Earth and Space Science

RESEARCH ARTICLE
10.1029/2021EA001908

Non-Gaussian Detection Using Machine Learning With Data Assimilation Applications

Michael R. Goodliff[1,2], **Steven J. Fletcher**[3], **Anton J. Kliewer**[3], **Andrew S. Jones**[3], and **John M. Forsythe**[3]

[1]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, CO, USA,
[2]National Oceanic and Atmospheric Administration (NOAA) Physical Sciences Laboratory (PSL), Boulder, CO, USA,
[3]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA

Key Points:

- Machine learning can be used to detect non-Gaussian distributions
- Given the machine learning detection, we can switch between optimal data assimilation schemes
- Data assimilation can be improved using machine learning techniques

Correspondence to:

M. R. Goodliff,
michael.goodliff@noaa.gov

Citation:

Goodliff, M. R., Fletcher, S. J., Kliewer, A. J., Jones, A. S., & Forsythe, J. M. (2022). Non-Gaussian detection using machine learning with data assimilation applications. *Earth and Space Science*, *9*, e2021EA001908. https://doi.org/10.1029/2021EA001908

Received 26 NOV 2021
Accepted 18 MAR 2022

Author Contributions:

Formal analysis: Michael R. Goodliff
Funding acquisition: Steven J. Fletcher
Methodology: Michael R. Goodliff
Supervision: Steven J. Fletcher
Visualization: Michael R. Goodliff
Writing – original draft: Michael R. Goodliff
Writing – review & editing: Michael R. Goodliff, Steven J. Fletcher, Anton J. Kliewer, Andrew S. Jones, John M. Forsythe

Abstract In most data assimilation and numerical weather prediction systems, the Gaussian assumption is prevalent for the behavior of the random variables/errors that are involved. At the Cooperative Institute for Research in the Atmosphere theory has been developed for different forms of variational data assimilation schemes that enables the Gaussian assumption to be relaxed. For certain variable types, a lognormally distributed random variable can be combined in a mixed Gaussian-lognormal distribution to better capture the interactions of the errors of different distributions. However, assuming that a distribution can change in time, then developing techniques to know when to switch between different versions of the data assimilation schemes becomes very important. By dynamically changing the formulation of the data assimilation system we are able to assimilate observations in a way that reflects the flow-dependent variability of their distribution.

In this paper, we present results with a machine learning technique (the support vector machine) to switch between data assimilation methods based on the detection of a change in the Lorenz 1963 model's z component's probability distribution. Given the machine learning technique's detection/prediction of a change in the distribution, then either a Gaussian or a mixed Gaussian-lognormal 3DVar based cost function is used to minimize the errors in this period of time. It is shown that switching from a Gaussian 3DVar to a lognormal 3DVar at lognormally distributed parts of the attractor improves the data assimilation analysis error compared to using one distribution type for the entire attractor.

1. Introduction

The assumption that variables, and their errors, are Gaussian distributed is commonplace in areas such as numerical weather prediction and modeling. Research such as that undertaken by Perron and Sura (2013) has shown that this assumption is generally false for atmospheric variables, and that Gaussian variables in the atmosphere are rare. The aforementioned statement was based on a 62 year long project from daily data taken from the National Centers for Environmental Prediction and the National Center for Atmospheric Research, using the Reanalysis I Project data set. Given this evidence, the need to be able to relax the Gaussian assumption for the errors involved in the data assimilation schemes becomes quite important if the analysis error is to be minimized. By doing so, we may deliver an improvement in the subsequent forecast.

Most of the current formulations of data assimilation, for example, variational methods such as 3DVar and 4DVar (which are based upon Bayes theorem Fletcher, 2017), and ensemble methods such as the Ensemble Kalman Filter (EnKF; Evensen & Van Leeuwen, 1996), the (local) Ensemble Transform Kalman Filter ((L)ETKF) Ott et al. (2004); Wang and Bishop (2003), which are based upon a control theory/weighted least squares approach using ensemble members to approximate the analysis mean and covariances, and the Maximum Likelihood Ensemble Kalman Filter, Zupanski (2005), which uses the Kalman filter equations combined with the 3DVar cost function, all assume that the errors involved are Gaussian distributed. Also in linear state estimation, the initial condition x_0 is also assumed to be Gaussian distributed. Other papers that look into non-Gaussian data assimilation methods are local particle filters, van Leeuwen et al. (2019), and Amezcua and Leeuwen (2014) who looked into Gaussian anamorphosis on the EnKF.

However, at the Cooperative Institute for Research in the Atmosphere (CIRA) at Colorado State University, a theory has been developed and tested, that allows for the Gaussian assumption for the distribution of the errors to be relaxed to a lognormal distribution. In Fletcher and Zupanski (2006a), the theory is presented for the case where there are lognormal observational errors in 3DVar. In Fletcher and Zupanski (2006b), a mixed Gaussian-lognormal distribution is presented, and an associated cost function that allows for the simultaneous minimization of Gaussian and lognormal errors is presented. The mixed approach was extended to the background term in Fletcher and Zupanski (2007), and then tested with the Lorenz 1963 model (Lorenz, 1963), Lorenz 63 hereafter, where it is shown here that the z component of this model is not Gaussian distributed. The mixed distribution theory was extended to a 4DVar type system in Fletcher (2010), and eventually shown for incremental 3DVar and 4DVar in Fletcher and Jones (2014). In Fletcher & Zupanski (2007) and Fletcher (2010), it is demonstrated that the lognormal variant of 3DVar and 4DVar provided improvements in the analysis accuracy over the traditional Gaussian, and a logarithmic transforms method applied to the z component, but that there was also improvement in the analysis error for the x and y components, where the errors associated with these components were assumed to be Gaussian distributed.

An important property of the mixed Gaussian-lognormal variational approach is that the mode of this distribution contains the covariances between the Gaussian components and the lognormal components, and as such this enables corrections to the lognormal component to impact the Gaussian component. An example of this can be found in Kliewer et al. (2016) where the mixed distribution is implemented in a microwave brightness temperature, temperature-mixing ratio retrieval system. It is shown in (Kliewer et al., 2016) that a better fit for the temperature channel can be obtained with the mixed approach than with the Gaussian-fits-all approach.

There are other techniques that have been proposed to handle non-Gaussian aspects, one of which is the localized particle filter in Poterjoy (2016); Poterjoy and Anderson (2016); Poterjoy et al. (2017), which utilizes an extension of the particle weights into vector quantities to reduce the influence of distant observations on the weight calculations via a localisation function. However, an advantage of the variational methods for non-Gaussian distributions developed by Fletcher et al. is that they are computational cheaper to run compared to the particle filter methods and, unlike anamorphosis methods, solve the estimation problem exactly as they find the mode of the correct posterior distribution.

As shown in the first part of Goodliff et al. (2020), the trajectory of the z component of the Lorenz 63 model changes distributions on different parts of the underlying attractor, and as such if the data assimilation is to be optimized then these changes need to be used to switch from the Gaussian to the mixed distribution-based cost functions. In the second part of Goodliff et al. (2020), a support vector machine and a neural network were tested with the Lorenz 63 model to detect non-Gaussian behavior. It was shown that these techniques were very capable of detecting skewness, and differences in descriptive statistics, in order to estimate, and predict, non-Gaussianity.

Recently, machine learning methods have become very popular in atmospheric sciences, especially in areas such as numerical weather prediction and modeling (Scher & Messori, 2018) to help find biases and correlations in data, and also to help reduce analysis and forecast errors. Pasini and Pelino (Pasini & Pelino, 2005) and Pasini (2008) used two Lorenz 63 attractors to analyze predictability. There have been many other studies using machine learning methods to try and improve weather forecasting and climate modeling and the reader is referred to Dueben and Bauer (2018); Rasp and Lerch (2018); Scher (2018); Scher and Messori (2019); Weyn et al. (2019) for some of these extra examples. In the realms of data assimilation, integrating machine learning has recently become a hot topic. An example of this can be seen in Penny et al. (2021). Here, they integrated machine learning within the data assimilation framework to generate data-driven state estimations using recurrent neural networks (RNNs). These RNNs were incorporated to replace key components of the data assimilation cycle, and results showed benefits for short term forecasts. Other recent examples of machine learning implementation within data assimilation can be found in Arcucci et al. (2021) and Brajard et al. (2021).

Given the progress made with machine learning techniques, and the need identified above to be able to inform a data assimilation scheme to switch between different versions of the cost function, this paper investigates a support vector machine, which is a supervised machine learning algorithm, to detect non-Gaussian probability density functions in the Lorenz 63 model. This approach is applied to the z component of the Lorenz attractor, where the skewness of said z variable is the target data, and the x and y components of the attractor are our predictors. We use this to then apply a "switch" to our data assimilation method to change between a Gaussian-fits-all cost function to a mixed Gaussian-lognormal based cost function, where the x and y components are assumed to have Gaussian error throughout the experiment, and that the z component is switching between a Gaussian and a lognormal distribution. This switch changes the data assimilation methodology from the traditional Gaussian

3DVar to a mixed Gaussian-lognormal variant of 3DVar Fletcher and Zupanski (2007) based on the distribution estimation given by the support vector machine.

The remainder of the paper is organized as follows: Section 2 will start with an overview of the Bayesian model for the variational data assimilation theories and the methods used. We will also discuss the machine learning methodology and how we use it with the data assimilation. Section 3 describes the Lorenz 63 model, and Section 4 shows the experimentation using the mixed DA/ML scheme on the Lorenz 63 model to improve forecasts. The paper is concluded in Section 5.

2. Methodology

In this section we shall present the different data assimilation and machine learning techniques that are used in the results presented later.

2.1. Traditional 3DVar (3DVar-G)

Variational data assimilation methods estimate the most probable state of the system, which is the mode of the posterior probability distribution function. In traditional 3DVar, this comes from Gaussian statistics and is a combination of the background and the likelihood, with the background term written as:

$$p(\mathbf{x}) = \frac{1}{|\mathbf{B}_c|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^{\top}\mathbf{B}_c^{-1}\left(\mathbf{x}-\mathbf{x}^b\right)\right), \quad (1)$$

where the background state, and initial state that is sought, are given by $\mathbf{x}^b$ and $\mathbf{x}$, respectively, $\mathbf{B}_c \in \mathcal{R}^{N\times N}$ is the background error covariance matrix, and N is the total number of background variables. The Gaussian likelihood is defined as

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{|\mathbf{R}|^{\frac{1}{2}}(2\pi)^{\frac{N_o}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{h}(\mathbf{x}))^{\top}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{h}(\mathbf{x}))\right), \quad (2)$$

where $\mathbf{y}$ is the observation, $\mathbf{h}(\cdot)$ is the (non)-linear observation operator, $\mathbf{R} \in \mathcal{R}^{N_o\times N_o}$ is the observational error covariance matrix, and N_o is the total number of observations. The next step is to substitute Equation 1 and Equation 2 into Bayes' theorem

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}), \quad (3)$$

and seek the state that maximizes the posterior PDF in Equation 3. However, it is quite often easier to work with the equivalent problem that seeks the state that minimizes the negative log-likelihood of Equation 3, which for the Gaussian definitions presented above results in the following cost function,

$$J(x) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^{\top}\mathbf{B}_c^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \frac{1}{2}(\mathbf{y}-\mathbf{h}(\mathbf{x}))^{\top}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{h}(\mathbf{x})), \quad (4)$$

that has to be minimized.

2.2. Mixed Gaussian-Lognormal 3DVar (3DVar-Mix)

The mixed Gaussian-lognormal 3DVar data assimilation scheme was first presented in Fletcher and Zupanski (2007) for both the background and likelihood components. This version of 3DVar uses a multivariate lognormal distribution based cost function for lognormal random variables that is derived through using a similar approach for Bayes theorem as presented above. Thus, for the lognormal approach the a priori probability density function is given by

$$p(\mathbf{x}) = \left(\prod_{i=1}^{N}\frac{1}{x_i}\right)\frac{1}{|\mathbf{B}_L|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}\left(\ln\mathbf{x}-\ln\mathbf{x}^b\right)^\top\mathbf{B}_L^{-1}\left(\ln\mathbf{x}-\ln\mathbf{x}^b\right)\right), \tag{5}$$

where $\mathbf{B}_L$ is the lognormal based background error covariance matrix, which is defined in terms of expectations of $\ln\mathbf{x}$ and not $\mathbf{x}$. The corresponding lognormal likelihood is given by

$$p(\mathbf{y}|\mathbf{x}) = \left(\prod_{i=1}^{N_o}\frac{(\mathbf{h}(\mathbf{x}))_i}{\mathbf{y}_i}\right)\frac{1}{|\mathbf{R}_L|^{\frac{1}{2}}(2\pi)^{\frac{N_o}{2}}} \exp\left(-\frac{1}{2}(\ln\mathbf{y}-\ln\mathbf{h}(\mathbf{x}))^\top\mathbf{R}_L^{-1}(\ln\mathbf{y}-\ln\mathbf{h}(\mathbf{x}))\right). \tag{6}$$

This then results in the lognormal 3DVar cost function given by

$$\begin{aligned} J(x) &= \frac{1}{2}\left(\ln\mathbf{x}-\ln\mathbf{x}^b\right)^\top\mathbf{B}_L^{-1}\left(\ln\mathbf{x}-\ln\mathbf{x}^b\right)+\left(\ln\mathbf{x}-\ln\mathbf{x}^b\right)^\top\mathbf{1}_N \\ &+\frac{1}{2}(\ln\mathbf{y}-\ln\mathbf{h}(\mathbf{x}))^\top\mathbf{R}_L^{-1}(\ln\mathbf{y}-\ln\mathbf{h}(\mathbf{x}))+(\ln\mathbf{y}-\ln\mathbf{h}(\mathbf{x}))^\top\mathbf{1}_{N_o}, \end{aligned} \tag{7}$$

where $\mathbf{1}_N$ denotes a vector of 1s of the length of the subscript. Minimizing this cost function gives us the solution to the lognormal 3DVar. For in depth information about lognormal 3DVar, refer to Fletcher and Zupanski (2007) and Fletcher (2010).

However, in the results that will be presented later in this paper, the mixed Gaussian-lognormal approach is utilized, which comes from considering the mixed Gaussian-lognormal probability density function derived in Fletcher and Zupanski (2006b), where the multivariate distribution used for the a priori distribution is given by

$$\begin{aligned} p(\mathbf{x}) = &\left(\prod_{i=p+1}^{N}\frac{1}{x_i}\right)\frac{1}{|\mathbf{B}_{mx}|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \\ &\times\exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{x}_p-\mathbf{x}_{bp}\\ \ln\mathbf{x}_q-\ln\mathbf{x}_{bq}\end{pmatrix}^\top\mathbf{B}_{mx}^{-1}\begin{pmatrix}\mathbf{x}_p-\mathbf{x}_{bp}\\ \ln\mathbf{x}_q-\ln\mathbf{x}_{bq}\end{pmatrix}\right), \end{aligned} \tag{8}$$

where p is the number of Gaussian random variables, q is the number of lognormal random variables, such that $N = p + q$. The vectors x_p and x_q are defined as the portions of the full state x which correspond to the Gaussian and lognormal variables respectively The mixed distribution error covariance matrix here is defined as

$$\mathbf{B}_{mx}\equiv\begin{pmatrix}\mathbf{B}_{\mathbf{pp}}^{\mathbf{GG}} & \mathbf{B}_{\mathbf{pq}}^{\mathbf{GL}}\\ \mathbf{B}_{\mathbf{qp}}^{\mathbf{LG}} & \mathbf{B}_{\mathbf{qq}}^{\mathbf{LL}}\end{pmatrix}. \tag{9}$$

where the different $\mathbf{B}$ are the sub-background error covariance matrices associated with the different combinations of Gaussian and lognormal errors, and the superscripts G and L denote the Gaussian and lognormal components. The mixed Gaussian-lognormal distribution that would be used for the likelihood of Gaussian and lognormal errors is given by

$$\begin{aligned} p(\mathbf{y}|\,\mathbf{x})\equiv&\left(\prod_{i=p+1}^{N_o}\frac{\mathbf{h}_i(\mathbf{x})}{\mathbf{y}_i}\right)\frac{1}{|\mathbf{R}_{\mathbf{mx}}|^{\frac{1}{2}}(2\pi)^{\frac{N_o}{2}}} \\ &\times\exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{y}_p-\mathbf{h}_p(\mathbf{x})\\ \ln\mathbf{y}_q-\ln\mathbf{h}_q(\mathbf{x})\end{pmatrix}^\top\mathbf{R}_{mx}^{-1}\begin{pmatrix}\mathbf{y}_p-\mathbf{h}_p(\mathbf{x})\\ \ln\mathbf{y}_q-\ln\mathbf{h}_q(\mathbf{x})\end{pmatrix}\right), \end{aligned} \tag{10}$$

where the observation covariance matrix is assumed to be diagonal, which is a current assumption made in several operational numerical weather prediction centers, but is being relaxed to allow for correlations between certain observations. The associated variances in these entries are calculated as per their distribution that they are associated with. Thus the associated mixed Gaussian-lognormal cost function is given by

$$\begin{aligned} J(\mathbf{x}) = &\frac{1}{2}\begin{pmatrix} \mathbf{x}_p - \mathbf{x}_{bp} \\ \ln \mathbf{x}_q - \ln \mathbf{x}_{bq} \end{pmatrix}^{\top} \mathbf{B}_{mx}^{-1} \begin{pmatrix} \mathbf{x}_p - \mathbf{x}_{bp} \\ \ln \mathbf{x}_q - \ln \mathbf{x}_{bq} \end{pmatrix} \\ &+ \begin{pmatrix} \mathbf{x}_p - \mathbf{x}_{bp} \\ \ln \mathbf{x}_q - \ln \mathbf{x}_{bq} \end{pmatrix}^{\top} \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \\ &+ \frac{1}{2}\begin{pmatrix} \mathbf{y}_p - \mathbf{h}_p(\mathbf{x}) \\ \ln \mathbf{y}_q - \ln \mathbf{h}_q(\mathbf{x}) \end{pmatrix}^{\top} \mathbf{R}_{mx}^{-1} \begin{pmatrix} \mathbf{y}_p - \mathbf{h}_p(\mathbf{x}) \\ \ln \mathbf{y}_q - \ln \mathbf{h}_q(\mathbf{x}) \end{pmatrix} \\ &+ \begin{pmatrix} \mathbf{y}_p - \mathbf{h}_p(\mathbf{x}) \\ \ln \mathbf{y}_q - \ln \mathbf{h}_q(\mathbf{x}) \end{pmatrix}^{\top} \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix}, \end{aligned} \tag{11}$$

where $\mathbf{0}_N$ is a vector of 0s of the length of the subscript.

It should be noted that it could well be the case that the number of state variables that are Gaussian or lognormal may not be the same as those of the observational errors. Another important feature to note here is that the mode of the mixed distribution is a function of the error sub-covariance matrices, as shown below for a generalized covariance matrix $\mathbf{\Sigma}$,

$$\begin{pmatrix} \mathbf{x}_p \\ \mathbf{x}_q \end{pmatrix}_{mode} = \begin{pmatrix} \boldsymbol{\mu}_p - \langle \mathbf{\Sigma}_{qp}, \mathbf{1}_p \rangle \\ \exp\left\{\boldsymbol{\mu}_q - \langle \mathbf{\Sigma}_{qq}, \mathbf{1}_q \rangle\right\} \end{pmatrix}, \tag{12}$$

(Fletcher, 2017). It is clear from Equation 12 that the Gaussian components become a function of the covariances with the lognormal components, which then enables a relationship between the Gaussian and lognormal components, which is not present in the mode of a Gaussian-fits-all approach.

2.3. Mixing Machine Learning Into Data Assimilation (3DVar-ML)

The machine learning method used in this study to classify our data is the Support Vector Machine (Nello & Shawe-Taylor, 2000). This method separates classified training data with a hyperplane. The support vector machine is a method of supervised learning where we supply a training set and a target set to train our model. In this experiment, we use the radial basis function kernel, and train the machine learning algorithm for 50,000 time steps of the Lorenz 63 model.

In this experiment, we predict the probability density function of the z component of the Lorenz 63 model based on the values of x and y components of the model. Using x and y as the training data and the z-score of the target data, we have shown (Goodliff et al., 2020) that we can be highly precise in our predictions of the distribution of z. The z-score (skewness statistic $\sqrt{\beta_1}$) is calculated and estimated by methods shown in D'agostino et al. (1990) from the standardized skewness:

$$\sqrt{\beta_1} = \frac{E(X-\mu)^3}{\sigma^3} \tag{13}$$

where μ and σ are the mean and standard deviation (of the assimilation window data), respectively. Here, a negative z-score represents a left (negative) skewed distribution, a positive z-score represents a right (positive) skewed distribution, and a z-score of zero refers to a symmetric distribution.

The window length in this study refers to the data around the observation point (e.g., a window length of 11 will be the data five points either side of the observation + the observation point). The z-score affects observation generation (see below) and which version of 3DVar the machine learning algorithm will choose at each observation point.

Through using the support vector machine to detect the probability density most suitable for the trajectory, we introduce this as a switch to decide which data assimilation method is best at the current point in time. The optimal data assimilation method is used with the machine learning prediction as:

$$\text{3DVar-ML} = \begin{cases} \text{3DVar-G}, & \text{if } z\text{-score} < 1, \\ \text{3DVar-Mix}, & \text{otherwise.} \end{cases} \tag{14}$$

3. Lorenz 63

As mentioned in the introduction, for the study that we shall present in the next section, we will be using the Lorenz 1963 model (Lorenz, 1963). This model is a good choice due to its simplicity for a dynamic model which also exhibits chaotic behavior. The model is very sensitive to the initial conditions from which it starts, and as such can give very different answers even by being out by a few decimal places from the true state, Fletcher (2017). These model equations are as given by.

$$\frac{dx}{dt} = -\sigma\,(x - y)\,, \tag{15}$$

$$\frac{dy}{dt} = \rho x - y - zx, \tag{16}$$

$$\frac{dz}{dt} = xy - \beta z, \tag{17}$$

where $x = x\,(t)$, $y = y\,(t)$, $z = z\,(t)$ are the state variables (where t is time) and $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$ are parameters. We start the machine learning training, the true run, and the data assimilation from different initial states on the attractor. Here, the true run (also known as a nature run) refers to the model trajectory without interference from data assimilation or machine learning methods.

4. Results

The experimentation starts with running the support vector machine algorithm on the Lorenz 63 model. We train the support vector machine on variables x and y, and the skewness of z as the output (target) data. The training of the machine learning method is performed over 50,000 time steps to obtain a somewhat robust fit for the system. This approach is based on the method in Goodliff et al. (2020).

To generate the observations in the z dimension, we use the machine learning fit to determine the probability density function at observation time. If the observation is on a positively skewed area of the attractor, that is to say that the z-score ≥ 1, the observation is generated using a lognormal distribution function, else, our observations are generated from a Gaussian distribution. Thus the observations for the three components of the Lorenz 63 model are of the form:

$$y_x = x_t + G_x\,(0, \sigma_{xx})\,, \tag{18}$$

$$y_y = x_t + G_y\,(0, \sigma_{yy})\,, \tag{19}$$

$$y_z = \begin{cases} x_t * \exp\left(G_z\,(0, \sigma_{zz})\right) & \text{if } z\text{-score} \geq 1 \\ x_t + G_z\,(0, \sigma_{zz}) & \text{else} \end{cases} \tag{20}$$

where $y_{x,y,z}$ are the observations, x_t is the truth, and $G_{x,y,z}\,(0, \sigma_{xx,yy,zz})$ is a Gaussian based random number generated with a standard deviation σ. The square of these standard deviations, the variance, will form the diagonal entries of the observational error covariance matrices, where we are assuming that the observations are uncorrelated, and as such the **R** matrices will only be diagonal. In this study, $\mathbf{R} = \mathbf{I}$.

We then run our three data assimilation schemes: 3DVar-G, 3DVar-Mix, and 3DVar-ML. Each method is run over 5,000 time steps, with 50 runs. Running the system for this long removes the effects of randomness (Goodliff et al., 2015). This is done over a mixture of observation periods to test different linearities, here we use $(4, 8, 12, 16, 20, 24, 28)$, and with different window lengths $(9, 13, 17, 21, 25, 29)$ for the machine learning skewness detection (Goodliff et al., 2020).

Throughout the development of the mixed distribution approach it became apparent that there was a sensitivity to the definition of the background error covariance matrix that impacted the ability for the mixed based approach to minimize the cost function. The reason for this problem is due to the property that the mode of the mixed distribution is a function (sum) of the covariances. To overcome this problem, a flow dependent approach was applied in Fletcher and Zupanski (2007) and all subsequent publications associated with the mixed distribution based data assimilation schemes. This flow dependency is achieved through using the averages and covariance averages from the differences between the previous background trajectory and the current trajectory through the time to the next cycle analysis time. This is given by

$$\mathbf{B} = \frac{1}{S}\sum_{i=1}^{S}\langle x^{b} - x^{f}, \left(x^{b} - x^{f}\right)^{T}\rangle, \tag{21}$$

where S is the total number of time steps between observations (Fletcher, 2017). The vectors x^b and x^f refer to the background and current trajectories, which need to be adjusted for use in Gaussian or mixed-lognormal frameworks such that,

$$x_{g}^{b/f} = (x, y, z), \tag{22}$$

$$x_{mx}^{b/f} = (x, y, ln(z)). \tag{23}$$

Here (x, y, z) refer to the L63 state variables, g and mx refer to the Gaussian and mixed-lognormal states respectively. This has been shown through the non-Gaussian development to help stabilise the mixed approach (Fletcher & Zupanski, 2007). To highlight the impact of not updating the background error covariance it can be seen in the results in Kliewer et al. (2016) that when the dynamics become more Gaussian, the Gaussian retrieval has a smaller root mean square error (RMSE) than the mixed approach, but when the dynamics appear more lognormal, then the mixed approach was optimal. This is an indicator that flow dependency helps improve the performance of the lognormal approach. However, because this was a retrieval system and not a model, there was no way to evolve the solution temporally from the previous retrieval time and hence a climatological background error covariance matrix was used.

In Figure 1, we compare the three data assimilation methods with an observation period of four time steps, and skewness window lengths of 9 points (left) and 29 points (right). It can be seen that the 3DVar-ML outperforms both the 3DVar-G and 3DVar-Mix in both scenarios. On the left plot, we see 3DVar-G also is more accurate (in terms of combined RMSE for x, y and z) than the 3DVar-Mix. On the right plot, we see the opposite, 3DVar-Mix outperforms 3DVar-G. Comparing both plots, the shorter skewness window length is more accurate than having a longer window length. This could be due to the skewness being accurate for the current observation, but as the skewness window length increases, more information from different parts of the attractor will be added to the distribution calculations.

By increasing the observation period to 28 time steps, it can be seen in Figure 2 how the methods work in a more nonlinear setting. On the left plot, with a skewness window length of 9 points, 3DVar-ML outperforms both other methods, this result is also the case in the right plot where the skewness window length is 29 points.

By comparing Figures 1 and 2, the common result is that 3DVar-ML outperforms both 3DVar-G and 3DVar-Mix. It is also seen that as we increase the observation period and skewness window length, the RMSE increases. The observation period correlation to increase RMSE values is due to the greater nonlinearity of the problem. As the data assimilation problem becomes more nonlinear, finding the minimum of the cost function becomes a more challenging problem (Goodliff et al., 2015).

In Figure 3 we compare the RMSE of each method at different window lengths. As the observation period increases, the RMSE increases. This is expected in data assimilation due to nonlinear problems being harder

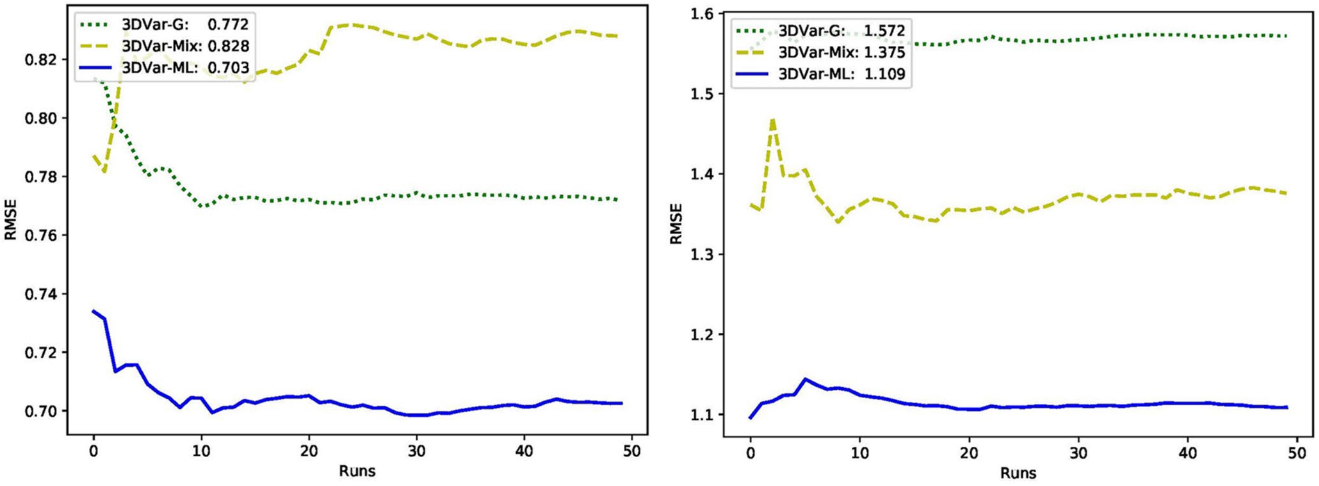

Figure 1. Plot comparing 3DVar-G (green), 3DVar-Mix (yellow) and 3DVar-ML (blue) with an observation period of four time steps, with skewness window lengths of 9 (left) and 29 (right) points. X-axis shows number of runs (each run has 5,000 observations) and the *y*-axis is RMSE, with a rounded cumulative RMSE for each method in the legend.

to solve for the data assimilation methods. Again, from the results shown in Figure 3, the common result for all window lengths is that 3DVar-ML outperforms 3DVar-G and 3DVar-Mix at all observation periods.

Figure 4 shows the percent improvement in RMSE comparing 3Dvar-G and 3DVar-ML over the all ranges set above, the improvement is higher in a more linear setting, with higher skewness window radii. A larger skewness window length improves the RMSE more than a shorter skewness window length. This could be due to the larger windows having more data, so that it is better able to describe the skewness statistic. Another reason to explain this is due to observations being generated from skewness data. With a short skewness window length, there is a higher probability of Gaussian observations (Goodliff et al., 2020). As such, with Gaussian observations, 3DVar-G and 3DVar-ML are equivalent. As the skewness window length increases, less Gaussian observation are generated, and so the 3DVar-ML outperforms 3DVar-G. As the observation period effects the Gaussianity of the background state distributions, and the skewness window length leads to a more (or less) Gaussian likelihood, changing these parameters determines which method should be chosen for optimal results.

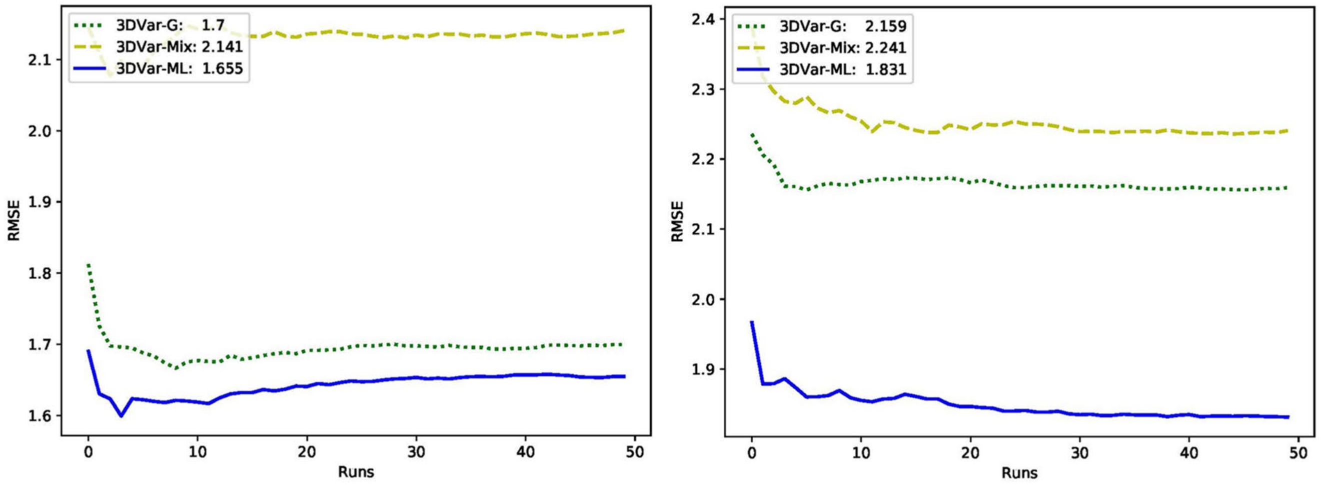

Figure 2. Plot comparing 3DVar-G (green), 3DVar-Mix (yellow) and 3DVar-ML (blue) with an observation period of 28 time steps, with skewness window lengths of 9 (left) and 29 (right) points. X-axis shows number of runs (each run has 5,000 observations) and the *y*-axis is RMSE, with a rounded cumulative RMSE for each method in the legend.

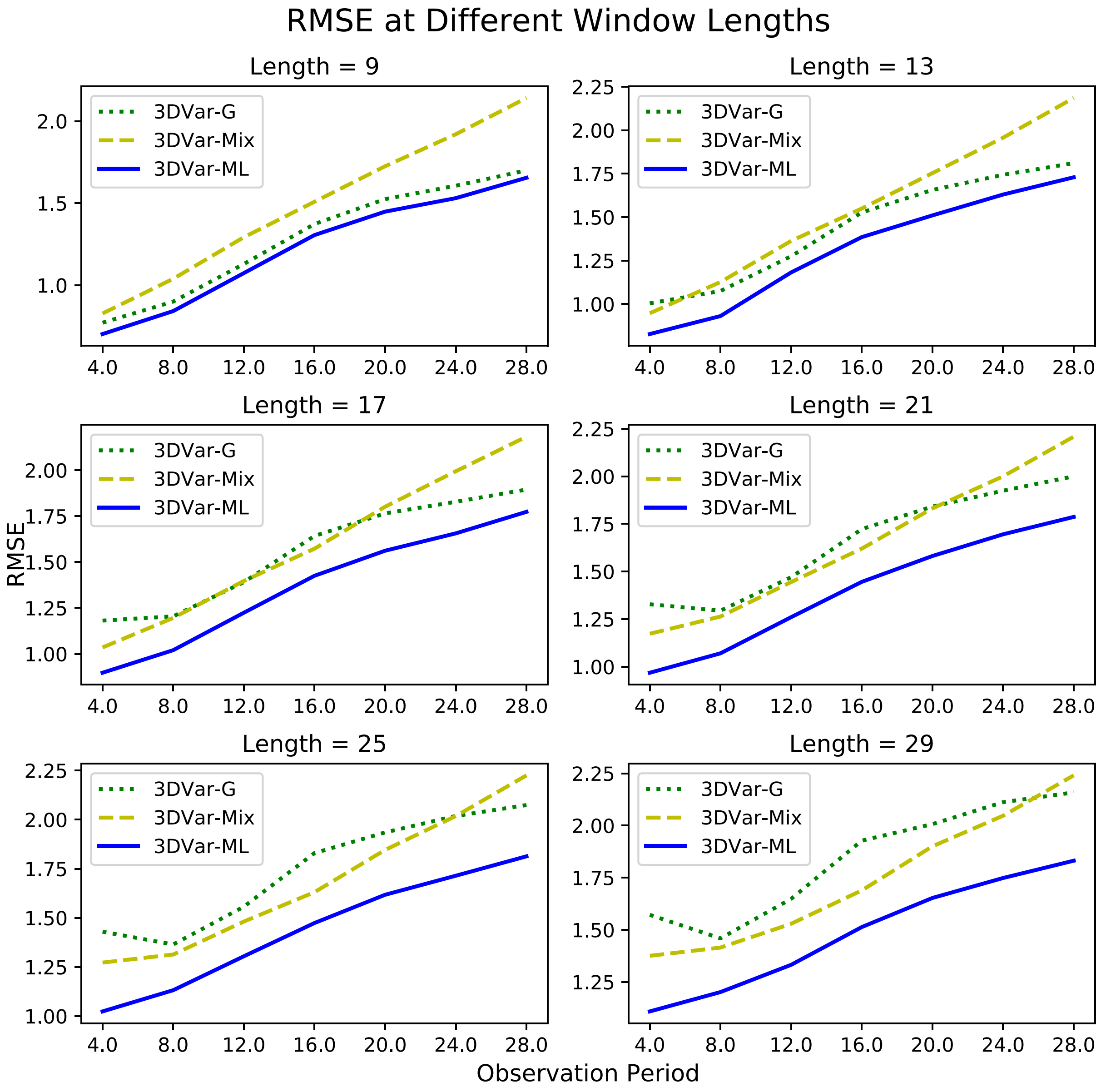

Figure 3. RMSE (Y-axis) of all methods with different skewness window lengths, as a function of different observation periods (X-axis), over 50 runs.

5. Conclusion

In this paper, we used a machine learning technique to improve the performance of 3DVar when there is a change in the underlying distribution for the background error distribution from Gaussian to lognormal and back to Gaussian again. This improvement was achieved through using a support vector machine to detect and predict non-Gaussian distributions on the z component of the Lorenz 63 model. This model was used due to its simplicity, and ability to exhibit chaotic behavior similar to that seen in the atmosphere. To determine the improvement through using the support vector machine approach three data assimilation methods were compared: a Gaussian-fits-all 3DVar, referred to as 3DVar-G, a mixed Gaussian-lognormal variant, which was referred to as 3DVar-Mix, and a version which used a support vector machine to switch between Gaussian and lognormal variants for the z component, where this formulation was referred to as 3DVar-ML.

The support vector machine approach showed promising results when used in conjunction with 3DVar. It has been shown before that certain areas of the Lorenz 63 attractor do better with a lognormal variant of 3DVar, Fletcher and Zupanski (2007), due to those areas being lognormally distributed. Here, we have shown that assimilating

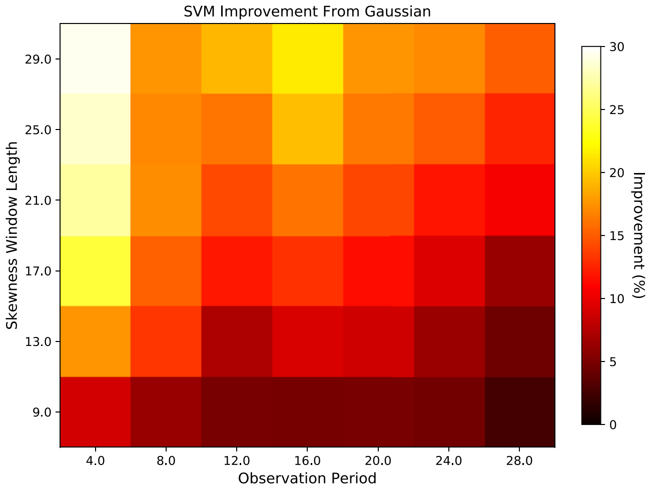

Figure 4. Skewness Window Length by Observation Period. This graph shows the improvement in RMSE from 3DVar-G to 3DVar-ML.

certain areas of the attractor, depending on their probability density function (either Gaussian or lognormal distributions), can show improvements with respect to the analysis RMSE.

For real world applications, applying the support vector machine learning method to choose different data assimilation types could be a way to relax the Gaussian assumption for the background and observational error distributions. These results could then imply that the most optimal assimilation method could be changing dynamically in time, to be consistent with the more physical behavior of the errors. By having this flexibility, we hypothesize that it may improve the forecast for non-Gaussian variables, as shown in Kliewer et al. (2016) for a temperature-mixing-ratio retrieval system, as well as in operational numerical weather prediction in the prediction of humidity and possible certain hydrometeors. Outside of the atmospheric sciences, areas that use the Gaussian assumption for data assimilation in non-Gaussian systems such as space weather (example: solar winds, Lang et al. (2017)) and ocean dynamics (e.g., ocean-biogeochemistry assimilation Goodliff et al. (2019)) could also benefit through changing the underlying cost function in their data assimilation systems. Implementation for this method into a high-dimensional geophysical application would be computationally low cost (except for training, which is usually done once, and offline). The switch would act in real time, giving the predicted optimal version of 4DVar for each variable. In future work, we shall apply non-Gaussian detection to augment data assimilation in numerical weather prediction models to evaluate sensitivities with respect to different training data choices as well as improvements in the corresponding analyses and forecasts.

Data Availability Statement

The data presented in this manuscript are available from: Goodliff (2021).

Acknowledgments

This work is supported by the National Science Foundation grant AGS-1738206 at Cooperative Institute for Research in the Atmosphere (CIRA)/Colorado State University (CSU).

References

Amezcua, J., & Leeuwen, P. J. V. (2014). Gaussian anamorphosis in the analysis step of the EnKF: A joint state-variable/observation approach. *Tellus A: Dynamic Meteorology and Oceanography*, *66*(1), 23493. https://doi.org/10.3402/tellusa.v66.23493

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

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. https://doi.org/10.1098/rsta.2020.0086

D'agostino, R. B., Belanger, A., & D'agostino, R. B., Jr (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, *44*(4), 316–321. https://doi.org/10.1080/00031305.1990.10475751

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on Machine Learning. *Geoscientific Model Development*, *11*(10), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018

Evensen, G., & Van Leeuwen, P. (1996). Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Monthly Weather Review*, *124*, 85–96. https://doi.org/10.1175/1520-0493(1996)124<0085:aogadf>2.0.co;2

Fletcher, S. J. (2010). Mixed Gaussian-lognormal four-dimensional data assimilation. *Tellus Series A-Dynamic Meteorology and Oceanography*, *62*(3), 266–287. https://doi.org/10.1111/j.1600-0870.2010.00439.x

Fletcher, S. J. (2017). *Data assimilation for the geosciences: From theory to application*. Elsevier.

Fletcher, S. J., & Jones, A. S. (2014). Multiplicative and additive incremental variational data assimilation for mixed lognormal–Gaussian errors. *Monthly Weather Review*, *142*(7), 2521–2544. https://doi.org/10.1175/MWR-D-13-00136.1

Fletcher, S. J., & Zupanski, M. (2006a). A data assimilation method for log-normally distributed observational errors. *The Quarterly Journal of the Royal Meteorological Society*, *132*, 2505–2519. https://doi.org/10.1256/qj.05.222

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

Fletcher, S. J., & Zupanski, M. (2007). Implications and impacts of transforming lognormal variables into normal variables in VAR, *Meteorologische Zeitschrift*, *16*, 755–765. https://doi.org/10.1127/0941-2948/2007/0243

Goodliff, M. (2021). Data sets associated with "Non-Gaussian Detection using Machine Learning with Data Assimilation Applications" [Data set]. University of Colorado Boulder. https://doi.org/10.25810/Y7RC-0571

Goodliff, M. R., Amezcua, J., & Van Leeuwen, P. J. (2015). Comparing hybrid data assimilation methods on the Lorenz 1963 model with increasing non-linearity. *Tellus Series A-Dynamic Meteorology and Oceanography*, *67*(1), 26928. https://doi.org/10.3402/tellusa.v67.26928

Goodliff, M. R., Bruening, T., Schwichtenberg, F., Li, X., Lindental, A., & Nerger, L. (2019). Temperature assimilation into a coastal ocean-biogeochemical model: Assessment of weakly and strongly coupled data assimilation. *Ocean Dynamics*, *69*, 1217–1237. https://doi.org/10.1007/s10236-019-01299-7

Goodliff, M. R., Fletcher, S. J., Kliewer, A. J., Forsythe, J. M., & Jones, A. S. (2020). Detection of non-Gaussian behavior using Machine Learning techniques: A case study on the Lorenz 63 model. *Journal of Geophysical Research: Atmospheres*, *125*(2), e2019JD031551. https://doi.org/10.1029/2019JD031551

Kliewer, A. J., Fletcher, S. J., Jones, A. S., & Forsythe, J. M. (2016). Comparison of Gaussian, logarithmic transform and mixed Gaussian-log-normal distribution based 1 DVAR microwave temperature-water-vapour mixing ratio retrievals. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 274–286. https://doi.org/10.1002/qj.2651

Lang, M., Browne, P., van Leeuwen, P. J., & Owens, M. (2017). Data assimilation in the solar wind: Challenges and first results. *Space Weather*, *15*(11), 1490–1510. https://doi.org/10.1002/2017SW001681

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of The Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2

Nello, C., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning method*. Cambridge University Press.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., & Yorke, J. A. (2004). A local ensemble transform Kalman filter for atmospheric data assimilation. *Tellus*, *56A*, 415–428. https://doi.org/10.3402/tellusa.v56i5.14462

Pasini, A. (2008). External forcings and predictability in Lorenz model: An analysis via neural network modelling. *Nuovo Cimento Della Societa Italiana di Fisica C-Colloquia on Physics*, *31*(3), 357–370. https://doi.org/10.1393/ncc/i2009-10312-1

Pasini, A., & Pelino, V. (2005). Can we estimate atmospheric predictability by performance of neural network forecasting? The toy case studies of unforced and forced Lorenz models. In *IEEE international conference on computational intelligence for measurement systems and applications* (pp. 69–74).

Penny, S. G., Smith, T. A., Chen, T.-C., Platt, J. A., Lin, H.-Y., Goodliff, M., & Abarbanel, H. D. I. (2021). *Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation*.

Perron, M., & Sura, P. (2013). Climatology of non-Gaussian atmospheric statistics. *Journal of Climate*, *26*(3), 1063–1083. https://doi.org/10.1175/JCLI-D-11-00504.1

Poterjoy, J. (2016). A localized particle filter for high-dimensional nonlinear systems. *Monthly Weather Review*, *144*, 59–79. https://doi.org/10.1175/mwr-d-15-0163.1

Poterjoy, J., & Anderson, J. L. (2016). Efficient assimilation of simulated observations in a high-dimensional geophysical system using a localized particle filter. *Monthly Weather Review*, *144*, 2007–2020. https://doi.org/10.1175/mwr-d-15-0322.1

Poterjoy, J., Sobash, A., & Anderson, J. L. (2017). Convective-scale data assimilation for the weather research and forecasting model using the local particle filter. *Monthly Weather Review*, *145*, 1897–1918. https://doi.org/10.1175/mwr-d-16-0298.1

Rasp, S., & Lerch, S. (2018). Neural networks for post processing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*(22), 12616–12622. https://doi.org/10.1029/2018GL080704

Scher, S., & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2830–2841. https://doi.org/10.1002/qj.3410

Scher, S., & Messori, G. (2019). Weather and climate forecasting with neural networks: Using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, *12*(7), 2797–2809. https://doi.org/10.5194/gmd-12-2797-2019

van Leeuwen, P. J., Künsch, H. R., Nesrger, L., Potthast, R., & Reich, S. (2019). Particle filters for high-dimensional geoscience applications: A review. *Quarterly Journal of the Royal Meteorological Society*, *145*(723), 2335–2365. https://doi.org/10.1002/qj.3551

Wang, X., & Bishop, C. H. (2003). A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *Journal of the Atmospheric Sciences*, *60*, 1140–1158. https://doi.org/10.1175/1520-0469(2003)060<1140:acobae>2.0.co;2

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500 hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. https://doi.org/10.1029/2019MS001705

Zupanski, M. (2005). Maximum likelihood ensemble filter. Part I: Theoretical aspects. *Monthly Weather Review*, *133*, 1710–1726. https://doi.org/10.1175/mwr2946.1