# *Ensemble Kalman filter in latent space using a variational autoencoder pair*



Pasmans, I. ORCID: https://orcid.org/0000-0001-5076-5421, Chen, Y. ORCID: https://orcid.org/0000-0002-2319-6937, Finn, T. S., Bocquet, M. and Carrassi, A. ORCID: https://orcid.org/0000-0003-0722-5600 (2025) Ensemble Kalman filter in latent space using a variational autoencoder pair. Quarterly Journal of the Royal Meteorological Society. ISSN 1477-870X doi: 10.1002/qj.70070 Available at https://centaur.reading.ac.uk/125425/

Received: 18 February 2025 | Revised: 30 September 2025 | Accepted: 21 October 2025

DOI: 10.1002/qj.70070

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

# Ensemble Kalman filter in latent space using a variational autoencoder pair

**Ivo Pasmans**[1] | **Yumeng Chen**[1] | **Tobias Sebastian Finn**[2] | **Marc Bocquet**[2] | **Alberto Carrassi**[1,3]

[1]Department of Meteorology, University of Reading, Reading, UK

[2]CEREA, ENPC, EDF R&D, Institut Polytechnique de Paris, Île-de-France, France

[3]Department of Physics "Augusto Righi", Universitá di Bologna, Bologna, Italy

**Correspondence**
Yumeng Chen, Department of Meteorology, University of Reading, Reading, RG6 6ET, UK.
Email: yumeng.chen@reading.ac.uk

**Present address**
Ivo Pasmans, European Centre for Medium-Range Weather Forecasts, Reading, UK

**Funding information**
Schmidt Sciences, Grant/Award Number: G-24-66154

**Abstract**

Popular (ensemble) Kalman filter data assimilation (DA) approaches assume that the errors in both the a priori estimate of the state and the observations are Gaussian. For constrained variables, for example, sea-ice concentration or stress, such an assumption does not hold. The variational autoencoder (VAE) is a machine-learning (ML) technique that allows us to map an arbitrary distribution to/from a latent space in which the distribution is supposedly closer to a Gaussian. We propose a novel hybrid DA–ML approach in which VAEs are incorporated in the DA procedure. Specifically, we introduce a variant of the popular ensemble transform Kalman filter (ETKF) in which the analysis is applied in the latent space of a single VAE or a pair of VAEs. In twin experiments with a simple circular model, whereby the circle represents an underlying submanifold to be respected, we find that the use of a VAE ensures that a posteriori ensemble members lie close to the manifold containing the truth. Furthermore, online updating of the VAE is necessary and achievable when this manifold varies in time, that is, when it is non-stationary. We demonstrate that introducing an additional second latent space for the observational innovations improves robustness against detrimental effects of non-Gaussianity and bias in the observational errors but lessens the performance slightly if observational errors are strictly Gaussian.

**KEYWORDS**

data assimilation, ensemble Kalman filter, machine learning, non-Gaussianity, variational autoencoder

## 1 | INTRODUCTION

Data assimilation (DA) aims to provide a more precise estimate of the true state of a system by combining a prior guess in the form of a probability distribution with observations (Carrassi *et al.*, 2018). Data assimilation is widely used in operational atmosphere, ocean, and sea-ice forecasting (Buehner *et al.*, 2025; De Rosnay *et al.*, 2022; Inverarity *et al.*, 2023; Qin *et al.*, 2023; Waters *et al.*, 2015) and will also be used by the neXtSIM$_{DG}$ sea-ice model (Jendersie *et al.*, 2024; Richter *et al.*, 2023), which is currently being developed as part of the Scale-Aware Sea Ice Project (SASIP).[1] In previous studies, we have explored the possibilities of tailoring DA to the discontinuous

Galerkin numerical core used by neXtSIM$_{DG}$ (Pasmans *et al.*, 2024) and inferring sea-ice parameters using an ensemble Kalman filter (EnKF: Chen *et al.*, 2024). Though intentionally fully conceptual in its nature, this work is also motivated by the challenges posed to DA by the complex physics of sea ice, in particular the presence of nonlinear relations and constraints in sea-ice dynamics. These issues are, however, not exclusive to sea-ice modelling, but instead pervasive in many other branches of climate and weather prediction at large; see, for example, the modelling of humidity in atmospheric models.

The EnKF (Evensen, 1994) is one of the most popular DA methodologies. In the EnKF, the probability distribution for the true state is assumed to be a Gaussian with mean and covariance estimated from an ensemble of model runs. During the update step, a correction is added to the ensemble members by taking into account the information from the observations. Following this, the ensemble members are propagated to the next time step with a dynamical model. In real scenarios, the use of the EnKF can be challenging. The computational demands of geophysical models can easily make running large ensembles prohibitive. The use of smaller, computationally more affordable, ensembles introduces sampling errors that need to be mitigated. Remediation of the effect of these errors requires the application of additional techniques such as ensemble inflation (Ehrendorfer, 2007; Whitaker & Hamill, 2012) and localisation (Ehrendorfer, 2007; Morzfeld & Hodyss, 2022). The EnKF assumes that the errors in the prior estimate, background errors, and observations are unbiased and Gaussian. Such assumptions generally do not hold. Finally, EnKF is derived by linear estimation theory, that is, apart from (structured) noise there is a linear relationship between observed differences between observations and model predictions (a.k.a. the innovation) on one hand and the DA correction on the other hand. This implies that there is an affine space, consisting of elements that can be decomposed as the initial guess plus a possible DA correction, in which all elements are model solutions as well. Although such spaces of possible solutions exist for linear models, their existence is not guaranteed for nonlinear models. This implies that, for nonlinear models, the DA process might produce physically non-realisable states. These assumptions are especially problematic when using sea-ice models with a brittle rheology (Dansereau *et al.*, 2016; Ólason *et al.*, 2022), such as neXtSIM$_{DG}$. These models contain strong nonlinear relationships between sea-ice damage and sea-ice viscosity and elasticity. Furthermore, errors and physically realisable sea-ice states are constrained by several bounds. While some of these are simple—for example, sea-ice concentration must lie between 0 and 1—other bounds imposed on sea-ice stresses by the Mohr–Coulomb relation are a nonlinear function of the sea-ice state themselves.

## 1.1 | DA in latent space

Over the last years, there has been a proliferation of works fusing DA with machine learning (ML). Some exemplary studies use corrections produced by DA to train neural networks to produce either model dynamics and/or model error corrections (Arcucci *et al.*, 2021; Bocquet *et al.*, 2020a, 2020b; Brajard *et al.*, 2020, 2021; Farchi *et al.*, 2021, 2023), or replace the code that produces the DA correction with neural networks (Bocquet *et al.*, 2024; Boudier *et al.*, 2023; Chinellato & Marcuzzi, 2024; McCabe & Brown, 2021). In other cases, ML is blended with DA in an attempt to address some of the aforementioned challenges of DA: the computational cost, the need for inflation and localisation, physical imbalances, and the violation of Gaussian or quasi-linear assumptions. In the following, we provide a very short and inevitably non-exhaustive overview of these attempts; they are reported schematically in Table 1. Our scope is to provide essential elements of the context within which our current study is rooted. Hence we focus primarily on studies in which the issues are addressed with the aid of a second, often lower-dimensional, space, called the latent space. The reader may find more extensive reviews in Buizza *et al.* (2022), Cheng *et al.* (2023), Bach *et al.* (2024), and Shlezinger *et al.* (2024).

One aim of the ML-for-DA schemes is to reduce the computational burden of DA. For instance, Maulik *et al.* (2022) reduce the computational cost of the forecast model by (1) reducing the dimensionality of the model states with the aid of a principal orthogonal decomposition (POD) and (2) training a recurrent neural network (RNN) to predict those coefficients. These coefficients are then corrected by four-dimensional variational data assimilation (Carrassi *et al.*, 2018), a variational DA method, exploiting automatic differentiation of the RNN. A similar concept is followed by Amendola *et al.* (2021), Peyron *et al.* (2021), and Akbari *et al.* (2023), but they replace the POD with either convolutional neural networks (CNN) or an autoencoder that maps the states to/from a low-dimensional latent space where the DA correction takes place. They propagate an ensemble of model trajectories in this latent space and apply an EnKF to it. However, this approach introduces another source of nonlinearity, since the observation operator becomes the composition of the operator on the original state space with the generally nonlinear decoder. To avoid or mitigate this problem, Cheng *et al.* (2022) introduce a second autoencoder for observations, while Pawar and San (2022)

**TABLE 1** Summary of various ML–DA studies leveraging on the small-dimension latent space. We list the ML algorithms used, the way states are propagated forward in time, the DA method used, the space to which it is applied, and additional information on how the observation operator that acts on a physical state (default operator) is modified to act on the latent space. Here VAE refers to the variational autoencoder, EnKF to the ensemble Kalman filter, and ETKF to the ensemble transform Kalman filter.

| Study | ML methods | Forward propagation | DA method | Observation operator |
|---|---|---|---|---|
| Canchumuni *et al.* (2019) | VAE | identity | latent ensemble smoother | default |
| Mack *et al.* (2020) | autoencoder | physical model | latent 3DVar | Penrose inverse, decoder, identity on the latent space |
| Amendola *et al.* (2021) | CNN, LSTM | latent model | latent EnKF | embedding |
| Grooms (2021) | VAE | none | physical EnOI | default |
| Peyron *et al.* (2021) | autoencoder, residual network | latent ensemble | latent ETKF | decoder, default |
| Bao *et al.* (2022) | VAE | identity | latent EnKF | decoder, default |
| Cheng *et al.* (2022) | autoencoder, LSTM | latent model | latent 3DVAR | identity in latent space |
| Maulik *et al.* (2022) | LSTM, POD | principal component model | 4DVAR | Penrose inverse—default |
| Pawar and San (2022) | LSTM | principal component model | DEnKF | Penrose inverse—identity latent space |
| Rozet and Louppe (2023) | diffusion model | physical model | deep Kalman filter | default |
| Finn *et al.* (2024a) | diffusion model | physical model | ETKF | default |
| Huang *et al.* (2024) | diffusion model | physical model | latent state nudging | embedding |
| Luk *et al.* (2024) | linear transformation | physical model | physical | default |
| Melinc and Zaplotnik (2024) | VAE | physical model | latent 3DVAR | decoder mean—default |
| Qu *et al.* (2024) | diffusion model | physical | score-function nudging | default |
| Si and Chen *et al.* (2024) | VAE, diffusion model | physical model | score-function nudging | identity on latent space |

replace the autoencoder with a linear projection on the principal components.

Machine-learning schemes capable of generating large ensembles cheaply have also been proposed to eliminate the need for inflation and localisation. One such scheme is the variational autoencoder (VAE: Kingma & Welling, 2019). Like the autoencoder, the VAE consists of an encoder–decoder pair, but now these functions output probability distributions instead of a single state. Examples of this approach can be found in Grooms (2021), in which ensemble members are drawn from probability distributions produced by applying the encoder–decoder to the a priori state. A similar approach, but with a denoising diffusion model instead of a VAE, can be found in Finn *et al.* (2024a). In contrast to Grooms (2021), Melinc and Zaplotnik (2024) applied DA in the latent space of a VAE. The posterior distribution in latent space is then mapped by the decoder back to the physical space. As the decoder has been trained to reproduce the climatological distribution of atmospheric states, the mapped physical states are expected to respect physical relations in the atmosphere. As the probability distribution depends on the a priori state, the generated ensemble will vary in time. However, the spread of the ensemble in latent space has to be specified by the data assimilator and hence the uncertainty in the prior and posterior states might not converge to its true value over time.

The presence of non-Gaussian errors in DA has traditionally been addressed by transforming realisations of the non-Gaussian distribution to realisations of a Gaussian distribution using anamorphosis, after which a conventional DA method can be applied. One common application of Gaussian anamorphosis is the transformation of strictly positive variables, such as

sea-ice concentration, from a log-normal to a normal or Gaussian distribution (Bocquet *et al.*, 2010; Fletcher & Zupanski, 2006; Polavarapu *et al.*, 2005; Simon & Bertino, 2012; Song *et al.*, 2012). Gaussian distributions can also be constructed from arbitrary distributions using quantile matching (Béal *et al.*, 2010; Bertino *et al.*, 2003; Grooms, 2022; Kotsuki *et al.*, 2017; Metref *et al.*, 2014; Simon & Bertino, 2012). However, the method is not applicable practically in all circumstances. The flexible quantile-matching method requires the availability of histograms of the background errors to infer the random variable distribution. Construction of such histograms necessitates the ergodic error assumption or large ensembles, in the case of, for example, the rank regression Kalman filter (Anderson, 2010, 2019). The former limits the amount of spatial variability of the distribution that can be captured by the transformation, and the latter is computationally costly. In high-dimensional applications, quantile matching is only practical for univariate variables. A ML alternative to quantile matching is provided by normalizing flows (Tabak & Turner, 2013). In these flows, the relation between the arbitrary probability distribution and the standard normal is constructed by neural networks performing a sequence of coordinate transformations. However, this approach could face its own challenges in high-dimensional applications, due to the need for multiple-determinant computations.

An alternative way to deal with non-Gaussianity is by applying DA in the latent space of a diffusion model. This is an approach followed by Amendola *et al.* (2021), Rozet and Louppe (2023), Huang *et al.* (2024), Qu *et al.* (2024), and Cheng *et al.* (2024). These models can deal with arbitrary background and observation-error distributions, but conditioning the model output generated on the observations is, however, non-trivial. Furthermore, the reverse denoising algorithm used to generate the corrected states can suffer from numerical instabilities (Qu *et al.*, 2024). Consequently, DA with diffusion models will not be pursued in this work.

### 1.2 | The double ETKF–VAE

In this work, we propose applying the ensemble transform Kalman filter (ETKF), a flavour of DA, in the latent space of a VAE. One of the reasons for doing so is that we want to address the concern around non-Gaussianity mentioned before. The variational autoencoder is trained to relate an arbitrary distribution to a standard normal distribution. Although the relationship is not perfect, we expect that the distribution of the ensemble members in latent space ends up closer to normal than the one in physical space. Hence, when the ETKF is applied in latent space, the Gaussian assumptions under which the ETKF solution is optimal are closer to being satisfied. Consequently, we hypothesise that this ETKF–VAE setup will outperform a conventional ETKF.

As shown in the last column of Table 1, the definition of the observation operator and observational error covariance is non-trivial in DA when using a latent space. In this work, we will also present our solution to this problem: a second VAE. The objective of this second VAE is twofold. First, it should bring the distribution of the ensemble members projected into the space of observations to be more Gaussian. Second, it should remove any non-Gaussianity in the observational errors. As such, we expect that, certainly when the observational errors are non-Gaussian, the double ETKF–VAE would outperform the setup in which the VAE is applied solely to the ensemble members.

Finally, we expect that, by mapping the corrected ensemble members from latent space back to physical space using the VAE, the post-DA ensemble members will remain physically consistent. In this aspect, we are building on the work by Canchumuni *et al.* (2019) and Bao *et al.* (2022), who also apply DA in the latent space of a VAE. However, we expand on their work in two directions: first, by adding the aforementioned second VAE, and, second, by using a cycling setup, propagating the DA correction forward in time, while Canchumuni *et al.* (2019) and Bao *et al.* (2022) try to find a correction for a single time only. To the best of our knowledge, this will be the first time the VAE–ETKF hybrid is used to correct a state that evolves over time.

The outline of the article is as follows. In Section 2, the mathematics behind the VAE is explained, our novel double VAE–DA method is introduced, and the necessary modifications to the classical EnKF algorithm are discussed. Section 3 contains the description of the ideal test-ground model and VAE architecture used in the experiments. The experiments themselves and their outcomes can be found in Section 4, while the discussion of the results, together with the conclusions drawn, is in Section 5.

## 2 | ENSEMBLE KALMAN FILTER-VARIATIONAL AUTOENCODER HYBRIDS

The hybrid DA–ML scheme introduced in this study is based on the cornerstone idea of applying a modified ensemble transform Kalman filter (ETKF) in the latent

space. In this section, we will first introduce the VAE in a general setting and then explain how it can be combined with the ETKF.

## 2.1 | Variational autoencoder

The VAE is a type of generative ML technique consisting of an encoder and a decoder. Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \subset \mathcal{X} \subseteq \mathbb{R}^{N_\mathcal{X}}$ be a set of realisations from an unknown probability distribution. In this section, no assumptions are imposed on the probability distribution, but from Section 2.3 onward it is assumed that the realisations will be either members of the forecast ensemble or members of the ensemble of innovations at the time at which a DA correction is calculated. The probability density function (PDF) for this distribution can be expressed by explanatory latent variables $\mathbf{z} \in \mathbb{R}^{N_\mathcal{Z}}$ with

$$p_\mathcal{X}(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z}) p_\mathcal{Z}(\mathbf{z})\, d\mathbf{z}.$$

Here, $p_\mathcal{Z}(\mathbf{z}) \overset{\text{def}}{=} \mathcal{N}(\mathbf{z}; \mathbf{0}_\mathcal{Z}, \mathbf{I}_\mathcal{Z})$ is the PDF of the standard normal distribution, with $\mathbf{0}_\mathcal{Z}$ and $\mathbf{I}_\mathcal{Z}$ being the zero vector and the identity matrix in $\mathbb{R}^{N_\mathcal{Z}}$ respectively. The decoder, $p_\theta(\mathbf{x}|\mathbf{z})$, provides the PDF of the transformation from a given latent state $\mathbf{z}$ to a state $\mathbf{x}$ in $\mathcal{X}$. Similarly, the encoder, $q_\phi(\mathbf{z}|\mathbf{x})$, provides a PDF for the transformation of a state $\mathbf{x}$ to a latent state $\mathbf{z}$. The VAE aims to find parameters such that the parameterised PDF $p_\theta(\mathbf{x})$ approximates the PDF $p_\mathcal{X}$ as well as possible by minimising the Kullback–Leibner (KL) divergence $KL\left[p_\mathcal{X}(\mathbf{x})||p_\theta(\mathbf{x})\right]$ (Kingma & Welling, 2019; Rezende *et al.*, 2014). The KL divergence is a positive measure that obtains its minimum of 0 if and only if $p_\mathcal{X} = p_\theta$ nearly everywhere. Expanding KL divergence and introducing the decoder gives

$$\begin{aligned} &KL\left[p_\mathcal{X}(\mathbf{x})||p_\theta(\mathbf{x})\right] \\ &\overset{\text{def}}{=} \int p_\mathcal{X}(\mathbf{x}) \ln \frac{p_\mathcal{X}(\mathbf{x})}{p_\theta(\mathbf{x})}\, d\mathbf{x} = \int p_\mathcal{X}(\mathbf{x}) \ln p_\mathcal{X}(\mathbf{x})\, d\mathbf{x} \\ &\quad + \int p_\mathcal{X}(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) \ln \frac{q_\phi(\mathbf{z}|\mathbf{x}) p_\theta(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) p_\theta(\mathbf{z}|\mathbf{x})}\, d\mathbf{z}\, d\mathbf{x}, \qquad (1a) \\ &= \int p_\mathcal{X}(\mathbf{x}) \ln p_\mathcal{X}(\mathbf{x})\, d\mathbf{x} \\ &\quad - \int p_\mathcal{X}(\mathbf{x}) KL[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})]\, d\mathbf{x} \\ &\quad - \int p_\mathcal{X}(\mathbf{x}) \mathcal{L}(\phi, \theta, \mathbf{x})\, d\mathbf{x}, \qquad (1b) \end{aligned}$$

with the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ defined as a parameterised PDF, and

$$\begin{aligned} \mathcal{L}(\phi, \theta, \mathbf{x}) &\overset{\text{def}}{=} \int q_\phi(\mathbf{z}|\mathbf{x}) \ln \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\, d\mathbf{z} \\ &= \int q_\phi(\mathbf{z}|\mathbf{x}) \ln p_\theta(\mathbf{x}|\mathbf{z})\, d\mathbf{z} \\ &\quad - \int q_\phi(\mathbf{z}|\mathbf{x}) \ln \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}\, d\mathbf{z} \qquad (1c) \end{aligned}$$

being the evidence lower bound (ELBO). In the second line of Equation (1a), we made use of the relation

$$\begin{aligned} \ln \frac{1}{p_\theta(x)} &= \ln \frac{1}{p_\theta(x)} \int q_\phi(\mathbf{z}|\mathbf{x})\, d\mathbf{z} \\ &= \int q_\phi(\mathbf{z}|\mathbf{x}) \ln \frac{1}{p_\theta(\mathbf{x})}\, d\mathbf{z}. \qquad (2) \end{aligned}$$

The weights of the VAE $\phi$ and $\theta$ are found by maximising the expectation value of the ELBO (last term in Equation 1b). Loosely stated, the maximisation of the expectation value of the ELBO will result in a decoder that approximately minimises the left-hand side of Equation (1b) while simultaneously producing an encoder $p_\phi(\mathbf{z}|\mathbf{x})$ that is an approximate "inverse" of the decoder $p_\theta(\mathbf{x}|\mathbf{z})$. Alternatively, one can view maximisation of the ELBO as maximisation of the reconstruction probability, $p_\theta(\mathbf{x}|\mathbf{z})$, in the first term on the right-hand side of Equation (1c), regularised by the need for the learned encoder, $q_\phi(\mathbf{z}|\mathbf{x})$, to stay close to the standard normal, $p_\theta(\mathbf{z})$.

As is conventional, the encoder and decoder distributions are assumed to be Gaussian with diagonal covariances. That is to say, $p_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \Sigma_\phi(\mathbf{x}))$ and $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z}))$, in which $\mu_\phi$, $\mu_\theta$, $\ln \Sigma_\phi$, and $\ln \Sigma_\theta$ are provided as outputs from neural networks. The exact architectures of these networks as used in this study are given in Section 4. When the expectation value of the ELBO (i.e., the last term in Equation 1b) is replaced by its single-sample Monte-Carlo approximation, it can be written as

$$\begin{aligned} &\mathbf{z}_\phi \sim \mathcal{N}(\mu_\phi(\mathbf{x}), \Sigma_\phi(\mathbf{x})), \qquad (3a) \\ &\ln \Sigma'_\theta \overset{\text{def}}{=} (1-\gamma) \ln \Sigma_\theta(\mathbf{z}_\phi) + \gamma \ln \Sigma_{\text{def}}, \qquad (3b) \\ &2 \int p_\mathcal{X}(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) \ln p_\theta(\mathbf{x}|\mathbf{z})\, d\mathbf{z}\, d\mathbf{x} \\ &\quad \approx -\ln \det(\Sigma'_\theta) - \|\Sigma_\theta'^{-1/2}(\mathbf{x} - \mu_\theta(\mathbf{z}_\phi))\|^2 \\ &\quad\quad - \|\ln \Sigma'_\theta - \ln \Sigma_\theta(\mathbf{z}_\phi)\|^2, \qquad (3c) \\ &2 \int p_\mathcal{X}(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x}) \ln \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})}\, d\mathbf{z}\, d\mathbf{x} \\ &\quad \approx \|\mu_\phi(\mathbf{x})\|^2 + \text{tr}(\Sigma_\phi(\mathbf{x})) - \ln \det(\Sigma_\phi(\mathbf{x})) - N_\mathcal{Z}, \qquad (3d) \end{aligned}$$

with the evaluation of the $\mathbf{z}$-integral in Equation (3d) taken from Zhang *et al.* (2023), Equation (3a) indicating that $\mathbf{z}_\phi$ is drawn from a Gaussian distribution with mean $\mu_\phi(\mathbf{x})$

and covariance $\Sigma_\phi(\mathbf{x})$, $\gamma$ an epoch-dependent regularisation factor that goes to zero as the number of epochs goes to infinity, and $\Sigma_{\text{def}}$ a reference covariance to be specified. The $\gamma$ regularization is there to avoid well-documented convergence issues with the ELBO (Dai & Wipf, 2019; Rezende & Viola, 2018), without having to resort to fixing the decoder variance $\Sigma_\theta$.

## 2.2 | Ensemble transform Kalman filter revisited

The Kalman filter (KF) is the analytic complete solution of the sequential Bayesian filter problem, under the assumption of Gaussian errors and linear dynamical and observation models. The KF estimates the unknown Gaussian PDF of the true state of the system of interest $\mathbf{x}^{\text{truth}} \sim \mathcal{N}(\mu_{\mathbf{X}}, \mathbf{P}_{\mathbf{X}})$, with $\mu_{\mathbf{X}}$ and $\mathbf{P}_{\mathbf{X}}$ being a function of both time and assimilated noisy observations $\mathbf{y}_t \sim \mathcal{N}(H(\mathbf{x}_t^{\text{truth}}), \mathbf{R})$, with $H = \mathbf{H}$ the linear observation operator. The posterior mean ($\mu_{\mathbf{X}}^{\text{a}}$) and covariance ($\mathbf{P}_{\mathbf{X}}^{\text{a}}$) are given by (section 6.1 of Evensen *et al.*, 2022)

$$\mu_{\mathbf{X}}^{\text{a}} = \mu_{\mathbf{X}}^{\text{f}} + \mathbf{P}_{\mathbf{X}}^{\text{f}}\mathbf{H}^{\text{T}}(\mathbf{H}\mathbf{P}_{\mathbf{X}}^{\text{f}}\mathbf{H}^{\text{T}} + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}\mu_{\mathbf{X}}^{\text{f}}), \tag{4a}$$

$$\mathbf{P}_{\mathbf{X}}^{\text{a}} = \mathbf{P}_{\mathbf{X}}^{\text{f}} - \mathbf{P}_{\mathbf{X}}^{\text{f}}\mathbf{H}^{\text{T}}(\mathbf{H}\mathbf{P}_{\mathbf{X}}^{\text{f}}\mathbf{H}^{\text{T}} + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}_{\mathbf{X}}^{\text{f}}. \tag{4b}$$

Here, $\cdot^{\text{f}}$ denotes the a priori, that is, the state just before the DA is applied, while $\cdot^{\text{a}}$ refers to the state directly after application of the DA, that is, the posterior state. The EnKF relaxes the requirement that the dynamical model and observation operator have to be linear and estimates the mean $\mu_{\mathbf{X}}$ and covariance $\mathbf{P}_{\mathbf{X}}$ from an ensemble of (possibly nonlinear) member runs that are collected as columns in a matrix $\mathbf{X}$, that is,

$$\mu_{\mathbf{X}} = \frac{1}{M}\mathbf{X}\mathbf{1}_M, \qquad \mathbf{P}_{\mathbf{X}} = \frac{1}{M-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\text{T}}, \quad \text{and} \quad \mathbf{Y} = \mathbf{H}\mathbf{X}.$$

Here, $\tilde{\cdot}$ indicates that the ensemble mean has been removed from each column to form the anomaly matrices, that is,

$$\tilde{\mathbf{X}} = \mathbf{X}\left(\mathbf{I} - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^{\text{T}}\right),$$

and $\mathbf{1}_M \in \mathbb{R}^M$ is a vector having ones as its elements. With these ensemble approximations, the equalities in Equation (4) can be satisfied, provided that

$$\mathbf{X}^{\text{a}} = \mathbf{X}^{\text{f}} + \tilde{\mathbf{X}}^{\text{f}}\mathbf{Y}^{\text{T}}\mathbf{U}\mathbf{S}^{-1}\mathbf{U}^{\text{T}}\left(\mathbf{y} - \frac{1}{M}\mathbf{H}\mathbf{X}^{\text{f}}\mathbf{1}_M\right)\mathbf{1}_M^{\text{T}} + \tilde{\mathbf{X}}^{\text{f}}\mathbf{Y}^{\text{T}}\mathbf{U}(\mathbf{I} - \mathbf{S}^{-1})^{1/2}, \tag{5a}$$

$$\frac{1}{M-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\text{T}} + \mathbf{R} \overset{\text{SVD}}{=} \mathbf{U}\mathbf{S}\mathbf{U}^{\text{T}}, \tag{5b}$$

$$\mathbf{X}^{\text{a}} = \frac{1}{M}\mathbf{X}^{\text{f}}\mathbf{1}_M\mathbf{1}_M^{\text{T}} + \tilde{\mathbf{X}}^{\text{f}}\tilde{\mathbf{Y}}^{\text{T}}\mathbf{C}^{-1}\left(\mathbf{y} - \frac{1}{M}\mathbf{H}\mathbf{X}^{\text{f}}\mathbf{1}_M\right)\mathbf{1}_M^{\text{T}} + \tilde{\mathbf{X}}^{\text{f}}\mathbf{U}\mathbf{S}^{1/2}\mathbf{U}^{\text{T}}, \tag{6a}$$

$$\mathbf{C} \overset{\text{def}}{=} \frac{1}{M-1}\tilde{\mathbf{Y}}^{\text{T}}\tilde{\mathbf{Y}} + \mathbf{R}, \tag{6b}$$

$$\mathbf{U}\mathbf{S}\mathbf{U}^{\text{T}} \overset{\text{def}}{=} \mathbf{I}_M - \tilde{\mathbf{Y}}^{\text{T}}\mathbf{C}^{-1}\tilde{\mathbf{Y}}, \tag{6c}$$

with $\mathbf{U}\mathbf{S}\mathbf{U}^{\text{T}}$ in Equation (6c) defining a singular-value decomposition (SVD). These filters, in which ensemble members undergo a deterministic transformation such that the first two statistical moments satisfy the ensemble approximation of the relations in Equation (4), are known as square-root filters. These filters produce more accurate analysis covariances as, in contrast to stochastically EnKF schemes, they do not use sampled observational errors to account for uncertainty in the observations (Evensen, 2004; Tippett *et al.*, 2003). Several square-root schemes have been conceived for DA; see Tippett *et al.* (2003) for an overview. The ETKF (Bishop *et al.*, 2001) and its localised equivalent the local ensemble transform Kalman filter (LETKF: Hunt *et al.*, 2007) have emerged as the most popular variants, due to the ability to assimilate all observations simultaneously, computational efficiency, as they require the eigensystem decomposition of only a $M \times M$ matrix, and ease of parallelisation. Consequently, they have been widely adapted by several numerical weather prediction agencies (Buehner *et al.*, 2025; Jun *et al.*, 2024; Matsugishi *et al.*, 2025; Vobig *et al.*, 2021). For these reasons we will use a variation on the ETKF in this work, which uses the same eigenvectors as ETKF but still calculates $\mathbf{C}^{-1}$ separately.

In anticipation of the work in Section 2.3, we modify the ETKF in Equation (6) slightly using

$$\mathbf{H}\mathbf{P}_{\mathbf{X}}^{\text{f}}\mathbf{H}^{\text{T}} + \mathbf{R} \approx \frac{1}{K-1}\tilde{\mathbf{D}}_K\tilde{\mathbf{D}}_K^{\text{T}}$$

from Desroziers and Ivanov (2001), with $\mathbf{D}_K \in \mathbb{R}^{N_{\text{y}} \times K}$ a matrix having perturbed innovations as ensemble members, that is, the $k$th column of $\mathbf{D}_K$ is given by $\mathbf{y} + \epsilon_{m_k}^{\text{y}} - \mathbf{H}\mathbf{x}_{m_k}$, with $K \gg M, N_y$ the number of observations, $\epsilon_{m_k}^{\text{y}}$ a realisation of the observational error distribution, and each $m_k$ chosen randomly from $\{1, 2, \ldots, M\}$. Simultaneously, we rewrite

$$\mathbf{H}\mathbf{P}_{\mathbf{X}}^{\text{f}} = \frac{1}{M-1}\mathbf{Y}\tilde{\mathbf{X}}^{\text{f}} = -\tilde{\mathbf{D}}_M\tilde{X}^{\text{f}},$$

with $\mathbf{D}_M = \mathbf{y}\mathbf{1}_M^{\text{T}} - \mathbf{H}\mathbf{X}^{\text{f}} \in \mathbb{R}^{N_y \times M}$, the matrix having the innovation of each ensemble member as columns. After making these substitutions, Equation (6) can

be rewritten as

$$\mathbf{X}^{\mathrm{a}} = \mathbf{X}^{\mathrm{f}} - \frac{1}{M}\tilde{\mathbf{X}}^{\mathrm{f}}\tilde{\mathbf{D}}_M^{\mathrm{T}}\mathbf{U}\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{D}_M\mathbf{1}_M\mathbf{1}_M^{\mathrm{T}} - \tilde{\mathbf{X}}^{\mathrm{f}}\tilde{\mathbf{D}}_M^{\mathrm{T}}\mathbf{U}(\mathbf{I}-\mathbf{S}^{-1})^{1/2}, \tag{7a}$$

$$\frac{1}{K-1}\tilde{\mathbf{D}}_K\tilde{\mathbf{D}}_K^{\mathrm{T}} \overset{\mathrm{SVD}}{=} \mathbf{U}\mathbf{S}\mathbf{U}^{\mathrm{T}}. \tag{7b}$$

## 2.3 | Single and double VAE–DA

In this section, the DA configurations that are going to be used in the experiments of Section 4 are formulated. In particular, we describe how the matrices $\mathbf{X}^{\mathrm{f}}$, $\mathbf{X}^{\mathrm{a}}$, $\mathbf{D}_K$, and $\mathbf{D}_M$, as well as the anomaly matrices $(\tilde{\mathbf{X}}^{\mathrm{f}}, \tilde{\mathbf{X}}^{\mathrm{f}}, \dots)$ introduced in Section 2.2, are replaced with equivalents lying in the latent space of one (*single ETKF–VAE* configuration) or two VAEs (*double ETKF–VAE* configuration). The description of the architecture and training procedures for these VAEs is postponed to Section 3.2. The main work hypothesis at the basis of the proposed approaches is that in the latent space the ensembles, that is, the column vectors in $\mathbf{X}^{\mathrm{f}}$, $\mathbf{X}^{\mathrm{a}}$, $\mathbf{D}_K$, and $\mathbf{D}_M$, respectively, are more Gaussian-distributed in the latent space than in the state space.

### 2.3.1 | Single ETKF–VAE

In the *single ETKF–VAE*, $\mathbf{X}^{\mathrm{f}}$ and $\mathbf{X}^{\mathrm{a}}$ in Equation (7a), which represent a model state directly before and directly after the application of DA, are replaced by their latent-space counterparts, $\mathbf{Z}^{\mathrm{f}} \in \mathbb{R}^{N_{z_1}\times M}$ and $\mathbf{Z}^{\mathrm{a}} \in \mathbb{R}^{N_{z_1}\times M}$, respectively, representing states in the latent space of the VAE just before and just after the application of the DA. With the necessary alterations, the same is happening with $\tilde{\mathbf{X}}^{\mathrm{f}}$ and $\tilde{\mathbf{X}}^{\mathrm{a}}$. Their vector columns, the ensemble members in the latent space, $\mathbf{z}_m^{\mathrm{f}}$ and $\mathbf{z}_m^{\mathrm{a}}$, are related to the original ensemble members in the physical space, the column vectors $\mathbf{x}_m^{\mathrm{f}}$ and $\mathbf{x}_m^{\mathrm{a}}$, via a VAE trained on model states (see Section 3 for details). Their statistical relations are $\mathbf{z}_m^{\mathrm{f}} \sim \mathcal{N}(\mu_{\phi_l}(\mathbf{x}_m^{\mathrm{f}}), \Sigma_{\phi_l}(\mathbf{x}_m^{\mathrm{f}}))$ and $\mathbf{x}_m^{\mathrm{a}} \sim \mathcal{N}(\mu_{\theta_l}(\mathbf{z}_m^{\mathrm{a}}), \Sigma_{\theta_l}(\mathbf{z}_m^{\mathrm{a}}))$. Next to this, the linear observation operator $\mathbf{H}$ appearing in Section 2.2 is replaced with a potentially nonlinear operator $H$. In the *single ETKF–VAE*, $\mathbf{D}_K$ and $\mathbf{D}_M$ remain as defined in Section 2.2. In particular, any non-Gaussianity present in $\mathbf{D}_K$ and $\mathbf{D}_M$ stemming from the ensemble contained in $\mathbf{X}^{\mathrm{f}}$ via $H$, the nonlinearity of $H$, or the non-Gaussianity of the observational errors remains unaddressed. This *single ETKF–VAE* is illustrated in the first row of Figure 1 and in Algorithm 2 in Appendix A.

### 2.3.2 | Double ETKF–VAE

In the *double ETKF–VAE*, a second VAE is trained on samples of the form $H(\mathbf{x}_{m_i}^{\mathrm{f}}) + \epsilon_{m_i}^{\mathbf{y}} - H(\mathbf{x}_{m_j}^{\mathrm{f}})$, with $m_i$, $m_j$ chosen randomly from $\{1, 2, \dots, M\}$ and $\epsilon_{(i)}^{\mathbf{y}}$ a realisation of the observational error probability distribution under

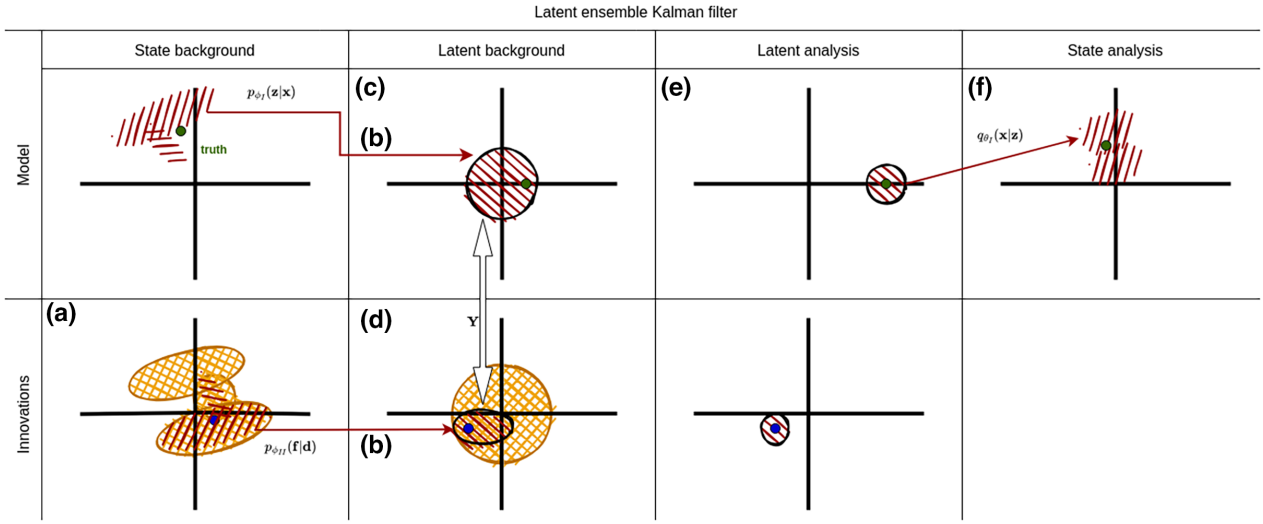


**FIGURE 1** Schematic overview of (top) *single ETKF–VAE* and (top+bottom) *double ETKF–VAE* approaches. (a) Alternative innovations (see Section 2.3.2) are generated by drawing ensemble members and adding realisations of the observational error, (b) the first and second VAE are trained on the forecast ensemble and alternative innovations, respectively, (c) the first encoder is used to sample one ensemble member in latent space for each ensemble member in state space, (d) the innovation encoder is used to sample $K$ perturbed innovations and $M$ unperturbed innovations in latent space, (e) the ETKF is performed using the ensembles of states, perturbed innovations, and innovations, and (f) for each member in the analysis ensemble, the first decoder samples a member in the state space.

the condition that $\mathbf{x}^{\text{truth}} = \mathbf{x}^{m_i}$. This assumes that the true observational error distribution is known. Given that in this study we defined the truth, this assumption holds trivially. In a realistic setup, estimating the observational error distribution is more difficult; see, for example, Tandeo *et al.* (2020), but this is no different from what is currently common in DA practice. These vectors represent possible innovations associated with different ensemble members in the absence of additional knowledge about the truth state. This includes any knowledge contained in the observation. Notice that, had the ensemble been infinite, the columns of $\mathbf{D}_M$ and $\mathbf{D}_K$ would be among the innovations created in this way.

The encoder part of this second VAE is then used to sample, for each of the columns in $\mathbf{D}_M$, a vector in the second latent space $\mathcal{Z}_2$. These are then collated as columns of $\mathbf{F}_M \in \mathbb{R}^{\mathcal{Z}_2 \times M}$. Similarly, a vector is sampled for each of the columns of $\mathbf{D}_K$, creating $\mathbf{F}_K \in \mathbb{R}^{N_{\mathcal{Z}_2} \times K}$. $\mathbf{F}_M$ and $\mathbf{F}_K$ replace $\mathbf{D}_M$ and $\mathbf{D}_K$ respectively in Equation (7a). In addition to this, $\mathbf{X}^{\text{f}}$ and $\mathbf{X}^{\text{a}}$ are replaced with $\mathbf{Z}^{\text{f}} \in \mathbb{R}^{\mathcal{Z}_1 \times M}$ and $\mathbf{Z}^{\text{a}} \in \mathbb{R}^{N_{\mathcal{Z}_1} \times M}$, respectively, as was already outlined in Section 2.3.1. Equation (7a) after these substitutions becomes

$$\mathbf{Z}^{\text{a}} = \mathbf{Z}^{\text{f}} - \frac{1}{M}\tilde{\mathbf{Z}}^{\text{f}}\tilde{\mathbf{F}}_M^{\text{T}}\mathbf{U}\mathbf{S}^{-1}\mathbf{U}^{\text{T}}\mathbf{F}_M\mathbf{1}_M\mathbf{1}_M^{\text{T}} - \tilde{\mathbf{Z}}^{\text{f}}\tilde{\mathbf{F}}_M^{\text{T}}\mathbf{U}(\mathbf{I} - \mathbf{S}^{-1})^{1/2}, \tag{8a}$$

$$\frac{1}{K-1}\tilde{\mathbf{F}}_K\tilde{\mathbf{F}}_K \overset{\text{SVD}}{=} \mathbf{U}\mathbf{S}\mathbf{U}^{\text{T}}. \tag{8b}$$

The procedure is visualised in Figure 1 and summarised as Algorithm 3 in Appendix A. A diagram with an overview of the steps in one iteration of the *ETKF*, *ETKF–VAE*$^{\text{single}}$, and *ETKF–VAE*$^{\text{double}}$ algorithms is included as Figure 2.

## 3 | EXPERIMENTAL SETUP

This section starts with a description of the model and VAE used in the experiments testing different scenarios. These are followed by the results produced in the experiments.

### 3.1 | The dynamical model

The *single* and *double ETKF–VAEs* are tested using a conceptual model, a discrete map that moves a point along a circle in a 2D plane ($\mathcal{X} = \mathbb{R}^2$). The position at time $t+1$ is given in polar coordinates by

$$r(t_{p+1}) = r(t_p) + \Delta t A\omega\cos(\omega t_p), \tag{9a}$$

$$\psi(t_{p+1}) = (1 + \alpha\Delta t)\psi(t_p), \tag{9b}$$

with $\psi \in [0, 2\pi)$, or, in Cartesian coordinates,

$$x(t_{p+1}) = x(t_p)\cos\alpha\Delta t\psi(\mathbf{x}(t_p)) - y(t_p)\sin\alpha\Delta t\psi(\mathbf{x}(t_p)) + \Delta t\frac{x(t_p)}{\|\mathbf{x}(t_p)\|}A\omega\cos(\omega t_p), \tag{10a}$$

$$y(t_{p+1}) = x(t_p)\sin\alpha\Delta t\psi(\mathbf{x}(t_p)) + y(t_p)\cos\alpha\psi(\mathbf{x}(t_p)) + \Delta t\frac{y(t_p)}{\|\mathbf{x}(t_p)\|}A\omega\cos(\omega t_p), \tag{10b}$$

$$\psi(\mathbf{x}(t_p)) = 2\arctan\frac{y(t_p)}{x(t_p) + \|\mathbf{x}(t_p)\|} \bmod 2\pi, \tag{10c}$$

with $\Delta t = 1$, $\alpha = 0.1$, $\omega = 2\pi/50$, $\mathbf{x}(t_p)$ the position in the 2D plane at time step $p$, time $t_p$, $x$ the position along the horizontal axis, $A$ the amplitude of the oscillation of the radius around 1, $y$ the position along the vertical axis, and $\psi(\mathbf{x})$ the polar coordinate of $\mathbf{x}$ between 0 and $2\pi$. From Equation (9), it is easy to see that the model has a set of stationary points given by $\{(r, \psi) : \psi = 0\}$ if $A = 0$, that is,

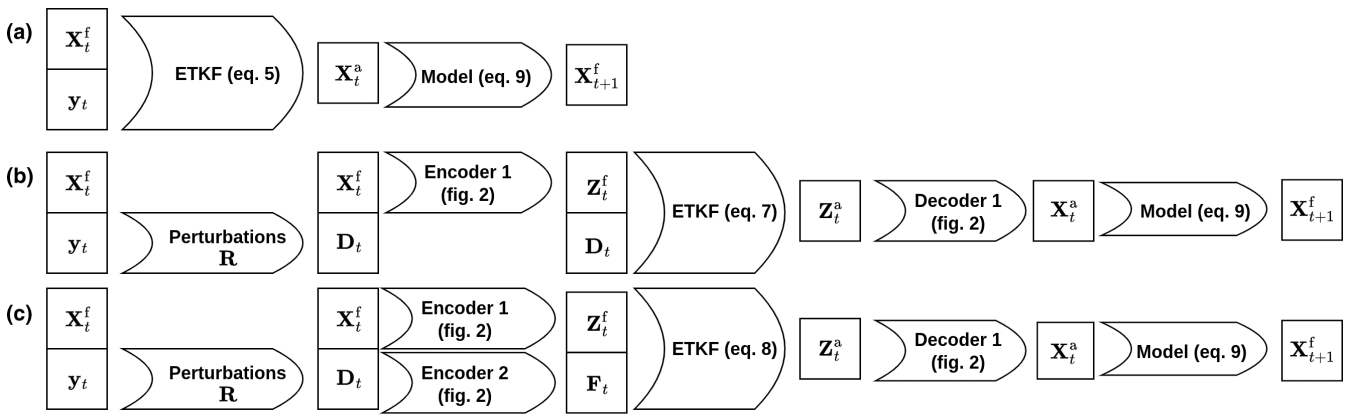


**FIGURE 2** Overview of the different steps in a single iteration of (a) *ETKF*, (b) *ETKF–VAE*$^{\text{single}}$, and (c) *ETKF–VAE*$^{\text{double}}$, together with references to the equations/architecture used in each step.

the positive $x$-axis. Furthermore, it is also a chaotic discrete mapping with a Lyapunov exponent $> 0$ (see Appendix B). Despite being simple, this model poses two difficulties for the EnKF that are exemplars of those it would face when applied to sea-ice models. First, if $A = 0$, the solution is constrained to a submanifold: the unit circle. Second, near $\psi \approx 0$ the ensemble has the potential to become bimodal, as positions with polar coordinate $\delta\psi$, $0 < \delta\psi \ll 2\pi$, are moved to $(1+\alpha)\delta\psi$, whilst those with coordinate $-\delta\psi$ are mapped to $-(1+\alpha)\delta\psi + 2\alpha\pi$. This last feature makes the model in the study more challenging than the commonly used Lorenz-63, for which trajectories also lie in a submanifold, but which does not possess a discontinuity. Because of these challenges, we opted for this model over Lorenz-63, for which it is already known that the conventional ETKF works well (Bocquet, 2011; Bowler, 2006; Sakov *et al.*, 2012).

## 3.2 | Network architecture

The means and logarithms of variance appearing in Equation (3a–3d) are produced by a multilayer perceptron. The architecture of this type of network is depicted in Figure 3. The encoder network consists of a single input layer accepting $\mathbf{x}$ and two sequences of six fully connected hidden layers with 32 nodes each, activations layers, an output layer, and a rescaling layer. One of the six-layer sequences renders $\mu_\phi$, the other $\Sigma_\phi$. The activation layers are leaky rectified linear units with a leakage factor of 0.1 and the diagonal elements $\Sigma'_{\text{def}}$ in Equation (3b) set equal to $0.05^2$. The latent space is chosen to be one-dimensional. The rescaling layer represents an affine transformation $\mathbf{z}^{\text{f}} \to a\mathbf{z}^{\text{f}} + b$. Prior to the training of weights $\phi_1$ and $\theta_1$, $\mathbf{x}^{\text{f}}_1, \ldots, \mathbf{x}^{\text{f}}_M$ are fed through the decoder and $a$, $b$ are chosen such that $\mathbf{z}^{\text{f}}_1, \ldots, \mathbf{z}^{\text{f}}_M$ have a sample mean of 0 and sample variance of 1. This rescaling layer is applied in an attempt to speed up this maximisation, setting the first two statistical moments of the distribution equal to that of the desired distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$. After fixing $a$, $b$, the ELBO is maximised to find $\phi_1$ and $\theta_1$. The decoder consists of a similar pair of six-layer networks. One member of the pair produces $\mu$ and the other $\ln\Sigma$. Before entering the six-layer network, the inverse affine transformation is applied to the latent state, that is, $\mathbf{z} \to (1/a)(\mathbf{z} - b)$. The architecture of the second VAE is the same; the number of input nodes for the encoder and output nodes for each six-layer network of the decoder is equal to $N_y = 1$.

Prior to any DA, the first VAE is trained on what will be called the climatology run. This climatology run is generated by running the model in Equation (10a)–(10c) with $A = 0$ for 10,000 steps. Each 10th time step is set

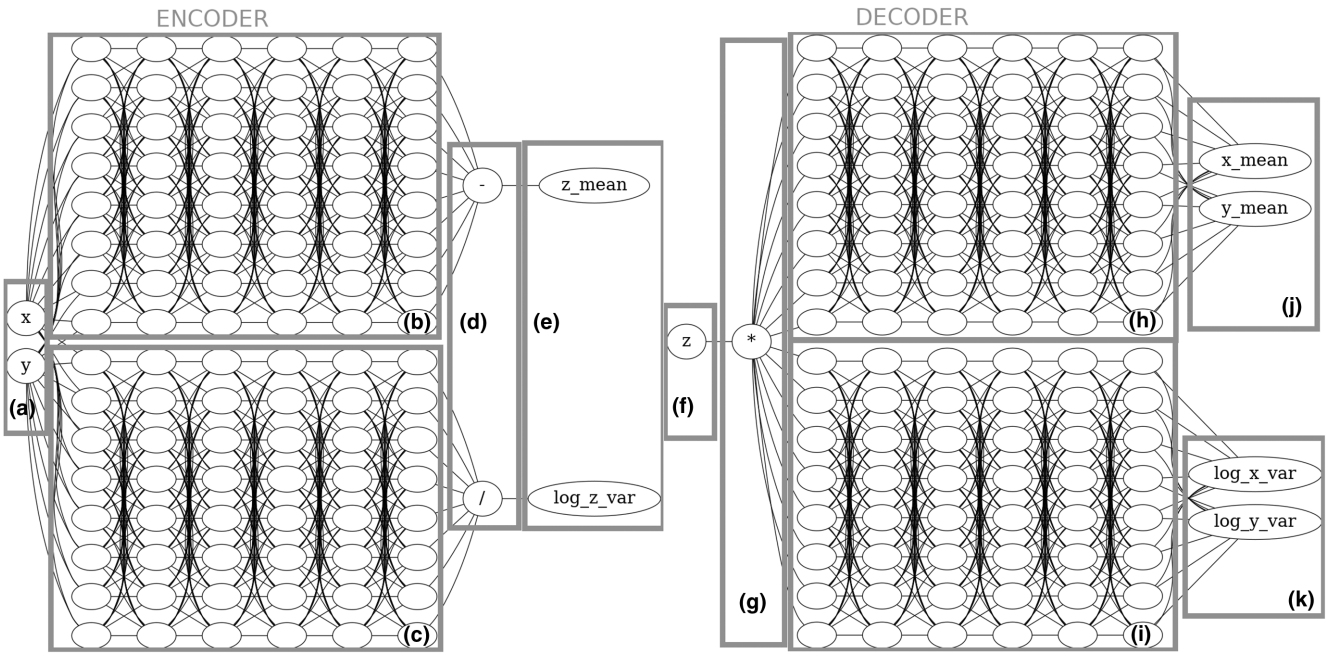


**FIGURE 3** Architecture of the first variational autoencoder. The encoder consists of (a) common input nodes, (b) six hidden layers for the encoder predicting the mean of the conditional distribution $\mu_\phi$, and (c) six layers for the encoder predicting its variance $\ln\Sigma_\phi$. Latent mean and log variance are (d) rescaled using an affine transformation (see text) and (e) outputted. The decoder (f) accepts a latent state as input, (g) applies the inverse of the latent affine transformation to the value and feeds the value to (h,i) a pair of six hidden layers outputting the (j) mean $\mu_\theta$ and (k) log of variance $\ln\Sigma_\theta$ in state space. For clarity, only eight of the 32 nodes in each hidden layer are shown.

aside and the VAE is trained on this following a He normal procedure (He *et al.*, 2015) to initialise $\phi_1$, $\theta_1$ before training. For minimisation of the ELBO Equation (1c), an Adaptive Momentum Estimation (ADAM) solver (Kingma & Ba, 2017) is used. The initial learning rate is $5.0 \times 10^{-3}$, but this is halved every time the reduction in the ELBO is smaller than 0.1 for two epochs in a row, until the learning rate reaches its minimum of $1.0 \times 10^{-6}$. Minimisation is ended when the reduction in ELBO is less than 0.1 for five epochs in a row or after 50 epochs, but not before 20 epochs have passed. ELBO values vary depending on the training data, but typically they are reduced from values between 10 and 100 to values between −10 and 0.

Before settling on an architecture with six hidden layers and 32 nodes per layer, alternative layers/nodes per layer were tried. It was found that architectures with 6–10 hidden dense layers and 32–64 nodes are capable of reconstructing the unit circle. To illustrate this capability of the VAE, 1000 samples $\mathbf{z}_i$ were drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})_{\mathcal{Z}}$. For each $i$, $\mathbf{x}_i \sim \mathcal{N}(\mu_{\theta_1}(\mathbf{z}_i), \Sigma_{\theta_1}(\mathbf{z}_i))$ was drawn and added as a dot to Figure 4a. The resulting point cloud is circular; however, somewhat spread out in a band around the circle. Next to this, a sample $\mathbf{z}_j \sim \mathcal{N}\left(\mu_{\phi_1}(\mathbf{x}_j), \Sigma_{\phi_1}(\mathbf{x}_j)\right)$ was taken for each sample $\mathbf{x}_j$ from the climatology run. The PDF estimated from these samples $\mathbf{z}_j$ is shown in Figure 4b. The figure testifies that the latent distribution approximates a standard normal, as expected based on the relation $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})\, d\mathbf{z}$ in Section 2, but it also shows that the relation is not perfect, as the estimated PDF is more strongly weighted towards $\mathbf{z} = 0$ than the standard normal.

## 3.3 | Configurations under study

For both the *single ETKF–VAE* and *double ETKF–VAE* architectures outlined in Section 2.3, two configurations are created, which differ in the type of training. In configurations with names that contain *clima*, the weights of the first VAE (i.e., the VAE concerned with the model states) are copied from the VAE trained on the climatology run (see Section 3.2). In particular, $\phi_1$ and $\theta_1$ are left unchanged during the execution of the experiments. This choice, intentionally referred to as *clima*, mimics a general situation whereby one interrogates a dataset representing an (assumed) stationary process. As mentioned in the previous section, the training is done using a 10,000-time-step long trajectory, sampled every 10 time steps.

On the other hand, in the *transfer* configurations, the weights of the first VAE are retrained at each analysis step using the ensemble-member current forecasts $\mathbf{x}_1^{\mathrm{f}}, \ldots, \mathbf{x}_M^{\mathrm{f}}$ as a training set. At each analysis time step, the weights, prior to training, are initialised by copying from the weights trained on the climatology run.

The second VAE is always trained at each analysis step in both the *clima* and *transfer* configurations. During the initialisation, weights are copied from the first VAE where possible, that is, for the nodes that have the same number of connections in the first and second VAEs. The remaining weights are initialised using He normalisation. An alternative approach, in which weights were initialised from $\phi_1$, $\theta_1$ obtained during the previous analysis time step, has been tried. This approach was found to be

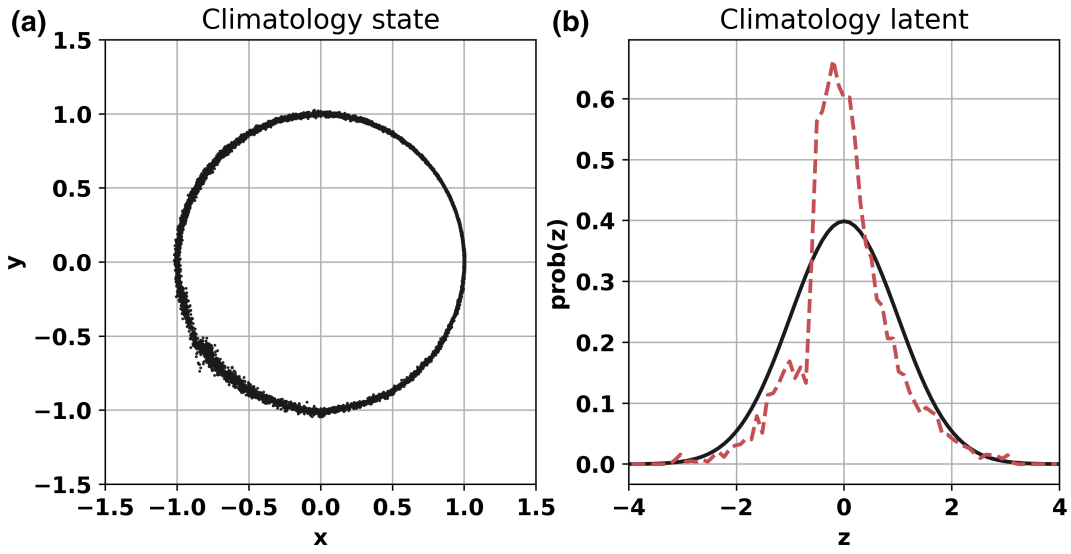


**FIGURE 4** (a) Climatology run in physical space generated by feeding samples from a standard normal distribution and drawing a sample from each of these using the first decoder. (b) Climatology run in latent space obtained by taking states from the climatology run, and for each state drawing a sample in latent space using the first encoder (red) together with the standard normal (black). The network architecture used six dense hidden layers and 32 nodes per hidden layer.

**TABLE 2** DA system configurations, together with (first column) states that are mapped to/from the latent space(s), and the datasets used for VAE training (second column) prior to the first DA step and (third column) prior to each DA step. Here Im stands for image, $\mathbf{X}$ contains the ensemble members in physical space, $\mathbf{Z}$ the ensemble members in the latent space of the first VAE, $\mathbf{D}$ the (perturbed) innovations in observation space, and $\mathbf{F}$ the (perturbed) innovations in the latent space of the second VAE.

| Configuration | VAE mappings | Offline training set | Online training set |
|---|---|---|---|
| No DA | – | – | – |
| ETKF | – | – | – |
| $ETKF\text{–}VAE^{\text{single}}_{\text{clima}}$ | $\mathbf{X}^{\text{f}} \rightarrow \mathbf{Z}^{\text{f}}, \mathbf{Z}^{\text{a}} \rightarrow \mathbf{X}^{\text{a}}$ | $\{\mathbf{x}^{\text{clima}}(t_p) : p \in P^{\text{clima}}\}$ | — |
| $ETKF\text{–}VAE^{\text{double}}_{\text{clima}}$ | $\mathbf{X}^{\text{f}} \rightarrow \mathbf{Z}^{\text{f}}, \mathbf{Z}^{\text{a}} \rightarrow \mathbf{X}^{\text{a}}, \mathbf{D}^{\text{f}} \rightarrow \mathbf{F}^{\text{f}}$ | $\{\mathbf{x}^{\text{clima}}(t_p) : p \in P^{\text{clima}}\}$ | $\text{Im}\,\mathbf{D}^{\text{f}}(t_p)$ |
| $ETKF\text{–}VAE^{\text{single}}_{\text{transfer}}$ | $\mathbf{X}^{\text{f}} \rightarrow \mathbf{Z}^{\text{f}}, \mathbf{Z}^{\text{a}} \rightarrow \mathbf{X}^{\text{a}}$ | $\{\mathbf{x}^{\text{clima}}(t_p) : p \in P^{\text{clima}}\}$ | $\text{Im}\,\mathbf{X}^{\text{f}}(t_p)$ |
| $ETKF\text{–}VAE^{\text{double}}_{\text{transfer}}$ | $\mathbf{X}^{\text{f}} \rightarrow \mathbf{Z}^{\text{f}}, \mathbf{Z}^{\text{a}} \rightarrow \mathbf{X}^{\text{a}}, \mathbf{D}^{\text{f}} \rightarrow \mathbf{F}^{\text{f}}$ | $\{\mathbf{x}^{\text{clima}}(t_p) : p \in P^{\text{clima}}\}$ | $\text{Im}\,\mathbf{X}^{\text{f}}(t_p)$, $\text{Im}\,\mathbf{D}^{\text{f}}(t_p)$ |

unstable, as over time the weights lost information about the shape of the circle and the members of the ensemble ended up spreading out through the domain, losing any information in it. Consequently, this approach was not pursued further.

In addition to these configurations, a configuration without DA (*no DA*) and one using the standard ETKF (*ETKF*) have been added to facilitate comparison and benchmarking. An overview of the six configurations can be found in Table 2.

**TABLE 3** Model setting in different experiments. Here $\mathcal{N}$ is the normal, or Gaussian, distribution, while $\mathcal{S}$ is the skewed normal distribution (see Equation 12).

| Experiment | Climate type | Obs. error |
|---|---|---|
| I | Stationary ($A = 0$) | Gaussian $x \sim \mathcal{N}(x^{\text{truth}}, 0.1^2)$ |
| II | Non-stationary ($A = 0.2$) | Gaussian $x \sim \mathcal{N}(x^{\text{truth}}, 0.1^2)$ |
| III | Stationary ($A = 0$) | Non-Gaussian $x \sim \mathcal{S}$ ($mode = x^{\text{truth}}, var = 0.1^2$) |

## 4 | NUMERICAL EXPERIMENTS AND RESULTS

Using the experimental setup in Section 3, four twin experiments are carried out. In each of these experiments, 65 model instances are run forward for 500 time steps. One of these model runs serves as the artificial truth, while the remaining 64 form the ensemble. Initially, all 65 members are located on the unit circle with polar angles drawn from a uniform distribution on $[-0.1\pi, 0.1\pi]$. A summary of the different model settings and observations can be found in Table 3. Each experiment is repeated 49 times using seven different long "climatology runs" used for training of the VAE. Each of these climatology runs differs only in the initial position of the particle, and for each of them the experiments are repeated using seven different initial ensembles and observations.

Initially we will use two evaluation metrics in Section 4.1: the root-mean-square error (RMSE) and the correlation between forecasts and truth. Evaluating DA performance by comparing the ensemble mean with the truth is standard practice for EnKFs. This is because, in an EnKF, the ensemble mean represents the most likely estimate for the truth state of the system. However, this equivalence between mean and mode does not hold if the a priori distribution is non-Gaussian, as is the case with the experiments in this work. Loosely formulated, when the distribution is Gaussian, the mean, as the most likely state, is indicative of what the true state can be, while for a non-Gaussian distribution the mean provides insufficient (or misleading) information about the true state. The mean may even be an unrealisable state if the distribution is non-Gaussian. Next to this, an accurate representation of the distribution itself could be of practical interest, for example, to assess the probability that extremes occur. Therefore, metrics based on the ensemble mean might not be the best measure to evaluate the performance of the DA system. Instead, we resort to the continuous rank probability score (CRPS) as our preferred metric of performance. The CRPS is a measure for the $L_2$ error in the cumulative probability distribution (CDF) and is defined as

$$\text{CRPS} = \int_{-\infty}^{\infty} \mathbb{E}[\left(\text{CDF}(w) - \mathcal{H}(w - w^{\text{truth}})\right)^2]\, dw, \quad (11)$$

where $\mathcal{H}$ is the Heaviside function and $w$ can be either the $x$-coordinate, $y$-coordinate, radius, or angle, the integral is approximated with a numerical Lebesgue integral using the ensemble values $w^{(m)}$ for $1 \leq m \leq M$, and

the expectation value is calculated over all times for which observations are available and over repetitions of the experiment. See Tödter and Ahrens (2012) for details.

## 4.1 | Stationary climate

In this experiment, the truth moves along the unit circle. The system is autonomous and stationary: the circle has a constant unit radius. Every 10 time steps, the $x$-coordinate is assimilated. Observations are drawn from the Gaussian distribution with a standard deviation of 0.1 and centred on the $x$-coordinate of the truth; they are therefore unbiased.

The $x$-coordinate, $y$-coordinate, radius (distance to origin), and polar angle of the mean of the forecast/analysis ensemble are calculated just before (after) the DA correction is applied. These mean values in the different DA configurations are compared with their true values. For each repetition of the experiment, the variance of the ensemble mean, that is, the forecast/analysis, over time and the correlation of the ensemble mean with the truth are computed. These are then averaged over all repetitions of the experiment. For the correlations, a weighted average is used, that is, the correlation for the forecast as shown in Figure 5 is produced as

$$\rho^{\mathrm{f}} = \frac{\sum_{j=1}^{999} \rho_j^{\mathrm{f}} \sigma_j^{\mathrm{truth}} \sigma_j^{\mathrm{f}}}{\sum_{j=1}^{999} \sigma_j^{\mathrm{truth}} \sigma_j^{\mathrm{f}}},$$

with $\rho_j^{\mathrm{f}}$ the correlation between the time series of a variable ($x$-coordinate/$y$-coordinate/radius/angle) obtained from the truth on one hand and the forecast mean for said variable of the $j$th bootstrap sample on the other hand, $\sigma_j^{\mathrm{truth}}$ the time-series standard deviation for the true value of the variable in the $j$th bootstrap sample, and $\sigma_j^{\mathrm{f}}$ the standard deviation of the forecast mean over time in the $j$th bootstrap sample. The same procedure is followed for $\rho^{\mathrm{a}}$. Results are shown in Figures 5 and 6. The confidence intervals for standard deviations and correlations, as well as other metrics in the subsequent sections, have been determined using bias-corrected and accelerated bootstrapping (Efron, 1987) with 999 bootstrap samples. They are shown as error bars in Figures 5 and 6. Some of them are so small that they disappear below the markers. Also shown in these diagrams, as dashed lines, are the average RMSEs between the forecast/analysis means and the truth.

Figure 5 shows that, in both $x$- and $y$-coordinates and in the polar angle, the *ETKF* and the *ETKF–VAE* have

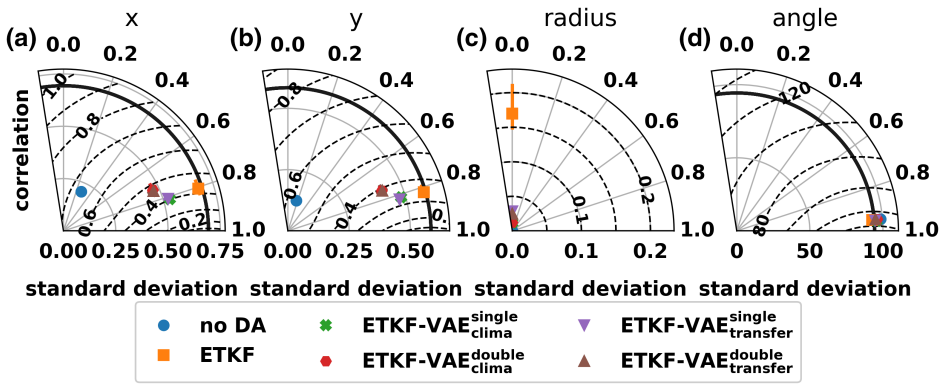


**FIGURE 5** Taylor diagrams for (a) $x$-coordinate, (b) $y$-coordinate, (c) radius, and (d) angle. The standard deviation of the time series of the forecast mean is shown along the radial, the correlation with the truth along the azimuthal, and the RMSE as dashed lines. Bars indicate the 90% confidence interval.

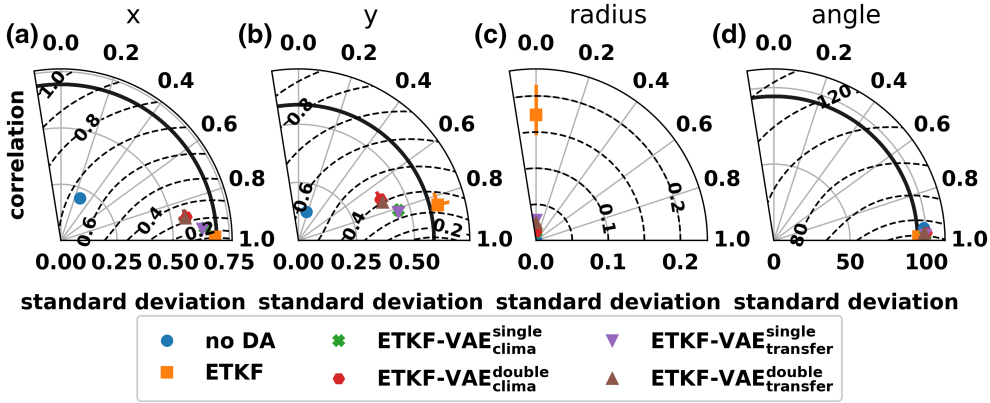


**FIGURE 6** As Figure 5, but now showing standard deviations and correlations with the mean of the analysis ensemble.

similar RMSEs, though the *ETKF–VAE* configurations all underestimate the variability in the *x*- and *y*-coordinates slightly. The situation is different for the radius. Note first that, for the truth and *No DA*, this value is constant, the standard deviation is zero, and it is not possible to calculate correlations between the different configurations and the truth. In the *ETKF* configuration, the standard deviation of the radius time series is $0.17 \pm 0.04$, indicating that in this experiment the reconstructed (forecast) radius is far from being a constant unit. The *ETKF–VAE* configurations perform better in this regard and have standard deviations smaller than $(2.8 \pm 0.4) \times 10^{-2}$. Overall, the *double ETKF–VAE* configurations perform worse, in terms of the RMSE, than their *single ETKF–VAE* counterparts. We attribute this to the fact that sampling the innovation vector in latent space adds an additional error. Part of this comes from the inherent stochastic nature of the encoder encapsulated in $\Sigma_\phi$ (see Appendix C for details), apart from the imperfect training of the VAE. These errors increase the spread in the latent innovation ensemble and consequently increase the covariance $\mathbf{H}\mathbf{P}_{\mathbf{X}}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}$ as well as the ensemble mean of the latent innovation vectors and consequently make smaller DA corrections. When we compare the forecast with the analysis metric in Figure 6, the performance of the different configurations is qualitatively the same. The main difference can be found in Figure 6a, which shows that all configurations, except *No DA*, exhibit higher correlations with the truth and smaller RMSEs. This is in line with expectation, as the *x*-coordinate is assimilated directly during the experiment.

The CRPS values for the different configurations and variables in this experiment are shown in Figure 7a. CRPS values for the *single ETKF–VAE* configurations are smaller (i.e., better) than those for the *ETKF* for all variables, though not significantly at the 90% confidence level for the angle. When using the *double ETKF–VAE* configurations, no CRPS reduction is achieved compared with *ETKF* for the *x*-coordinate, *y*-coordinate, and angle. Measured by the CRPS of the radius, all ETKF–VAE configurations perform better than the *ETKF*. The reason for the improved performance for the radius is clearly visible in Figure 7a. In this figure the truth, together with the forecast and analysis ensembles, is shown for time 360. In the *ETKF*, ensemble

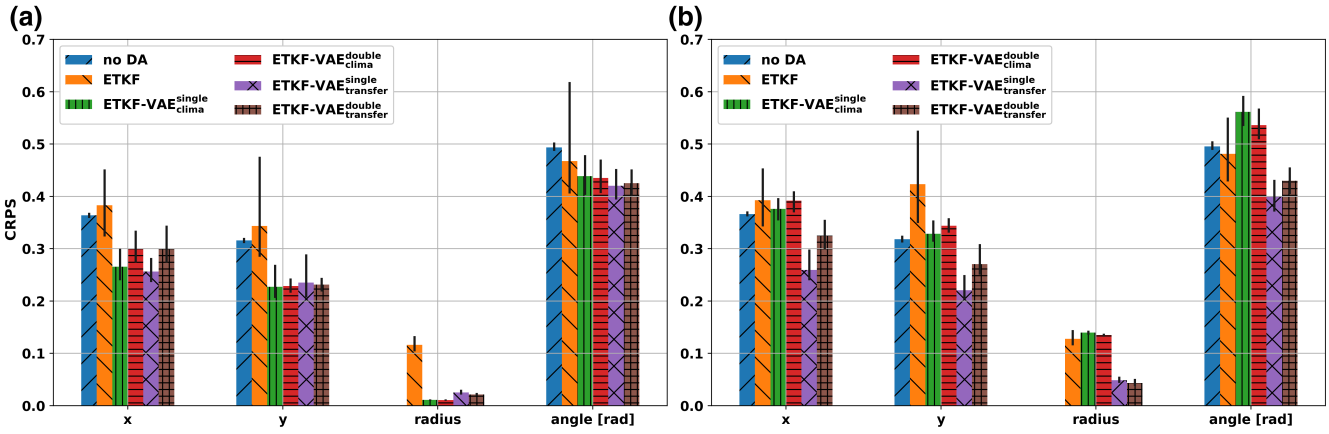


**FIGURE 7** CRPS of the *x*-coordinate, *y*-coordinate, radius, and angle for the experiments with (a) stationary climatology run and (b) non-stationary climatology run. Black lines indicate the 90% confidence interval.

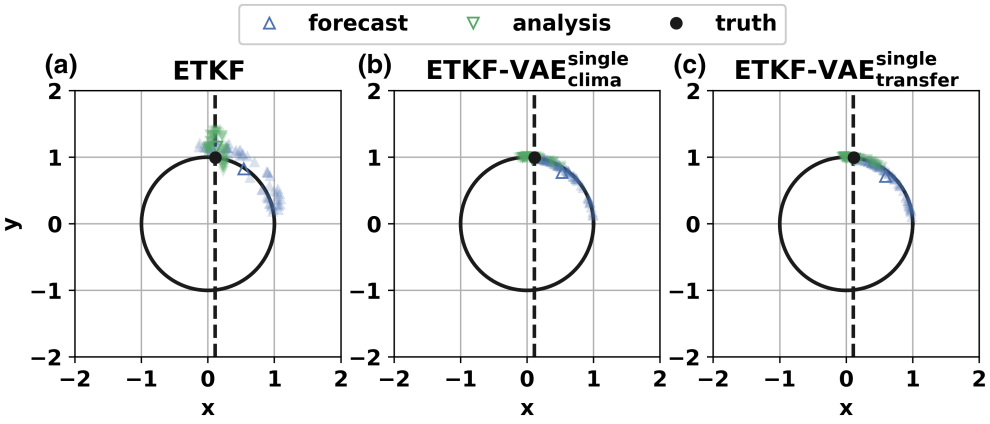


**FIGURE 8** Truth (black), forecast ensemble (blue), and analysis ensemble at time 360 (green) for (a) *ETKF*, (b) *ETKF–VAE*$^{\text{single}}_{\text{clima}}$, and (c) *ETKF–VAE*$^{\text{single}}_{\text{transfer}}$. Ensemble means are depicted as triangles. The assimilated value of the *x*-coordinate is depicted as a dashed black line.

members are moved very close to the truth, sometimes even closer than in the ETKF–VAE configurations. This reduces the RMSE effectively, but in doing so they move off the circle. On the other hand, in the ETKF–VAE configurations, the analysis lies in the image of the first decoder and, if properly trained, this ensures that the analysis members end up with the correct radius close to the circle. Therefore the benefit of using the VAE lies mainly in its ability to restrict the analysis ensemble to the manifold of physically possible solutions.

## 4.2 | Non-stationary climate

In the previous section, it was shown that the use of the ETKF–VAE improves the CRPS for the radius drastically compared with *ETKF*, as the use of a decoder for the states ensures that ensemble members stay close to the circle. This raises the question of what would happen if the submanifold containing the model solution, in our case the circle, were to change over time. This is done to mimic the effects of, for example, the change in seasons in a sea-ice model or the effects of climate change in a general circulation model. To this end, we set $A = 0.2$ in Equation (10), so that the radius of the truth will vary slowly between 0.8 and 1.2 in this experiment.

Results relative to the CRPS from this non-stationary experiment are shown in Figure 7b. By comparing it with Figure 7a, we see that the CRPS for the $x$-coordinate in the *ETKF* is not impacted by the variation in radius of the truth, but the forecasting capacity for the non-observed $y$-coordinate is diminished. A larger increase in CRPS for all variables can be observed for the $ETKF\text{–}VAE^{\text{single}}_{\text{clima}}$ and $ETKF\text{–}VAE^{\text{double}}_{\text{clima}}$ configurations, while CRPS values for $ETKF\text{–}VAE^{\text{single}}_{\text{transfer}}$ and $ETKF\text{–}VAE^{\text{double}}_{\text{transfer}}$ are not significantly different at the 90% confidence level. We can deduce the cause of this difference by comparing Figures 8 and 9. When the submanifold is stationary ($A = 0$), *ETKF* places analysis members outside the submanifold (Figure 8a), while the first VAE ensures that, in $ETKF\text{–}VAE^{\text{single}}_{\text{clima}}$ (Figure 8b) and $ETKF\text{–}VAE^{\text{single}}_{\text{transfer}}$ (Figure 8c), the members stay on the circle. In the $ETKF\text{–}VAE^{\text{single}}_{\text{clima}}$ configuration (see Figure 9b), DA brings the ensemble members closer to the truth. However, in this non-stationary experiment the VAE in this configuration was trained on a climatology run in which the radius was fixed to 1 and it is therefore unaware of the fact that the radius of the truth changes during the model run. Consequently, the first decoder creates analysis ensemble members with a radius of 1 instead of the radius equal to that of the truth at that specific analysis time (1.12). In $ETKF\text{–}VAE^{\text{single}}_{\text{transfer}}$ (Figure 9c), the weights of the VAE are updated at each analysis time using the forecast ensemble. This results in the analysis ensemble members being positioned at similar polar angles to those in $ETKF\text{–}VAE^{\text{single}}_{\text{clima}}$, but now at the appropriate radius. Hence we conclude that online training of the VAE is essential if the submanifold holding the model solutions changes over time. The solution we adopted here is based on the use of the ensemble members that are already at our disposal whenever running any ensemble-based DA, thus making the method very versatile.

To test the robustness of our transfer learning approach, we have studied its performance against the

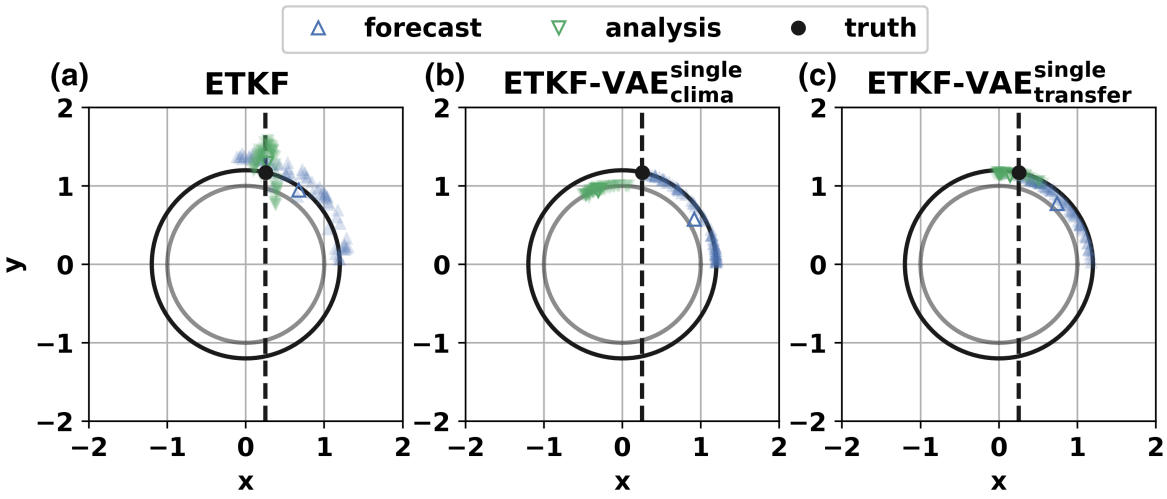


**FIGURE 9** Truth (black), forecast ensemble (blue), and analysis ensemble (green) with (dashed line) observed $x$-coordinate at time 360 for the same configurations as in Figure 9. Shown are the results for the model in Equation (10) with $A = 0.2$ instead of $A = 0.0$. A circle with the same radius as the truth at time 360 is depicted in black. In (b), the first VAE is trained on a climatological run in which the particle moves over the unit circle. This circle has been added in grey for reference.

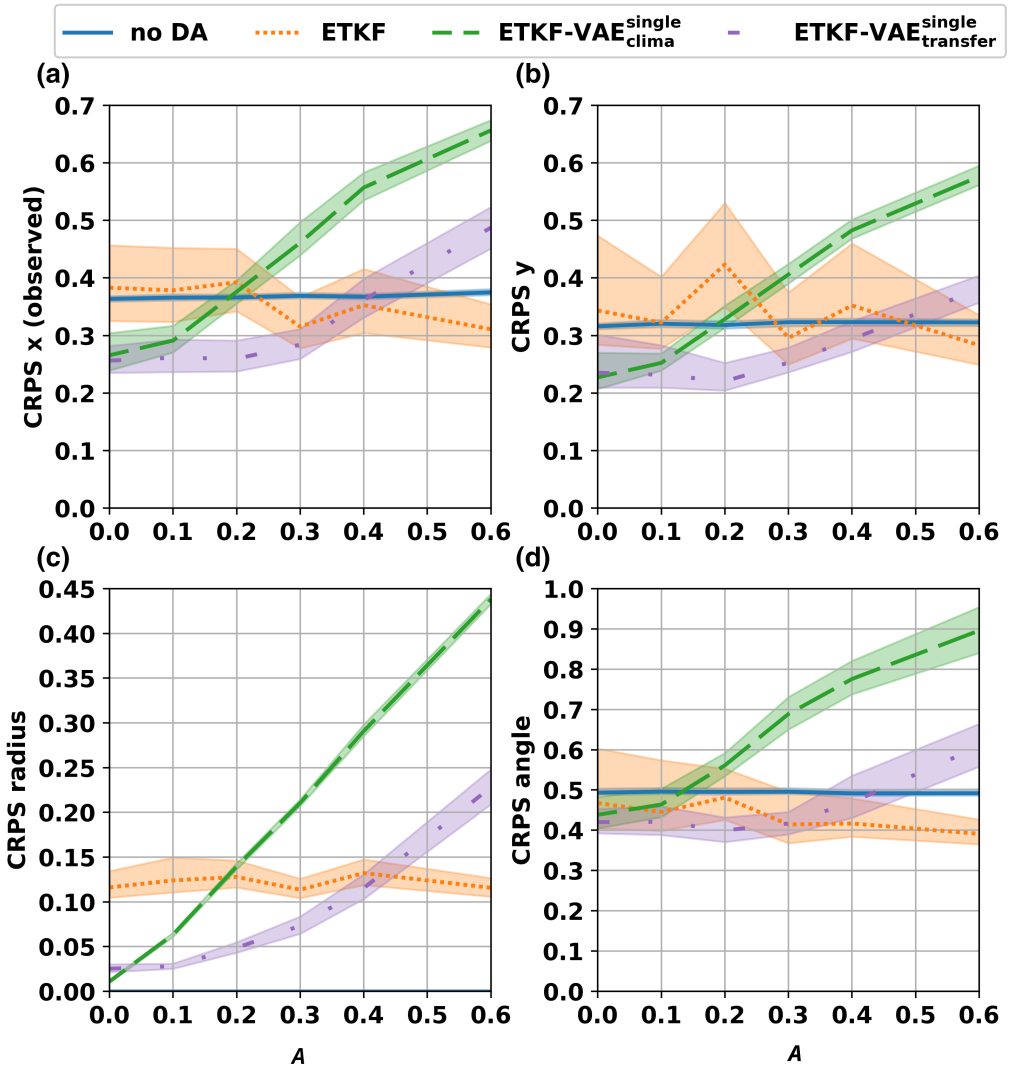


**FIGURE 10** CRPS for (a) observed $x$-coordinate, (b) $y$-coordinate, (c) radius, and (d) polar angle as a function of the circle radius rate of change $A$ in Equation (10). Results are shown for the *no DA*, *ETKF*, *ETKF–VAE*$^{\text{single}}_{\text{clima}}$, and *ETKF–VAE*$^{\text{single}}_{\text{transfer}}$ configurations. Bands indicate the 90% confidence interval of the CRPS.

parameter $A$ in Equation (10), which modulates the rate at which the radius of the circle changes over time. The results for the CRPS in different variables are shown in Figure 10. For the sake of clarity, *ETKF–VAE*$^{\text{double}}_{\text{clima}}$ and *ETKF–VAE*$^{\text{double}}_{\text{transfer}}$ are not shown, as the ratio of their CRPSs (see Figure 7b) is qualitatively comparable with the ratio of the CRPSs for *ETKF–VAE*$^{\text{single}}_{\text{clima}}$ and *ETKF–VAE*$^{\text{single}}_{\text{transfer}}$. The ETKF is unaware of the submanifold in which the truth moves and, consequently, also any changes in this manifold. Hence, the *ETKF* exhibits little dependence on $A$. The CRPS of *ETKF–VAE*$^{\text{single}}_{\text{clima}}$, on the other hand, increases sharply with increasing $A$ for all variables. This is consistent with the idea that, as $A$ increases, the placement of the analysis ensemble member near the unit circle places them further and further away from the truth. On the other hand, *ETKF–VAE*$^{\text{single}}_{\text{transfer}}$ succeeds in keeping the CRPS almost insensitive to $A$ for $A \leq 0.3$. It grows for $A > 0.3$, but keeps its values well below *ETKF–VAE*$^{\text{single}}_{\text{clima}}$. The deterioration of the *ETKF–VAE*$^{\text{single}}_{\text{transfer}}$ skill for $A > 0.3$ suggests that updating the weights of the first VAE using the ensemble fails when the actual radius of the truth is very different from the climatology run. In that case, the analysis ensemble members are placed in the vicinity of the unit circle (not shown).

## 4.3 | Non-Gaussian observations

So far, non-Gaussianity was limited to the ensemble members and, indirectly, to the projection of those ensemble members into the space of observations via the observation operator. Observation errors were assumed to be Gaussian with zero mean. In this section, we relax this assumption and instead assume that the observation errors are realisations from a skew normal distribution (Henze, 1986; O'Hagan & Leonard, 1976):

$$S(x) = \frac{1}{\pi\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \int_{-\infty}^{\lambda x} e^{-\frac{(\xi-\lambda\mu)^2}{2\sigma^2}}\, d\xi, \tag{12}$$

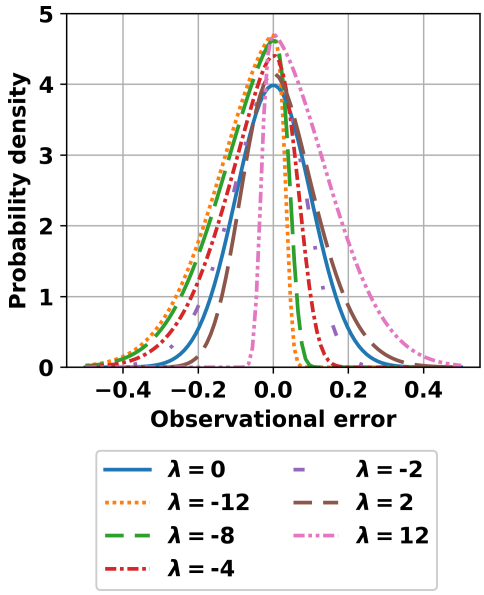


**FIGURE 11** Skew normal probability distributions for the observational errors for different values of the skewness parameter.

where $\lambda$ is the skewness parameter and $\mu$ and $\sigma$ are chosen such that the mode of the distribution is zero and the standard deviation equal to 0.1. By doing so, the skew normal distribution reverts to the Gaussian distribution used in Sections 4.1 and 4.2 if $\lambda = 0$. On the other hand, whenever $\lambda \neq 0$, the distribution Equation (12) is asymmetric with non-zero mean. The shape of the skew normal distribution for various values of the skewness parameter $\lambda$ is illustrated in Figure 11. The use of a skew normal violates the assumptions behind the KF, but it is not unlike a situation one encounters when assimilating, for example, satellite radiances (Saunders *et al.*, 2013; Zhu *et al.*, 2014). We are interested in exploring the capabilities of our ETKF–VAE approaches to cope with this scenario.

Given that our focus here is on the impact of non-Gaussian observational errors, in this section we consider only configurations in which the training is solely on the climatology run (no transfer learning) and the true circle's radius is fixed to one. CRPS values for the $x$-coordinate, $y$-coordinate, radius, and polar are shown in Figure 12 as functions of the skewness parameter $\lambda$. The figure shows that the performance of *ETKF* in all variables deteriorates as long as the observational error distribution becomes more skewed and consequently more biased. This agrees with the findings of Dee (2005), Lea *et al.* (2008), and Huang *et al.* (2020) that the EnKF performance can degenerate if observational errors are biased and no bias-correction scheme is applied. *ETKF–VAE*$^{\text{single}}_{\text{clima}}$ follows a similar pattern, but the impact of the bias is less pronounced. We believe that this is due to the decoder effectively limiting the extent to which the bias can impact the position of the ensemble members, given that it restricts them to the circle. The situation is radically different when looking at *ETKF–VAE*$^{\text{double}}_{\text{clima}}$. The observational error bias is contained in the synthetic innovations on which the second VAE, part of the *ETKF–VAE*$^{\text{double}}_{\text{clima}}$ configuration, is trained. When the biased distribution is mapped by the encoder to a standard normal in latent space, the bias is removed. Consequently, *ETKF–VAE*$^{\text{double}}_{\text{clima}}$ does not exhibit a dependence on skewness and outperforms *ETKF–VAE*$^{\text{single}}_{\text{clima}}$ whenever the absolute value of the skewness parameter exceeds 5.

To see how this compares with a conventional bias-correction scheme, CRPS scores were also compared with an ETKF in which the observation part of the cost function $\frac{1}{2}(\mathbf{y}-\mathbf{Hx})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{Hx})$ is replaced with $\frac{1}{2}(\mathbf{y}-\bar{\epsilon}_{\mathbf{y}}-\mathbf{Hx})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\bar{\epsilon}_{\mathbf{y}}-\mathbf{Hx})$. Here, $\bar{\epsilon}_{\mathbf{y}}$ is the exact mean of the observational error distribution. The CRPS scores for this configuration have been added to Figure 12 as *ETKF-BC*. *ETKF-BC* outperforms the *ETKF* and, for very skewed observational errors, also *ETKF–VAE*$^{\text{single}}_{\text{clima}}$ when looking at the CRPS scores for the $x$-, $y$-coordinates and the angle. However, *ETKF–VAE*$^{\text{double}}_{clima}$ still renders lower CRPS scores than *ETKF-BC*. For the radius, all *ETKF–VAE* configurations produce significantly lower CRPS scores than both *ETKF* and *ETKF-BC*. Here the latter behaves more like the *ETKF* and yields scores that are many times worse than those produced by the *ETKF–VAE* configuration. Overall, the results in Figure 12 indicate two main aspects: (1) only a portion of the improvement brought by *ETKF–VAE*$^{\text{double}}$ is due to its ability to estimate and remove the observational error bias, and (2) the bias correction does not help the ETKF to confine the analysis to the circle manifold, a benefit unique to the ETKF–VAE.

It is worth noting that in these experiments the observational error distribution is known exactly. This knowledge is used in *ETKF-BC* to provide the mean used in the scheme and in *ETKF–VAE*$^{\text{double}}_{\text{clima}}$ to generate the synthetic innovations used to form $\mathbf{D}_K$. If the actual observation-error distribution used to train the second VAE differs significantly from the true distribution, the risk exists that the second encoder will send the ensemble of innovations to the "wrong" part of latent space. This may cause an incorrect rescaling of the error distribution and the benefits of using a second VAE may be small or, in the worst scenario, null. A similar issue can emerge with the bias-correction scheme if the specified mean observational error is erroneous.

In Figure 4b, it was shown that the VAE is not capable of transforming the distribution into a perfect Gaussian, although it gets a unimodal and almost symmetric one. To

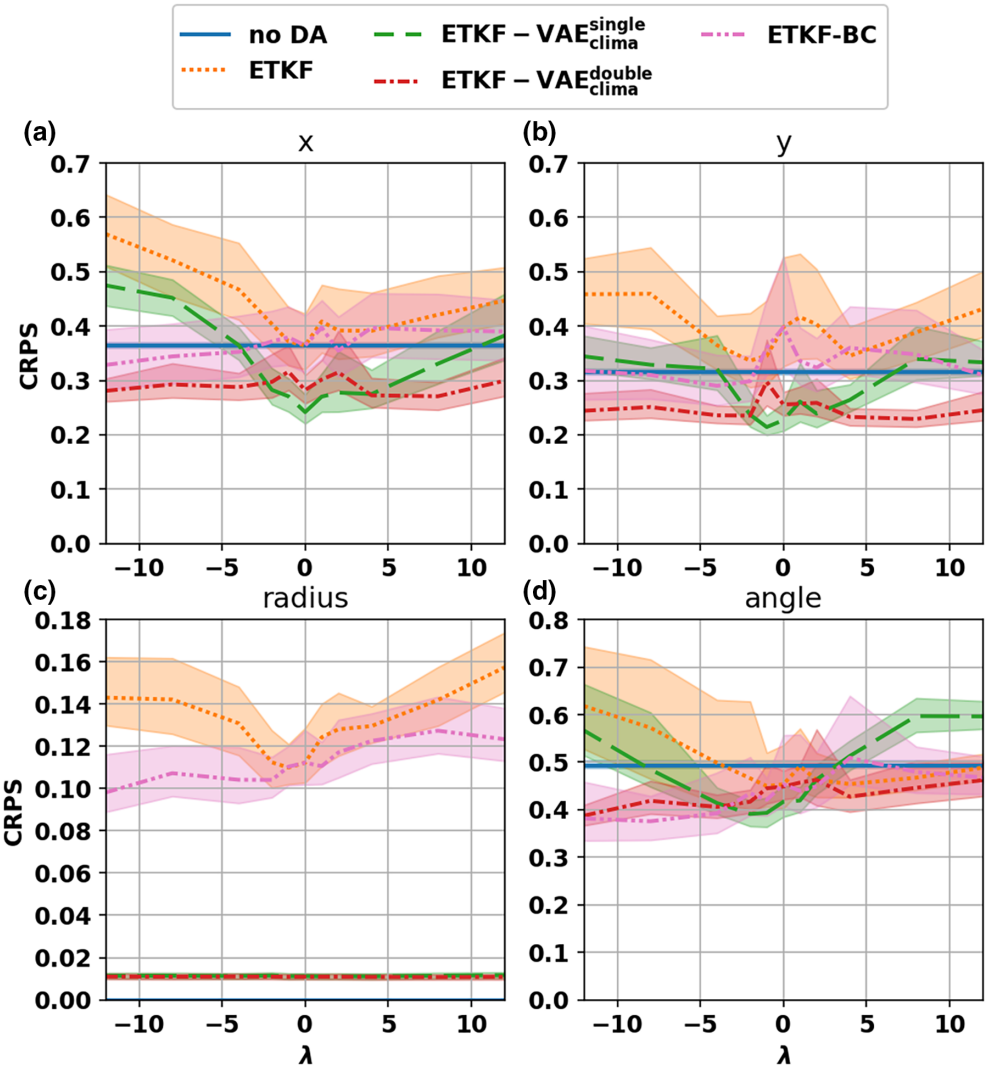


**FIGURE 12** CRPS for (a) the observed $x$-coordinate, (b) $y$-coordinate, (c) radius, and (d) polar angle as function of the skewness parameter $\lambda$ for the observational error distribution as appearing in Equation (12). Results are shown for the *no DA*, *ETKF*, *ETKF–VAE*$^{single}_{clima}$, and *ETKF–VAE*$^{double}_{clima}$ configurations. Bands indicate the 90% confidence interval of the CRPS. In the *No DA* configuration, CRPS is zero, as it is unaltered by the model when $A = 0$.

quantify the "degree of Gaussianity", we have calculated the Anderson–Darling statistic for a Gaussian distribution for the forecast ensemble at each assimilation time in *ETKF–VAE*$^{single}_{clima}$, *ETKF–VAE*$^{double}_{clima}$, and *ETKF*. The square of this metric is given by

$$A^2 = -M - \sum_{m=1}^{M} \frac{2m-1}{M} [\ln(F(z_m)) - \ln(1 - F(z_{M+1-m}))]. \tag{13}$$

In the former two experiments, the $z_m$ in Equation (13) are the latent values produced by the first VAE of the different ensemble members sorted in ascending order, while in the latter $z_m$ are the sorted values of either the $x$-coordinates or the $y$-coordinates and $F$ is the cumulative distribution of a Gaussian with mean and variance equal to that of the ensemble. The values for the Anderson–Darling statistic (Stephens, 1979) are binned and the relative occurrence of each value is shown in Figure 13a when the observational errors are unbiased and in Figure 13b when the errors are drawn from a skew normal distribution with $\lambda = -1.2$. Also shown in the figure, as the vertical dashed line, is the value for the statistic above, where the null hypothesis that the ensemble is drawn from a Gaussian distribution is rejected at the 95% confidence level. The figure shows that the latent ensembles are more likely to lie below this confidence level. In particular, for the case $\lambda = 0$, 852 ensembles have an Anderson–Darlington statistic below this level in *ETKF–VAE*$^{double}_{clima}$, which is more than the 594 ensembles in *ETKF–VAE*$^{single}_{clima}$ and the 425 for *ETKF*. This indicates that the latent ensemble, though not a perfect Gaussian distribution itself, satisfies the Gaussian assumption underlying the Kalman filter better than the ETKF. When the observation errors are not normally distributed but are taken from a skew normal distribution with skewness parameter $\lambda = -1.2$, somehow, counterintuitively, the introduction of non-Gaussian observation errors increases the number of ensembles below the 95% confidence threshold (except for *ETKF–VAE*$^{double}_{clima}$). Nevertheless, even in this case, the *ETKF–VAE* experiments produce ensembles that are closer to a Gaussian distribution.

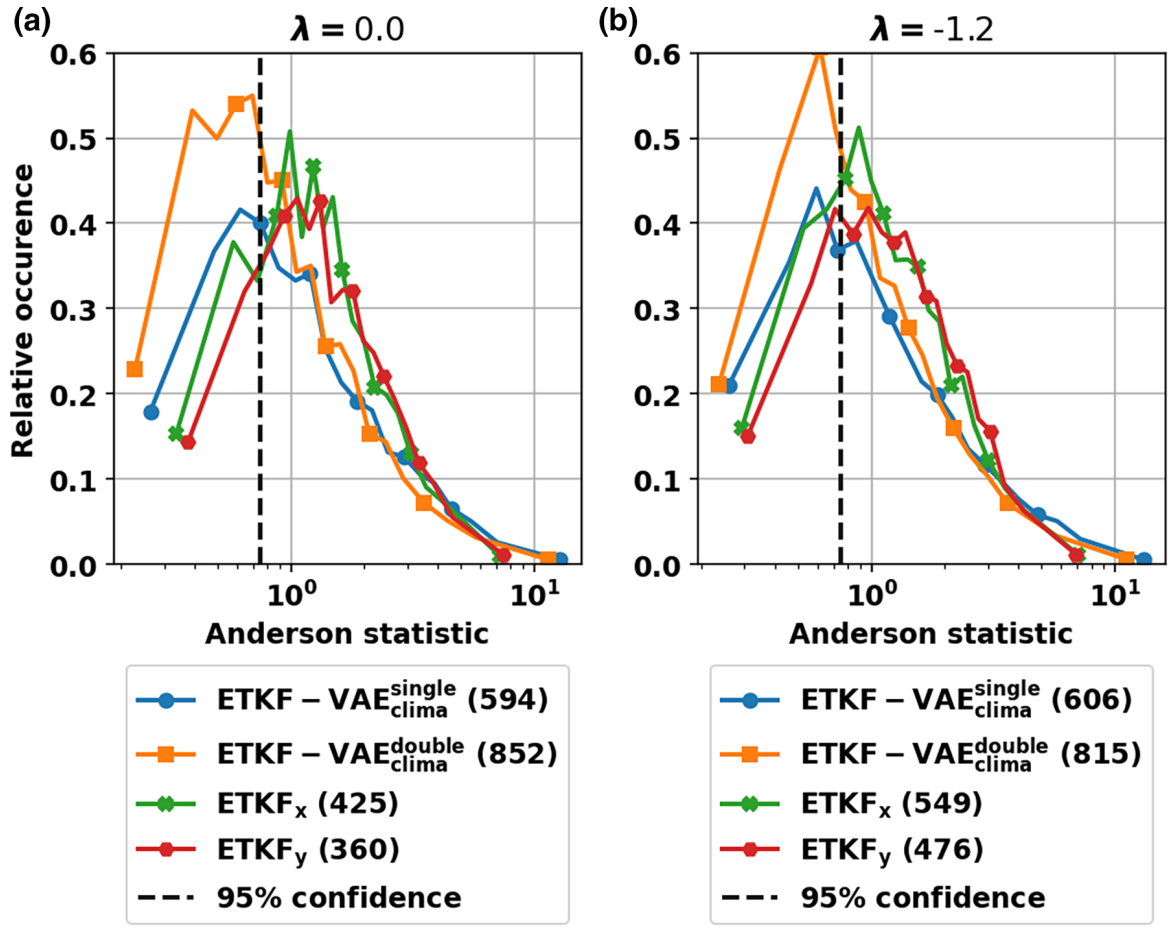


**FIGURE 13** Left: probability distribution of the Anderson statistic for the different forecast ensembles of the latent variable in experiment *ETKF–VAE*$^{single}_{clima}$ (blue), the latent variable in *ETKF–VAE*$^{double}_{clima}$ (yellow), the *x*-coordinate in *ETKF* (green), and the *y*-coordinate in *ETKF* (red) in the case in which the observational error is not skewed. Right: same but now with the observational errors drawn from a skewed Gaussian distribution with skewness parameter $\lambda = -1.2$. Also shown is the Anderson value above which the ensemble is different from Gaussian at 95% confidence and, between brackets, the number of times the Anderson metric of the ensemble is below this level.

## 5 | DISCUSSION AND CONCLUSIONS

Models in the geosciences, including the new neXtSIM$_{DG}$ sea-ice model, generally do not satisfy the conditions underlying the KF. Background and observational errors are non-Gaussian and model and observation operators are nonlinear. EnKFs can operate successfully when weak nonlinearity and/or small non-Gaussianity are present. In that case, however, filter updates will be suboptimal (Fowler, 2019; Lei *et al.*, 2010), that is, the a posteriori distribution of the ensemble members does not reflect $p(\mathbf{x}|\mathbf{y})$. In this work, we investigated the use of VAE to map ensemble members onto realisations sampled from Gaussian distributions. This work falls under the umbrella of studies aimed at using ML to mitigate or, in the best case, correct weaknesses of the DA procedure. We focus specifically on ensemble-based DA and, as a prototype of the latter, we choose to work with the ETKF, among the most celebrated and widely used ensemble-based approaches (Bishop *et al.*, 2001; Majumdar *et al.*, 2002; Wei *et al.*, 2006). We introduced two novel formulations of the ETKF, in which either one or two VAEs are merged into the ETKF's workflow. In the *single ETKF–VAE*, the goal is to tackle the non-Gaussianity in the physical model outputs and their physical balance. The latter is identified as a phase-space submanifold, to which all realisations of the model must be confined. In this work, this submanifold is the unit circle. With the *double ETKF-VAE*, we aimed also to address the non-Gaussianity in the observational errors.

The *single* and *double ETKF–VAE* configurations feature either an offline (climatological-like) or an online approach, in which weights are retrained using transfer learning. The latter allowed us to study the impact of, and the skill against, a time-varying physical submanifold for the model states. Hence, in this work

**TABLE 4** Overview of the different DA algorithms in this study, together with order-of-magnitude estimates of the number of floating-point operations necessary for each step. Operations are expressed as number of ensemble members $M$, dimension of the state space $N_{\mathcal{X}}$, observation-space dimension $N_{\mathcal{Y}}$, number of hidden layers in the encoder/decoder $N_L$, number of nodes per layer $N_N$, maximum number of epochs used during training $N_e$, and the dimension of both latent spaces $N_{\mathcal{Z}}$.

| ***ETKF*** | | $\boldsymbol{ETKF\text{–}VAE^{\text{single}}_{\text{clima}}}$ | | $\boldsymbol{ETKF\text{–}VAE^{\text{single}}_{\text{transfer}}}$ | |
|---|---|---|---|---|---|
| **Step** | **Order** | **Step** | **Order** | **Step** | **Order** |
| $\mathbf{X}^{\text{f}} \to \tilde{\mathbf{Y}}$ | $MN_{\mathcal{Y}}$ | $\mathbf{X}^{\text{f}} \to \tilde{\mathbf{D}}_M$ | $MN_{\mathcal{Y}}$ | $\mathbf{X}^{\text{f}} \to \tilde{\mathbf{D}}_M$ | $MN_{\mathcal{Y}}$ |
| $\mathbf{Y} \to \mathbf{R}^{1/2}\mathbf{Y}$ | $N_{\mathcal{Y}}M$ | $\tilde{\mathbf{D}}_M \to \tilde{\mathbf{D}}_K\tilde{\mathbf{D}}_K^{\text{T}}$ | $KN_{\mathcal{Y}}^2$ | $\tilde{\mathbf{D}}_M \to \tilde{\mathbf{D}}_K\tilde{\mathbf{D}}_K^{\text{T}}$ | $KN_{\mathcal{Y}}^2$ |
| SVD $\mathbf{R}^{-1/2}\mathbf{Y}$ | $MN_{\mathcal{Y}}\min(N_{\mathcal{Y}}, M)$ | $\mathbf{X}^{\text{f}} \to \mathbf{Z}^{\text{f}}$ | $MN_LN_N^2$ | Retrain first VAE | $N_LN_N^2MN_e$ |
| ETKF $\to \mathbf{X}^{\text{a}}$ | $\max(N_{\mathcal{X}}, M)M^2$ | SVD Equation (7b) | $N_{\mathcal{Y}}^3$ | $\mathbf{X}^{\text{f}} \to \mathbf{Z}^{\text{f}}$ | $MN_LN_N^2$ |
| | | ETKF Equation (7a) | $N_{\mathcal{Z}}N_{\mathcal{Y}}M$ | SVD Equation (7b) | $N_{\mathcal{Y}}^3$ |
| | | $\mathbf{Z}^{\text{a}} \to \mathbf{X}^{\text{a}}$ | $MN_LN_N^2$ | ETKF Equation (7a) | $N_{\mathcal{Z}}N_{\mathcal{Y}}M$ |
| | | | | $\mathbf{Z}^{\text{a}} \to \mathbf{X}^{\text{a}}$ | $MN_LN_N^2$ |
| $\boldsymbol{ETKF\text{–}VAE^{\text{double}}_{\text{clima}}}$ | | $\boldsymbol{ETKF\text{–}VAE^{\text{double}}_{\text{transfer}}}$ | | | |
| **Step** | **Order** | **Step** | **Order** | | |
| $\mathbf{X}^{\text{f}} \to \tilde{\mathbf{D}}_M$ | $MN_{\mathcal{Y}}$ | $\mathbf{X}^{\text{f}} \to \tilde{\mathbf{D}}_M$ | $MN_{\mathcal{Y}}$ | | |
| Retrain 2nd VAE | $MN_LN_N^2N_e$ | Retrain 2nd VAE | $MN_LN_N^2N_e$ | | |
| $\mathbf{D}_{N_{\mathcal{Y}}} \to \mathbf{F}_K$ | $KN_LN_N^2$ | $\mathbf{D}_{N_{\mathcal{Y}}} \to \mathbf{F}_K$ | $KN_LN_N^2$ | | |
| $\mathbf{F}_K \to \tilde{\mathbf{F}}_K\tilde{\mathbf{F}}_K^{\text{T}}$ | $KN_{\mathcal{Z}}^2$ | $\mathbf{F}_K \to \tilde{\mathbf{F}}_K\tilde{\mathbf{F}}_K^{\text{T}}$ | $KN_{\mathcal{Z}}^2$ | | |
| $\mathbf{X}^{\text{f}} \to \mathbf{Z}^{\text{f}}$ | $MN_LN_N^2$ | Retrain first VAE | $MN_LN_N^2N_e$ | | |
| SVD Equation (8b) | $N_{\mathcal{Z}}^3$ | $\mathbf{X}^{\text{f}} \to \mathbf{Z}^{\text{f}}$ | $MN_LN_N^2$ | | |
| ETKF Equation (8a) | $N_{\mathcal{Z}}^2M$ | SVD Equation (8b) | $N_{\mathcal{Z}}^3$ | | |
| $\mathbf{Z}^{\text{a}} \to \mathbf{X}^{\text{a}}$ | $MN_LN_N^2$ | ETKF Equation (8a) | $N_{\mathcal{Z}}^2M$ | | |
| | | $\mathbf{Z}^{\text{a}} \to \mathbf{X}^{\text{a}}$ | $MN_LN_N$ | | |

we have tested our approaches on non-autonomous dynamics with explicit dependence on time, such as one would encounter in a scenario with climate change.

We tested these setups in a conceptual model in which a point rotates around a circle, the *de facto* submanifold in these experiments. We find that the main advantage of application of the ETKF in the latent space of the VAE is that the posterior ensemble members stay close to the circle manifold, whereas the conventional ETKF places members at "unphysical" positions in and outside the circle. This means that the VAE is able to identify the existence of the submanifold containing the physically possible model states. The ensemble also provides a more accurate representation for the distribution of the truth. To quantify this, we have used the CRPS. All the ETKF–VAE configurations yield better CRPS values than the standard ETKF. Our work has also shown that updating weights for the first VAE using the ensemble members and transfer learning is essential if the manifold changes over time. Fine-tuning the VAE in this way is, in our view, a viable way of doing this. Its success in the idealistic setup employed in this study encourages its application and testing in more complex and higher-dimensional scenarios.

The benefits of using a second VAE for the innovations, aimed at coping with non-Gaussian observational errors, are mixed. When the observational error is Gaussian or only weakly non-Gaussian, the second VAE introduces additional variability in the innovations, weakening the correlations between innovations and ensemble members and thus reducing performance compared with the offline "clima" configurations. On the other hand, when the non-Gaussian character of the observational error and its bias are predominant, our findings highlight the benefit of using a second VAE for the innovations.

Motivated by the large and continuously growing amount of available satellite data, compression methods that make the number of assimilated observations computationally affordable (while simultaneously maintaining high informational content) have received increased interest in recent years (Cheng *et al.*, 2021b; Pasmans

*et al.*, 2024; Xu, 2011). In these studies, special attention was paid to how the presence of correlations in observational errors (Cheng *et al.*, 2021a; Fowler, 2019) impacts the compression that can be achieved, that is, what percentage of observations can be removed without deteriorating the DA performance significantly. Although not pursued among the specific goals here (and made impossible by the low dimensionality of our model), we believe that in higher dimensions the second VAE will achieve effective data compression by using a latent space of smaller dimension than the number of observations. The key advantage of our approach, in which innovations, instead of observations (see Cheng *et al.*, 2024), are mapped to the latent space, is that the compression will depend on both the forecast and the observational error statistics. This is important, because, according to the findings of Fowler (2019), the covariance reduction by the DA as a function of the number of observations depends on both the correlation length-scales in the observational errors and the scales in the forecast errors. Extending this study to the presence of correlated observations is one of the venues left for future work.

The ensemble-size-versus-dimensionality ratio in this work is 32, much higher than the $\ll 1$ ratios typical of operational forecasting systems. Therefore, future work should also investigate how the proposed approach scales to higher-dimensional models. In particular, the ability to train the VAE weights using the ensemble members and transfer learning might falter for more realistic model setups. One possible way to mitigate this would be the inclusion of time-lagged or time-shifted ensemble members in the training dataset (Lorenc, 2017), or somewhat shifting the physical fields in space: an approach similar to the application of covariance localisation using convolution (Berre & Desroziers, 2010; Courtier *et al.*, 1998; Gaspari & Cohn, 1999). This, in combination with a switch to a convolutional variational auto-encoder in which the forward neural networks are replaced with convolutional networks (Sohn *et al.*, 2015), could make it feasible to retrain the VAE at each analysis time using an ensemble that is relatively small compared with the model's dimension. This is because in a convolutional network the number of neural-network nodes does not scale with the size of the model state, but with the size of the convolution kernel, usually $3 \times 3$, and the number of convolution kernels used. Alternatively, one could use model emulators, for example, like the partial neXtSIM simulator convolutional emulator developed by Durand *et al.* (2024), to generate larger background ensembles in a computationally effective way (Chattopadhyay *et al.*, 2022), or draw the ensemble members using diffusion models explicitly trained to represent the background probability distribution correctly (Finn *et al.*, 2024a; Li *et al.*, 2024; Price *et al.*, 2025). This would allow one to generate an arbitrary amount of training data. However, this would not reduce the size of the neural networks used in the VAE, so this method might be more memory-intensive, which would increase training times. The latter would be an issue for $ETKF\text{–}VAE_{\text{transfer}}$ especially.

Another point that could not be addressed in our low-dimensional setup is the tendency of VAEs to over-smooth the small scales (see, e.g., fig. 2 of Finn *et al.*, 2024b). This could lead to an underestimation of the smaller scales in the analysis ensemble. For example, in a sea-ice model, deformation, a process rich in small scales in areas in which the ice is damaged, might be underestimated. If a plastic rheology is used, the stress depends on the deformation rate of the sea ice. This deformation rate is in turn a function of the spatial derivatives of the sea-ice velocity field. Therefore, underestimation of the small scales in the velocity will result in an overly smooth field in which the magnitudes of the derivatives, and consequently the deformation rate and stress, will be too small. In elastic rheologies, time changes in sea-ice stress, instead of the stress itself, depend on the deformation rate. Consequently, a consistent underestimate of the magnitude of the deformation will gradually build up into an underestimate of the magnitude of the stress. The straightforward mitigation fix for this effect is to increase the latent space's dimension (Finn *et al.*, 2024b).

The training of VAEs is notably highly computationally demanding. Here, we found that training in the transfer configurations can take up a considerable amount of time ($\approx 0.3$ hour for a single realisation on a laptop with a NVIDIA RTX A2000) compared with running the ETKF ($\approx 1$ minute per realisation). In order to understand how our approach might scale to higher-dimensional systems, estimates of the computational cost, expressed as the number of floating-point operations using big-O or Bachmann–Landau notation, have been included as Table 4. The cost for running the model has not been included in these estimates, as this cost is the same for all configurations. Neither has the cost for training the offline training on the climatology run been included, as this is a one-time cost, which becomes irrelevant in the limit in which time goes to infinity. When the dimensions of observational and state spaces are small compared with the number of nodes in the VAE encoder/decoder, as is the case in this study, the *ETKF* is considerably cheaper than the *ETKF–VAE*. As training takes place twice per DA step in the $ETKF\text{–}VAE^{\text{double}}_{\text{transfer}}$ configuration, this is by far the most costly. For large numbers of observations,

the cost of the SVD in Equation (7b) could overtake that of the VAE. In this situation, the *ETKF–VAE*$^{\text{double}}$ might actually become the better choice, even performance-wise, as the SVD in Equation (8b) scales with the dimension of the latent space, which might be set to be smaller than the number of observations. Therefore, for high-dimensional systems the ETKF–VAE might actually scale very well, although ultimately this has to be tested in future work.

## ACKNOWLEDGEMENTS

This work is part of the Scale-Aware Sea Ice Project (SASIP) and is supported by grant G-24-66154 of Schmidt Sciences, LLC—a philanthropy that propels scientific knowledge and breakthroughs towards a thriving world. CEREA is a member of Institut Pierre-Simon Laplace. We also thank the three anonymous reviewers for their constructive and insightful comments.

## DATA AVAILABILITY STATEMENT

Code for the VAEs, DA, and figures has been embedded in the DAPPER DA framework (Raanes *et al.*, 2023) and is available as the *ETKF_VAE* branch of doi.org/10.5281/zenodo.7339457.

## ENDNOTE

[1]https://sasip-climate.github.io/.

## ORCID

*Ivo Pasmans* https://orcid.org/0000-0001-5076-5421
*Yumeng Chen* https://orcid.org/0000-0002-2319-6937
*Tobias Sebastian Finn* https://orcid.org/0000-0001-9585-8349
*Marc Bocquet* https://orcid.org/0000-0003-2675-0347

## REFERENCES

Akbari, S., Dabaghian, P.H. & San, O. (2023) Blending machine learning and sequential data assimilation over latent spaces for surrogate modeling of Boussinesq systems. *Physica D: Nonlinear Phenomena*, 448, 133711. Available from: https://doi.org/10.1016/j.physd.2023.133711

Amendola, M., Arcucci, R., Mottet, L., Casas, C.Q., Fan, S., Pain, C. et al. (2021) Data assimilation in the latent space of a convolutional autoencoder. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J. & Sloot, P.M. (Eds.) *Computational science – ICCS 2021*. Cham: Springer International Publishing, pp. 373–386. Available from: https://doi.org/10.1007/978-3-030-77977-1.30

Anderson, J.L. (2010) A non-Gaussian ensemble filter update for data assimilation. *Monthly Weather Review*, 138(11), 4186–4198. Available from: https://doi.org/10.1175/2010MWR3253.1

Anderson, J.L. (2019) A nonlinear rank regression method for ensemble Kalman filter data assimilation. *Monthly Weather Review*, 147(8), 2847–2860. Available from: https://doi.org/10.1175/MWR-D-18-0448.1

Arcucci, R., Zhu, J., Hu, S. & Guo, Y.-K. (2021) Deep data assimilation: integrating deep learning with data assimilation. *Applied Sciences*, 11(3), 1114. Available from: https://doi.org/10.3390/app11031114

Bach, E., Baptista, R., Sanz-Alonso, D. & Stuart, A. (2024) Inverse problems and data assimilation: a machine learning approach. 10.48550/arXiv.2410.10523

Bao, J., Li, L. & Davis, A. (2022) Variational autoencoder or generative adversarial networks? A comparison of two deep learning methods for flow and transport data assimilation. *Mathematical Geosciences*, 54(6), 1017–1042. Available from: https://doi.org/10.1007/s11004-022-10003-3

Béal, D., Brasseur, P., Brankart, J.-M., Ourmières, Y. & Verron, J. (2010) Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis. *Ocean Science*, 6(1), 247–262. Available from: https://doi.org/10.5194/os-6-247-2010

Berre, L. & Desroziers, G. (2010) Filtering of background error variances and correlations by local spatial averaging: a review. *Monthly Weather Review*, 138(10), 3693–3720. Available from: https://doi.org/10.1175/2010MWR3111.1

Bertino, L., Evensen, G. & Wackernagel, H. (2003) Sequential data assimilation techniques in oceanography. *International Statistical Review*, 71(2), 223–241. Available from: https://doi.org/10.1111/j.1751-5823.2003.tb00194.x

Bishop, C.H., Etherton, B.J. & Majumdar, S.J. (2001) Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Monthly Weather Review*, 129(3), 420–436. Available from: https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2

Bocquet, M. (2011) Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, 18(5), 735–750. Available from: https://doi.org/10.5194/npg-18-735-2011

Bocquet, M., Pires, C.A. & Wu, L. (2010) Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, 138(8), 2997–3023. Available from: https://doi.org/10.1175/2010MWR3164.1

Bocquet, M., Brajard, J., Carrassi, A. & Bertino, L. (2020a) Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *Foundations of Data Science*, 2(1), 55–80. Available from: https://doi.org/10.3934/fods.2020004

Bocquet, M., Farchi, A. & Malartic, Q. (2020b) Online learning of both state and dynamics using ensemble Kalman filters. *Foundations of Data Science*, 3(3), 305–330. Available from: https://doi.org/10.3934/fods.2020015

Bocquet, M., Farchi, A., Finn, T.S., Durand, C., Cheng, S., Chen, Y. et al. (2024) Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble. *Chaos*, 34(9), 091104. Available from: https://doi.org/10.1063/5.0230837

Boudier, P., Fillion, A., Gratton, S., Gürol, S. & Zhang, S. (2023) Data assimilation networks. *Journal of Advances in Modeling Earth Systems*, 15(4), e2022MS003. Available from: https://doi.org/10.1029/2022MS003353

Bowler, N.E. (2006) Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus A*, 58(5), 538–548. Available from: https://doi.org/10.1111/j.1600-0870.2006.00197.x

Brajard, J., Carrassi, A., Bocquet, M. & Bertino, L. (2020) Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model. *Journal of Computational Science*, 44, 101171. Available from: https://doi.org/10.1016/j.jocs.2020.101171

Brajard, J., Carrassi, A., Bocquet, M. & Bertino, L. (2021) Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 20200086. Available from: https://doi.org/10.1098/rsta.2020.0086

Buehner, M., Caron, J.-F., Lapalme, E., Caya, A., Du, P., Rochon, Y. et al. (2025) The modular and integrated data assimilation system at environment and climate change Canada (MIDAS v3.9.1). *Geoscientific Model Development*, 18, 1–18. Available from: https://doi.org/10.5194/gmd-18-1-2025

Buizza, C., Quilodrán Casas, C., Nadler, P., Mack, J., Marrone, S., Titus, Z. et al. (2022) Data learning: integrating data assimilation and machine learning. *Journal of Computational Science*, 58, 101525. Available from: https://doi.org/10.1016/j.jocs.2021.101525

Canchumuni, S.W.A., Emerick, A.A. & Pacheco, M.A.C. (2019) Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother. *Computers & Geosciences*, 128, 87–102. Available from: https://doi.org/10.1016/j.cageo.2019.04.006

Carrassi, A., Bocquet, M., Bertino, L. & Evensen, G. (2018) Data assimilation in the geosciences: an overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5), e535. Available from: https://doi.org/10.1002/wcc.535

Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E. & Kashinath, K. (2022) Towards physics-inspired data-driven weather forecasting: integrating data assimilation with a deep spatial-transformer-based U-NET in a case study with ERA5. *Geoscientific Model Development*, 15(5), 2221–2237. Available from: https://doi.org/10.5194/gmd-15-2221-2022

Chen, Y., Smith, P., Carrassi, A., Pasmans, I., Bertino, L., Bocquet, M. et al. (2024) Multivariate state and parameter estimation with data assimilation applied to sea-ice models using a Maxwell elasto-brittle rheology. *The Cryosphere*, 18(5), 2381–2406. Available from: https://doi.org/10.5194/tc-18-2381-2024

Cheng, S., Argaud, J.-P., Iooss, B., Lucor, D. & Ponçot, A. (2021a) Error covariance tuning in variational data assimilation: application to an operating hydrological model. *Stochastic Environmental Research and Risk Assessment*, 35(5), 1019–1038. Available from: https://doi.org/10.1007/s00477-020-01933-7

Cheng, S., Lucor, D. & Argaud, J.-P. (2021b) Observation data compression for variational assimilation of dynamical systems. *Journal of Computational Science*, 53, 101405. Available from: https://doi.org/10.1016/j.jocs.2021.101405

Cheng, S., Prentice, I.C., Huang, Y., Jin, Y., Guo, Y.-K. & Arcucci, R. (2022) Data-driven surrogate model with latent data assimilation: application to wildfire forecasting. *Journal of Computational Physics*, 464, 111302. Available from: https://doi.org/10.1016/j.jcp.2022.111302

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P. et al. (2023) Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review. *IEEE/CAA Journal of Automatica Sinica*, 10(6), 1361–1387. Available from: https://doi.org/10.1109/JAS.2023.123537

Cheng, S., Zhuang, Y., Kahouadji, L., Liu, C., Chen, J., Matar, O.K. et al. (2024) Multi-domain encoder–decoder neural networks for latent data assimilation in dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 430, 117201. Available from: https://doi.org/10.1016/j.cma.2024.117201

Chinellato, E. & Marcuzzi, F. (2024) State estimation of partially unknown dynamical systems with a deep Kalman filter. In: Franco, L., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J. & Sloot, P.M.A. (Eds.) *Computational science – ICCS 2024*. Cham: Springer Nature Switzerland, pp. 307–321. Available from: https://doi.org/10.1007/978-3-031-63775-9.22

Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A. et al. (1998) The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550), 1783–1807. Available from: https://doi.org/10.1002/qj.49712455002

Dai, B. & Wipf, D. (2019) Diagnosing and enhancing VAE models. 10.48550/arXiv.1903.05789

Dansereau, V., Weiss, J., Saramito, P. & Lattes, P. (2016) A Maxwell elasto-brittle rheology for sea ice modelling. *Cryosphere*, 10(3), 1339–1359. Available from: https://doi.org/10.5194/tc-10-1339-2016

De Rosnay, P. et al. (2022) Coupled data assimilation at ECMWF: current status, challenges and future developments. *Quarterly Journal of the Royal Meteorological Society*, 148(747), 2672–2702. Available from: https://doi.org/10.1002/qj.4330

Dee, D.P. (2005) Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613), 3323–3343. Available from: https://doi.org/10.1256/qj.05.137

Desroziers, G. & Ivanov, S. (2001) Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 127(574), 1433–1452. Available from: https://doi.org/10.1002/qj.49712757417

Durand, C., Finn, T.S., Farchi, A., Bocquet, M., Boutin, G. & Ólason, E. (2024) Data-driven surrogate modeling of high-resolution sea-ice thickness in the Arctic. *Cryosphere*, 18(4), 1791–1815. Available from: https://doi.org/10.5194/tc-18-1791-2024

Efron, B. (1987) Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185. Available from: https://doi.org/10.1080/01621459.1987.10478410

Ehrendorfer, M. (2007) A review of issues in ensemble-based Kalman filtering. *Meteorologische Zeitschrift*, 16, 795–818. Available from: https://doi.org/10.1127/0941-2948/2007/0256

Evensen, G. (1994) 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. Available from: https://doi.org/10.1029/94JC00572

Evensen, G. (2004) Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 54(6), 539–560. Available from: https://doi.org/10.1007/s10236-004-0099-2

Evensen, G., Vossepoel, F.C. & Van Leeuwen, P.J. (2022) *Data assimilation fundamentals: a unified formulation of the state and parameter estimation problem*. Cham, Switzerland: Springer Nature.

Farchi, A., Laloyaux, P., Bonavita, M. & Bocquet, M. (2021) Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, 147(739), 3067–3084. Available from: https://doi.org/10.1002/qj.4116

Farchi, A., Chrust, M., Bocquet, M., Laloyaux, P. & Bonavita, M. (2023) Online model error correction with neural networks in the incremental 4D-Var framework. *Journal of Advances in Modeling Earth Systems*, 15(9), e2022MS003474. Available from: https://doi.org/10.1029/2022MS003474

Finn, T.S., Disson, L., Farchi, A., Bocquet, M. & Durand, C. (2024a) Representation learning with unconditional denoising diffusion models for dynamical systems. *Nonlinear Processes in Geophysics*, 31(3), 409–431. Available from: https://doi.org/10.5194/npg-31-409-2024

Finn, T.S., Durand, C., Farchi, A., Bocquet, M. & Brajard, J. (2024b) Towards diffusion models for large-scale sea-ice modelling. https://doi.org/10.48550/arXiv.2406.18417

Fletcher, S.J. & Zupanski, M. (2006) A hybrid multivariate normal and lognormal distribution for data assimilation. *Atmospheric Science Letters*, 7(2), 43–46. Available from: https://doi.org/10.1002/asl.128

Fowler, A. (2019) Data compression in the presence of observational error correlations. *Tellus A: Dynamic Meteorology and Oceanography*, 71(1), 1634. Available from: https://doi.org/10.1080/16000870.2019.1634937

Gaspari, G. & Cohn, S.E. (1999) Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554), 723–757. Available from: https://doi.org/10.1002/qj.49712555417

Grooms, I. (2021) Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders. *Quarterly Journal of the Royal Meteorological Society*, 147(734), 139–149. Available from: https://doi.org/10.1002/qj.3910

Grooms, I. (2022) A comparison of nonlinear extensions to the ensemble Kalman filter. *Computational Geosciences*, 26(3), 633–650. Available from: https://doi.org/10.1007/s10596-022-10141-x

He, K., Zhang, X., Ren, S. & Sun, J. (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. New York, NY, USA: IEEE, pp. 1026–1034. Available from: https://doi.org/10.1109/ICCV.2015.123

Henze, N. (1986) A probabilistic representation of the skew-Normal distribution. *Scandinavian Journal of Statistics*, 13(4), 271–275.

Huang, Y., Jia, G., Chen, B. & Zhang, Y. (2020) A new robust Kalman filter with adaptive estimate of time-varying measurement bias. *IEEE Signal Processing Letters*, 27, 700–704. Available from: https://doi.org/10.1109/LSP.2020.2983552

Huang, L., Gianinazzi, L., Yu, Y., Dueben, P.D. & Hoefler, T. (2024) DiffDA: a diffusion model for weather-scale data assimilation. 10.48550/arXiv.2401.05932

Hunt, B.R., Kostelich, E.J. & Szunyogh, I. (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230(1–2), 112–126. Available from: https://doi.org/10.1016/j.physd.2006.11.008

Inverarity, G.W., Tennant, W.J., Anton, L., Bowler, N.E., Clayton, A.M., Jardak, M. et al. (2023) Met Office MOGREPS-G initialisation using an ensemble of hybrid four-dimensional ensemble variational (En-4DEnVar) data assimilations. *Quarterly Journal of the Royal Meteorological Society*, 149(753), 1138–1164. Available from: https://doi.org/10.1002/qj.4431

Janjić, T., Bormann, N., Bocquet, M., Carton, J.A., Cohn, S.E., Dance, S.L. et al. (2018) On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144(713), 1257–1278. Available from: https://doi.org/10.1002/qj.3130

Jendersie, R., Lessig, C. & Richter, T. (2025) A GPU parallelization of the neXtSIM-DG dynamical core (v0.3.1). *Geoscientific Model Development*, 18, 3017–3040. Available from: https://doi.org/10.5194/gmd-18-3017-2025

Jun, S., Seol, K., Kwon, Y., Kwon, I.-H. & Koo, M.-S. (2024a) Development of soil moisture assimilation based on KIM-LETKF System. In: *IGARSS 2024 - 2024 IEEE International geoscience and remote sensing symposium*. New York, NY, USA: IEEE, pp. 5581–5585. Available from: https://doi.org/10.1109/IGARSS53475.2024.10642475

Kingma, D.P. & Ba, J. (2017) Adam: a method for stochastic optimization. 10.48550/arXiv.1412.6980

Kingma, D.P. & Welling, M. (2019) An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4), 307–392. Available from: https://doi.org/10.1561/2200000056

Kotsuki, S., Miyoshi, T., Terasaki, K., Lien, G.-Y. & Kalnay, E. (2017) Assimilating the global satellite mapping of precipitation data with the nonhydrostatic icosahedral atmospheric model (NICAM). *Journal of Geophysical Research: Atmospheres*, 122(2), 631–650. Available from: https://doi.org/10.1002/2016JD025355

Lea, D.J., Drecourt, J.-P., Haines, K. & Martin, M.J. (2008) Ocean altimeter assimilation with observational- and model-bias correction. *Quarterly Journal of the Royal Meteorological Society*, 134(636), 1761–1774. Available from: https://doi.org/10.1002/qj.320

Lei, J., Bickel, P. & Snyder, C. (2010) Comparison of ensemble Kalman filters under non-Gaussianity. *Monthly Weather Review*, 138(4), 1293–1306. Available from: https://doi.org/10.1175/2009MWR3133.1

Li, L., Carver, R., Lopez-Gomez, I., Sha, F. & Anderson, J. (2024) Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13), eadk4489. Available from: https://doi.org/10.1126/sciadv.adk4489

Lorenc, A.C. (2017) Improving ensemble covariances in hybrid variational data assimilation without increasing ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 143(703), 1062–1072. Available from: https://doi.org/10.1002/qj.2990

Luk, E., Bach, E., Baptista, R. & Stuart, A. (2024) Learning optimal filters using variational inference. 10.48550/arXiv.2406.18066

Mack, J., Arcucci, R., Molina-Solana, M. & Guo, Y.-K. (2020) Attention-based convolutional autoencoders for 3D-variational data assimilation. *Computer Methods in Applied Mechanics and Engineering*, 372, 113,291. Available from: https://doi.org/10.1016/j.cma.2020.113291

Majumdar, S.J., Bishop, C.H., Etherton, B.J. & Toth, Z. (2002) Adaptive sampling with the ensemble transform Kalman filter. Part II: field program implementation. *Monthly Weather Review*, 130(5), 1356–1369. Available from: https://doi.org/10.1175/1520-0493(2002)130<1356:ASWTET>2.0.CO;2

Matsugishi, S., Chen, Y.-.W., Terasaki, K., Kanemaru, K., Kotsuki, S., Yashiro, H. et al. (2025) NICAM–LETKF JAXA research analysis (NEXRA) version 2.0. *Geoscience Data Journal*, 12(3), e70011. Available from: https://doi.org/10.1002/gdj3.70011

Maulik, R., Rao, V., Wang, J., Mengaldo, G., Constantinescu, E., Lusch, B. et al. (2022) Efficient high-dimensional variational data assimilation with machine-learned reduced-order models. *Geoscientific Model Development*, 15(8), 3433–3445. Available from: https://doi.org/10.5194/gmd-15-3433-2022

McCabe, M. & Brown, J. (2021) Learning to assimilate in chaotic dynamical systems. In: M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang and J. W. Vaughan (Eds.) *Advances in Neural Information Processing Systems*, 34, pp. 12237–12250, Red Hook, NY: Curran Associates Inc.

Melinc, B. & Zaplotnik, Ž. (2024) 3D-Var data assimilation using a variational autoencoder. *Quarterly Journal of the Royal Meteorological Society*, 150(761), 2273–2295. Available from: https://doi.org/10.1002/qj.4708

Metref, S., Cosme, E., Snyder, C. & Brasseur, P. (2014) A non-Gaussian analysis scheme using rank histograms for ensemble data assimilation. *Nonlinear Processes in Geophysics*, 21(4), 869–885. Available from: https://doi.org/10.5194/npg-21-869-2014

Morzfeld, M. & Hodyss, D. (2023) A theory for why even simple covariance localization is so useful in ensemble data assimilation. *Monthly Weather Review*, 151(3), 717–736. Available from: https://doi.org/10.1175/MWR-D-22-0255.1

O'Hagan, A. & Leonard, T. (1976) Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1), 201–203. Available from: https://doi.org/10.1093/biomet/63.1.201

Ólason, E., Boutin, G., Korosov, A., Rampal, P., Williams, T., Kimmritz, M. et al. (2022) A new brittle rheology and numerical framework for large-scale sea-ice models. *Journal of Advances in Modeling Earth Systems*, 14(8), e2021MS002685. Available from: https://doi.org/10.1029/2021MS002685

Pasmans, I., Chen, Y., Carrassi, A. & Jones, C.K.R.T. (2024) Tailoring data assimilation to discontinuous Galerkin models. *Quarterly Journal of the Royal Meteorological Society*, 150(762), 2820–2847. Available from: https://doi.org/10.1002/qj.4737

Pawar, S. & San, O. (2022) Equation-free surrogate modeling of geophysical flows at the intersection of machine learning and data assimilation. *Journal of Advances in Modeling Earth Systems*, 14(11), e2022MS003170. Available from: https://doi.org/10.1029/2022MS003170

Peyron, M., Fillion, A., Gürol, S., Marchais, V., Gratton, S., Boudier, P. et al. (2021) Latent space data assimilation by using deep learning. *Quarterly Journal of the Royal Meteorological Society*, 147(740), 3759–3777. Available from: https://doi.org/10.1002/qj.4153

Polavarapu, S., Ren, S., Rochon, Y., Sankey, D., Ek, N., Koshyk, J. et al. (2005) Data assimilation with the Canadian middle atmosphere model. *Atmosphere-Ocean*, 43(1), 77–100. Available from: https://doi.org/10.3137/ao.430105

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., el-Kadi, A., Masters, D. et al. (2025) Probabilistic weather forecasting with machine learning. *Nature*, 637(8044), 84–90. Available from: https://doi.org/10.1038/s41586-024-08252-9

Qin, Y., Yu, Q., Wan, L., Liu, Y., Mo, H., Wang, Y. et al. (2023) A global-ocean-data assimilation for operational oceanography. *Journal of Marine Science and Engineering*, 11(12), 2255. Available from: https://doi.org/10.3390/jmse11122255

Qu, Y., Nathaniel, J., Li, S. & Gentine, P. (2024) Deep generative data assimilation in multimodal setting. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New York, NY, USA: IEEE, pp. 449–459. Available from: https://doi.org/10.1109/CVPRW63382.2024.00050

Raanes, P.N., Chen, Y., Grudzien, C., Tondeur, M. & Dubois, R. (2023) DAPPER, Nansen Environmental and Remote Sensing Center.

Rezende, D.J. & Viola, F. (2018) Taming VAEs. https://doi.org/10.48550/arXiv.1810.00597

Rezende D.J., S. Mohamed, D. Wierstra (2014) Stochastic backpropagation and approximate inference in deep generative models. In: E.P. Xing and T. Jebara (Eds.) *Proceedings of the 31st international conference on machine learning*, 32(2), pp. 1278–1286, Cambridge, MA: JMLR.

Richter, T., Dansereau, V., Lessig, C. & Minakowski, P. (2023) A dynamical core based on a discontinuous Galerkin method for higher-order finite-element sea ice modeling. *Geoscientific Model Development*, 16(13), 3907–3926. Available from: https://doi.org/10.5194/gmd-16-3907-2023

Rozet, F. & Louppe, G. (2023) Score-based data assimilation, Advances in Neural Information Processing Systems, 36, 40,521–40,541.

Sakov, P., Oliver, D.S. & Bertino, L. (2012) An iterative EnKF for strongly nonlinear systems. *Monthly Weather Review*, 140(6), 1988–2004. Available from: https://doi.org/10.1175/MWR-D-11-00176.1

Saunders, R.W., Blackmore, T.A., Candy, B., Francis, P.N. & Hewison, T.J. (2013) Monitoring satellite radiance biases using NWP models. *IEEE Transactions on Geoscience and Remote Sensing*, 51(3), 1124–1138. Available from: https://doi.org/10.1109/TGRS.2012.2229283

Shlezinger, N. et al. (2024) AI-aided Kalman filters. https://doi.org/10.48550/arXiv.2410.12289

Si, P. & Chen, P. (2024) Latent-EnSF: a latent ensemble score filter for high-dimensional data assimilation with sparse observation data. https://doi.org/10.48550/arXiv.2409.00127

Simon, E. & Bertino, L. (2012) Gaussian anamorphosis extension of the DEnKF for combined state parameter estimation: application to a 1D ocean ecosystem model. *Journal of Marine Systems*, 89(1), 1–18. Available from: https://doi.org/10.1016/j.jmarsys.2011.07.007

Sohn, K., Lee, H. & Yan, X. (2015) Learning structured output representation using deep conditional generative models. In: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.) *Advances in neural information processing systems*, Vol. 28, p. 9. Red Hook, NY: Curran Associates Inc.

Song, H., Edwards, C.A., Moore, A.M. & Fiechter, J. (2012) Incremental four-dimensional variational data assimilation of positive-definite oceanic variables using a logarithm transformation. *Ocean Modelling*, 54–55, 1–17. Available from: https://doi.org/10.1016/j.ocemod.2012.06.001

Stephens, M.A. (1979) *The anderson-darling statistic, Technical report TR-39*. Stanford, CA: Department of Statistics, Stanford University, p. 23.

Tabak, E.G. & Turner, C.V. (2013) A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2), 145–164. Available from: https://doi.org/10.1002/cpa.21423

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M. et al. (2020) A review of innovation-based methods to jointly

estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*, 148(10), 3973–3994. Available from: https://doi.org/10.1175/MWR-D-19-0240.1

Tippett, M.K., Anderson, J.L., Bishop, C.H., Hamill, T.M. & Whitaker, J.S. (2003) Ensemble Square root filters. *Monthly Weather Review*, 131(7), 1485–1490. Available from: https://doi.org/10.1175/1520-0493(2003)131<1485:ESRF>2.0.CO;2

Tödter, J. & Ahrens, B. (2012) Generalization of the ignorance score: continuous ranked version and its decomposition. *Monthly Weather Review*, 140(6), 2005–2017. Available from: https://doi.org/10.1175/MWR-D-11-00266.1

Vobig, K., Stephan, K., Blahak, U., Khosravian, K. & Potthast, R. (2021) Targeted covariance inflation for 3D-volume radar reflectivity assimilation with the LETKF. *Quarterly Journal of the Royal Meteorological Society*, 147(740), 3789–3805. Available from: https://doi.org/10.1002/qj.4157

Waters, J., Lea, D.J., Martin, M.J., Mirouze, I., Weaver, A. & While, J. (2015) Implementing a variational data assimilation system in an operational 1/4 degree global ocean model. *Quarterly Journal of the Royal Meteorological Society*, 141(687), 333–349. Available from: https://doi.org/10.1002/qj.2388

Wei, M., Toth, Z., Wobus, R., Zhu, Y., Bishop, C.H. & Wang, X. (2006) Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus A: Dynamic Meteorology and Oceanography*, 58(1), 28–44. Available from: https://doi.org/10.1111/j.1600-0870.2006.00159.x

Whitaker, J.S. & Hamill, T.M. (2012) Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, 140(9), 3078–3089. Available from: https://doi.org/10.1175/MWR-D-11-00276.1

Xu, Q. (2011) Measuring information content from observations for data assimilation: spectral formulations and their implications to observational data compression. *Tellus Series A: Dynamic Meteorology and Oceanography*, 63(4), 793–804. Available from: https://doi.org/10.1111/j.1600-0870.2011.00524.x

Zhang, Y., Pan, J., Li, L.K., Liu, W., Chen, Z., Liu, X. et al. (2023) In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. & Levine, S. (Eds.) *On the Properties of Kullback-Leibler divergence between multivariate Gaussian distributions in advances in neural information processing systems*, Vol. 36. Red Hook, NY, USA: Curran Associates Inc., pp. 58152–58165.

Zhu, Y., Derber, J., Collard, A., Dee, D., Treadon, R., Gayno, G. et al. (2014) Enhanced radiance bias correction in the National Centers for environmental Prediction's gridpoint statistical interpolation data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 140(682), 1479–1492. Available from: https://doi.org/10.1002/qj.2233

**How to cite this article:** Pasmans, I., Chen, Y., Sebastian Finn, T., Bocquet, M. & Carrassi, A. (2025) Ensemble Kalman filter in latent space using a variational autoencoder pair. *Quarterly Journal of the Royal Meteorological Society*, e70070. Available from: https://doi.org/10.1002/qj.70070

## APPENDIX A. ETKF–VAE ALGORITHMS

This Appendix contains pseudocode running the total DA system using either *single ETKF–VAE* or *double ETKF–VAE* configurations. The $\mathcal{U}(a,b)$ appearing in the algorithms stands for a uniform probability distribution on the integers between $a$ up to and including $b$ (Algorithm 1).

**Algorithm 1.** Clima training

Draw $\mathbf{x} \sim p_0(\mathbf{x})$
**for** $t = 0, 1, \ldots$ **do**
  **if** $\mathbf{y}_t \neq []$ **then**
    Add $\mathbf{x}$ as column to $\mathbf{X}$.
  **end if**
  $\mathbf{x} \leftarrow M_{t\to t+1}\mathbf{x}$
**end for**
Find $\phi_1, \theta_1 = \underset{\phi_1,\theta_1}{\operatorname{argmax}}\, \mathbb{E}_{m\sim\mathcal{U}(1,\mathrm{rank}\mathbf{X})}$
$[\mathcal{L}(\phi_1, \theta_1, \mathbf{X}\ \text{column}\ m)]$

**Algorithm 2.** Single ETKF–VAE

Find $\phi_1, \theta_1$ using Algorithm 1
**for** $m = 1, \ldots, M$ **do**
  Draw $\mathbf{x}^{\mathrm{f}}_m \sim p_0(\mathbf{x}^{\mathrm{f}})$
**end for**
**for** $t = 0, 1, \ldots$ **do**
  **if** $\mathbf{y}(t) \neq []$ and *transfer* configuration **then**
    Find $\phi_1, \theta_1 = \underset{\phi_1,\theta_1}{\operatorname{argmax}}\, \mathbb{E}_{m\sim\mathcal{U}(1,M)}[\mathcal{L}(\phi_1, \theta_1, \mathbf{x}^{\mathrm{f}}_m)]$
  **end if**
  **if** $\mathbf{y}(t) \neq []$ **then**
    **for** $k = 1, \ldots, K$ **do**
      $\mathbf{D}_K$ column $k \leftarrow \mathbf{y} + \epsilon^{\mathbf{y}}_k - H(\mathbf{x}^{\mathrm{f}}_{m_k})$
    **end for**
    **for** $k = 1, \ldots, M$ **do**
      $\mathbf{D}_M$ column $m \leftarrow \mathbf{y} - H(\mathbf{x}^{\mathrm{f}}_m)$
    **end for**
    **for** $m = 1, \ldots, M$ **do**
      $\mathbf{Z}^{\mathrm{f}}$ column $m \leftarrow\sim \mathcal{N}(\mu_{\phi_1}(\mathbf{x}^{\mathrm{f}}_m), \Sigma_{\phi_1}(\mathbf{x}^{\mathrm{f}}_m))$
    **end for**
    Obtain $\mathbf{Z}^{\mathrm{a}}$ from Equation (7a) using $\mathbf{Z}^{\mathrm{f}}$, $\mathbf{D}_K$ and $\mathbf{D}_M$.
    **for** $m = 1, \ldots, M$ **do**
      $\mathbf{z}^{\mathrm{a}}_m \leftarrow \mathbf{Z}^{\mathrm{a}}$ column $m$
      Draw $\mathbf{x}^{\mathrm{a}}_m \sim \mathcal{N}(\mu_{\theta_1}(\mathbf{z}^{\mathrm{a}}_m), \Sigma_{\theta_1}(\mathbf{z}^{\mathrm{a}}_m))$
    **end for**
  **end if**
  **for** $m = 1, \ldots, M$ **do**
    $\mathbf{x}^{\mathrm{f}}_m \leftarrow M_{t\to t+1}\mathbf{x}^{\mathrm{a}}_m$
  **end for**
**end for**

**Algorithm 3.** Double ETKF–VAE

Find $\phi_1, \theta_1$ using Algorithm 1
**for** $m = 1, \dots, M$ **do**
  Draw $\mathbf{x}^{\mathrm{f}}_m \sim p_0(\mathbf{x}^{\mathrm{f}})$
**end for**
**for** $t = 0, 1, \dots$ **do**
  **if** $\mathbf{y}(t) \neq []$ and *transfer* configuration **then**
    Find $\phi_1, \theta_1 = \underset{\phi_1,\theta_1}{\operatorname{argmax}} \mathbb{E}_{m\sim\mathcal{U}(1,M)}[\mathcal{L}(\phi_1, \theta_1, \mathbf{x}^{\mathrm{f}}_m)]$
  **end if**
  **if** $\mathbf{y}(t) \neq []$ **then**
    **for** $k = 1, \dots, K$ **do**
      $\mathbf{d}_{ij} \leftarrow H(\mathbf{x}^{\mathrm{f}}_{m_i}) + \epsilon^{\mathbf{y}}_k - H(\mathbf{x}^{\mathrm{f}}_{m_j})$
    **end for**
    Find $\phi_2, \theta_2 = \underset{\phi_2,\theta_2}{\operatorname{argmax}} \mathbb{E}_{i,j\sim\mathcal{U}(1,M)}[\mathcal{L}(\phi_2, \theta_2, \mathbf{d}_{ij})]$
    **for** $k = 1, \dots, K$ **do**
      $\mathbf{d} \leftarrow \mathbf{y} + \epsilon^{\mathbf{y}}_k - H(\mathbf{x}^{\mathrm{f}}_{m_k})$
      $\mathbf{D}_K$ column $k \leftarrow\sim \mathcal{N}(\mu_{\phi_2}(\mathbf{d}), \Sigma_{\phi_2}(\mathbf{d}))$
    **end for**
    **for** $k = 1, \dots, M$ **do**
      $\mathbf{d} \leftarrow \mathbf{y} - H(\mathbf{x}^{\mathrm{f}}_m)$
      $\mathbf{D}_M$ column $m \leftarrow\sim \mathcal{N}(\mu_{\phi_2}(\mathbf{d}), \Sigma_{\phi_2}(\mathbf{d}))$
    **end for**
    **for** $m = 1, \dots, M$ **do**
      $\mathbf{Z}^{\mathrm{f}}$ column $m \leftarrow\sim \mathcal{N}(\mu_{\phi_1}(\mathbf{x}^{\mathrm{f}}_m), \Sigma_{\phi_1}(\mathbf{x}^{\mathrm{f}}_m))$
    **end for**
    Obtain $\mathbf{Z}^{\mathrm{a}}$ from Equation (7a) using $\mathbf{Z}^{\mathrm{f}}$, $\mathbf{D}_K$ and $\mathbf{D}_M$.
    **for** $m = 1, \dots, M$ **do**
      $\mathbf{z}^{\mathrm{a}}_m \leftarrow \mathbf{Z}^{\mathrm{a}}$ column $m$
      Draw $\mathbf{x}^{\mathrm{a}}_m \sim \mathcal{N}(\mu_{\theta_1}(\mathbf{z}^{\mathrm{a}}_m), \Sigma_{\theta_1}(\mathbf{z}^{\mathrm{a}}_m))$
    **end for**
  **end if**
  **for** $m = 1, \dots, M$ **do**
    $\mathbf{x}^{\mathrm{f}}_m \leftarrow M_{t\to t+1}\mathbf{x}^{\mathrm{a}}_m$
  **end for**
**end for**

## APPENDIX B. LYAPUNOV EXPONENT OF THE MODEL

Let $(x(t_{p+1}), y(t_{p+1}) = M(x(t_p, y_p))$, with $M$ the homogeneous ($A = 0$) mapping outlined in Equation (10). For a discrete map, the Lyapunov spectrum is defined as the eigenvalues of

$$\lim_{p\to\infty} \frac{1}{2p} \sum_{p'=0}^{p-1} \log \Lambda_{p'}, \tag{B1}$$

with $\Lambda_p$ the eigenvalues of $\mathbf{D}M|_{\mathbf{x}(t_p)}\mathbf{D}M|^{\mathrm{T}}_{\mathbf{x}(t_p)}$ and $\mathbf{D}M$ the derivative of $M$ at $\mathbf{x}(t_p)$. Since $\mathbf{D}M(t_p)\mathbf{D}M(t_p)^{\mathrm{T}}$ is a constant in the homogeneous model, the limit in Equation (B1) is equal to half the log of the eigenvalues of $\mathbf{D}M|_{\mathbf{x}(t_p)}\mathbf{D}M|^{\mathrm{T}}_{\mathbf{x}(t_p)}$. The numerically calculated values of the largest Lyapunov exponents are shown for different angular accelerations $\alpha$ in Figure B1a. The figure shows that, in all cases shown, including the $\alpha = 0.1$ case used in this study, the maximum Lyapunov exponent is larger than zero. This indicate that it is a map that grows perturbations over time, making it chaotic.

To test the ergodicity of the system, 4000 angles were selected randomly and their associated positions on the unit circle were used as initial conditions for a 8000-time-step long model runs. After discarding the first 4000 time steps as spin-up, the 4000 states at time 4000 were used to estimate the CDFs for the $x$- and $y$-coordinates. These ensemble CDFs are shown in Figure B1b. For each of the 4000 ensemble members, a time CDF has also been calculated based on the last 4000 time steps of the ensemble member model trajectory. The range of these time CDFs is depicted in Figure B1b as a shaded region. In a ergodic system, the ensemble-based and time-based CDFs should match in the limit in which the number of time steps and ensemble members go to infinity. Though not a formal proof, Figure B1b shows that this, to a good approximation, is satisfied and that the system at least behaves like an ergodic one.

## APPENDIX C. KALMAN EQUATIONS IN LATENT SPACE

In order to gain some insight into what the KF in the latent spaces of VAEs looks like, we expand the expressions for the states in the latent space and their mean and covariance around the truth as functions of the observational and forecast errors. We will use a linear approximation, that is, terms that are second order or higher in the errors are neglected. The following derivation holds for the *double ETKF–VAE* configurations. However, the equivalent for the *single ETKF–VAE* can be obtained by setting $\mu_{\phi_2}(\mathbf{d}) = \mathbf{d}$ and $\Sigma_{\phi_2}(\mathbf{d}) = \mathbf{0}$.

For a state $\mathbf{x} \in \mathcal{X}$ and observation $\mathbf{y} \in \mathcal{Y}$, we define

$$\mathbf{z} = \mu_{\phi_1}(\mathbf{x}) + \epsilon^{\mathbf{z}} \approx \mu_{\phi_1}(\mathbf{x}^{\mathrm{truth}}) + \mathbf{D}\mu_{\phi_1}\epsilon^{\mathbf{x}} + \epsilon^{\mathbf{z}}, \tag{C1a}$$

$$\mathbf{f} = \mu_{\phi_2}(\mathbf{y} + \epsilon^{\mathbf{y}} - H(\mathbf{x})) + \epsilon^{\mathbf{d}} \approx \mu_{\phi_2}(\mathbf{y} - H(\mathbf{x}^{\mathrm{truth}})) - \mathbf{D}\mu_{\phi_2}\mathbf{H}\epsilon^{\mathbf{x}} + \mathbf{D}\mu_{\phi_2}\epsilon^{\mathbf{y}} + \epsilon^{\mathbf{d}}, \tag{C1b}$$

$$\mathbf{g} = \mu_{\phi_2}(\mathbf{y} - H(\mathbf{x})) \approx \mu_{\phi_2}(\mathbf{y} - H(\mathbf{x}^{\mathrm{truth}})) - \mathbf{D}\mu_{\phi_2}\mathbf{H}\epsilon^{\mathbf{x}} + \epsilon^{\mathbf{d}}, \tag{C1c}$$

with $\mathbf{x}^{\mathrm{truth}}$ the unknown true state of the model, $\mathbf{H}$ the derivative of the potentially nonlinear observation operator $H$, $\epsilon^{\mathbf{y}}$ the observational error, $\epsilon^{\mathbf{x}}$ the forecast error in state $\mathbf{x}$, $\mu_{\phi_1}$ the function for the conditional mean in the

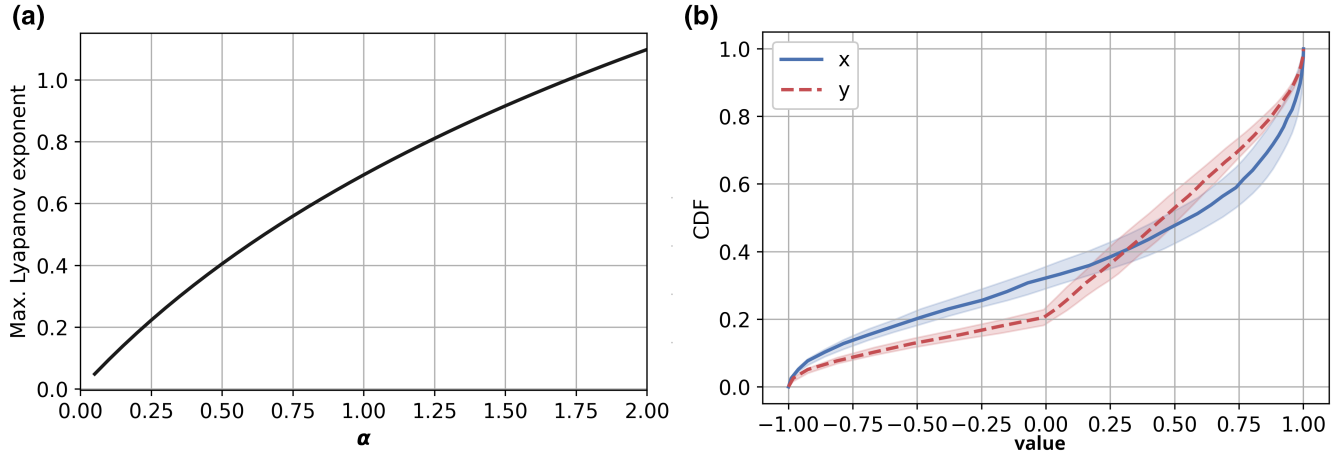


**FIGURE B1** (a) Largest Lyapunov exponent as a function of the angular acceleration $\alpha$ (see Equation 9). (b) Cumulative probability distribution of the $x$- (blue) and $y$-coordinate (red) estimated for a 4000-member ensemble (line) and 4000 time steps of a model trajectory (shading).

first VAE, $\mu_{\phi_2}$ the function for the conditional mean in the second VAE,

$$\mathbf{D}\mu_{\phi_1} = \frac{d\mu_{\phi_1}}{d\mathbf{x}}(\mathbf{x}^{\text{truth}}), \qquad \mathbf{D}\mu_{\phi_2} = \frac{d\mu_{\phi_2}}{d\mathbf{d}}(\mathbf{y} - H(\mathbf{x}^{\text{truth}}))$$

indicating their derivatives, $\epsilon^{\mathbf{z}}$ a realisation from a Gaussian with zero mean and covariance $\Sigma_{\phi_1}(\mathbf{x})$, and $\epsilon^{\mathbf{d}}$ a realisation from a Gaussian with zero mean and covariance $\Sigma_{\phi_2}(\mathbf{d})$.

If $M, K \gg 1$, the matrix products appearing in Equation (8) can now be approximated as

$$\frac{1}{M-1}\tilde{\mathbf{Z}}^{\text{f}}(\tilde{\mathbf{Z}}^{\text{f}})^{\text{T}} \approx \mathbb{E}[\mathbf{z}\mathbf{z}^{\text{T}}] = \mathbf{D}\mu_{\phi_1}\mathbf{P}^{\text{f}}_{\mathbf{X}}\mathbf{D}\mu^{\text{T}}_{\phi_1} + \Sigma_{\phi_1}, \tag{C2a}$$

$$\frac{1}{K-1}\tilde{\mathbf{F}}_K\tilde{\mathbf{F}}_K^{\text{T}} \approx \mathbb{E}[\tilde{\mathbf{d}}\tilde{\mathbf{d}}^{\text{T}}] = \mathbf{D}\mu_{\phi_2}\mathbf{R}\mathbf{D}\mu^{\text{T}}_{\phi_2} + \mathbf{D}\mu_{\phi_2}\mathbf{H}\mathbf{P}^{\text{f}}_{\mathbf{X}}\mathbf{H}^{\text{T}}\mathbf{D}\mu^{\text{T}}_{\phi_2} + \Sigma_{\phi_2}, \tag{C2b}$$

$$-\frac{1}{M-1}\tilde{\mathbf{Z}}\tilde{\mathbf{F}}_M^{\text{T}} \approx \mathbb{E}[\tilde{\mathbf{z}}(-\tilde{\mathbf{g}})^{\text{T}}] = \mathbf{D}\mu_{\phi_1}\mathbf{P}^{\text{f}}_{\mathbf{X}}\mathbf{H}^{\text{T}}\mathbf{D}\mu^{\text{T}}_{\phi_2}, \tag{C2c}$$

$$\frac{1}{M}\mathbf{F}_M\mathbf{1}_M \approx \mathbb{E}[\mathbf{g}] = \mu_{\phi_2}(\mathbf{y} - H(\mathbf{x}^{\text{truth}})) - \mathbf{D}\mu_{\phi_2}\mathbf{H}\mathbb{E}[\epsilon^{\mathbf{x}}], \tag{C2d}$$

$$\frac{1}{M}\mathbf{Z}\mathbf{1}_M \approx \mathbb{E}[\mathbf{z}] = \mu_{\phi_1}(\mathbf{x}^{\text{truth}}) + \mathbf{D}\mu_{\phi_1}\mathbb{E}[\epsilon^{\mathbf{x}}]. \tag{C2e}$$

Here it is assumed for simplicity that observational errors are unbiased ($\mathbb{E}[\epsilon^{\mathbf{y}}] = \mathbf{0}$) and, as is conventional in DA, that $\epsilon^{\mathbf{x}}$, $\epsilon^{\mathbf{y}}$, $\epsilon^{\mathbf{z}}$, and $\epsilon^{\mathbf{d}}$ are statistically independent.

The post-DA ensemble mean ($\mu^{\text{a}}_{\mathbf{z}}$) and covariance ($\mathbf{P}^{\text{a}}_{\mathbf{z}}$) in the latent space are given by $\mathbf{X}$ with $\mathbf{Z}$ in Equation (4), with $\cdot_{\mathbf{x}}$ replaced by $\cdot_{\mathbf{z}}$. Substitution of $\mu_{\mathbf{Z}}$ and $\mathbf{P}_{\mathbf{Z}}$ with their ensemble estimates based on $\mathbf{Z}$, $\mathbf{F}_M$, and $\mathbf{F}_K$ as defined in Section 2.3.2 then gives, after inserting the approximations in Equation (C2a–e),

$$\mathbf{K} \overset{\text{def}}{=} \mathbf{D}\mu_{\phi_1}\mathbf{P}^{\text{f}}_{\mathbf{X}}\mathbf{H}^{\text{T}}\mathbf{D}\mu^{\text{T}}_{\phi_2}\left(\mathbf{D}\mu_{\phi_2}\mathbf{H}\mathbf{P}^{\text{f}}_{\mathbf{X}}\mathbf{H}^{\text{T}}\mathbf{D}\mu^{\text{T}}_{\phi_2} + \mathbf{D}\mu_{\phi_2}\mathbf{R}\mathbf{D}\mu^{\text{T}}_{\phi_2} + \Sigma_{\phi_2}\right)^{-1}, \tag{C3a}$$

$$\mu^{\text{a}}_{\mathbf{Z}} \approx \mu_{\phi_1}(\mathbf{x}^{\text{truth}}) + \mathbf{D}\mu_{\phi_1}\mathbb{E}[\epsilon^{\mathbf{x}}] + \mathbf{K}(\mu_{\phi_2}(\mathbf{y} - H(\mathbf{x}^{\text{truth}})) - \mathbf{D}\mu_{\phi_2}\mathbf{H}\mathbb{E}[\epsilon^{\mathbf{x}}]), \tag{C3b}$$

$$\mathbf{P}^{\text{a}}_{\mathbf{z}} \approx \mathbf{D}\mu_{\phi_1}\mathbf{P}^{\text{f}}_{\mathbf{X}}\mathbf{D}\mu^{\text{T}}_{\phi_1} + \Sigma_{\phi_1} - \mathbf{K}\mathbf{H}\mathbf{P}^{\text{f}}_{\mathbf{X}}\mathbf{D}\mu^{\text{T}}_{\phi_1}. \tag{C3c}$$

Based on Equation (C3a–c), the following two observations can be made. First, if we write down the SVD of $\mathbf{D}\mu_{\phi_1}$ and $\mathbf{D}\mu_{\phi_2}$, for example, $\mathbf{D}\mu_{\phi_1} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^{\text{T}}$, then we can see that the right singular vectors (the columns of $\mathbf{V}_i$) of $\mathbf{D}\mu_{\phi_i}$ with $i \in \{1, 2\}$ basically act as feature selectors determining which features in the state space dominate the uncertainty in the latent space. Here, the singular values on the diagonal of $\mathbf{S}_i$ emphasize or de-emphasize the importance of the different features. Second, the VAEs introduce additional uncertainty in the ensemble. The additional uncertainty from the first VAE, $\Sigma_{\phi_1}$, takes on the same role as the model error covariance in the conventional KF. The uncertainty introduced by the second VAE, $\Sigma_{\phi_2}$, behaves as a representativeness error (Janjić *et al.*, 2018).