DEEP BAYESIAN FILTER FOR BAYES-FAITHFUL DATA ASSIMILATION

Anonymous authors
Paper under double-blind review

ABSTRACT

State estimation for nonlinear state space models (SSMs) is a challenging task. Existing assimilation methodologies predominantly assume Gaussian posteriors on physical space, where true posteriors become inevitably non-Gaussian. We propose Deep Bayesian Filtering (DBF) for data assimilation on nonlinear SSMs. DBF constructs new latent variables h_t in addition to the original physical variables z_t and assimilates observations o_t. By (i) constraining the state transition on the new latent space to be linear and (ii) learning a Gaussian inverse observation operator $r(h_t|o_t)$, posteriors remain Gaussian. Notably, the structured design of test distributions enables an analytical formula for the recursive computation, eliminating the accumulation of Monte Carlo sampling errors across time steps. DBF trains the Gaussian inverse observation operators $r(h_t|o_t)$ and other latent SSM parameters (e.g., dynamics matrix) by maximizing the evidence lower bound. Experiments demonstrate that DBF outperforms model-based approaches and latent assimilation methods in tasks where the true posterior distribution on physical space is significantly non-Gaussian.

1 INTRODUCTION

Data assimilation (DA) is a crucial technique across various scientific domains. Its primary objective is to estimate the trajectory and current state of a system by integrating an imperfect model with partially informative observations. Specifically, given a series of observations T time steps $o_{1:T}$, the goal is to infer the posterior distribution of the system's physical variables z_t: $p(z_t|o_{1:t})$. DA has been widely applied in fields such as weather forecasting (Hunt et al., 2007; Lorenc, 2003; Andrychowicz et al., 2023), ocean research analysis (Ohishi et al., 2024), sea surface temperature prediction (Larsen et al., 2007), seismic wave analysis (Alfonzo & Oliver, 2020), multi-sensor fusion localization (Bach & Ghil, 2023), and visual object tracking (Awal et al., 2023).

A key challenge in DA arises from the non-Gaussian nature of the posterior distributions $p(z_t|o_{1:t})$, which results from the inherent nonlinearity in both the system dynamics and observation models. Despite this, many operational DA systems, such as those used in weather forecasting, rely on methods like the ensemble Kalman Filter (EnKF) (Evensen, 1994; Bishop et al., 2001) for sequential state filtering (i.e., $p(z_t|o_{1:t})$) and the four-dimensional variational method (4D-Var) for retrospective state analysis (i.e., $p(z_t|o_{1:T}), t < T$). These approaches assume Gaussianity in their test distributions $q(z_t|o_{1:t})$ or $q(z_t|o_{1:T})$, a simplification driven by computational constraints. While exact methods such as bootstrap Particle Filters (PF) or sequential Monte Carlo (SMC) (Chopin & Papaspiliopoulos, 2020; Daum & Huang, 2007; Hu & van Leeuwen, 2021) could compute the true posterior, their performance degrades significantly when the number of particles is insufficient (Beskos et al., 2014). This issue is exacerbated in high-dimensional systems, making SMC approaches impractical for many physical problems.

To address these limitations, we propose a novel variational inference approach called Deep Bayesian Filtering (DBF) for posterior estimation. Our strategy consists of two main components: (i) constraining the test distribution to remain Gaussian to ensure computational tractability, and, in cases where the original dynamics are nonlinear, (ii) leveraging a nonlinear mapping to enhance the expressive capability of the test distribution. The DBF methodology diverges into two paths depending on the nature of the system dynamics, whether linear Gaussian or nonlinear:

Linear dynamics When the system's dynamics $p(z_{t+1}|z_t)$ are linear, DBF assumes Gaussianity in the original space, similar to traditional methods. However, DBF introduces the concept of the inverse observation operator (IOO; see also Frerix et al. 2021) to construct Gaussian test distributions $q(z_t|o_{1:t})$. The IOO, along with any unknown system parameters, are trained to minimize the Kullback-Leibler divergence between the test distribution $q(z_t|o_{1:t})$ and the true posterior $p(z_t|o_{1:t})$. The IOO and the system parameters are trained without teacher signals z_t.

Nonlinear dynamics In the more common case of nonlinear dynamics, DBF operates in a latent space, assuming Gaussianity in the latent variables h_t. The original physical variables are recovered through a nonlinear mapping function ϕ, implemented via neural networks (NNs). This nonlinear mapping allows for a more flexible representation of the test distribution $q(z_t|o_{1:t})$. The IOO and other parameters are trained in a supervised manner (i.e., z_t is used during training).

For state space models (SSMs) with nonlinear dynamics, DBF functions as a variational autoencoder (VAE) that adheres to the Markov property. Posterior distributions of the latent variables h_t are expressed in a Bayesian framework. This approach is closely related to dynamical VAEs (DVAEs, Girin et al. 2021 for a review), which use VAEs to model time-series data. However, DBF distinguishes itself by its posterior design. Unlike DVAEs, where Monte Carlo sampling is required for inference (see Sec. 2.6.1), DBF allows for the analytical computation of the prediction step, recursively computing posteriors through closed-form expressions.

When applied to problems with nonlinear or unknown dynamics, DBF can be interpreted as learning the Koopman operator (Koopman, 1931) using NNs. The discovery of such latent spaces and operators through machine learning has been extensively studied (Takeishi et al., 2017; Lusch et al., 2018; Azencot et al., 2020) and will be experimentally validated through the handling of nonlinear filtering tasks involving chaotic dynamics.

Key contributions of the proposed DBF methodology include:

- DBF is the first VAE-based model for time-series data that maintains a posterior structure faithful to the Markov property in SSMs.
- For systems with linear dynamics, DBF extends the Kalman Filter (KF) to handle nonlinear observations through learnable NNs. The training process enables the model to infer unknown system parameters directly from data (see Section 3.1).
- For nonlinear dynamics, DBF constructs a new latent space for data assimilation, allowing for the analytical integration of time steps and preventing the accumulation of Monte Carlo sampling errors. This is accomplished through the application of Koopman operator theory, which ensures that the model's representational power is maintained, as long as the latent space is sufficiently high-dimensional (see Sections 3.2 and 3.3).
- As a generative model, DBF estimates the uncertainty of the physical variables z_t, in contrast to regression methods that yield only point estimates (see Section 3.2 and Fig. 3).
- The linear constraint on dynamics stabilizes the training process, which is known to be unstable in standard recurrent NNs (see Section 3.3 and Fig. 6).

DBF has demonstrated superior performance over classical DA algorithms and latent assimilation methods in scenarios with highly non-Gaussian posteriors, particularly in the presence of strongly nonlinear observation operators or large observation noise.

2 Method

2.1 Inference of physical variables in a state-space model

A physical system is defined by variables z_t, with its evolution described by the dynamics model $p(z_{t+1}|z_t) = \mathcal{N}(z_{t+1}; f(z_t), Q)$, where $\mathcal{N}(x|\mu, \Sigma)$ denotes a Gaussian whose mean and covariance are μ and Σ. The nonlinear function f is the dynamics operator and Q is the system covariance. The Markov property holds, as z_{t+1} depends only on z_t. An observation model $p(o_t|z_t) = \mathcal{N}(o_t; h(z_t), R)$ relates observations to physical variables via the observation operator h and covariance R. The panel (a) of Fig. 1 shows the system's graphical model. The objective of sequential DA is to compute the posterior of z_t given $o_{1:t}$.

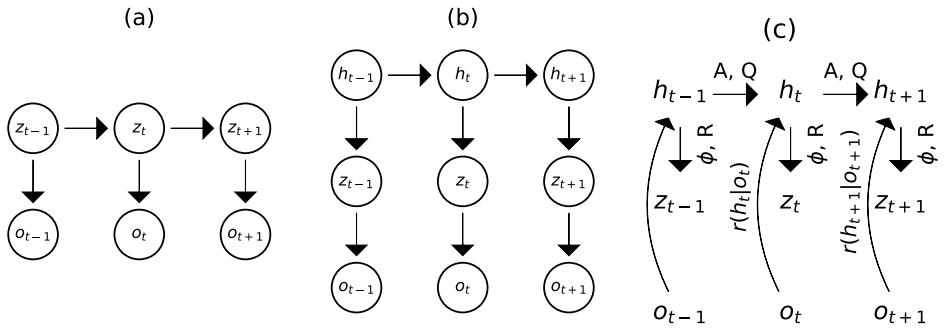


Figure 1: Panel (a) shows the graphical model for the simplest SSM. If the dynamics of the original SSM is linear, DBF assimilates on that space. Panel (b) shows the graphical model for the SSM assumed for SSM with nonlinear dynamics. Panel (c) shows the inference structure of our methodology for SSM with nonlinear dynamics.

2.2 KF FOR LINEAR DYNAMICS, LINEAR OBSERVATIONS

In the KF, the dynamics and observation models are both linear Gaussian, with the mean being linear with respect to the conditional variable and a constant covariance matrix. Given that the dynamics and observation operators f, h are linear, we can represent them using matrices A and C, respectively. All matrices (A, C, Q, and R) are constant. The filter distribution $p(z_t|o_{1:t})$ remains Gaussian, provided that the initial distribution $p(z_1)$ is Gaussian. We can recursively compute the posterior parameters (means μ_t and covariance matrices Σ_t) using the following equations:

$$\mu_t = \Sigma_t(A\Sigma_{t-1}A^T + Q)^{-1}A\mu_{t-1} + K_t(o_t - H\mu_t), \tag{1}$$

$$\Sigma_t^{-1} = (A\Sigma_{t-1}A^T + Q)^{-1} + HR^{-1}H^T, \tag{2}$$

where $K_t = (A\Sigma_{t-1}A^T + Q)H^T(H(A\Sigma_{t-1}A^T + Q)H^T + R^{-1})^{-1}$ is the Kalman Gain.

2.3 DBF FOR LINEAR DYNAMICS, NONLINEAR OBSERVATIONS

In this scenario, Gaussianity of the test distribution is lost during the KF update step. We introduce an inverse observation operator (IOO) $r(z_t|o_t)$ (see also Frerix et al. 2021):

$$p(z_t|o_{1:t}) = \frac{p(o_t|z_t)p(z_t|o_{1:t-1})}{p(o_t|o_{1:t-1})} \propto \frac{r(z_t|o_t)}{\rho(z_t)}p(z_t|o_{1:t-1}), \tag{3}$$

where $r(z_t|o_t) = \frac{p(o_t|z_t)\rho(z_t)}{\int p(o_t|z_t)\rho(z_t)dz_t}$ and $\rho(z_t)$ is a virtual prior. By approximating both the IOO and the virtual prior as Gaussians, $r(z_t|o_t) = \mathcal{N}(f_\theta(o_t), G_\theta(o_t))$ and $\rho(z_t) = \mathcal{N}(m, V)$, respectively, the posterior $q(z_t|o_{1:t})$ can be analytically computed as a Gaussian, where the mean μ_t and covariance Σ_t are given as:

$$\mu_t = \Sigma_t(A\Sigma_{t-1}A^T + Q)^{-1}A\mu_{t-1} + G_\theta(o_t)^{-1}f_\theta(o_t) - V^{-1}m, \tag{4}$$

$$\Sigma_t^{-1} = (A\Sigma_{t-1}A^T + Q)^{-1} + G_\theta(o_t)^{-1} - V^{-1}, \tag{5}$$

where $f_\theta(o_t)$ and $G_\theta(o_t)$ are NNs with parameters θ, and m and V are constants set to $m = 0$ and $V = 10^8 I$. These values bias the NNs' outputs without affecting performance.

The recursive formula for the exact posterior (Equation 3) requires no approximation. Thus, DBF computes the exact posterior when the true IOO $r_{\text{true}}(h_t|o_t)$ is Gaussian, i.e., the SSM is a LGSS. In that case, the posterior update formula agrees with the KF (see Equations 1, 2 and 4, 5). The key difference is that nonlinear functions are applied to both the mean, $f_\theta(o_t)$, and the covariance, $G_\theta(o_t)$. In the KF, $f_\theta(o_t)$ is linear, and $G_\theta(o_t)$ is a constant covariance matrix (see Equations 1 and 2). The dependence of $G_\theta(o_t)$ on observations provides flexibility in adjusting the impact of the new observation on the state estimation. The importance of adjusting the internal state updates based on observations has also been discussed in recent SSM-based approaches (Gu & Dao, 2023).

2.4 DBF FOR NONLINEAR DYNAMICS, LINEAR/NONLINEAR OBSERVATIONS

In this scenario, the Gaussianity of the test distribution is lost during the predict step, making it impossible to apply the original dynamics over the physical variables z_t. Therefore, we introduce a new set of latent variables h_t and assume a dynamics model over h_t: $p(h_{t+1}|h_t) = \mathcal{N}(h_{t+1}|Ah_t, Q)$ (see panel (b) in Fig. 1). The IOO maps observations into the latent variables h_t: $r(h_t|o_t)$. The recursive formula follows Equations 4 and 5. To retrieve the distribution of the original physical variables z_t, we introduce an emission model $p(z_t|h_t) = \mathcal{N}(z_t; \phi(h_t), R)$, where ϕ is represented by a NN. By marginalizing over h_t with this emission model, a trained DBF can generate samples of z_t that follow the test distribution $q(z_t|o_{1:t})$ given observations $o_{1:t}$.

Although the dynamics operator A for the latent variables h_t is linear, it is able to express any nonlinear dynamics if the latent space is sufficiently high-dimensional. The Koopman operator (Koopman, 1931) provides a framework for representing nonlinear systems by mapping observables—functions of the system's state—into a higher-dimensional space where the dynamics become linear. Mathematically, for a dynamical system $z_{t+1} = f(z_t)$, the Koopman operator $\mathcal{K}$ is a linear operator acting on a set of observables $g(z)$, such that $\mathcal{K}g(z_t) = g(f(z_t))$. This allows the system to be represented as $h_{t+1} = Ah_t$ in the latent space h_t, where A is the dynamics matrix learned by DBF. While the physical dynamics $f(z)$ are nonlinear, the Koopman operator ensures the existence of an embedding that linearizes the dynamics, enabling recursive computation of test distributions. Discovering such embeddings in finite dimensions has been widely studied (Takeishi et al., 2017; Lusch et al., 2018; Azencot et al., 2020). In high-dimensional simulations, the true degrees of freedom are often far fewer than the simulated variables, making surrogate modeling with the Koopman operator a promising approach to reducing computational costs.

2.5 TRAINING

When assimilating in the physical space (i.e., when the dynamics are linear), we train the IOO (i.e., f_θ and G_θ) by optimizing the evidence lower bound (ELBO) without using the teacher signal z_t:

$$\begin{aligned} \log p(o_{1:T}) &= \sum_{t=1}^{T} \log p(o_t|o_{1:t-1}) \geq -\mathcal{L}_{\text{ELBO}}, \\ \mathcal{L}_{\text{ELBO}} &= -\sum_{t=1}^{T} \int q(h_t|o_{1:t}) \log p(o_t|h_t) dh_t + KL[q(h_t|o_{1:t})||q(h_t|o_{1:t-1})], \end{aligned} \tag{6}$$

where $KL[p||q]$ denotes the Kullback-Leibler divergence between distributions p and q. If the SSM contains any unknown parameters, we can train these parameters as well.

For SSMs with nonlinear or unknown dynamics, we have two approaches:

Strategy 1 Pretrain the Koopman operator, which consists of the nonlinear mapping from z_t to h_t, the linear dynamics between h_t and h_{t+1} represented by matrix A, and the reverse nonlinear mapping from h_t to z_t denoted by ϕ. With these components (A and ϕ) of the Koopman operator, the method designed for linear dynamics can be applied. For pretraining, we require samples of z_t or the SSM for the physical variables to generate these samples. Pairs of z_t and o_t are not necessary, as the training for the linear dynamics (A and ϕ) and the IOO ($r(h_t|o_t)$) can be performed separately.

Strategy 2 Train all components (the matrix A, the stochastic mapping $p(z_t|h_t) = \mathcal{N}(z_t; \phi(h_t), \text{diag}[\sigma^2])$, and the IOO) simultaneously. In this case, samples of (z_t, o_t) pairs or the SSM for both physical and observation variables to generate these sample pairs are required. The parameters are optimized by maximizing a joint ELBO, $\mathcal{L}_{\text{ELBO,joint}}$, via supervised training:

$$\begin{aligned} \log p(o_{1:T}, z_{1:T}) &= \sum_{t=1}^{T} \log p(o_t, z_t|o_{1:t-1}, z_{1:t-1}) \geq -\mathcal{L}_{\text{ELBO,joint}}, \\ \mathcal{L}_{\text{ELBO,joint}} &= -\sum_{t} \int q(h_t|o_{1:t}) \log p(z_t|h_t) dh_t + KL[q(h_t|o_{1:t})||q(h_t|o_{1:t-1})]. \end{aligned} \tag{7}$$

Here, we have neglected $\log p(o_t|z_t)$ from $\log p(o_t, z_t|h_t) = \log p(o_t|h_t, z_t) + \log p(z_t|h_t) = \log p(o_t|z_t) + \log p(z_t|h_t)$. Since $\log p(o_t|z_t)$ does not depend on h_t and its distributions, it is

sufficient to include $\log p(z_t|h_t)$. We have replaced $q(h_t|o_{1:t}, z_{1:t})$ with its special case $q(h_t|o_{1:t})$ as our objective is to give the best estimate of z_t given observations $o_{1:t}$.

2.6 Related works

We discuss related works in the order of relevance.

2.6.1 Dynamical variational autoencoders

DVAEs (see Girin et al. 2021 for a review) are models closely related to DBF for nonlinear dynamics. Both incorporate time-series architecture in a VAE, but there are two key differences: (i) the posterior design and realization of the dynamics step, and (ii) the loss function.

posterior design Our strategy for the test distribution is to incorporate an appropriate architecture that reflects the Markov property in the time dimension of the test distribution. The IOO, $r(h_t|o_t)$, and the linear dynamics model serve as key instruments in constructing the test posterior distributions. A distinguishing feature of our methodology is that each component's role is defined with respect to the Markov property of the state-space model (SSM) and is clearly differentiated from other components involved in posterior construction. For example, the IOO influences only the update step and does not affect the prediction step. We refer to this methodology as "Bayes-Faithful" due to its tailored design for SSMs that exhibit the Markov property.

In contrast, the test posterior distributions in DVAEs are constructed using RNNs. The complexity of the transition model prevents the analytical computation of latent variables across time steps. As a result, these values can only be estimated via Monte Carlo sampling. Consequently, during inference, successive Monte Carlo sampling ("cascade trick"; Girin et al. 2021) becomes unavoidable.

loss function DBF takes the ELBO from factorized density $\log p(o_t|o_{1:t-1})$ in $\log p(o_{1:T}) = \sum_t \log p(o_t|o_{1:t-1})$:

$$\log p(o_{1:T}) \quad \geq \quad \sum_{t=1}^{T} (E_{q(h_t|o_{1:t})}[\log p(o_t|h_t)] - KL[q(h_t|o_{1:t})|q(h_t|o_{1:t-1})]). \tag{8}$$

On the other hand, DVAEs take the ELBO from probability density with all the observations at once.

$$\log p(o_{1:T}) \quad \geq \quad E_{q(h_{1:T}|o_{1:T})}[\log p(h_{1:T}, o_{1:T}) - \log q(h_{1:T}|o_{1:T})]. \tag{9}$$

Therefore, DBF seeks for the filtered distributions $q(h_t|o_{1:t})$ whereas DVAEs model the smoother distributions $q(h_t|o_{1:T})$. Again, for DVAEs, to evaluate the expected values in Equation 9, we need to undergo successive Monte-Carlo sampling over T variables ($h_{1:T}$).

Assuming linear Gaussian dynamics and a Gaussian IOO, DBF allows for the analytical integration of $q(h_t|o_{1:t-1})$, resulting in a structured encoder. This structured posterior enables the recursive computation of the filtered distribution $q(h_t|o_{1:t})$ without relying on Monte Carlo sampling, setting it apart from other DVAEs. By constraining the dynamics to be linear, DBF ensures exact integration without the accumulation of Monte Carlo sampling errors across time steps.

Moreover, the linear assumption helps DBF mitigate the instability issues commonly faced when training standard RNNs. SSMs are increasingly favored for modeling long-range dependencies (Gu & Dao, 2023). S4 (Gu et al., 2022) utilizes linear dynamics in the latent space and learns the dynamics, proposing an efficient computation algorithm that outperforms transformers on datasets with long-range dependencies. LS4 (Zhou et al., 2023) extends S4 by introducing stochasticity through a VAE-like structure. Both LS4 and DBF employ linear SSMs and Gaussian posterior approximations, but DBF updates the mean and covariance using a recursive formula based on Bayes' rule, while LS4 replaces recurrence with convolutions, forgoing a recursive approach.

2.6.2 KF-based methods

Various approaches have been explored to address LGSS limitations, including linearizing the model via first-order approximations like the extended Kalman Filter (EKF), approximating populations with a Gaussian distribution in the ensemble Kalman Filter (EnKF; Evensen 1994), and using NNs

to approximate the Kalman gain (Revach et al., 2022). The EnKF and its variants (e.g., ETKF; Bishop et al. 2001) are commonly used in real-time data assimilation for weather forecasting. However, these methods rely on the KF's posterior update equations, limiting the expressivity of the distributions they can represent. Additionally, computations for covariance matrices become challenging in high-dimensional spaces, requiring specialized techniques for computational efficiency.

2.6.3 Sampling-based methods

The Particle Filter is a popular method for assimilating any posterior. However, achieving adequate particle density in high-dimensional state spaces poses significant challenges. Insufficient density of particles leads to particle degeneracy, where few particles explain the observed data (Beskos et al., 2014). In contrast, DBF directly learns to position density through the IOO, offering advantages for high-dimensional tasks. The Particle Flow Filter (PFF; Daum & Huang 2007; Hu & van Leeuwen 2021) addresses particle degeneracy by moving particles according to gradient flow and effectively scales to nonlinear SSMs with hidden state dimensions up to 1000 (Hu & van Leeuwen, 2021).

2.6.4 Approximate MAP estimation method

MAP estimation is used to identify the high-density point of the posterior in high-dimensional space, such as in weather forecasting Lorenc (2003); Frerix et al. (2021). Even if the computation of the posterior $p(h_t|o_{1:t})$ is intractable, we can optimize $\log p(h_t|o_{1:t}) = \log p(o_t|h_t) + \log p(h_t|o_{1:t-1})$ if we can describe $p(o_t|h_t)$ and $p(h_t|o_{1:t-1}) = \int p(h_t|h_{t-1})p(h_{t-1}|o_{1:t-1})dh_{t-1}$ explicitly. In practice, we cannot access $p(h_{t-1}|o_{1:t-1})$ and therefore the integral $\int p(h_t|h_{t-1})p(h_{t-1}|o_{1:t-1})dh_{t-1}$, so we only compute the mean. The downside is that sequential computation of the covariance matrix of $p(h_t|o_{1:t-1})$ is impossible.

2.6.5 NN-based PDE surrogate

Recently, there have been attempts to approximate partial differential equations (PDEs) using NNs. We also tried one of the latest methods, PDE-refiner (Lippe et al., 2023), in the experiments of this study. However, its performance was poor, and we decided not to include it in the experiments. We suspect this is because PDE-refiner was designed for constructing PDE surrogates and did not account for noisy observations, making it susceptible to noise. We confirmed that it produces good predictions under noiseless observations.

3 Experiments

We evaluate the performance of DBF on three tasks: a linear dynamics problem (moving MNIST) and two nonlinear dynamics problems (double pendulum and Lorenz96).

Linear dynamics: moving MNIST In the moving MNIST task, the goal is to identify the images, positions, and velocities of two handwritten digits as they move within the observed frames. While the dynamics of these digit images and their observation processes are provided, the actual images, positions, and velocities are not available, making supervised learning impossible. DBF assimilates directly in the physical space. It trains the parameters of the IOO, f_θ and G_θ, which are represented by NNs, along with the pixel values of the embedded images. The results are compared against conventional DA methods (EnKF, ETKF, PF) that also operate in the physical space.

Nonlinear dynamics: double pendulum and Lorenz96 For nonlinear dynamics problems, such as the double pendulum and Lorenz96, DBF constructs a new latent space in addition to the original physical space. Here, we took Strategy 2 in Sec.2.5 for the training: we simultaneously train NNs for the IOO, nonlinear observation operator ϕ, the dynamics matrix A, and the emission model's standard deviation. We compare the performance of DBF with the classical DA algorithms (EnKF, ETKF, PF), state-of-the-art assimilation methodologies (PFF Daum & Huang 2007; Hu & van Leeuwen 2021, KalmanNet Revach et al. 2022), and DVAE-based approaches (deep Kalman Filter; DKF, Krishnan et al. 2015; 2016, variational recurrent neural network; VRNN, Chung et al. 2015, and stochastic recurrent neural network; SRNN, Fraccaro et al. 2016). DBF and other DVAEs are trained by optimizing the evidence lower bound (ELBO), as described in Sec. 2.5.

For all the experiments, we generate random initial conditions and let them evolve via the dynamics of the problem. Synthetic observation data are generated by applying the observation operator and additive noise. Noise levels and observation operators are further explained in the section for each problem and in the appendix. Training details, including hyperparameters and architectures, are also provided in the appendix.

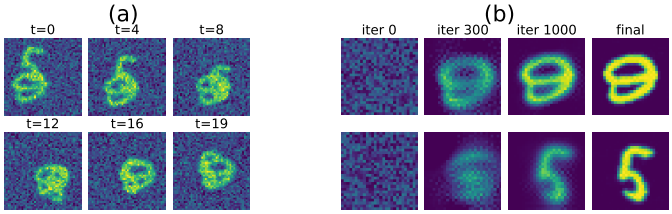
3.1 Linear dynamics: two-body moving MNIST

This experiment demonstrates DBF's ability to handle linear dynamics where key parameters of the observation operator are unknown. The dataset consists of 2D figures containing two embedded images, each moving at a constant speed and bouncing off frame edges. The system's physical state is described by eight variables: the positions (x, y) and velocities (v_x, v_y) of the two embedded images. The dynamics matrix is block-diagonal, composed of four (two-body times two dimensions) translation matrices, A_{tr}: $A_{tr} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, z_t = \begin{pmatrix} x_t \\ v_{x_t} \end{pmatrix}$. Observations are corrupted by additive Gaussian noise with a standard deviation of $\sigma = 50$ per pixel, where the original pixel values range from 0 to 255 (see panel (a) of Fig. 2 for an example of the data provided).

The aim is to show that DBF can track the linear dynamics while estimating unknown system parameters. DBF learns the pixel values of the embedded images from noisy observations, while maintaining consistency with physical motion. The observation model contains $1{,}568$ unknown parameters, corresponding to the number of pixels in the images. In classical DA algorithms, it is not possible to train unknown system parameters. However, it may be possible to infer these parameters by incorporating them as new physical dimensions. We have adopted this strategy for classical DA algorithms (EnKF, ETKF, and PF). We tested at three different noise levels (σ_{sys} of 1, 0.1, and 0.01). DVAEs are not comparable as they need to undergo supervised training. We were unable to compare with KalmanNet, as its input dimensionality, $x^2_{\text{dim}} = (44 \times 44)^2$, exceeded the available GPU memory.

Fig. 2 summarizes the experiment. Panel (a) shows an example from the test set, illustrating the challenges posed by strong noise and overlapping images. Panel (b) presents the DBF learning process. In the rightmost table, we compare the success rates of DBF against model-based approaches (EnKF, ETKF, PF). We define success as achieving a root-mean-square error (RMSE) of less than 1.0 for both position (x_1, y_1, x_2, y_2) and velocity $(v_{x_1}, v_{y_1}, v_{x_2}, v_{y_2})$ of the two digits over the final ten steps. DBF successfully performs assimilation without explicit knowledge of the images, while all the other model-based approaches fail. The KF-inspired approaches (EnKF, ETKF) failed because of very strong non-Gaussianity in the observation process and the high system dimension. Similarly, PF underperformed because the number of particles ($10{,}000$) was insufficient for the problem dimension ($z_{\text{dim}} = 8$ and two digits images $2 \times 28 \times 28 = 1{,}568$). Figures for visualizing the assimilation results for all the algorithms are given in the appendix (Fig. 7).

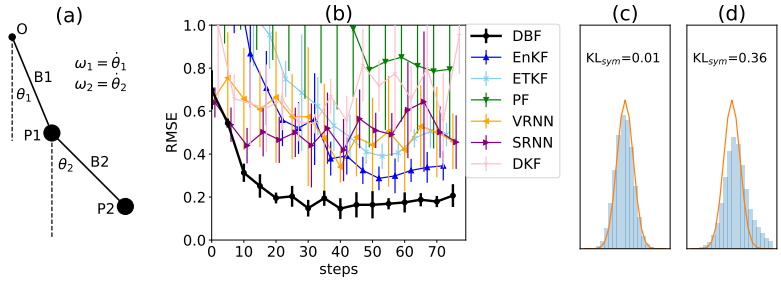
Panel (b) of Fig. 2 illustrates the evolution of the estimated figures. Initially, DBF assumes two random shapes. As training progresses, it first identifies one of the numbers ("9") and subsequently detects the second shape ("5"). By the end of the training process, DBF nearly perfectly estimates the parameters of the observation model, including the positions of the figures, which is crucial for adjusting their reflective behavior. The ability to learn system parameters through gradient descent sets DBF apart from traditional model-based approaches.



Method	Success rate
DBF	100% (50/50)
EnKF	0% (0/50)
ETKF	0% (0/50)
PF	0% (0/50)

Figure 2: Figures from the two-body Moving MNIST experiments. Panel (a) displays examples of the observation data. Panel (b) illustrates the evolution of the observation model parameters (the embedded images) during training. The table compares the success rates of four methodologies.

3.2 NONLINEAR DYNAMICS 1: DOUBLE PENDULUM



$\sigma = 0.1$	KL_{sym}
DBF	0.02
EnKF	10.2
ETKF	0.12

Figure 3: A schematic figure (panel a) and results for double pendulum experiments. Panel (b) shows the RMSE evolution during assimilation. Panels (c) and (d) show example histograms for normalized errors in DBF and ETKF samples compared against the unit Gaussian $\mathcal{N}(x; \mu = 0, \sigma^2 = 1)$. The small table compares the Jeffreys divergence of normalized errors and the unit Gaussian between DBF, EnKF, and ETKF predictions.

Table 1: RMSE at the final ten steps of assimilation in double pendulum experiments.

	$\sigma = 0.1$		$\sigma = 0.3$		$\sigma = 0.5$	
	θ	ω	θ	ω	θ	ω
DBF	**0.03 ± 0.01**	**0.21 ± 0.04**	**0.05 ± 0.02**	**0.26 ± 0.05**	**0.06 ± 0.01**	**0.36 ± 0.04**
EnKF	0.05 ± 0.00	0.33 ± 0.07	0.14 ± 0.01	0.71 ± 0.09	0.24 ± 0.01	1.17 ± 0.22
ETKF	0.05 ± 0.01	0.46 ± 0.08	0.22 ± 0.05	1.41 ± 0.41	0.36 ± 0.08	2.70 ± 1.25
PF	0.05 ± 0.00	0.63 ± 0.24	0.21 ± 0.14	1.41 ± 1.30	0.32 ± 0.08	2.36 ± 2.29
PFF	1.27 ± 0.29	1.04 ± 0.15	NA	5.99 ± 1.09	5.88 ± 0.67	NA
KNet	NA	NA	NA	NA	NA	NA
VRNN	0.04 ± 0.01	0.44 ± 0.19	0.06 ± 0.02	0.35 ± 0.14	0.08 ± 0.04	0.40 ± 0.16
SRNN	0.05 ± 0.02	0.52 ± 0.18	0.06 ± 0.02	0.44 ± 0.08	0.08 ± 0.03	0.52 ± 0.22
DKF	0.12 ± 0.02	2.70 ± 0.28	0.17 ± 0.03	2.61 ± 0.74	0.23 ± 0.04	2.61 ± 0.56

This section presents our experiments with a double pendulum system, selected for its nonlinear and chaotic behavior. The pendulum consists of two 1 kg masses, P1 and P2, connected by two 1 meter bars, B1 and B2. One end of the bar B1 is fixed at the origin ("O"), with the other end attached to P1. Mass P2 is connected to P1 via bar B2. A schematic of the setup is shown in panel (a) of Fig. 3.

We use the angles θ_1 and θ_2, and the two angular velocities, ω_1 and ω_2, as target physical variables. The latent dimension for DBF, VRNN, SRNN, and DKF is set to 50. Observation data consists of the two-dimensional spatial positions of masses P_1 and P_2, corrupted by Gaussian noise. The observation operator combines trigonometric functions for θ_1 and θ_2, creating a highly nonlinear relationship. Experiments are conducted with noise levels of $\sigma = 0.1, 0.3$, and 0.5 meters, with a time step of 0.03 seconds between observations. In the emission model $p(z_t|h_t)$, we assume von Mises distributions for θ_1 and θ_2, while ω_1 and ω_2 follow Gaussian distributions.

Table 1 presents the RMSE between the physical variables and the mean of the filtered distribution. Training for KalmanNet was unsuccessful under all conditions. For the DVAEs, we exclude failed initial conditions when calculating the RMSE. DBF outperforms both model-based and latent assimilation methods across all settings, showing significant improvements in estimating ω, which cannot be inferred from a single observation. Fig. 3 (b) illustrates an example of RMSE evolution during assimilation, where DBF consistently outperforms the other methods. The assimilation of ω occurs within the first ~ 20 steps, maintaining an excellent estimation accuracy throughout the experiment.

A key feature of DBF is its ability to generate samples of z_t and assess the uncertainty in state estimates. To evaluate this capability, we analyze the distributions of normalized errors defined as $\epsilon_{norm,t,i} = (z_{t,sample,i} - z_{t,i})/\delta_i$, where $z_{t,i}$ represents the true value of dimension i at time t, and δ_i is the standard deviation of $z_{t,sample,i}$. We collect $\epsilon_{norm,t,i}$ across all time steps, focusing on $i = \omega_1$ and $i = \omega_2$, since θ_1 and θ_2 follow von Mises distributions. If the uncertainty estimates are accurate, $\epsilon_{norm,t,i}$ should approximate a Gaussian distribution with a standard deviation

of one. To quantify the accuracy, we compute the symmetric KL divergence (Jeffreys divergence) $KL_{sym}[p, q] = (KL[p||q] + KL[q||p])/2$ between the histogram of $\epsilon_{norm,t,i}$ and a unit Gaussian. DBF exhibits very low KL_{sym} values, indicating accurate error estimation. Panels (c) and (d) display example histograms of $\epsilon_{norm,t,i}$ for DBF and ETKF.

3.3 Nonlinear dynamics 2: Lorenz96

In the final experiment, we focus on state estimation in the Lorenz96 model (Lorenz, 1995), a benchmark for testing data assimilation algorithms on noisy, nonlinear observations. The Lorenz96 model describes the evolution of a one-dimensional array of variables, each representing a physical quantity over a spatial domain, like an equilatitude circle. The dynamics are governed by the following coupled ordinary differential equations:

$$\frac{dz_i}{dt} = (z_{i+1} - z_{i-2})z_{i-1} - z_i + F, \quad i = 1, \ldots, N, \quad (10)$$

where z_i is the value at grid point i, N is the number of grid points, and F is external forcing. For our experiments, we take $F = 8$ and $N = 40$.

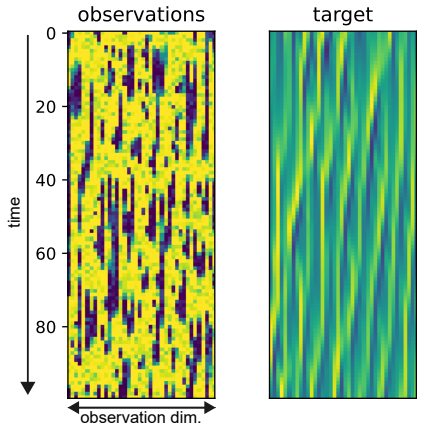


Figure 4: A Hovmöller diagram for one of data in the test set. The observation operator is nonlinear, $o_{t,j} = min(z_{t,j}^4, 10) + \epsilon$.

We consider two observation operators. The first adds Gaussian noise to direct observations: $o_{t,j} = z_{t,j} + \epsilon$, with noise levels $\sigma = 1, 3, 5$. The second uses a nonlinear operator: $o_{t,j} = \min(z_{t,j}^4, 10) + \epsilon$, with the same noise levels. The dynamic range of $z_{t,j}$ is around ± 10, and observations are capped at 10 when $z_{t,j}$ exceeds 1.8. This makes it highly challenging for classical DA methods, as each observation offers limited information. The filter must integrate data over long timesteps, where nonlinear dynamics distort the probability distribution. Fig. 4 illustrates observations and target values. All models use 80 observation steps with a 0.03 time interval. The latent dimension for DBF, VRNN, SRNN, and DKF is set to 800. For further details for the experiment, see Sec. A.3.

Table 2: RMSE at the final ten steps of assimilation in Lorenz96 experiments.

	direct observation			nonlinear observation		
	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$
DBF	0.53 ± 0.04	**0.82 ± 0.03**	**1.16 ± 0.07**	**1.08 ± 0.15**	**1.29 ± 0.18**	**1.65 ± 0.17**
EnKF	0.31 ± 0.01	0.83 ± 0.10	1.73 ± 0.12	4.69 ± 0.14	3.93 ± 0.08	3.81 ± 0.07
ETKF	**0.30 ± 0.01**	1.06 ± 0.15	2.42 ± 0.11	4.57 ± 0.25	4.28 ± 0.04	4.23 ± 0.07
PF	2.80 ± 0.04	3.12 ± 0.06	3.62 ± 0.13	6.05 ± 0.16	4.95 ± 0.12	4.58 ± 0.14
PFF	0.60 ± 0.02	1.00 ± 0.05	2.20 ± 0.09	3.75 ± 0.09	3.85 ± 0.04	3.83 ± 0.11
KNet	0.60 ± 0.02	1.81 ± 0.05	3.02 ± 0.09	2.97 ± 0.21	3.47 ± 0.17	3.99 ± 0.25
VRNN	3.67 ± 0.06	3.67 ± 0.06	3.67 ± 0.06	3.69 ± 0.04	2.51 ± 0.79	3.67 ± 0.06
SRNN	3.08 ± 0.56	3.63 ± 0.05	3.40 ± 0.29	3.30 ± 0.81	3.62 ± 0.41	2.96 ± 0.32
DKF	3.70	NA	NA	NA	NA	NA

Table 2 presents the assimilation performance across different noise levels and observation settings. DBF outperforms existing methods in direct observations with $\sigma = 3, 5$, and across all noise levels for nonlinear observation cases. In the $\sigma = 1$ setting with direct observation, traditional algorithms like EnKF and ETKF outperform DBF.

The superior performance of EnKF and ETKF with direct observations at the lowest noise level can be attributed to the minimal non-Gaussianity in the posteriors within physical space. Non-Gaussianity can originate from both the dynamics model (predict step) and the observation model (update step). In this setting, the linearity of the observation operator prevents non-Gaussianity from being introduced during the update step, provided that the prior $q(z_t|o_{1:t-1})$ is Gaussian. Additionally, state estimation from each observation is highly accurate due to small noise. As a result, the prior $p(z_{t+1}|o_{1:t})$ remains close to a Gaussian distribution, as the locally linear approximation of

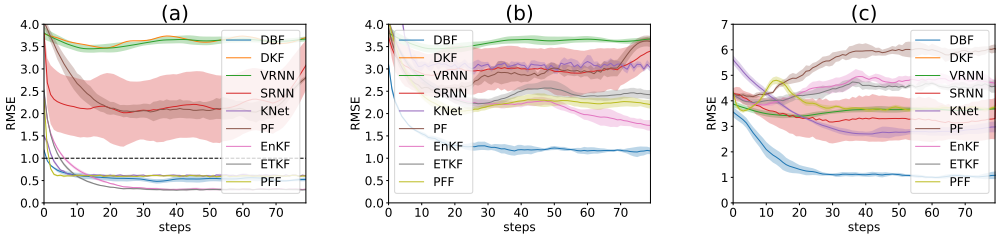
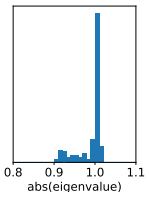


Figure 5: RMSE results for Lorenz96 experiments. Panels (a), (b) show results for direct observation with $\sigma = 1$ and $\sigma = 5$. Panel (c) shows results for nonlinear observation with $\sigma = 1$.

the dynamics adequately captures the time evolution of probability distributions. The poorer performance of EnKF and ETKF in the $\sigma = 5$ experiment is attributed to the increased non-Gaussianity introduced during each predict step. Similarly, when the observation operator is nonlinear, each update step introduces substantial non-Gaussianity. This results in a significant drop in performance for traditional filtering methods across all noise levels. In these scenarios, DBF consistently maintains an advantage over classical DA algorithms.

We observe that training DVAE-based methods is highly unstable, while that for DBF exhibits stability. Dynamics in DVAEs are modeled by RNNs, which often suffer from unstable training due to exploding or vanishing gradients. In contrast, DBF employs matrix multiplication for dynamics. If the eigenvalues of the matrix exceed one by a large margin, the model predictions, and consequently the loss function, would explode irrespective of inputs. Fig. 6 shows the histogram of the absolute values of eigenvalues at the end of training, which are distributed around or below one, indicating stable training.



setting	max[abs(eig)]
D, $\sigma = 1$	1.016 ± 0.002
D, $\sigma = 3$	1.014 ± 0.002
D, $\sigma = 5$	1.011 ± 0.001
N, $\sigma = 1$	1.012 ± 0.003
N, $\sigma = 3$	1.008 ± 0.004
N, $\sigma = 5$	1.004 ± 0.001

Figure 6: Histogram of 800 eigenvalues of the dynamics matrix in Lorenz96. D for direct and N for nonlinear observations.

4 Limitation

DBF's learning of IOO requires a training phase, unlike classical model-based data assimilation methods. Specifically, when dealing with nonlinear dynamics, DBF requires either: (i) a pair of (z_t, o_t) generated from the original SSM, (ii) a pair of (z_t, o_t) obtained via, e.g., retrospective reanalysis (ERA5; Hersbach et al. 2020 in weather forecasting), or (iii) a pretrained Koopman operator and observed data o_t.

In the Lorenz96 experiment, DBF's performance with direct observation with $\sigma = 1$ falls short compared to EnKF and ETKF. In this setting, the non-Gaussianity of posteriors is weak, resulting in minor approximation errors due to Gaussian assumptions. Consequently, a model-based approach may be more advantageous in such situations, as it leverages complete SSM knowledge without introducing training biases.

5 Conclusion

We propose DBF, a novel DA method. DBF is a NN-based extension of the KF designed to handle nonlinear observations. While constraining the test distributions to remain Gaussian, DBF enhances their representational capacity by leveraging nonlinear transform expressed by a NN. DBF is the first "Bayes-Faithful" amortized variational inference methodology, constructing test distributions that mirror the inference structure of a SSM with the Markov property. This structured inference enables analytical computation of test distributions, preventing the accumulation of Monte Carlo sampling errors over time steps. DBF exhibits superior performance over existing methods in scenarios where posterior distributions become highly non-Gaussian, such as in the presence of nonlinear observation operators or significant observation noise.

Reproducibility Statement We have provided the source code to reproduce the experiments for double pendulum (Sec. 3.2) and the Lorenz96 (Sec. 3.3) in the supplementary material. The hyperparameters for the training are provided in Table 3 in the appendix. Generation method of the training and test dataset, the dynamics model, the observation model, and the architectures are detailed in the appendix: Sec. A.1, A.2, and A.3.

REFERENCES

Miguel Alfonzo and Dean S. Oliver. Seismic data assimilation with an imperfect model. *Computational Geosciences*, 24(2):889–905, 2020. Marine Environmental Monitoring and Prediction.

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *ArXiv*, abs/2306.06079, 2023. URL https://api.semanticscholar.org/CorpusID:259129311.

Md Abdul Awal, Md Abu Rumman Refat, Feroza Naznin, and Md Zahidul Islam. A particle filter based visual object tracking: A systematic review of current trends and research challenges. *International Journal of Advanced Computer Science and Applications*, 14(11), 2023. doi: 10.14569/IJACSA.2023.01411131. URL http://dx.doi.org/10.14569/IJACSA.2023.01411131.

Omri Azencot, N. Benjamin Erichson, Vanessa Lin, and Michael Mahoney. Forecasting sequential data using consistent koopman autoencoders. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 475–485. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/azencot20a.html.

Eviatar Bach and Michael Ghil. A multi-model ensemble kalman filter for data assimilation and forecasting. *Journal of Advances in Modeling Earth Systems*, 15(1):e2022MS003123, 2023. doi: https://doi.org/10.1029/2022MS003123. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003123. e2022MS003123 2022MS003123.

Alexandros Beskos, Dan Crisan, and Ajay Jasra. On the stability of sequential Monte Carlo methods in high dimensions. *The Annals of Applied Probability*, 24(4):1396 – 1445, 2014. doi: 10.1214/13-AAP951. URL https://doi.org/10.1214/13-AAP951.

Craig H. Bishop, Brian J. Etherton, and Sharanya J. Majumdar. Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Wea. Rev.*, 129(3):420–436, March 2001. ISSN 0027-0644, 1520-0493. doi: 10.1175/1520-0493(2001)129⟨0420:ASWTET⟩2.0.CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to Sequential Monte Carlo*. Springer series in statistics. Springer, 2020. URL https://ci.nii.ac.jp/ncid/BC03234800.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A Recurrent Latent Variable Model for Sequential Data. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/b618c3210e934362ac261db280128c22-Abstract.html.

Fred Daum and Jim Huang. Nonlinear filters with log-homotopy. In Oliver E. Drummond and Richard D. Teichgraeber (eds.), *Signal and Data Processing of Small Targets 2007*, volume 6699, pp. 669918. International Society for Optics and Photonics, SPIE, 2007. doi: 10.1117/12.725684. URL https://doi.org/10.1117/12.725684.

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994. ISSN 2156-2202. doi: 10.1029/94JC00572. URL https://onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/94JC00572.

Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2207–2215, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

Thomas Frerix, Dmitrii Kochkov, Jamie Smith, Daniel Cremers, Michael Brenner, and Stephan Hoyer. Variational Data Assimilation with a Learned Inverse Observation Operator. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3449–3458. PMLR, July 2021. URL https://proceedings.mlr.press/v139/frerix21a.html. ISSN: 2640-3498.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends® in Machine Learning*, 15(1-2):1–175, 2021. ISSN 1935-8237. doi: 10.1561/2200000089. URL http://dx.doi.org/10.1561/2200000089.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Chih-Chi Hu and Peter Jan van Leeuwen. A particle flow filter for high-dimensional system applications. *Quarterly Journal of the Royal Meteorological Society*, 147(737):2352–2374, 2021. doi: https://doi.org/10.1002/qj.4028. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4028.

Brian R. Hunt, Eric J. Kostelich, and Istvan Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform kalman filter. *Physica D: Nonlinear Phenomena*, 230(1):112–126, 2007. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2006.11.008. URL https://www.sciencedirect.com/science/article/pii/S0167278906004647. Data Assimilation.

B. O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315. URL https://www.pnas.org/doi/abs/10.1073/pnas.17.5.315.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters, 2015. URL https://arxiv.org/abs/1511.05121.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models, 2016. URL https://arxiv.org/abs/1609.09869.

J. Larsen, J.L. Høyer, and J. She. Validation of a hybrid optimal interpolation and kalman filter scheme for sea surface temperature assimilation. *Journal of Marine Systems*, 65(1):122–133, 2007. ISSN 0924-7963. doi: https://doi.org/10.1016/j.jmarsys.2005.09.013. URL https://www.sciencedirect.com/science/article/pii/S0924796306002880. Marine Environmental Monitoring and Prediction.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E Turner, and Johannes Brandstetter. PDE-Refiner: Achieving Accurate Long Rollouts with Temporal Neural PDE Solvers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Qv6468llWS.

Andrew C. Lorenc. Modelling of error covariances by 4d-var data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 129(595):3167–3182, 2003. doi: https://doi.org/10.1256/qj.02.131. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.02.131.

E.N. Lorenz. *Predictability: a problem partly solved*. PhD thesis, Shinfield Park, Reading, 1995 1995.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9:4950, November 2018. doi: 10.1038/s41467-018-07210-0.

Shun Ohishi, Takemasa Miyoshi, Takafusa Ando, Tomohiko Higashiuwatoko, Eri Yoshizawa, Hiroshi Murakami, and Misako Kachi. Letkf-based ocean research analysis (lora) version 1.0. *Geoscience Data Journal*, n/a(n/a), 2024. doi: https://doi.org/10.1002/gdj3.271. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/gdj3.271.

Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adrià López Escoriza, Ruud J. G. van Sloun, and Yonina C. Eldar. Kalmannet: Neural network aided kalman filtering for partially known dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022. doi: 10.1109/TSP.2022.3158588.

Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3a835d3215755c435ef4fe9965a3f2a0-Paper.pdf.

Linqi Zhou, Michael Poli, Winnie Xu, Stefano Massaroli, and Stefano Ermon. Deep latent state space models for time-series generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42625–42643. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhou23i.html.

A Settings and additional results for experiments

parametrization of the dynamics matrix We have parametrized the dynamics matrix A following Lusch et al. (2018): we consider that $h_{dim}/2$ complex eigenvalues $\lambda_i (0 \leq i < h_{dim}/2)$ characterize A. Namely, A is a block-diagonal matrix of $h_{dim}/2$ blocks. Each block consists of 2×2 matrix, whose components are:

$$A_{block} = \exp(\rho_i) \begin{pmatrix} \cos(\omega_i) & -\sin(\omega_i) \\ \sin(\omega_i) & \cos(\omega_i) \end{pmatrix}, \tag{11}$$

where $\rho_i = \mathrm{Re}[\lambda_i]$ and $\omega_i = \mathrm{Im}[\lambda_i]$. In contrast to Lusch et al. (2018), we apply the same dynamics matrix at any positions on the latent space. We consider that this representation is sufficiently expressive, as it can express any matrix on a complex number field that is diagonalizable.

One key advantage of DBF is that augmenting the latent dimension only results in a linear increase in computational demand. This scaling is due to the efficient parametrization of the dynamics matrix, where the block-diagonal structure allows operations to scale linearly with the latent dimension. In contrast, methods such as Sequential Monte Carlo (SMC) suffer from exponential increases in computational demand as the latent space grows, assuming that the same density of particles must be maintained to capture posterior distributions. This makes DBF particularly well-suited for high-dimensional systems where traditional methods struggle with computational complexity.

Computational resources We conduct experiments on a cluster of V100 GPUs. Each GPU has memory of 32GB.

hyperparameters for training For all experiments, we have used Adam optimizer with default parameters. Table 3 shows hyperparameters employed in our experiments. Trainings for moving MNIST and double pendulum are conducted with one GPU, while that for Lorenz96 is with eight GPUs.

Table 3: Hyperparameters for training

	lr	batch size	h_{dim}	$N_{data,train}$	Epochs	train time per model
moving MNIST	10^{-3}	64	8	480,000	2	3hr× 1GPU
double pendulum	10^{-3}	256	50	1.0×10^7	1	6hr× 1GPU
Lorenz96	3×10^{-3}	64	800	2.6×10^7	1	15hr× 8GPUs

A.1 MOVING MNIST

Dataset: The dataset consists of a series of 2D images, where each pixel has a dynamic range from 0 to 255. The training set contains $480{,}000$ initial conditions, while the test set consists of ten initial conditions, with both datasets comprising 20 time steps each. The number of training samples and epochs is sufficiently large to ensure that the training converges effectively. A Gaussian noise with a standard deviation of $\sigma = 50$ is added to all pixels. The MNIST images of the digits "9" (data point 5740) and "5" (data point 5742) move at constant speeds until they reach the edges, where reflection occurs.

Training: The network weights for G_θ are fixed during the first epoch to facilitate the learning of f_θ and the image tensor for the observation model. Subsequently, G_θ is trained during the second epoch. In total, DBF undergoes training for two epochs.

Dynamics model: Constant velocity model. The exact dynamics matrix we have used is:

$$z_{t+1} = F z_t \tag{12}$$

$$F = \begin{pmatrix} 1 & 0 & 0 & 0 & dt & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & dt & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & dt & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & dt \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, z_t = \begin{pmatrix} x_{1,t} \\ y_{1,t} \\ x_{2,t} \\ y_{2,t} \\ v_{x_{1,t}} \\ v_{y_{1,t}} \\ v_{x_{2,t}} \\ v_{y_{2,t}} \end{pmatrix}, \tag{13}$$

and true observation model:

$$\tilde{x}_t = \begin{cases} (x_t \bmod 16) & \text{if x//16 is even} \\ 9 - (x_t \bmod 16) & \text{if x//16 is odd} \end{cases}, \text{ same for } y \tag{14}$$

$$o_t = h(z_t), \dim(o_t) = 44 \times 44 \text{ , a } 28 \times 28 \text{ image is embedded at}(\tilde{x}_t, \tilde{y}_t). \tag{15}$$

The formulation above addresses image reflection through the observation operator, resulting in linear dynamics while permitting multiple solutions for each observed figure. This approach presents significant challenges for the EnKF, which assumes a single-peak Gaussian distribution in the assimilating space. To ensure a fair comparison, we revise the dynamics and observation models to allow for a single solution for each figure. This adjustment notably enhances the performance of the EnKF if the image is provided. However, even with this modification, the EnKF fails to accurately estimate the position, velocity, and the embedded image.

Table 4: List of hyperparameters for moving MNIST experiment.

parameter	value
R	diag[e^6]
Q	diag[e^{-4}]

Network architecture: f_θ: Two-dimension convolutional NNs. Below is the list of layers.

- conv1: nn.Conv2d(1, 2, kernel_size=3, stride=2, padding=1)
- conv2: nn.Conv2d(2, 4, kernel_size=3, stride=2, padding=1)
- conv3: nn.Conv2d(4, 4, kernel_size=3, stride=1, padding=1)
- conv4: nn.Conv2d(4, 4, kernel_size=3, stride=1, padding=1)
- fc: nn.Linear($11 \times 11 \times 4$, 8)

The input image, sized 44×44, is sequentially processed by convolutional layers (conv1, conv2, conv3, and conv4). The output is then flattened to serve as the input for the fully connected layer (fc). Ultimately, this process yields eight variables for $f_\theta(o_t)$. The network G_θ follows the same architecture as f_θ, but it produces only the diagonal components of $G_\theta(o_t)$ through the NN.

Example figures: In Fig. 8, we show example images for observations and all the algorithms in image-informed setting.

Table 5: The success rates of different methodologies in the two-body moving MNIST problem. For the model-based approaches, we used the same dynamics and observation models that generated the data. For DBF, the model was initialized with random image tensors and trained solely on the data.

Method	Success rate
DBF	100% (50/50)
EnKF	58% (29/50)
ETKF	0% (0/50)
PF	0% (0/50)

A.2 Double pendulum

Dataset: The dataset consists of 2D coordinates representing the positions of two weights. The training set includes $10,240,000$ initial conditions, while the test set contains 10 initial conditions. The number of training samples is sufficiently large to ensure that the training converges. During DVAE training, we observed that some initial conditions resulted in training failure due to instability; however, we maintained the total number of training samples since the training was successful for at least one initial condition. Both datasets comprise 80 time steps. Numerical integration is performed using the `solve_ivp` function in SciPy, with relative tolerance (`rtol`) set to 10^{-2} and absolute tolerance (`atol`) set to 10^{-2}.

A schematic figure explaining the problem setting is presented in panel (a) of Fig. 3 in the main text.

Dynamics model is described in https://matplotlib.org/stable/gallery/animation/double_pendulum.html. The length of the bars is 1 [m], and the positions of the two pendulum weights are observable with Gaussian noise of $\sigma = 0.1, 0.3$, or 0.5 [m]. The observation interval is 0.03 [s]. The task is to predict the positions of the two weights in the successive ten frames.

Network architecture: f_θ: A sequence of ten "linear blocks" composed of fully connected layers, layer normalizations, and skip connections. Namely, each linear block has three components:

- fc: (input dimension)$\times$ (output dimension) linear layer,
- norm: layer normalization,

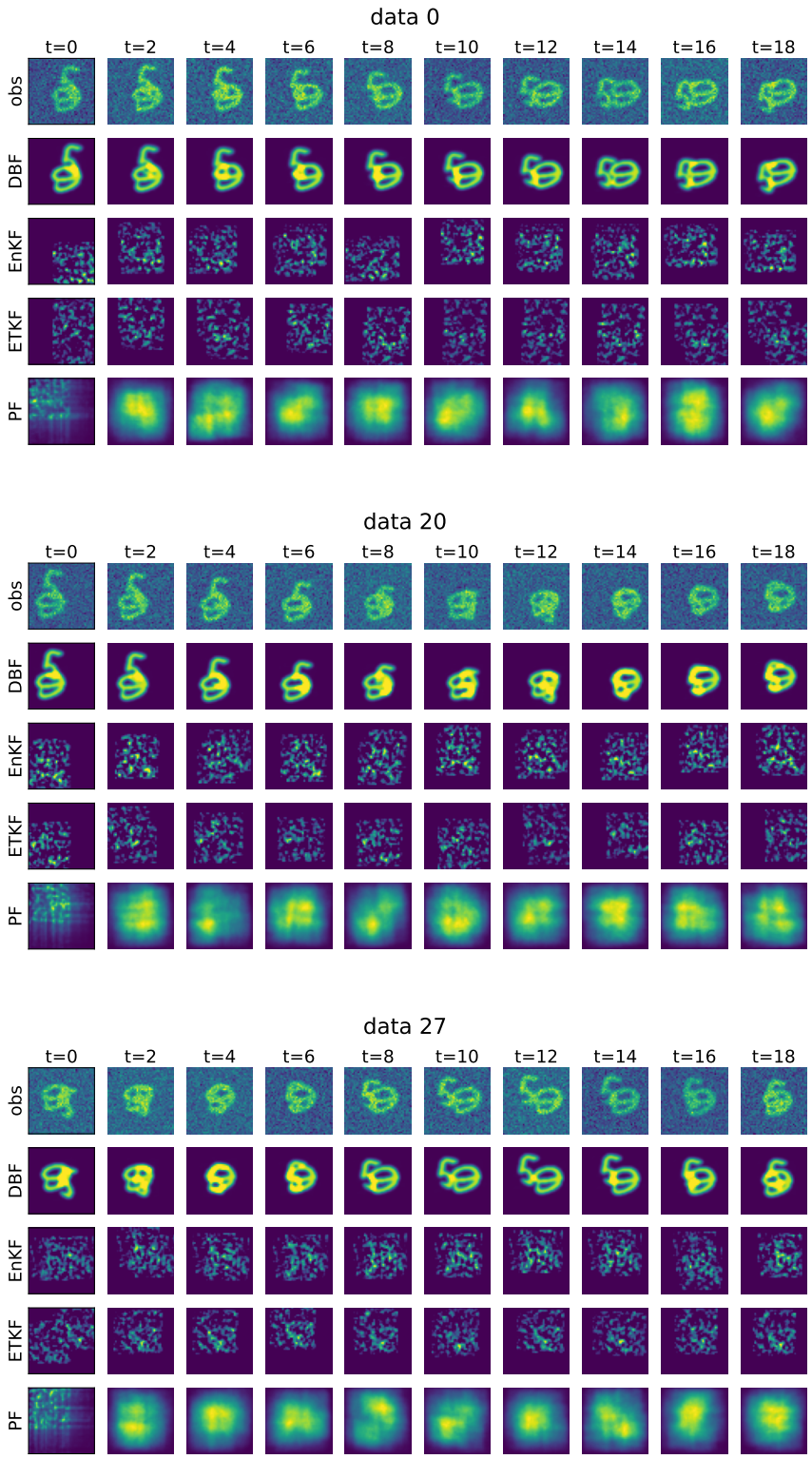


Figure 7: Example figures for two-body moving MNIST experiment. This is the setting explained in the main text. For all algorithms, the two embedded images are not explicitly informed: algorithms need to deal with many unknown parameters in the observation model.

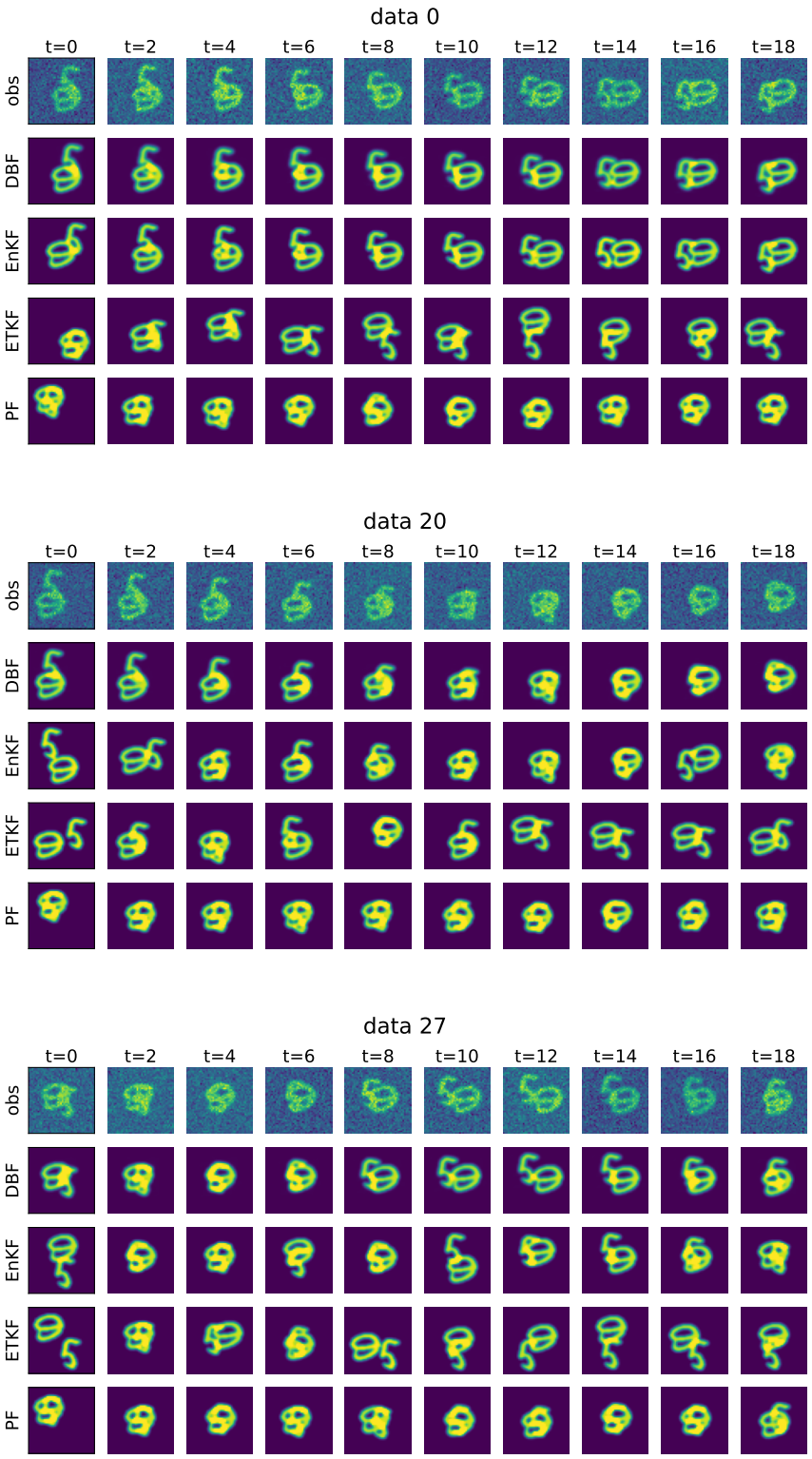


Figure 8: Example figures for two-body moving MNIST experiment. For model-based approaches (EnKF, ETKF, PF), contrary to the experiment reported in the main text, the true images are informed. In data 0, both DBF and EnKF successfully determine and follow the position of the two images. On the other hand, in data 20 and 27, EnKF estimate becomes unstable soon after the two letters overlap. Even in that situation, DBF stably follows the positions of the embedded images.

Table 6: List of hyperparameters for double pendulum experiment.

parameter	value
R_{init}	diag[1]
Q	diag[e^{-6}]
initial concentration parameter	e^5

- skip: skip connection.

Taking four observation variables as input, the first linear block expands the dimensionality to 100. The intermediate linear blocks maintain these 100-dimensional variables. The final linear block reduces the 100-dimensional input to a 50-dimensional output, representing 50 latent space variables. The ReLU activation function is applied throughout the network. The structure of G_θ mirrors that of f_θ, while ϕ_θ serves as the inverse of f_θ. The initial eigenvalues are randomly sampled from the range between e^0 and $e^{0.01}$.

Training: All training variables (network weights for the IOO (f_θ, G_θ), the emission model operator ϕ, eigenvalues λ for the dynamics matrix A, Gaussian noise parameter σ for angular velocity ω, and the concentration parameter for Von Mises distribution used for angular coordinate θ) are trained together.

Examples: Here, we show examples for assimilated θ and ω in Fig. 9. Also, we give an additional figure for the RMSE of θ for various methods.

A.3 LORENZ96

Dataset: The dataset consists of physical and observed variables sampled at 40 grid points. The training set includes 25,600,000 initial conditions, while the test set contains 10 initial conditions. The number of training samples is sufficiently large to ensure that the training converges in most cases. The original datasets comprise 80 time steps. Numerical integration is performed using the `solve_ivp` function in SciPy, with a relative tolerance $\texttt{rtol} = 10^{-2}$ and an absolute tolerance of $\texttt{atol} = 10^{-2}$. Gaussian noise with standard deviations of $\sigma = 1, 3$, or 5 is added to all measurements.

For KalmanNet, we attempted to train with 25,600,000 and 400,000 initial conditions; however, the process was terminated due to memory limitations. Consequently, we report results using a dataset size of 120,000. For DKF, VRNN, and SRNN, we also tried training with 25,600,000 conditions, but all models encountered a RuntimeError due to instability during the backward computation. To obtain results, we reduced the number of training samples to 512,000. With this adjustment, both SRNN and VRNN successfully completed the training procedure for some initial conditions.

A physical quantity z_j is defined at each grid point $j (1 \leq j \leq 40)$. The time evolution of this quantity is described by the following set of differential equations:

$$\frac{dz(t)_j}{dt} = (z_{j+1} - z_{j-2})z_{j-1} - z_j + F, (1 \leq j \leq 40) \tag{16}$$

In this equation, the driving term F is set to 8. The first term models the advection of the physical quantity, while the second term represents its diffusion along a fixed latitude. With these parameters, the evolution of the physical quantity exhibits chaotic behavior.

Network architecture: The NN f_θ consists of ten convolutional blocks followed by a fully connected layer. Each convolutional block comprises a 1D convolution, layer normalization, and a skip connection:

- conv1d: nn.Conv1d(c_{in}, c_{out}, kernel_size=5, padding=2, padding_mode="circular",)
- norm: layer normalization,

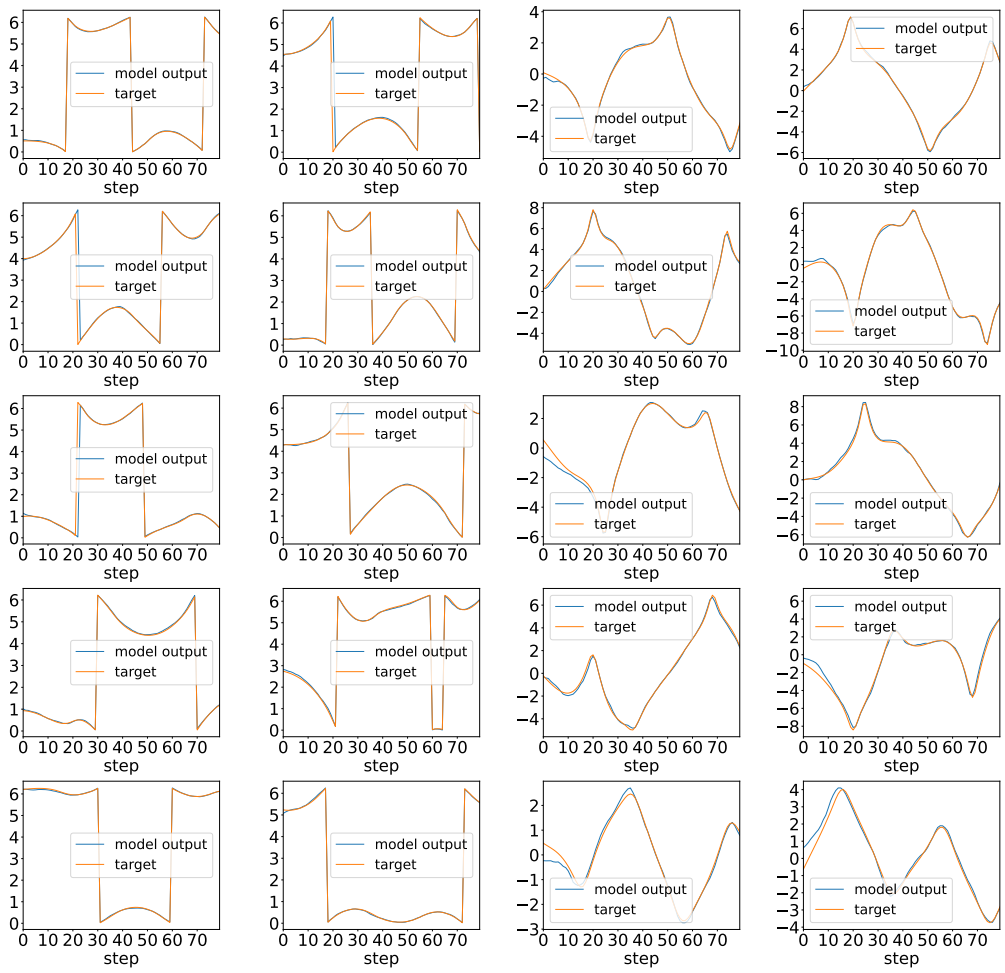


Figure 9: Results for five example data in test set. Two left columns show evolution of θ_1 and θ_2 (rad) (, therefore, the values are cyclic with the period of $2\pi \simeq 6.3$, and we corrected for those periodic shifts) and the two right columns show ω_1 and ω_2 (rad/s).

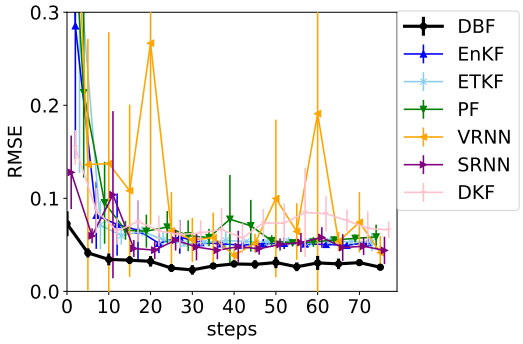


Figure 10: Assimilation results for the angle variable θ. All models successfully determine the angle coordinate in spite of the strong nonlinearity in the observation (trigonometric function). Among these, performance of DBF is the best.

Table 7: List of hyperparameters for Lorenz96 experiment.

parameter	value
R_{init}	diag[1]
Q	diag[e^{-8}]

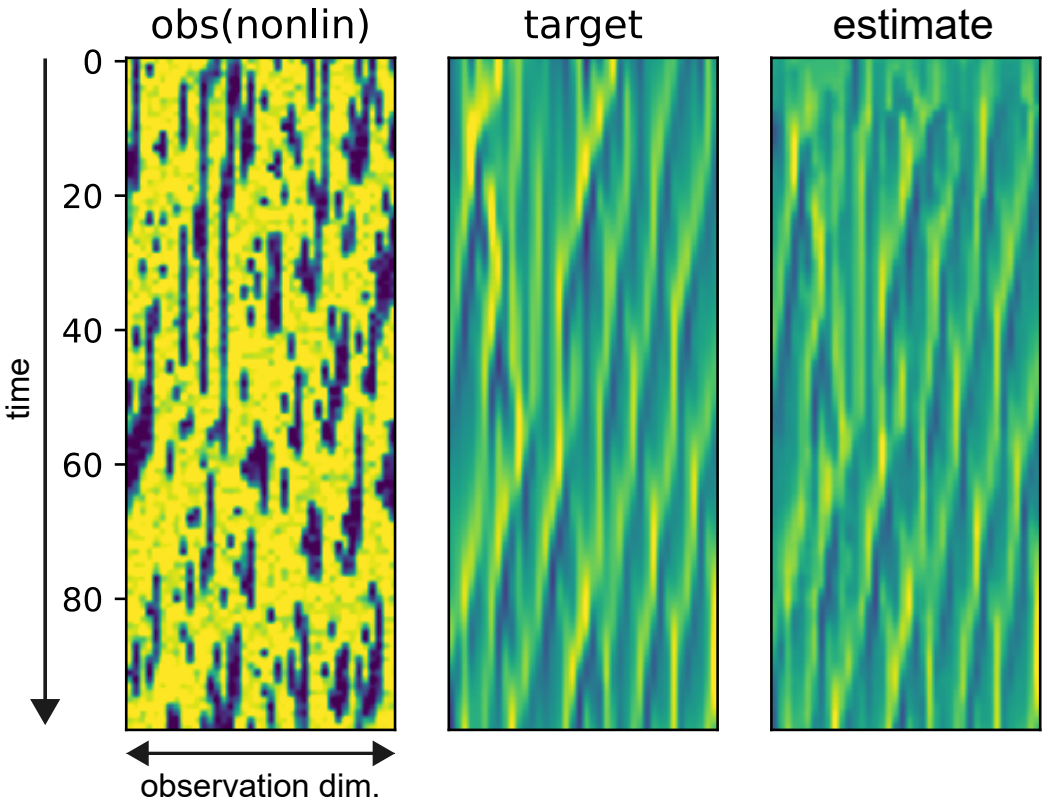


Figure 11: An example of assimilation output in the experiment with nonlinear observation operator. The observation is not very informative due to low threshold for saturation in the observation operator ($o_{t,j} = min(z_{t,j}^4, 10) + \epsilon$,, all cells with $z_{t,j} > 1.8$ are just observed as $10 + \epsilon$). In the first 20 steps, the model output resembles little with the target. However, as the step proceeds, the estimated state begins to capture features of the true state. Even with such a poor observation operator, DBF finds a latent space representation that captures the evolution of the true state.

- skip: skip connection.

The first convolutional block has $c_{\text{in}} = 1$ and $c_{\text{out}} = 20$, expanding the input by a factor of 20 in the channel dimension. The subsequent eight layers maintain 20 channels. Finally, the 20 channels and 40 physical dimensions are flattened into 800-dimensional variables, which are then fed into a fully connected layer of size 800×800. For all layers, the activation function used is ReLU. The function G_θ is structured identically to f_θ, while ϕ_θ represents the inverse of f_θ.

Training: All training variables, including the network weights for the inverse observation operator f_θ and G_θ, the emission model operator ϕ, the eigenvalues λ for the dynamics matrix A, and the Gaussian noise parameter σ, are trained concurrently.

Examples: We show an example figure for assimilation experiment with DBF in Fig. 11.

B TRAINING STABILITY

We observe that the training of our proposed method is stable compared to RNN-based models. Fig. 12 shows the evolution of the real parts of eigenvalues. Although we do not impose constraints on the real parts of eigenvalues, the values only marginally exceed one. Therefore, long-time dynamics is stable during training.

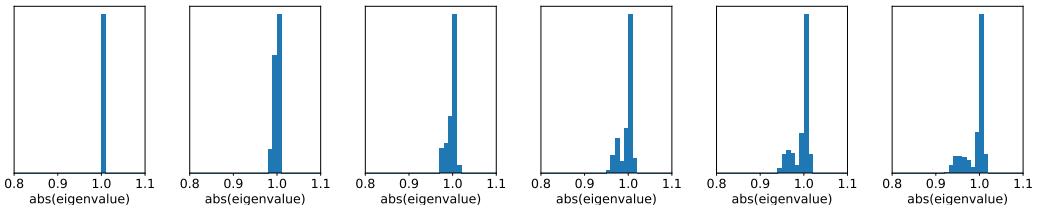


Figure 12: Evolution of histograms for the real parts of 800 complex eigenvalues in Lorenz96 experiment. Initially, eigenvalues are taken as one. As the model learns the dynamics, eigenvalues lower than 1.0 appear. However, the largest eigenvalue λ_{max} mostly remains less than 1.02.