the computation of a maximum likelihood estimator, and this computation is valid for any (H, y) pair for the same neural network. At prediction time, the trained neural networks $\mu(z;\theta), P(z;\theta)$ may be used to assimilate a new observation y obtained from an arbitrary observation process H by solving (3.2).

Training Given a dataset $(x^{(i)}, y^{(i)}, z_0^{(i)})$ consisting of signals $x^{(i)}$ and partial observations $y^{(i)}$ obtained from different observation processes $H^{(i)}$, supplemented with coarse estimates $z^{(i)}$ of the signal, the neural prior (3.1) is trained to minimize the reconstruction error with the following objective:

$$\begin{aligned} \underset{\theta \in \mathbb{R}^n}{\text{minimize}} \quad & \sum_{i=1}^{N} \|A(z^{(i)}, y^{(i)}; \theta) - x^{(i)}\|^2 \\ \text{with} \quad & A(z, y; \theta) = x_{\mathrm{MAP}}(y; \mu(z;\theta), P(z;\theta)). \end{aligned} \tag{3.3}$$

We train our model by minimizing (3.3) using stochastic gradient descent, with the ADAM optimizer (Kingma & Ba, 2015). This training objective takes the form of a bilevel optimization problem. Solving the inner optimization problem involves computing the optimal interpolation (2.3), which is computed solving a linear system of size m. We need to propagate the gradients with respect to θ through this no-trivial operation during training. This may be handled by implicit differentiation, allowing to compute the gradients of the solution with respect to θ, without explicitly inverting the system's matrices (Johnson, 2012).

This training objective is similar to that of (Fablet et al., 2021), where a neural interpolator called 4DVarNet is used to learn both the global prior $U(x)$ and the minimization algorithm of (2.1), rather than a local operator $A(z, y) \mapsto x$. In our case, however, the inner optimization problem (2.1) can be solved explicitly because the cost is quadratic. In contrast, it is only approximately solved in the case of 4DVarNet, due to the non-convexity of the inner cost.

Computational cost As we mentioned, the large size of the targeted physical systems requires carefully considering the computational cost of the data assimilation methods. We model P as a band matrix, hence limiting both the memory storage to a $\mathcal{O}(d)$ cost and the computational complexity of solving the linear system in (2.3) using the Thomas algorithm (Datta, 2010). This structure also imposes a temporal structure in the signal.

3.2. Incremental neural data assimilation

Since our assimilation operator is trained to reconstruct the signal from a coarse approximation, a one-shot reconstruction is likely to yield blurry results. To improve reconstruction, we may iterate this operator, with the aim of progressively improving the reconstruction signal. Building on the recent advances of cold diffusion (Bansal et al., 2024), we propose an iterative strategy aiming at reconstructing the signal in a coarse-to-fine fashion. We introduce a scalar temperature parameter $0 \leq s \leq 1$ modeling the coarseness of the reconstruction, and we allow our neural prior to depend on s as $\mu(z; \theta, s)$, $P(z; \theta, s)$. Intuitively, the prior should be coarser for larger values of s, and become sharper and more local as $s \to 0$. We provide estimates z_k at different temperature levels $\{s_1 \geq \cdots \geq s_\ell\}$ as linear interpolations between z_0 and $z_\ell := x$:

$$z_k^{(i)} = s_k z_0^{(i)} + (1 - s_k) z_\ell^{(i)}. \tag{3.4}$$

Our training objective is adapted as

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{k=1}^{\ell} \sum_{i=1}^{N} \| A(z_k^{(i)}, y^{(i)}; \theta, s_k) - x^{(i)} \|^2. \tag{3.5}$$

At prediction time, the signal is reconstructed by iteratively applying $A(z, y; \theta, s)$ following the sampling algorithm introduced in (Bansal et al., 2024). We provide a detailed description of our iterative reconstruction method in Algorithm 2.

4. Experiments on physical systems

In order to evaluate the performances of our data assimilation algorithm, we experiment on two simulated dynamical systems: the pendulum and the Lorenz 63 dynamical systems. We train our neural model on a dataset generated from the dynamical system with different trajectories x sampled from random initial conditions, and different observation processes, leading to various (x, y) pairs for the same x. Our JAX implementation of our neural assimilation algorithm is available online at `https://github.com/MB-29/assimilation`.

Architecture We take for $\mu(z; \theta, s)$ and $P(z; \theta, s)$ two fully-connected neural networks of depth 4 and width 32. The dependence with respect to s is implemented as a positional embedding. The $d \times d$ matrix P is modeled as a band matrix with bandwidth $b = 2\varphi$, with φ the phase space dimension.

Baselines We compare our neural assimilation algorithm with various baseline. Each method starts from a first guess estimate z_0 of the signal, computed by performing a Gaussian interpolation from the observations (see below). We implement the weak-constraint 4D-Var algorithm as a Levenberg-Marquardt Gauss-Newton Algorithm using the JAXopt implementation (Blondel et al., 2021), and the Diffrax library for differentiating through differential equation solvers (Kidger, 2021). As an ablation, an "unconditional" cold diffusion model is trained to restore the signal by minimizing objective (3.5) without the information provided by the observations. It is then applied following Algorithm 2 just as our neural assimilation algorithm, without using y. The resulting reconstructed signal depends on the observations only through the first estimate z_0, which is computed to match y, but the neural network is trained to compute the next iterates by increasing only the prior term $U(x)$ in (2.1), not the observation likelihood.

4.1. Pendulum

We start with the pendulum, which is arguably one of the simplest nonlinear physical systems. Importantly, the pendulum is simple enough to be decently approximated by linear dynamics. It can be shown that a linear dynamical model with Gaussian model noise yields a Gaussian prior distribution for the trajectory x. Therefore, a natural first guess for the pendulum consists in the quadratic least-squares estimator $z_0 := x_{\text{MAP}}(y; \mu_0, P_0)$, where μ_0 and P_0 can be computed analytically as a function of the initial condition distribution and the pendulum's linear model. Starting from this estimate, we run the baselines and our neural assimilation algorithm.

Data We generate discrete trajectories $x^{(i)}$ of $T = 100$ time steps from the nonlinear pendulum dynamics with random initial conditions sampled in phase space, which is of dimension 2, hence $d = 2 \times 100 = 200$. The observations are generated by observing the pendulum's position at sparse time steps, with Gaussian observation noise $\xi \sim \mathcal{N}(0, \rho^2 I_m)$, with $\rho = 0.01$.

Experimental setup We train an adaptation operator to reconstruct the signal in one shot from z_0, following (3.3). At prediction time, we apply the trained neural assimilation map $A(z; y; \theta)$ to z_0 on a separate independent dataset.

Results Reconstruction samples are presented in Figure 1. While the linear model fails at reconstructing the trajectories outside the linearization zone (angle and momentum close to 0), one application of our neural assimilation operator accurately reconstructs the signal. The performances of the various methods are shown in Table 1. Although the pendulum is simple enough for all the methods to accurately reconstruct the signal, we see that the computational gain offered by a train neural network is considerable with respect to computing the physical model.

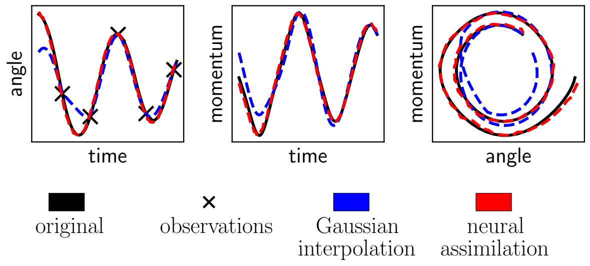

Figure 1. Reconstructed trajectories for the pendulum.

4.2. Lorenz 63

We now turn to a more complex system. The Lorenz system is a simplified physical model for for atmospheric convection (Lorenz, 1963). Three variables are governed by the following set of coupled nonlinear ordinary differential equations:

$$\begin{aligned} \frac{\mathrm{d}u_1}{\mathrm{d}t} &= \sigma(u_2 - u_1) \\ \frac{\mathrm{d}u_2}{\mathrm{d}t} &= \rho u_1 - u_2 - u_1 u_3 \\ \frac{\mathrm{d}u_3}{\mathrm{d}t} &= u_1 u_2 - \beta u_3. \end{aligned} \tag{4.1}$$

We set $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, values for which the system is known to exhibit chaotic solutions. We sample the initial conditions in the system's stationary distribution, following the experimental setup of (Rozet & Louppe, 2023).

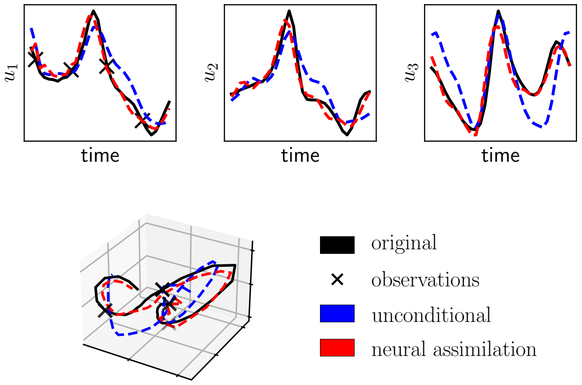

Figure 2. Reconstructed trajectories for the Lorenz 63 system.

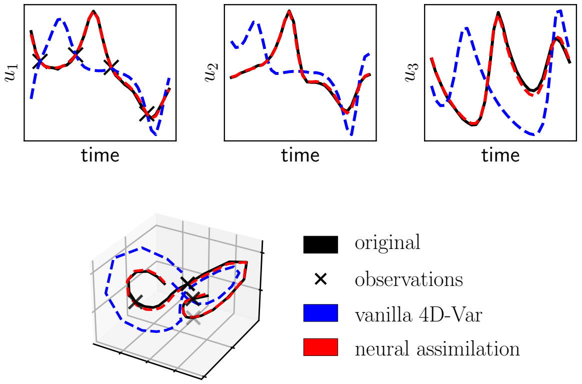

Figure 3. Output of 4D-Var from various initializations.

Data We generate datasets of trajectories by integrating (4.1) between time steps of length $\mathrm{d}t = 0.025$, and adding a small amount of Gaussian noise $\eta \sim \mathcal{N}(0, \mathrm{d}t I_3)$ at each time step. The number of time steps is $T = 32$, hence $d = 96$. We normalize each component of the trajectory to have zero mean and unit variance. The observations are sparse samples from the first component u_1 only, with observation noise of size 0.05. We take for the initial state estimate $z^{(i)}$ the maximum likelihood interpolation of $y^{(i)}$ under the moment-matching Gaussian distribution of $x^{(i)}$, which is the coarse Gaussian approximate of $p(x)$. More precisely, $z_0^{(i)} = x_{\mathrm{MAP}}(y^{(i)}; \hat{\mu}, \hat{P})$, with $\hat{\mu}$ and $\hat{P}$ the empirical mean and the empirical covariance of $\{x^{(i)}\}$. We define $\{z_k^{(i)}\}$ as in (3.4) with regular spacing $s_k = 1 - k/(\ell + 1)$. We take $\ell = 5$.

Experimental setup We train our neural assimilation operator to reconstruct the signal at different temperatures following (3.5). At prediction time, we apply Algorithm 2

Table 1. Performances of the various approaches. The computational time unit is the run-time of the fastest of the algorithms at prediction time.

Method	4D-Var	Cold diffusion	Neural assimilation
Pendulum error	0.1	0.15	0.12
Lorenz 63 error	0.9	1.1	0.5
Computational time	10	1	2

for the neural methods, along with the 4D-Var algorithm (Algorithm 1). Furthermore, in order to establish a link between our new neural method and traditional assimilation methods, we investigate how the output of the neural method, which is a priori not interpretable, may be transformed into a plausible physical signal. To do this, we correct these estimates with several iterations of 4D-Var on top of the neural estimate of the signal, until the objective function (2.1) becomes lower than 0.05. As a result, the new output is constrained to satisfy the physical model, but potentially at a lower cost than if we had started from scratch because the initialization that we provided is already close to the true signal.

Results Figure 2 shows reconstruction samples from the baselines and from our method, and Table 1 shows the average reconstruction error for the various methods. We can see that our neural data assimilation algorithm can reconstruct the signal while staying close to the observations. In contrast, the unconditional baseline cannot efficiently improve both the signal likelihood and the data fidelity. Compared to 4D-Var, our neural approach offers considerable computational gains, and good accuracy in these experiments. Further, we compare the reconstructed signals corrected by 4D-Var for an observation sample in Figure 3, where fixed number of 4D-Var iterations are applied to two different initializations: the Gaussian first-guess and the neural reconstruction of our algorithm. The initialization provided by our method allows to recover the original signal with very high accuracy by running few steps of 4D-Var on top of the neural estimate, while the 4D-Var algorithm with Gaussian initialization ("vanilla") leads to an inaccurate local minimum. Importantly, the improvement with respect to a Gaussian initialization is significant, both in terms of reconstruction error and in terms of number of iterations, as the 4D-var algorithm converged after 4 iterations from the neural initialization and 23 iterations from the Gaussian initialization. We further discuss the comparison between deep learning data assimilation approaches and 4D-Var in Section 6.

5. Related work

The state of the art methods for data assimilation are the 4D-Var algorithm (Le Dimet & Talagrand, 1986; Trémolet, 2007) and the ensemble Kalman filter (Evensen, 2003; Bocquet et al., 2014). The statistical component in these approaches lies in the definition of covariance matrices for the background state estimates, for the model and for the observations. The numerical cost of computing the physical model and its linear tangent local approximation may be considerable for large systems.

In recent years, several deep learning algorithms have been proposed for the data assimilation problem. Building on diffusion models (Ho et al., 2020), (Rozet & Louppe, 2023) propose a data assimilation method based on score-based diffusion. This approach proceeds in a plug-and-play fashion, and sampling from the posterior distribution relies on an approximation that is computed on the trained model. Among "end-to-end" deep learning approaches for data assimilation, the one that is closest related to ours is the 4DVarNet algorithm of (Fablet et al., 2021), which aims to directly train a neural network to minimize the reconstruction error. The complex prior modeled by the neural network is non-Gaussian, and estimating the maximum a posteriori reconstruction in this framework relies on non-convex optimization.

6. Conclusion

In this work, we have shown how deep learning methods may be applied to the data assimilation problem. Our neural method models in a coarse-to-fine fashion and is trained to minimize the reconstruction error. Importantly, we have shown how such a deep learning method may be used in combination with a traditional data assimilation method to enhance the reconstruction accuracy and reduce the computational time.

We believe that deep learning methods alone might not be accurate enough to completely outperform traditional physics-based approaches such as 4D-Var. While our neural approach had good reconstruction results on the presented simulated physical systems, it should be noted that the small size of these systems allows for the neural network to learn the stationary distribution from a reasonably small dataset. For real-life systems, it is unlikely that a neural network can accurately generalize the learned signal outside a training dataset, where the physics may be complex and fairly different from what the model has seen. In contrast, physics-based approaches are far more general, as the simulated physical laws are accurate everywhere in the state space. Therefore, using a deep learning algorithm to provide an approximate solution, and using it as an input to 4D-Var to reduce the number of iterations seems like a good trade-off benefitting the best of both words.

In future work, it would be interesting to apply our method to physical systems of larger scale, and to explore how the computational burden of data assimilation may be further reduced on such high-dimensional systems. Another important aspect that is crucial for data assimilation is uncertainty quantification, for which there has been recent progress in the deep learning community (Arcucci et al., 2021; Corso et al., 2022).

References

Arcucci, R., Zhu, J., Hu, S., and Guo, Y.-K. Deep data assimilation: integrating deep learning with data assimilation. *Applied Sciences*, 11(3):1114, 2021.

Bai, Y., Chen, W., Chen, J., and Guo, W. Deep learning methods for solving linear inverse problems: Research directions and paradigms. *Signal Processing*, 177:107729, 2020.

Bansal, A., Borgnia, E., Chu, H.-M., Li, J., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., and Goldstein, T. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36, 2024.

Bell, B. and Cathey, F. The iterated kalman filter update as a gauss-newton method. *IEEE Transactions on Automatic Control*, 38(2):294–297, 1993. doi: 10.1109/9.250476.

Ben Bouallègue, Z., Clare, M. C., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T., et al. The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 2024.

Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-López, F., Pedregosa, F., and Vert, J.-P. Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*, 2021.

Bocquet, M. et al. Introduction to the principles and methods of data assimilation in geosciences. *Notes de cours, École des Ponts ParisTech*, 2014.

Bouttier, F. and Courtier, P. Data assimilation concepts and methods march 1999. *Meteorological training course lecture series. ECMWF*, 718:59, 2002.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F., Fisher, M., and Pailleux, J. The ecmwf implementation of three-dimensional variational assimilation (3d-var). i: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

Datta, B. N. *Numerical linear algebra and applications*. SIAM, 2010.

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

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

Fisher, M., Leutbecher, M., and Kelly, G. On the equivalence between kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 131(613):3235–3246, 2005.

Fisher, M., Trémolet, Y., Auvinen, H., Tan, D., and Poli, P. *Weak-constraint and long-window 4D-Var*. ECMWF Reading, UK, 2012.

Fletcher, R. and Reeves, C. M. Function minimization by conjugate gradients. *The computer journal*, 7(2):149–154, 1964.

Gauss, C. F. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. FA Perthes, 1877.

Gratton, S., Lawless, A. S., and Nichols, N. K. Approximate gauss–newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106–132, 2007.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoyer, S., Yuval, J., Kochkov, D., Langmore, I., Norgaard, P., Mooers, G., and Brenner, M. P. Neural general circulation models for weather and climate. *AGU23*, 2023.

Jazwinski, A. H. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

Johnson, S. G. Notes on adjoint methods for 18.335. *Introduction to Numerical Methods*, 2012.

Kidger, P. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

Laumont, R., Bortoli, V. D., Almansa, A., Delon, J., Durmus, A., and Pereyra, M. Bayesian imaging using plug & play priors: when langevin meets tweedie. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022.

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

Liang, S. *Quantitative remote sensing of land surfaces*. John Wiley & Sons, 2005.

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

Mandel, J., Bergou, E., and Gratton, S. 4dvar by ensemble kalman smoother. *arXiv preprint arXiv:1304.5271*, 2013.

Ménard, R. and Daley, R. The application of kalman smoother theory to the estimation of 4dvar error statistics. *Tellus A*, 48(2):221–237, 1996.

Mukherjee, S., Carioni, M., Öktem, O., and Schönlieb, C.-B. End-to-end reconstruction meets data-driven regularization for inverse problems. *Advances in Neural Information Processing Systems*, 34:21413–21425, 2021.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9):4349–4383, 2021.

Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer, 1999.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Rangayyan, R. M. and Krishnan, S. *Biomedical signal analysis*. John Wiley & Sons, 2024.

Rozet, F. and Louppe, G. Score-based data assimilation. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 40521–40541. Curran Associates, Inc., 2023.

Talagrand, O. and Courtier, P. Variational assimilation of meteorological observations with the adjoint vorticity equation. i: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328, 1987.

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