A Machine Learning Augmented Data Assimilation Method for High-Resolution Observation

Lucas Howard[1], Aneesh Subramanian[1], and Ibrahim Hoteit[2]

[1]University of Colorado, Boulder
[2]King Abdullah University of Science and Technology

April 20, 2023

Abstract

The accuracy of initial conditions is an important driver of the forecast skill of numerical weather prediction models. Increases in the quantity of available measurements, in particular high-resolution remote sensing observational data products from satellites, are valuable inputs for improving those initial condition estimates. However, the data assimilation methods used for integrating observations into forecast models are computationally expensive. This makes incorporating dense observations into operational forecast systems challenging, and it is often prohibitively time-consuming. As a result, large quantities of data are discarded and not used for state initialization. We demonstrate, using the Lorenz-96 system for testing, that a simple machine learning method can be trained to assimilate high-resolution data. Using it to do so improves both initial conditions and forecast accuracy. Compared to using the Ensemble Kalman Filter with high-resolution observations ignored, our augmented method has an average root-mean-squared error reduced by 15%. Ensemble forecasts using initial conditions generated by the augmented method are more accurate and reliable at up to 10 days of forecast lead time.

A Machine Learning Augmented Data Assimilation Method for High-Resolution Observations

Lucas J. Howard[1], Aneesh Subramanian[1], Ibrahim Hoteit[2]

[1]University of Colorado, Boulder. Department of Atmospheric and Oceanic Science.
[2]King Abdullah University of Science and Technology

Key Points:

- Machine learning augmented data assimilation of high-resolution observations improves the analysis in a nonlinear dynamical model.
- Explainable Artificial Intelligence identifies system covariances to guide neural network training for analysis state reproduction.
- Short-term forecasts from the analysis generated by the machine learning augmented data assimilation are more accurate and more reliable.

Corresponding author: Lucas J. Howard, `Lucas.Howard@colorado.edu`

Abstract

The accuracy of initial conditions is an important driver of the forecast skill of numerical weather prediction models. Increases in the quantity of available measurements, in particular high-resolution remote sensing observational data products from satellites, are valuable inputs for improving those initial condition estimates. However, the data assimilation methods used for integrating observations into forecast models are computationally expensive. This makes incorporating dense observations into operational forecast systems challenging, and it is often prohibitively time-consuming. As a result, large quantities of data are discarded and not used for state initialization. We demonstrate, using the Lorenz-96 system for testing, that a simple machine learning method can be trained to assimilate high-resolution data. Using it to do so improves both initial conditions and forecast accuracy. Compared to using the Ensemble Kalman Filter with high-resolution observations ignored, our augmented method has an average root-mean-squared error reduced by 15%. Ensemble forecasts using initial conditions generated by the augmented method are more accurate and reliable at up to 10 days of forecast lead time.

Plain Language Summary

Weather forecasts are highly sensitive to the estimate of the current state of the atmosphere, known as initial conditions. The atmosphere is chaotic, meaning that small errors in this estimate can grow quickly as the forecast model predicts events further into the future. The satellite era has contributed to large improvements in weather forecasts by providing additional data that allow for more accurate estimates of initial conditions. However, current methods for generating initial conditions are computationally time-consuming, and as a result, large fractions of available measurements are not used for this purpose. In a proof-of-concept study using a simplified representation of the atmosphere for testing, we train a machine learning method to replicate the results of a traditional method. Once trained, machine learning models are usually very fast. Applying the trained model exclusively to measurements that would otherwise be too time-consuming to use produces better initial conditions and more accurate forecasts.

1 Introduction

The accuracy of operational weather forecast models is highly dependent on the quality of the initial conditions provided to the model (Bauer et al., 2015). To correct drift and maintain the robustness of forecasts, model initial conditions are regularly updated based on measurements (Bannister, 2017; Edwards et al., 2015). These measurements include both in-situ data, such as weather station measurements, and remotely sensed data (Zhang & Tian, 2021; Choi et al., 2017). Observations are noisy and may not be aligned with the model grid or state variables; the task of identifying optimal initial conditions consistent with all available information is therefore challenging and computationally expensive. Data assimilation (DA) methods are employed to do this. With the proliferation of high-resolution datasets, often at resolutions higher than that of the forecast models, otherwise useful data is regularly ignored and not assimilated into operational models due to time or computational constraints (Eyre et al., 2022; Kumar et al., 2022). Assimilation of a subset of available satellite data has improved forecasts, making it likely that leveraging currently unused data could generate further improvements (Eyre et al., 2022).

DA techniques can be generally categorized as variational methods or sequential methods (Bannister, 2017; Edwards et al., 2015). Variational methods use numerical optimization, finding the initial condition that minimizes an error metric or cost function. Sequential methods nowadays are some variant of the Ensemble Kalman Filter (EnKF), in which a set of model realizations are simulated to quantify covariance structures before assimilating observations (Evensen, 2003; Hoteit et al., 2018). Both methods require running multiple simulations of the full forecast model, a step that is computationally expensive and time-consuming. To capture the information of a high-resolution measurement, the model itself must at least match the resolution of the measurement – further increasing the cost of this step as model run time scales with resolution. In addition to the necessity of using a higher resolution model (run multiple times), the physics of smaller scale dynamics create more complex correlation structures, and a larger ensemble size is required to actually improve the forecast (Miyoshi et al., 2015).

An additional cost associated with assimilating high-resolution observations is the observation operator. Both variational and sequential methods assess the error of a model forecast for a given initial condition in observation space. This requires, for each observation, explicitly mapping between forecast model space and observation space. For some observations, this is straightforward. For others, particularly for remotely sensed data such as high-resolution satellite measurements, this calculation itself is a physics-based model that can be a computational bottleneck (Eyre et al., 2022). In some operational models, the trade-off between the speed and accuracy of these observation operators is already an important avenue of research for improving the performance of their DA systems even before currently unusable high-resolution data is considered (Shahabadi & Buehner, 2021). The observation operator calculation must be performed for each data point and so also scales with the number of discrete observations, again increasing its cost.

When possible, assimilation of these remotely sensed observations can and has improved forecasts, especially since in-situ observations of large portions of the atmosphere and surface are sparse (Bannister, 2017). Efforts to incorporate satellite and other remotely sensed observations into assimilation systems have been effective at improving model initialization and forecast accuracy (Shahabadi & Buehner, 2021). However, as a result of the expense associated with assimilation much of the potential of these high-resolution measurements for improving state initialization in forecast models has not been realized. Currently employed DA methods are simply not efficient enough to sufficiently quickly ingest this data to be useful in an operational setting. Machine Learning (ML) methods may provide a potential solution.

ML techniques have been increasingly used in earth science applications, including DA (Sonnewald et al., 2021; Abarbanel et al., 2018; Bonavita et al., 2021; Penny et al., 2022). They are appealing for this particular problem mostly due to their speed. While the training process is often expensive, once trained ML methods are very fast compared to weather forecast models. Since many of the bottlenecks in traditional DA methods are related to computational efficiency, as described above, much of the effort in employing ML to improve DA has been targeted at the most computationally expensive parts of the process.

One obvious place to look is at the model simulations themselves. Attempts to use deep learning, in which neural networks comprised of many layers are used to capture complex structures, have proven successful. The basic approach is to train the ML model on the output of a traditional physics-based model (Kim et al., 2019; Pathak et al., 2017). The result is more computationally efficient but at the cost of accuracy. In the context of DA, it has been demonstrated that model surrogates can be successfully trained iteratively using DA state estimates (Brajard et al., 2020; Gottwald & Reich, 2021).

Extending this approach beyond demonstrating that ML can capture the dynamics of complex and chaotic systems, augmented approaches that use model surrogates only to represent scales unresolved by the physical model (Brajard et al., 2021) have shown that the integration of ML as a model surrogate can generate improvements over traditional DA methods. Other work has demonstrated the utility of using ML model surrogates to increase the ensemble size beyond what would be otherwise practical (Yang & Grooms, 2021; Wu et al., 2021). Related to the issue of unresolved scale and model resolution, ML has been employed to successfully generate parameterizations used to capture unresolved physics, making the generation of the tangent linear models needed in many variational DA methods more efficient (Hatfield et al., 2021).

The observation operator, which can be another computational bottleneck, has also been targeted using ML (Jung et al., 2010; J. Liang et al., 2023; Wang et al., 2022; Geer, 2021; X. Liang et al., 2022; Stegmann et al., 2022). Other efforts have used ML methods to identify regions of tropical cyclone activity to target high-resolution modeling in a subdomain (Lee et al., 2019), to perform bias correction of model forecasts before they are fed into the assimilation algorithm (Arcucci et al., 2021; Chen et al., 2022), and apply time-varying localization to the covariance structure of the system (Lacerda et al., 2021).

In contrast, relatively limited efforts have been directed toward using ML to perform the assimilation step directly. Rather than using ML to replace pieces of the DA process, we propose an augmented DA method in which a ML model is trained offline to assimilate high-resolution measurements. Convolutional neural networks (CNN) are particularly good candidates for assimilating spatial data and learning the spatial correlation structure of the system of interest, as their design and main demonstrated use cases rely on their ability to identify spatial patterns (Dong et al., 2021; Mallat, 2016).

In a real-world scenario, computational resource bottlenecks require some high-resolution observations to be either thinned before being assimilated or discarded entirely. As a proof-of-concept demonstration for our proposed method, we use a synthetic system with observations available at regular time intervals. The observations are alternatively high-resolution or low-resolution. Low-resolution observations are always assimilated using the EnKF, and in the augmented method, high-resolution observations are assimilated using the trained CNN.

This study will explore these two hypotheses:

1. A shallow CNN can be successfully trained to reproduce the analysis of the EnKF offline
2. When used online to assimilate otherwise ignored high-resolution data, with the traditional EnKF used for low-resolution data, assimilation performance will be improved

with the chaotic Lorenz-96 model as the test system. Yet the relevance is broader with applications in weather and climate predictions. Section 2 describes the Lorenz-96 modeling system, the machine learning augmentation of EnKF framework and the explainable AI methodology. Section 3 presents the results from the experiments and analyses performed. Section 4 summarizes the results and discussion and section 5 concludes.

2 Methods

2.1 Lorenz-96 System

The Lorenz-96 system is described by a set of N discrete differential equations, designed to mimic some behaviors of the atmosphere (Lorenz, 2005). It is commonly used for testing data assimilation methods. It is defined as a 1-D analog of an atmospheric state variable at discrete points evenly spaced in the zonal direction, with its dynamics governed by:

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-2} - x_i + F \tag{1}$$

for $i \in$[1,N] and F a constant forcing term. The system is cyclically symmetric with gird point $i = N + 1$ equal to grid point $i = 1$. We use $F = 8$, a value for which the system is known to be chaotic, and $N = 40$, a typical value for testing DA methods.

To generate a reference trajectory for our experiments, we numerically integrated Equation 1 forward. A 5^{th} order Runge-Kutta method was used, with a variable time step to control error assuming 4^{th}-order accuracy (Dormand & Prince, 1980), as implemented in the SciPy package (Virtanen et al., 2020). The maximum allowable relative error was set to 0.001 and the maximum allowable absolute error to 10^{-6}. The system is in an unstable equilibrium when all variables are equal to F; initial conditions were set by perturbing one of the variables to a value of 8.01. The system was integrated out to $t = 2000$ and output was generated at time intervals of $\Delta t = 0.05$ generating data for a 40 variable vector at 40,000 time steps, or 800,000 data points representing the true time evolution of the system. Synthetic observations were generated by adding normally distributed random noise with a standard deviation equal to 30% of the standard deviation of the reference state. This level of observation noise is consistent with other work done using the Lorenz-96 system for testing DA methods. (Hatfield et al., 2018; Brajard et al., 2020; Hoteit et al., 2008).

2.2 The Ensemble Kalman Filter

Data assimilation is used to combine model forecasts and observations and solves the filtering problem. Formally, the filtering problem is to generate a minimum-variance estimate of a state vector, $\vec{x}$, conditional on a noisy forecast and a noisy observation. The state vector evolves in time with model dynamics represented by a forward operator M. The time evolution of the system is defined iteratively; the system states at times t_i and t_{i+1}, $\vec{x}_i$ and $\vec{x}_{i+1}$ are related via:

$$M(\vec{x}_i) = \vec{x}_{i+1} + \vec{\mu} \tag{2}$$

where μ is the assumed model forecast error. Observations y_i at time t_i are related to the state vector via an observation operator, H:

$$y_i = H(\vec{x}_i) + \vec{\nu} \tag{3}$$

where ν is the assumed observation error.

The solution to the filtering problem is referred to as the analysis. When forecast and observation errors are unbiased, normally distributed, and independent, and the forecast model and observation operator are both linear, the Kalman filter (KF) provides the closed-form optimal solution of the filtering problem(Kalman, 1960).

In earth system applications, the system's time evolution and thus the forecast models are often non-linear. The Ensemble Kalman filter (EnKF) is an extension of the KF that accommodates nonlinear models at the cost of being an approximate, rather than exact, solution to the filtering problem by using an ensemble of model forecasts (Evensen, 2003). The EnKF analysis equation is:

$$X^a = X^f + CH^T(HCH^T + R)^{-1}(Y - HX^f). \tag{4}$$

Here, X^a is a matrix whose column vectors are analysis ensemble members. X^f is a matrix whose column vectors are individual forecasts. C is the sample covariance of the ensemble forecast, X^f, used as an approximate representation of the true covariance/ R is the observation error covariance matrix, and Y is a matrix whose columns are the observation vector y_i. To ensure that the covariance does not systematically underrepresent the true error, random Gaussian noise with covariance R is added to the observation matrix Y (Evensen, 2003).

The EnKF assumes normality for the forecast and observation errors, $\vec{\mu}$ and $\vec{\nu}$, although has been demonstrated to be somewhat robust to non-Gaussian distributions (Reichle et al., 2002). Also relevant for earth system models in which the state space is very large, the EnKF can be effective even when the number of ensemble members is much smaller than the size of the state space. This is in contrast to more exact methods such as particle filters, which

Run Name	Observation Error StDev (fraction)	Ensemble Size	Inflation Factor	Localization Distance
Base	0.3	100	1	5
s1	0.4	100	1	5
s2	0.2	100	1	5
s3	0.3	33	1	5
s4	0.3	1000	1	5
s5	0.3	100	1.01	5
s6	0.3	100	1.05	5
s7	0.3	100	1.1	5
s8	0.3	100	1	3
s9	0.3	100	1	7

Table 1. Sensitivity settings for the EnKF runs. Observation error is presented as observation noise standard deviation as a fraction of total system standard deviation.

often exhibit stability problems in such situations (Farchi & Bocquet, 2018; Hoteit et al., 2008).

The EnKF is known to be vulnerable to issues associated with the fact that it approximates a PDF with samples represented by a finite number of ensemble members. These issues include spurious correlations as well as a covariance collapse, in which the ensemble becomes sharply clustered at a point in state space far away from the true state. To address this localization is often used, a technique that has been shown to improve performance by reducing the impact of spurious correlations of the system state at distant grid points (Evensen, 2003). For this application we used a step function to localize the covariance, setting any covariance between variables greater than five grid points apart equal to zero.

Covariance inflation, in which all covariance values are multiplied by a factor greater than 1 before computing equation 4, is another technique used for improving the stability and performance of the EnKF. Our base settings did not include covariance inflation as our initial experiments did not show significant improvements employing it. Both approaches can improve performance in some circumstances by preventing covariance collapse (Evensen, 2003). However, since both address issues created by the finite size of the ensemble, they become less necessary with larger ensemble sizes and must be tuned (Miyoshi et al., 2015). As such they are appropriate parameters to vary in our sensitivity analysis in order to identify optimal values.

We use the EnKF here for two purposes: to generate training data for a CNN and as a benchmark to evaluate the performance of our augmented method. After assimilating the synthetic observations with the EnKF using settings described above, at all 40,000 time steps, the following data are available: the true model state, the model forecast, the synthetic observation, and the EnKF analysis. Other combinations of settings were also used to assess sensitivity. These are outlined in Table 1. Observation error is specified as the standard deviation of the added noise used to generate the synthetic observations, as a fraction of the system standard deviation.

2.3 CNN Architecture and Training

The machine learning model consists of a convolutional neural network with two hidden layers. Its architecture is shown in Figure 1. The input layer has two channels: the model ensemble forecast mean and the difference between the forecast mean and the observation (the innovation). A CNN is defined by the following parameters for each layer: the filter size, the number of feature maps, and the activation function (Alzubaidi et al., 2021). We use a filter size of 3 for all convolutions. The weights of each convolutional filter are independent of space and are applied uniformly across the domain. 5 feature maps are used in both hidden layers, with each feature map assigned different filter weights and a constant bias weight. The ReLu activation function is used for both hidden layers and no activation function is applied to the output, which is then a linear weighted sum of the activation values in the second hidden layer.

Output from a convolution has a smaller dimension than its input, since locations on the edge of the domain don't have a neighboring point on one side. In image recognition and other similar tasks, zero-padding is used to address this issue by artificially adding zeros to the input on the edges of the domain. Here, with a cyclically symmetric system, zero-padding is not an appropriate solution. Instead, we implemented cyclic padding such that the neighboring spatial nodes for $i = 1$ are $i = N$ and $i = 2$. Similarly, the neighboring spatial nodes for $i = N$ are $i = 1$ and $i = N - 1$. Applying three convolutions with a filter size of three will reduce the domain size by 6. The data from spatial locations $i = [1, 3]$ were concatenated to the end of the spatial domain, and the data from spatial locations $i = [N - 2, N]$ were concatenated to the beginning of the spatial domain. This maintains the cyclic nature of the Lorenz-96 system and ensures that the size of the CNN output matches the dimensions of the system. The size of the input to the neural network is 46x2. Its output, the analysis, is 40x1. The model has 131 trainable weights.

The training data is comprised of the first half of the EnKF analysis states, for times $t = 0$ to 1,000 covering 20,000 individual time steps. The dataset from the best-performing EnKF sensitivity run settings described in Table 1 was used for training. A stochastic gradient descent optimizer (Sutskever et al., 2013) was used to train the model, using 20 training epochs and 100 batches per epoch.

2.4 Augmented Method and Experimental Setup

The augmented method is designed to be applicable to a scenario in which high-resolution observations are available but not assimilated (Figure 2). To create an analog of this scenario with the Lorenz-96 system, we created a set of low-resolution observations at every other time step ($\Delta t = 0.1$) for the second half of the time series ($t = 1000.05$ to $t = 2000$) by randomly selecting 25% of the variables to measure. The EnKF using base settings was then used to assimilate these observations. This run is the baseline against which the augmented method will be compared. This method will subsequently be referred to as "EnKF SparseObs", with the method used to generate the training data in which 100% of variables were measured at every time step referred to as "EnKF AllObs".

The augmented method uses the EnKF to assimilate low-resolution observations. On alternating time steps, a "high-resolution" observation is available that includes observations of 100% of the variables. For EnKF SparseObs these observations are assumed to be prohibitively computationally expensive to assimilate and are therefore ignored. The forecast continues on to the next time step where a low-resolution observation is available and assimilated by the EnKF. In the augmented method, the CNN is used to assimilate the high-resolution observations that cannot be assimilated by the EnKF.

The CNN takes the ensemble forecast mean and the observation as input and returns a single analysis as output. At this stage, an ensemble must be re-created to generate an ensemble forecast for the next time step (where the EnKF will be applied to the low-resolution observation). To be consistent with the analysis generated by the CNN, the new ensemble mean must be equal to the vector analysis produced by the CNN. We create such an ensemble by computing the vector difference between the CNN analysis value and the ensemble forecast mean, $\vec{\delta} = \vec{x}^{f}_{mean} - \vec{x}^{a}_{cnn}$. To generate the new initial ensemble, $\vec{\delta}$ is subtracted from each member of the ensemble forecast. The new ensemble mean then by definition is equal to the vector analysis generated by the CNN with the same spread as the ensemble forecast.

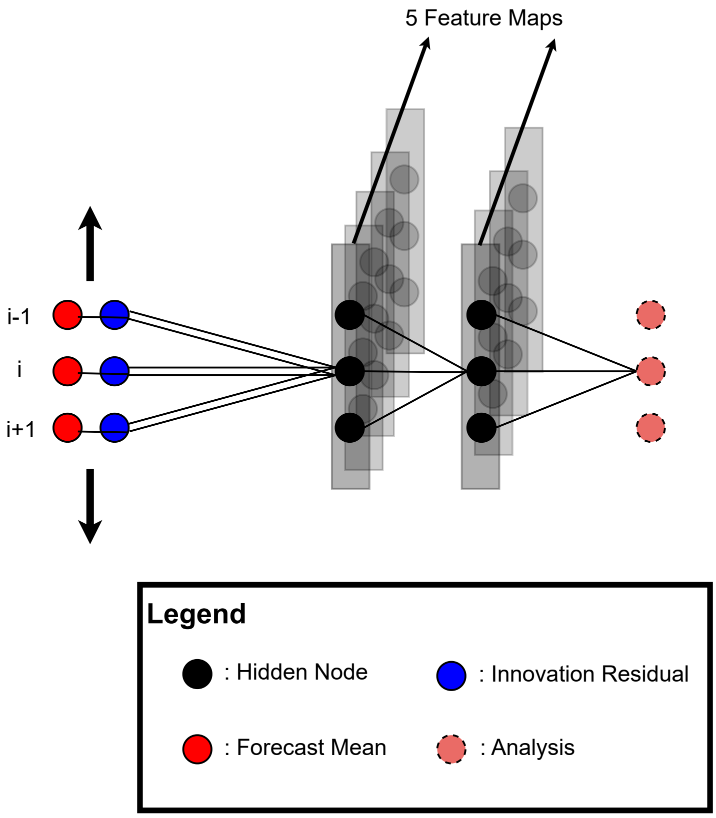

Figure 1. Architecture of the CNN trained to emulate the EnKF analysis step with observations of all variables. Forecast mean and observations are provided in separate collocated input channels. Two hidden convolutional layers each contain 5 feature maps, with different filter weights. Analysis mean is generated as output.

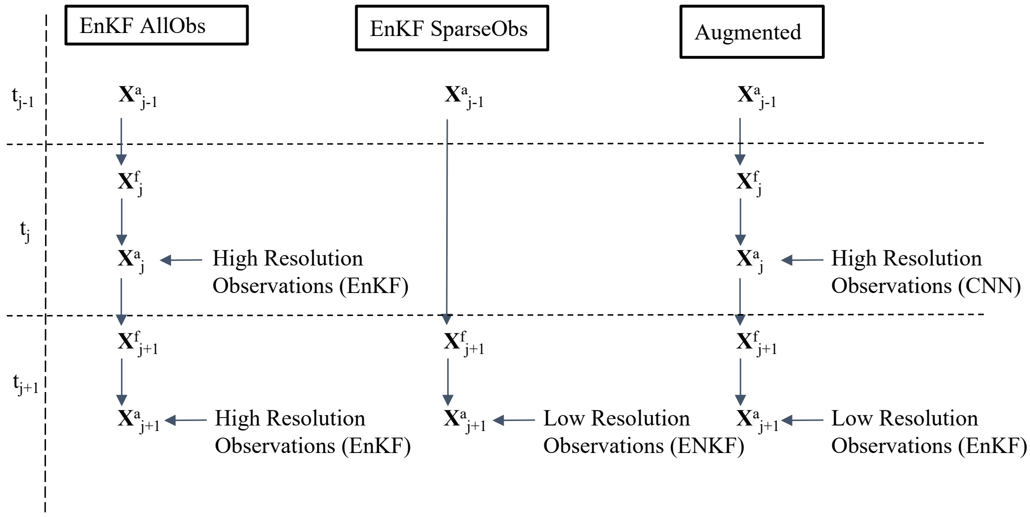

Figure 2. Flow chart of the experimental setup and augmented method. EnKF AllObs is provided with full observations at every time step. EnKF SparseObs is provided with observations 25% of the variables at every other time step. The augmented method is identical to EnKF SparseObs but is additionally provided with observations of 100% of the variables on alternating time steps and uses the trained CNN to assimilate these.

2.5 Explainable AI

Machine learning models are fast to run and accurate when sufficient training data are available. In many earth system science applications, the computational efficiency of traditional tools is a significant bottleneck and available training data is voluminous. These models have a major drawback, however: models are a black box and it is therefore often not clear how they are generating their predictions (Gevaert, 2022). Using testing and validation datasets can provide some level of confidence in the models by demonstrating their level of accuracy on data not used for training. If they are to be deployed in something like an operational weather forecast system, however, such demonstrations may not provide a sufficient level of confidence. Out-of-sample input data cannot be guaranteed never to occur, and no guarantees can be made about the behavior of the machine learning model when presented with such data.

A variety of tools are available to make otherwise opaque data-driven models more transparent, collectively referred to as Explainable AI (XAI) methods (Linardatos et al., 2021). Shapely Additive Explanations, or SHAP, is one such tool. SHAP quantifies the impact of a specific input variable on the output generated by a model. The method is model-agnostic and is equally applicable to a simple linear regression model or a deep neural network with millions of trainable parameters. Full details and a formal definition can be found in Lundberg and Lee (2017). Heuristically, a SHAP value is the anomaly in an output variable attributable to the anomaly in an input variable. It provides a way of apportioning the deviation from the mean in the output to each input variable. This information can increase confidence that the trained model is reliable as well as provide insights into the structure of the underlying system.

We apply it here to analyze how the trained CNN generates analyses from forecasts and innovations. In a DA context, the behavior of the CNN should be predictable and consistent with our understanding of the dynamics of the Lorenz-96 system; it should not, for example, heavily weight forecasts from highly spatially distant nodes. Evaluating the

Run Name	Observation Error (StDev fraction)	Ensemble Size	Inflation Factor	Localization Distance	RMSE (% of Observation StDev)
Base	0.3	100	1	5	20.3059
s1	0.2	100	1	5	20.1084
s2	0.4	100	1	5	20.6152
s3	0.3	33	1	5	29.5916
s4	0.3	1000	1	5	19.6315
s5	0.3	100	1.01	5	20.3089
s6	0.3	100	1.05	5	20.3404
s7	0.3	100	1.1	5	20.4216
s8	0.3	100	1	3	22.6087
s9	0.3	100	1	7	19.4958

Table 2. EnKF-only sensitivity results. RMSE is presented as a fraction of the observation standard deviation to allow for comparison between different observation error settings.

CNN in this way can provide confidence in its ability to perform well when presented with new data.

3 Results

3.1 Sensitivity and Training

The results of the base run and 9 sensitivity runs using the EnKF are outlined in Table 2. These runs are intended to identify optimal settings for generating training data, with the EnKF assimilating all observations at every time step (i.e. the high-resolution observation at every time step). The performance for each run is evaluated as the mean analysis root-mean-squared error (RMSE) divided by the standard deviation of the observation error. For all 10 cases, the EnKF analysis has a lower error than the observation error, as expected, with all runs achieving better than 24% on this metric. As the EnKF approximates the optimal solution by using the first two moments of the forecast ensemble to represent a normal distribution, errors caused by the finite size of the ensemble are expected to decrease with ensemble size. This is evident in our results, which show larger ensemble sizes producing lower errors.

Localization and covariance inflation can improve EnKF performance by mitigating errors related to finite ensemble size but can be detrimental for larger ensemble sizes as such errors become less important. We expect the performance to be dependent on localization and inflation settings but it is not clear a priori which values will be optimal. The best-performing combination of settings was run s9 with localization of 7 grid spaces and covariance inflation factor of 1 (i.e. no inflation). These are the EnKF settings used for training the CNN and used in the augmented method.

The results from the training process are shown in Figure 3. The training targets were the EnKF analyses produced using observations of all variables as described in section 2.3. CNN error can be considered in terms of how well CNN output matches the EnKF analyses as well as its deviation from the true state. The RMSE with respect to these training targets is shown across 20 training iterations. Additionally, the error with respect to true states in the second half of the time series, i.e. the set not used for training, is included for validation. Over-fitting would be indicated by an increase in validation error even as the training error remained flat or decreased. This is not evident here, demonstrating that our trained CNN is not overfitting and produces reliable predictions when presented with data from outside its training set (Ying, 2019).

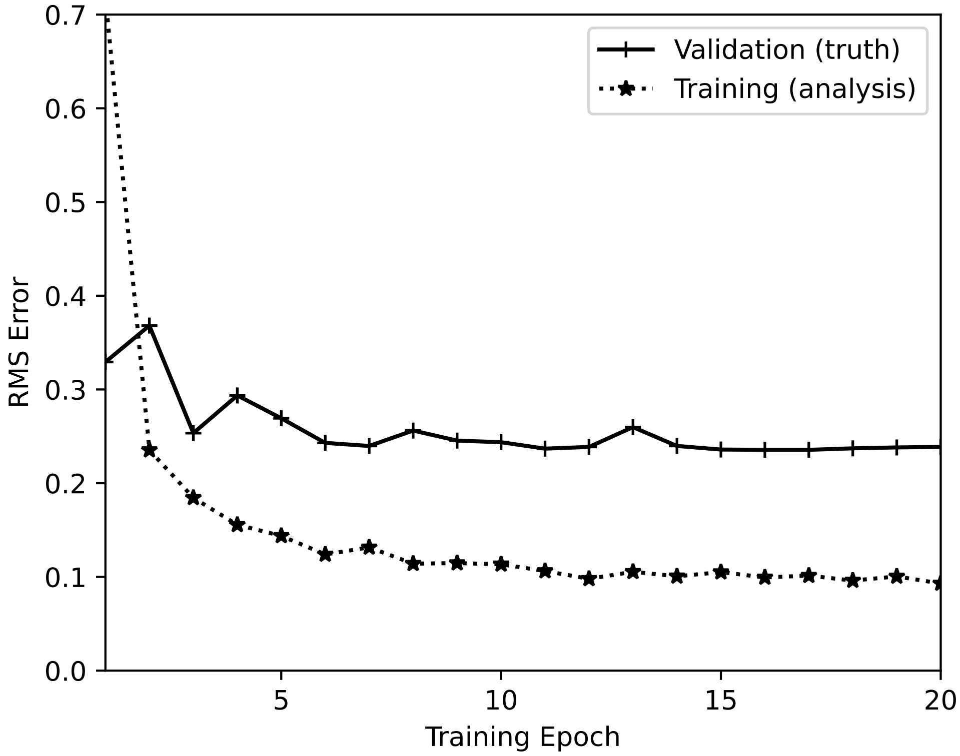

Figure 3. Root mean squared error of the CNN through training epochs, with the error shown based on both training targets (dashed line) and validation error (solid line). Training error is with respect to EnKF analysis, validation error is with respect to truth.

Run Name	Observation Error (StDev fraction)	Ensemble Size	Inflation Factor	Localization Distance	Augmented RMSE	SparseObs RMSE
Base	0.3	100	1	5	0.750	0.877
s1	0.2	100	1	5	n/a	n/a
s2	0.4	100	1	5	n/a	n/a
s3	0.3	33	1	5	0.782	1.453
s4	0.3	1000	1	5	0.7371	0.8243
s5	0.3	100	1.01	5	0.751	0.883
s6	0.3	100	1.05	5	0.754	0.876
s7	0.3	100	1.1	5	0.759	0.882
s8	0.3	100	1	3	0.873	0.989
s9	0.3	100	1	7	0.728	0.899

Table 3. Sensitivity results for the augmented and EnKF SparseObs methods. RMSE for both is presented as a fraction of the observation error standard deviation. For all runs, the augmented method outperforms EnKF SparseObs.

Another check on the trained CNN is to compare its RMSE on the validation dataset to the observation error. To generate an improvement in the state estimate the CNN must perform better than observations alone. Using EnKF AllObs forecasts and residuals, the final validation RMSE with respect to the true state in Figure 3 is 23% of the observation standard deviation.

The sensitivity of the augmented method to different settings was tested. These results are shown in Table 3. For all runs, the CNN trained on results using the s9 run settings was used. The s1 and s2 run settings were not tested; the CNN was trained on data with the observation error specified by s9 settings and the s1 and s2 settings are therefore not applicable for the augmented method. For those cases that are applicable, the augmented method performance was compared to the EnKF assimilating only the low-resolution observations (EnKF SparseObs). For all sensitivity settings, the augmented method outperforms EnKF SparseObs (Table 3).

3.2 Performance Comparison

Having shown that the trained CNN does not over-fit and that its error is 23% of observation error, we can now assess the performance of the augmented method using this CNN. In addition to the augmented method and EnKF SparseObs, the performance of EnKF AllObs assimilating observations of all variables at every time step is included for comparison. The analysis RMSE for all three methods is shown in Figure 4. Time is shown as earth-years with 1 model time unit is equivalent to 5 real days (Lorenz, 2005). EnKF AllObs performs best with an average RMSE of 0.22. Considering only the density and frequency of assimilated observations, this comparative overperformance is unsurprising as EnKF AllObs assimilates more data than the other methods. More interestingly, the augmented method performs better than EnKF SparseObs, with an average RMSE of 0.75 compared to 0.88, representing an improvement of 14.5%.

The other thing to note is the variability of errors between methods. The time series on the left is smoothed, and in this rolling average the augmented method consistently outperforms sparse obs at nearly all time steps. The histogram on the right shows the distribution of errors at all time steps for EnKF AllObs, EnKF SparseObs, and the augmented method. There is substantial overlap in the distribution of errors; using unsmoothed data, the augmented method outperforms in 30% of coincident time steps. The EnKF SparseObs errors have a notably fatter tail in the histogram, however. These spikes are periods where the

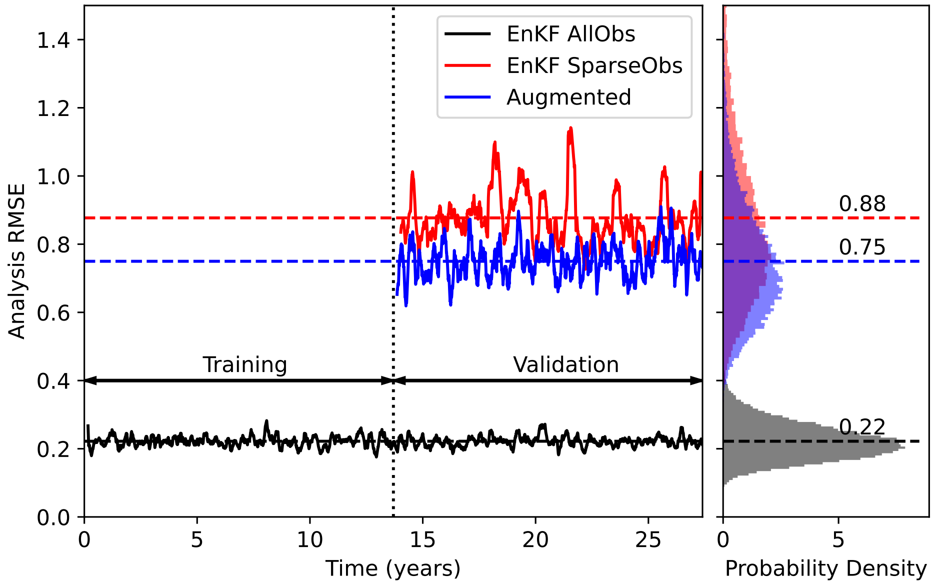

Figure 4. A time series (left) and histogram (right) of the RMSE of the EnKF AllObs, EnKF SparseObs, and augmented methods. The mean error for each method is represented by a horizontal dashed line. The first half of the time series includes only results for EnKF AllObs, which is used for training the CNN. The second half of the time series includes EnKF SparseObs and augmented as well, with both initialized using the last analysis produced by the EnKF AllObs. The time series data is smoothed with a moving window of 60 days for readability. On the right, a histogram of the distribution of RMSE for all three methods is shown using the same axis as the time series plot. This plot uses unsmoothed data and as a result, the tails extend beyond the range of time series traces.

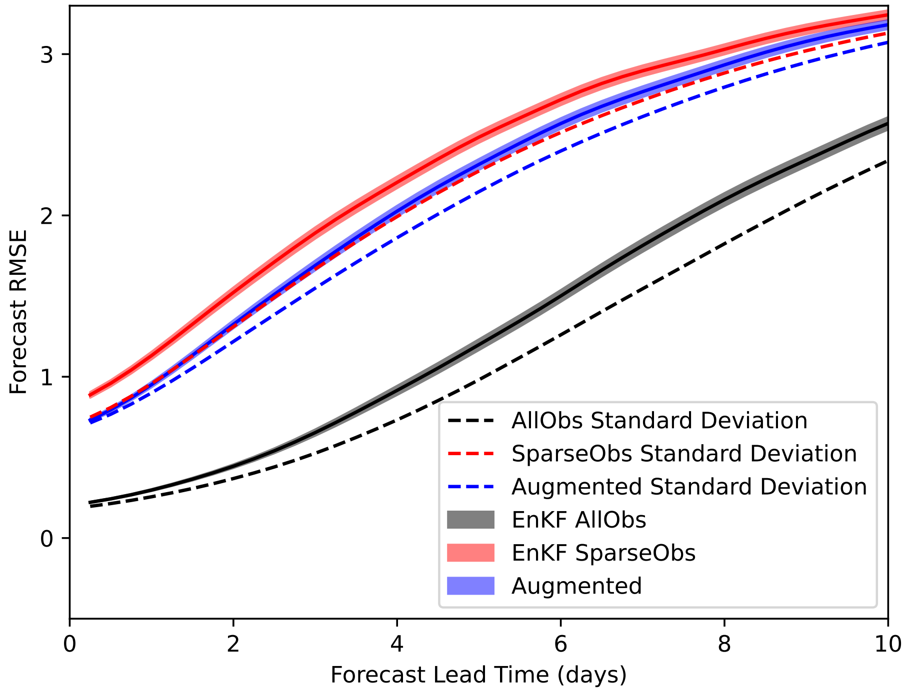

Figure 5. The forecast accuracy out to 10 days using initial conditions produced by the augmented method as well as the EnKF (sparse and all obs). 95% confidence intervals are included for all three methods based on the RMSE standard deviation across 1,000 randomly selected initial conditions. The mean ensemble standard deviations are also shown as dashed lines.

state estimate diverged from reality, generating instabilities in the EnKF that resulted in large errors.

Despite having a similar error distribution to EnKF SparseObs, the augmented method does not have the same fat tail. It is better at maintaining the state estimate in the vicinity of the true state, preventing instabilities and periodic spikes in the analysis error. This accounts for the consistent over-performance in the smoothed time series. The improved stability is an important factor in evaluating the relative performance and suggests that the augmented method is more reliable in excess of what would be otherwise assessed based on the fact that it only outperforms EnKF SparseObs in 30% of time steps. The improved stability and reduced mean RMSE are clear benefits of exploiting all available data in assimilation using an efficient but possibly sub-optimal technique (the CNN) compared to ignoring a subset of observations.

Another way of assessing the performance of the three methods is to generate forecasts using their analyses as initial conditions. Forecast skill over time can then be compared. These results are shown in Figure 5. The mean error of ensemble forecasts from a sample of initial conditions is plotted out to 10 days of lead time. As with the results in Figure 4, EnKF AllObs performs best, generating better forecasts for all lead times. The augmented method again outperforms EnKF SparseObs. Out to 5 days of forecast lead time, the RMSE of forecasts generated using initial conditions from the augmented method is statistically

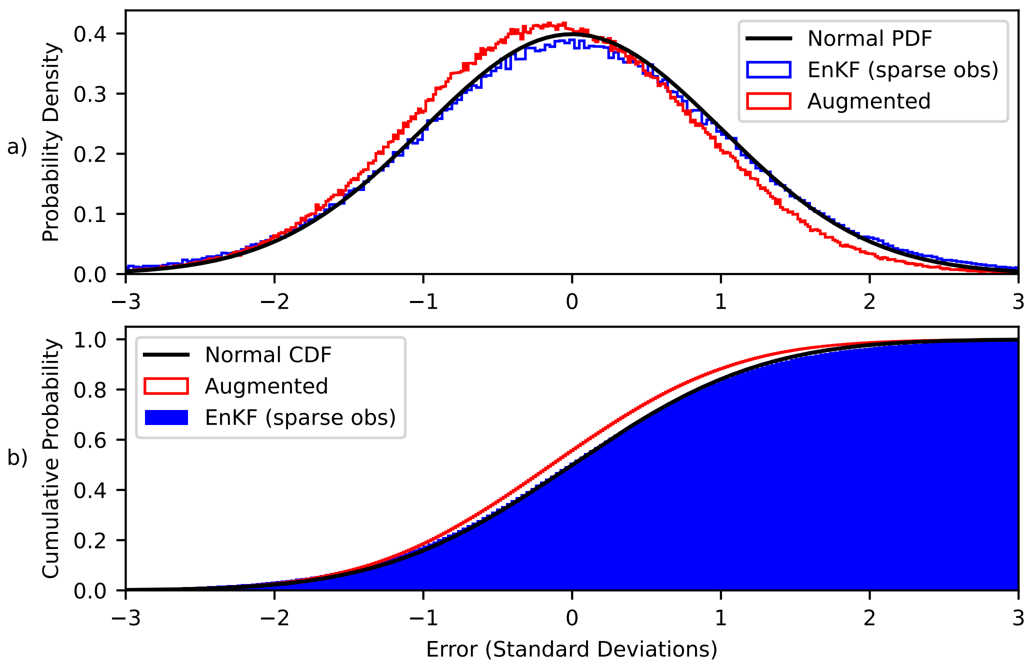

Figure 6. The distribution of augmented (red) and EnKF SparseObs (blue) analysis errors normalized by ensemble standard deviation as probability density plots (a) and cumulative probability plots (b). A reference unit normal distribution is also included in both plots in black.

significantly better at the $p<0.05$ level. Graphically, this is immediately evident as the 95% confidence interval bands do not initially overlap.

Figure 5 also shows another metric for evaluating the reliability of the ensemble forecasts. By definition, when errors are unbiased the standard deviation of the error and the RMSE are equivalent. If the ensemble spread is reflective of the true error, then the actual RMSE should equal the ensemble standard deviation (Leutbecher & Palmer, 2008; Gneiting & Katzfuss, 2014). If the ensemble is overprecise with estimated errors smaller than actual, it is said to be underdispersive. If the ensemble is under-precise, with its spread overestimating actual errors and precision, it is said to be overdispersive.

Here, the ensemble standard deviation for all three methods is generally less than the RMSE indicating underdispersive ensemble forecasts that do not adequately represent the true forecast error. For the first day, however, the standard deviation of the augmented analysis error is within the 95% confidence interval of its RMSE. Beyond this, the augmented method is underdispersive but less so than EnKF SparseObs with the difference between its RMS and standard deviation smaller for several more days. This is another indication of the improved reliability of the augmented method compared to EnKF SparseObs. In addition to avoiding large spikes in RMSE shown in Figure 4, ensemble forecasts using the augmented method analyses as initial conditions produce both more accurate forecasts as well as uncertainty estimates that more closely match the true statistics of forecast errors.

For a more detailed examination of the reliability of ensemble state estimates, we examine the distribution of actual vs. expected analysis errors. These results (as opposed to the distributions of RMSE) can indicate if forecasts are biased or otherwise not well distributed.

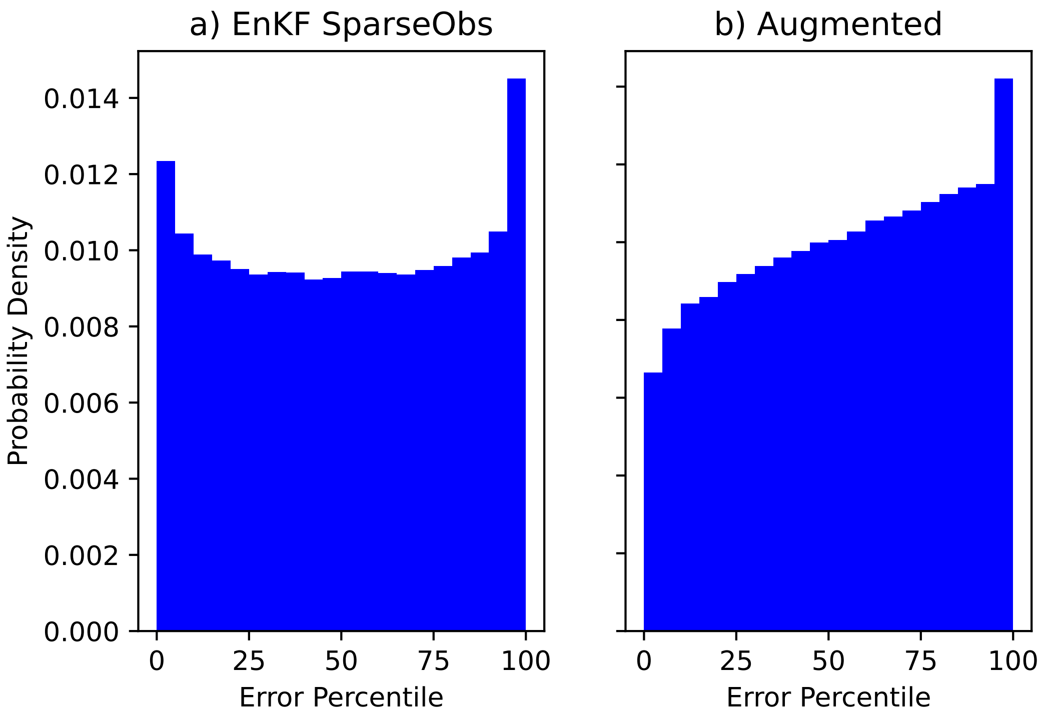

Figure 7. Rank histograms of the observations with respect to the ensemble of analyses for EnKF SparseObs (left) and augmented (right) methods. The percentile in which an observation falls is on the x-axis, with the normalized frequency on y-axis.

If errors are assumed to be Gaussian, then if scaled by the ensemble standard deviation the error distribution should follow a unit normal with a mean of zero. Conversely, if the errors are not well represented by a normal distribution, or the ensemble standard deviations don't reflect the true analysis error statistics, the scaled distribution will diverge from the reference unit normal. This comparison is shown in Figure 6.

Figure 6a suggests that the augmented method has a slight bias with its PDF shifted left compared with the reference unit normal. It also suggests that EnKF SparseObs is slightly underdispersive, with its peak lower and its tails higher than the reference unit normal PDF. These features are also apparent in Figure 6b, which displays the same data but as a CDF instead. For negative errors, EnKF SparseObs is higher than the reference distribution; for positive errors, it is lower. Consistent with the results presented in Figure 5, despite its bias the augmented method better represents the true statistics of its error than EnKF SparseObs, with a standard deviation of 0.96 compared to 1.18, while a perfectly dispersive ensemble would have a standard deviation of 1.

Rank histograms are an alternative way of visualizing the dispersion of ensemble forecasts or analyses (Hatfield et al., 2018; Candille & Talagrand, 2005; Weigel, 2011; Hamill, 2001). For each forecast (or analysis), the percentile of the true value within the ensemble is calculated. When the distribution of the percentile values is plotted, a uniform distribution indicates a well-dispersed forecast. Errors of a given size occur as frequently as would be expected if the ensemble spread represents the true error statistics. A U shape is underdispersive, with errors outside the range of the ensemble over-represented. A tilt indicates a biased ensemble forecast, with positive errors more or less likely to occur than negative errors.

Figure 7 includes rank histograms for both methods. It is more visually apparent here that the EnKF SparseObs produces underdispersive state estimates. Small and large percentile frequencies are clearly larger than frequencies at or around the 50^{th} percentile. Conversely, while the dispersion of the augmented method state estimates is not as visually clear, the bias evident in Figure 6 is also visible here. The augmented method produces more accurate state estimates and more stability but with a slight bias compared with EnKF SparseObs.

3.3 Explainable AI: SHAP Values

We now return to the behavior the CNN in producing state estimates from forecasts and innovations. In an operational setting, allowing black-box operators to produce new initial conditions is not tenable. There must be some confidence that the system won't generate unrealistic results when presented with out-of-sample data, and some understanding of how it is producing its state estimates. Here we use SHAP values to estimate the impact of input variables on the outputs generated by the CNN (Figure 8).

Figure 8a shows the mean absolute SHAP values in decreasing order. The largest contributors to the state estimate of a variable are the forecast and the innovation of that variable. This is an excellent first sanity check on the CNN. In estimating a state variable, it weights the forecast and observation of that variable more heavily than forecasts or observations of nearby variables.

Figure 8b, identical to Figure 8a but with the first input variable not shown, suggests that the next two most heavily weighted inputs for generating a state estimate are the forecasts at 1 and 3 spatial lags, followed by observations and forecasts at 2 spatial lags. The long-term spatial correlation structure of the Lorenz-96 system is important to note at this point. Since the system is symmetric, without loss of generality we can consider the dynamics and correlation between locations only in terms of absolute spatial lag. The dynamics of a state variable are nonlinearly dependent on the state variables at spatial lags of 1 and 2 as described in Equation 1.

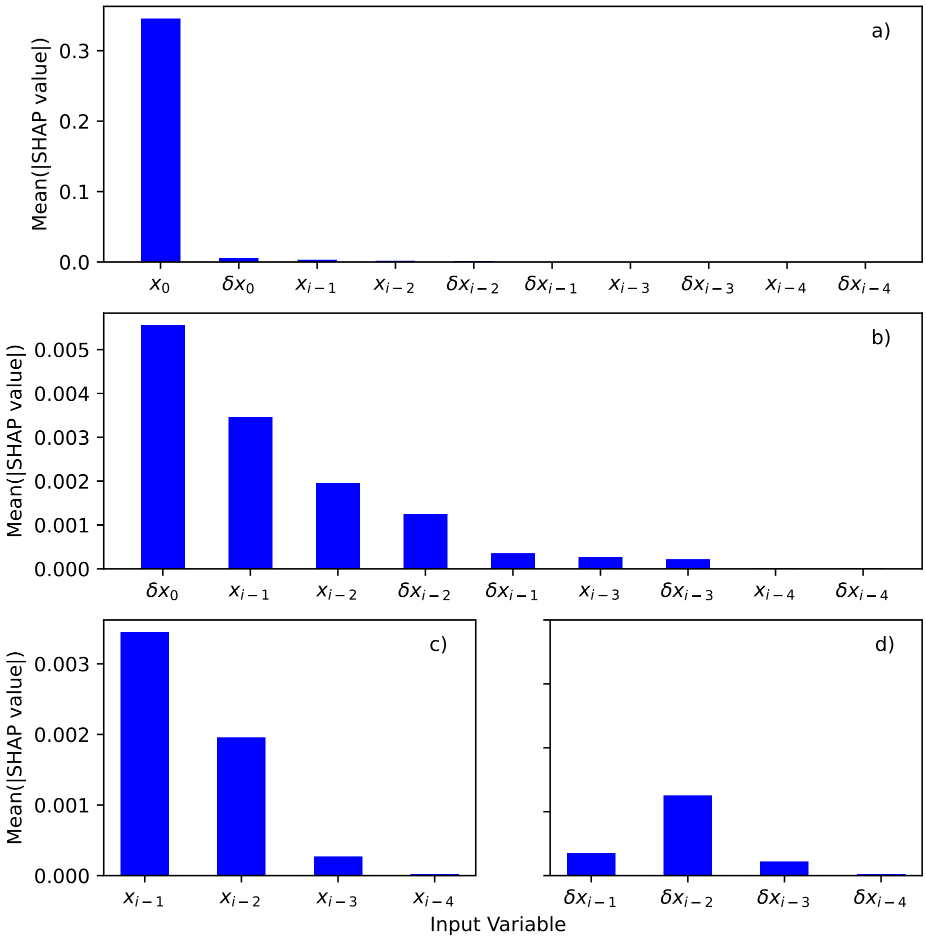

Figure 8. Estimated mean absolute SHAP values at a given location for the forecast(x) and innovation (δx) at spatial lags of 0 to 4. In panel a), input variables are sorted from largest to smallest mean absolute SHAP value. In panel b), panel a) is replicated without the first value (the analysis at a spatial lag of zero). Panel c) includes SHAP values for the forecast at spatial lags of 1-4, and panel d) for innovations at spatial lags of 1-4.

This dependence defines the temporal derivative. In non-differential terms, considering the variable values rather than rates of change, the correlation extends further than two grid points. At spatial lags of 1-4, the long-term absolute correlation coefficients are 0.05, 0.33, 0.11, and 0.03. Variables at a distance of 2 and 3 grid points removed from one another are more highly correlated than immediately adjacent variables one grid point removed (Lorenz, 2005). It is also important to note that the convolutional layers will tend to create a binomial distribution here: larger lags have fewer paths of influence. This means that we should expect the SHAP results at increasing spatial lags to be a combination of a binomial distribution and the Lorenz-96 correlations. Figure 8c demonstrates monotonically decreasing SHAP values with increasing spatial lag consistent with the binomial influence of the CNN. Figure 8d shows a peak SHAP value at a spatial lag of 2, consistent with the correlation structure of Lorenz-96. In both cases, at a spatial lag of 4 the SHAP value is essentially zero. This is another important check on the results: the structure of the CNN means that the impact of data 4 spatial lags away cannot impact the output. The fact that the SHAP results reflect the known behavior of the CNN at a lag of 4 provides confidence that the other SHAP values have meaning.

4 Discussion

The results of using the augmented method outlined above are encouraging, and clearly show that it outperforms a traditional approach using only the EnKF. However, considering the training process of the CNN makes some limitations apparent. First: the network is trained using only the ensemble mean, rather than the entire ensemble, as input. As a result, it can only learn the covariance structure to the extent that the covariance is dependent on location in the state space. Other factors, most obviously the time since the last observation and analysis, will impact forecast uncertainty. The trained CNN cannot include such factors.

Even within the confines of the experiment we have set up, the limitations of the training data have an impact on performance. When employed online in the augmented method the CNN is provided as input a forecast initialized by assimilating only 25% of the variables. In comparison, all variables are observed in the EnKF configuration used to generate training data. In the online setting, therefore, the initial condition error will be larger and the forecast precision lower than in the training set. In the experimental setup, the augmented method has an RMS of 0.74, nearly three times worse than the RMSE when it is simply applied to forecasts from the validation time period generated offline. This partly reflects the fact that the augmented method is simply assimilating less data. On a time-averaged basis, it is observing 62.5% of the observations assimilated in the training set, but that does not fully explain the 3-fold increase in RMSE. The remaining decrease in accuracy is attributable to the smaller forecast errors in the training data compared to forecast errors in the online setting.

While the reduced performance of the CNN applied to an online setting as opposed to input data generated offline is unavoidable to some extent, future opportunities for improvement may be found by allowing the CNN to better approximate forecast accuracy. Providing additional input to the network, such as ensemble standard deviation at the last time step combined with time since observations were last assimilated, is one option. Our results here provide no indication either way whether a neural network would be able to learn effectively from other input data, or how complex the network would have to be, but it is a potential avenue of further exploration.

The results from the SHAP analysis provide additional insights into the possible extensions of the approach. Localization is widely used to improve the performance of many assimilation systems. The SHAP values demonstrate that the trained CNN has applied localization to the forecast. The CNN also has learned the long-term correlation structure (teleconnections) of the system, applying a localization structure to the innovations consistent with that of the Lorenz-96 system. These are both reasons to think it is plausible that in future extensions convolutional layers may be able to generate spatial estimates that blend forecasts and observations in a way that is both reliable and skillfully reflects underlying system behavior and dynamics.

5 Conclusion

This study demonstrated a proof-of-concept augmented assimilation methodology in which machine learning was used to directly assimilate high-resolution observations for potential improvement of the performance of an assimilation scheme. Significant quantities of observational data, particularly from remote-sensing platforms, go unused in operational forecast models due to the computational cost and time required for incorporating them into the model. The potential viability of training a machine learning model offline to assimilate this data could have a significant impact – improved state initialization has real and notable impacts on forecast quality, and the ability to use the vast amounts of newly available observational data products to that end is of clear benefit.

As a demonstration of the potential feasibility of such an approach, we trained a 2-layer convolutional neural network to replicate the results generated by the Ensemble Kalman Filter on synthetic observations. Using the EnKF on low-resolution observations and the trained CNN on the high-resolution observations outperformed an EnKF assimilating only low-resolution data. More specifically, in an experimental setting using the Lorenz-96 model, the analyses generated by the augmented method have a mean RMSE 14.5% lower than using the EnKF on only low-resolution observations. Forecasts using analyses generated by the augmented method as initial conditions produce lower RMSE up to a forecast lead time of 10 days. Ensemble forecasts using initial conditions from the augmented method were also found to be less underdispersive, with ensemble standard deviations that more closely reflect true forecast error.

Additionally, using an explainable AI method, we demonstrate that the trained CNN effectively both applies localization as well as learns the correlation structure (teleconnections) of the underlying system via training. Distant observations do not impact its estimates. The natural tendency of convolutional layers to exploit local spatial correlations in this way is encouraging for potential extensions to more realistic applications. It also generates confidence that such a method would both be reliable and generate physically realistic results when presented with new data.

Further studies are needed to demonstrate the ability of this approach to work in more complex systems and at scale. Testing using a quasigeostrophic model and more realistic observational data would be a logical next step. The demonstrated feasibility of the general approach in this proof-of-concept study will hopefully encourage additional efforts to address the large quantity of data that is currently unusable in an operational forecast setting using machine learning approaches.

6 Open Research

Code for generating the data used in this study as well as code for generating the plots in this paper (and the processed data used in the plots) can be accessed at `https://github.com/climprocpred/machine_learning_DA_part_1`.

Acknowledgments

Support for this work was provided by the Virtual Red Sea Initiative funded by KAUST CRG grant number ORA-2021-CRG10-4649.2. This study was also supported by the U.S. Army Corps of Engineers (USACE) as part of Forecast Informed Reservoir Operations (FIRO) under grant W912HZ-15-2-0019 and by the California Department of Water Resources AR Program (grant: 4600013361). This project was additionally funded by the Climate Change

AI Innovation Grants program, hosted by Climate Change AI with the support of the Quadrature Climate Foundation, Schmidt Futures, and the Canada Hub of Future Earth.

References

Abarbanel, H. D. I., Rozdeba, P. J., & Shirman, S. (2018, August). Machine Learning: Deepest Learning as Statistical Data Assimilation Problems. *Neural Computation*, *30*(8), 2025–2055. Retrieved 2021-08-14, from https://doi.org/10.1162/neco_a_01094 doi: 10.1162/neco_a_01094

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... Farhan, L. (2021, March). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. Retrieved 2023-04-05, from https://doi.org/10.1186/s40537-021-00444-8 doi: 10.1186/s40537-021-00444-8

Arcucci, R., Zhu, J., Hu, S., & Guo, Y.-K. (2021, January). Deep Data Assimilation: Integrating Deep Learning with Data Assimilation. *Applied Sciences*, *11*(3), 1114. Retrieved 2021-10-13, from https://www.mdpi.com/2076-3417/11/3/1114 (Number: 3 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/app11031114

Bannister, R. N. (2017). A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 607–633. Retrieved 2022-02-07, from http://onlinelibrary.wiley.com/doi/abs/10.1002/qj.2982 (_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2982) doi: 10.1002/qj.2982

Bauer, P., Thorpe, A., & Brunet, G. (2015, September). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. Retrieved 2022-12-15, from https://www.nature.com/articles/nature14956 (Number: 7567 Publisher: Nature Publishing Group) doi: 10.1038/nature14956

Bonavita, M., Arcucci, R., Carrassi, A., Dueben, P., Geer, A. J., Saux, B. L., ... Raynaud, L. (2021, April). Machine Learning for Earth System Observation and Prediction. *Bulletin of the American Meteorological Society*, *102*(4), E710–E716. Retrieved 2021-09-02, from https://journals.ametsoc.org/view/journals/bams/102/4/BAMS-D-20-0307.1.xml (Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society) doi: 10.1175/BAMS-D-20-0307.1

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2020, July). 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. Retrieved 2022-10-11, from https://www.sciencedirect.com/science/article/pii/S1877750320304725 doi: 10.1016/j.jocs.2020.101171

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021, April). 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. Retrieved 2021-08-14, from https://royalsocietypublishing-org.colorado.idm.oclc.org/doi/10.1098/rsta.2020.0086 (Publisher: Royal Society) doi: 10.1098/rsta.2020.0086

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, *131*(609), 2131–2150. Retrieved 2023-02-23, from https://onlinelibrary.wiley.com/doi/abs/10.1256/qj.04.71 (_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.04.71) doi: 10.1256/qj.04.71

Chen, T.-C., Penny, S. G., Whitaker, J. S., Frolov, S., Pincus, R., & Tulich, S. (2022). Correcting Systematic and State-Dependent Errors in the NOAA FV3-GFS Using Neural Networks. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003309. Retrieved 2022-12-16, from http://onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003309 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003309) doi: 10.1029/2022MS003309

Choi, Y., Cha, D.-H., Lee, M.-I., Kim, J., Jin, C.-S., Park, S.-H., & Joh, M.-S. (2017). Satellite radiance data assimilation for binary tropical cyclone cases over the western North Pacific. *Journal of Advances in Modeling Earth Systems*, *9*(2), 832–853. Retrieved 2022-12-15, from http://onlinelibrary.wiley.com/doi/abs/10.1002/2016MS000826 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016MS000826) doi: 10.1002/2016MS000826

Dong, S., Wang, P., & Abbas, K. (2021, May). A survey on deep learning and its applications. *Computer Science Review*, *40*, 100379. Retrieved 2022-12-20, from https://www.sciencedirect.com/science/article/pii/S1574013721000198 doi: 10.1016/j.cosrev.2021.100379

Dormand, J. R., & Prince, P. J. (1980, March). A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, *6*(1), 19–26. Retrieved 2022-12-14, from https://www.sciencedirect.com/science/article/pii/0771050X80900133 doi: 10.1016/0771-050X(80)90013-3

Edwards, C. A., Moore, A. M., Hoteit, I., & Cornuelle, B. D. (2015). Regional Ocean Data Assimilation. *Annual Review of Marine Science*, *7*(1), 21–42. Retrieved 2023-03-14, from https://doi.org/10.1146/annurev-marine-010814-015821 (_eprint: https://doi.org/10.1146/annurev-marine-010814-015821) doi: 10.1146/annurev-marine-010814-015821

Evensen, G. (2003, November). The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. Retrieved 2022-04-25, from https://doi.org/10.1007/s10236-003-0036-9 doi: 10.1007/s10236-003-0036-9

Eyre, J. R., Bell, W., Cotton, J., English, S. J., Forsythe, M., Healy, S. B., & Pavelin, E. G. (2022). Assimilation of satellite data in numerical weather prediction. Part II: Recent years. *Quarterly Journal of the Royal Meteorological Society*, *148*(743), 521–556. Retrieved 2022-12-16, from http://onlinelibrary.wiley.com/doi/abs/10.1002/qj.4228 (_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4228) doi: 10.1002/qj.4228

Farchi, A., & Bocquet, M. (2018). Review article: Comparison of local particle filters and new implementations. *Nonlinear Processes in Geophysics*, *25*(4), 765–807. Retrieved from https://npg.copernicus.org/articles/25/765/2018/ doi: 10.5194/npg-25-765-2018

Geer, A. J. (2021, April). Learning earth system models from observations: machine learning or data assimilation? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200089. Retrieved 2021-09-02, from https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0089 (Publisher: Royal Society) doi: 10.1098/rsta.2020.0089

Gevaert, C. M. (2022, August). Explainable AI for earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102869. Retrieved 2023-02-16, from https://www.sciencedirect.com/science/article/pii/S1569843222000711 doi: 10.1016/j.jag.2022.102869

Gneiting, T., & Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. Retrieved 2023-02-23, from https://doi.org/10.1146/annurev-statistics-062713-085831 (_eprint: https://doi.org/10.1146/annurev-statistics-062713-085831) doi: 10.1146/annurev-statistics-062713-085831

Gottwald, G. A., & Reich, S. (2021, October). Combining machine learning and data assimilation to forecast dynamical systems from noisy partial observations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *31*(10), 101103. Retrieved 2022-10-

11, from https://aip.scitation.org/doi/full/10.1063/5.0066080 (Publisher: American Institute of Physics) doi: 10.1063/5.0066080

Hamill, T. M. (2001, March). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, *129*(3), 550–560. Retrieved 2023-02-23, from https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/1520-0493(2001)129⟨0550:IORHFV⟩2.0.CO;2

Hatfield, S., Chantry, M., Dueben, P., Lopez, P., Geer, A., & Palmer, T. (2021). Building Tangent-Linear and Adjoint Models for Data Assimilation With Neural Networks. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002521. Retrieved 2022-02-07, from http://onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002521 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002521) doi: 10.1029/2021MS002521

Hatfield, S., Subramanian, A., Palmer, T., & Düben, P. (2018, January). Improving Weather Forecast Skill through Reduced-Precision Data Assimilation. *Monthly Weather Review*, *146*(1), 49–62. Retrieved 2022-12-05, from https://journals.ametsoc.org/view/journals/mwre/146/1/mwr-d-17-0132.1.xml (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/MWR-D-17-0132.1

Hoteit, I., Luo, X., Bocquet, M., Kohl, A., & Ait-El-Fquih, B. (2018). Data assimilation in oceanography: Current status and new directions. *New frontiers in operational oceanography*, 465–512.

Hoteit, I., Pham, D.-T., Triantafyllou, G., & Korres, G. (2008, January). A New Approximate Solution of the Optimal Nonlinear Filter for Data Assimilation in Meteorology and Oceanography. *Monthly Weather Review*, *136*(1), 317–334. Retrieved 2023-03-14, from https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr1927.1.xml (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/2007MWR1927.1

Jung, Y., Xue, M., & Zhang, G. (2010, February). Simultaneous Estimation of Microphysical Parameters and the Atmospheric State Using Simulated Polarimetric Radar Data and an Ensemble Kalman Filter in the Presence of an Observation Operator Error. *Monthly Weather Review*, *138*(2), 539–562. Retrieved 2021-08-24, from http://www.proquest.com/docview/198117168/abstract/A47AEA8713CD4562PQ/1 (Num Pages: 24 Place: Washington, United States Publisher: American Meteorological Society)

Kalman, R. E. (1960, March). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. Retrieved 2023-02-09, from https://doi.org/10.1115/1.3662552 doi: 10.1115/1.3662552

Kim, B., Azevedo, V. C., Thuerey, N., Kim, T., Gross, M., & Solenthaler, B. (2019). Deep Fluids: A Generative Network for Parameterized Fluid Simulations. *Computer Graphics Forum*, *38*(2), 59–70. Retrieved 2022-11-07, from http://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13619 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13619) doi: 10.1111/cgf.13619

Kumar, S., Kolassa, J., Reichle, R., Crow, W., de Lannoy, G., de Rosnay, P., … Ruston, B. (2022). An Agenda for Land Data Assimilation Priorities: Realizing the Promise of Terrestrial Water, Energy, and Vegetation Observations From Space. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003259. Retrieved 2022-12-15, from http://onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003259 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003259) doi: 10.1029/2022MS003259

Lacerda, J. M., Emerick, A. A., & Pires, A. P. (2021, June). Using a machine learning proxy for localization in ensemble data assimilation. *Computational Geosciences*, *25*(3), 931–944. Retrieved 2021-08-14, from https://doi.org/10.1007/s10596-020-10031-0 doi: 10.1007/s10596-020-10031-0

Lee, Y.-J., Hall, D., Stewart, J., & Govett, M. (2019). Machine Learning for Targeted Assimilation of Satellite Data. In U. Brefeld et al. (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 53–68). Cham: Springer International Publishing. doi: 10.1007/978-3-030-10997-4_4

Leutbecher, M., & Palmer, T. N. (2008, March). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. Retrieved 2023-02-23, from `https://www.sciencedirect.com/science/article/pii/S0021999107000812` doi: 10.1016/j.jcp.2007.02.014

Liang, J., Terasaki, K., & Miyoshi, T. (2023). A Machine Learning Approach to the Observation Operator for Satellite Radiance Data Assimilation. *Journal of the Meteorological Society of Japan. Ser. II*, *101*(1), 79–95. doi: 10.2151/jmsj.2023-005

Liang, X., Garrett, K., Liu, Q., Maddy, E. S., Ide, K., & Boukabara, S. (2022). A Deep-Learning-Based Microwave Radiative Transfer Emulator for Data Assimilation and Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 8819–8833. (Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing) doi: 10.1109/JSTARS.2022.3210491

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021, January). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. Retrieved 2023-02-16, from `https://www.mdpi.com/1099-4300/23/1/18` (Number: 1 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/e23010018

Lorenz, E. N. (2005, May). Designing Chaotic Models. *Journal of the Atmospheric Sciences*, *62*(5), 1574–1587. Retrieved 2022-06-07, from `https://journals.ametsoc.org/view/journals/atsc/62/5/jas3430.1.xml` (Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences) doi: 10.1175/JAS3430.1

Lundberg, S. M., & Lee, S.-I. (2017, December). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777). Red Hook, NY, USA: Curran Associates Inc.

Mallat, S. (2016, April). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150203. Retrieved 2022-12-20, from `http://royalsocietypublishing.org/doi/10.1098/rsta.2015.0203` (Publisher: Royal Society) doi: 10.1098/rsta.2015.0203

Miyoshi, T., Kondo, K., & Terasaki, K. (2015, November). Big Ensemble Data Assimilation in Numerical Weather Prediction. *Computer*, *48*(11), 15–21. (Conference Name: Computer) doi: 10.1109/MC.2015.332

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., & Ott, E. (2017, December). Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *27*(12), 121102. Retrieved 2021-09-07, from `https://aip.scitation.org/doi/10.1063/1.5010300` (Publisher: American Institute of Physics) doi: 10.1063/1.5010300

Penny, S. G., Smith, T. A., Chen, T.-C., Platt, J. A., Lin, H.-Y., Goodliff, M., & Abarbanel, H. D. I. (2022). Integrating Recurrent Neural Networks With Data Assimilation for Scalable Data-Driven State Estimation. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002843. Retrieved 2022-11-09, from `http://onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002843` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002843) doi: 10.1029/2021MS002843

Reichle, R. H., McLaughlin, D. B., & Entekhabi, D. (2002, January). Hydrologic Data Assimilation with the Ensemble Kalman Filter. *Monthly Weather Review*, *130*(1), 103–114. Retrieved 2023-02-09, from `https://journals.ametsoc.org/view/journals/mwre/130/1/1520-0493_2002_130_0103_hdawte_2.0.co_2.xml` (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/1520-0493(2002)130⟨0103:HDAWTE⟩2.0.CO;2

Shahabadi, M. B., & Buehner, M. (2021, November). Toward All-Sky Assimilation of

Microwave Temperature Sounding Channels in Environment Canada's Global Deterministic Weather Prediction System. *Monthly Weather Review*, *149*(11), 3725–3738. Retrieved 2022-12-15, from `https://journals.ametsoc.org/view/journals/mwre/149/11/MWR-D-21-0044.1.xml` (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/MWR-D-21-0044.1

Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., & Balaji, V. (2021, July). Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, *16*(7), 073008. Retrieved 2023-04-14, from `https://dx.doi.org/10.1088/1748-9326/ac0eb0` (Publisher: IOP Publishing) doi: 10.1088/1748-9326/ac0eb0

Stegmann, P. G., Johnson, B., Moradi, I., Karpowicz, B., & McCarty, W. (2022, April). A deep learning approach to fast radiative transfer. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *280*, 108088. Retrieved 2023-02-12, from `https://www.sciencedirect.com/science/article/pii/S0022407322000255` doi: 10.1016/j.jqsrt.2022.108088

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013, May). On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 1139–1147). PMLR. Retrieved 2023-02-05, from `https://proceedings.mlr.press/v28/sutskever13.html` (ISSN: 1938-7228)

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... van Mulbregt, P. (2020, March). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. Retrieved 2022-12-14, from `https://www.nature.com/articles/s41592-019-0686-2` (Number: 3 Publisher: Nature Publishing Group) doi: 10.1038/s41592-019-0686-2

Wang, Y., Shi, X., Lei, L., & Fung, J. C.-H. (2022, August). Deep Learning Augmented Data Assimilation: Reconstructing Missing Information with Convolutional Autoencoders. *Monthly Weather Review*, *150*(8), 1977–1991. Retrieved 2023-02-23, from `https://journals.ametsoc.org/view/journals/mwre/150/8/MWR-D-21-0288.1.xml` (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/MWR-D-21-0288.1

Weigel, A. P. (2011). Ensemble Forecasts. In *Forecast Verification* (pp. 141–166). John Wiley & Sons, Ltd. Retrieved 2023-02-23, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119960003.ch8` (Section: 8 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119960003.ch8) doi: 10.1002/9781119960003.ch8

Wu, P., Chang, X., Yuan, W., Sun, J., Zhang, W., Arcucci, R., & Guo, Y. (2021, April). Fast data assimilation (FDA): Data assimilation by machine learning for faster optimize model state. *Journal of Computational Science*, *51*, 101323. Retrieved 2021-08-12, from `https://www.sciencedirect.com/science/article/pii/S187775032100020X` doi: 10.1016/j.jocs.2021.101323

Yang, L. M., & Grooms, I. (2021, October). Machine learning techniques to construct patched analog ensembles for data assimilation. *Journal of Computational Physics*, *443*, 110532. Retrieved 2021-08-14, from `https://www.sciencedirect.com/science/article/pii/S0021999121004277` doi: 10.1016/j.jcp.2021.110532

Ying, X. (2019, February). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022. Retrieved 2023-02-07, from `https://dx.doi.org/10.1088/1742-6596/1168/2/022022` (Publisher: IOP Publishing) doi: 10.1088/1742-6596/1168/2/022022

Zhang, H., & Tian, X. (2021). Evaluating the Forecast Impact of Assimilating ATOVS Radiance With the Regional System of Multigrid NLS-4DVar Data Assimilation for Numerical Weather Prediction (SNAP). *Journal of Advances in Modeling Earth Systems*, *13*(7), e2020MS002407. Retrieved 2022-12-15, from `http://onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002407` (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002407) doi: 10.1029/2020MS002407