ORIGINAL ARTICLE

Data Assimilation

Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders

Ian Grooms

Department of Applied Mathematics, University of Colorado, Boulder

[1]Department of Applied Mathematics, University of Colorado, Boulder

Correspondence
Ian Grooms PhD, Department of Applied Mathematics, University of Colorado, Boulder, Colorado, 80309, USA
Email: ian.grooms@colorado.edu

Funding information
US National Science Foundation, Division of Mathematical Sciences, Grant/Award Number: 1821074

It is proposed to use analogs of the forecast mean to generate an ensemble of perturbations for use in ensemble optimal interpolation (EnOI) or ensemble variational (EnVar) methods. A new method of constructing analogs using variational autoencoders (VAEs; a machine learning method) is proposed. The resulting analog methods using analogs from a catalog (AnEnOI), and using constructed analogs (cAnEnOI), are tested in the context of a multiscale Lorenz-'96 model, with standard EnOI and an ensemble square root filter for comparison. The use of analogs from a modestly-sized catalog is shown to improve the performance of EnOI, with limited marginal improvements resulting from increases in the catalog size. The method using constructed analogs (cAnEnOI) is found to perform as well as a full ensemble square root filter, and to be robust over a wide range of tuning parameters.

Keywords — Ensemble Optimal Interpolation, analogs, machine learning, variational autoencoder

1 | INTRODUCTION

Data assimilation methods are widely used in geophysics for a variety of purposes. Workhorse methods include the Ensemble Kalman Filter (EnKF) and its many variants (Evensen, 1994; Houtekamer and Mitchell, 1998; Burgers et al., 1998), and 3D-Var and 4D-Var (Talagrand, 2010). Traditional variational methods suffer from the use of a time-independent background covariance,

Abbreviations: EnKF, Ensemble Kalman Filter; EnOI, Ensemble Optimal Interpolation; ESRF, Ensemble Square Root Filter; VAE, Variational Autoencoder

whereas the drawbacks of the EnKF include the sometimes high cost of generating ensemble members and less accurate treatment of nonlinearity and non-Gaussianity. A variety of hybrids exist between ensemble and variational methods that aim to combine the strengths of the different methods (Bannister, 2017). Ensemble optimal interpolation (EnOI; Oke et al., 2002; Evensen, 2003) is in some sense a less expensive and less accurate version of the EnKF. It uses a time-independent background covariance that is generated from a time-independent ensemble of perturbations. In EnOI a single model simulation is required for each assimilation cycle to propagate the mean state. EnOI uses a gain matrix to compute the increment between the forecast and analysis means.

The ensemble of perturbations used in EnOI usually comes from a catalog of model states from a long-running simulation. Since the ensemble of perturbations used in EnOI is static, EnOI suffers from the same drawbacks as early implementations of variational methods, but has the benefit that only a single forecast is required for each assimilation cycle. The goal of this investigation is to explore a way of improving the performance of EnOI by generating a time-dependent ensemble of perturbations from a large catalog. The premise is that an ensemble of model states chosen as a subset from the catalog that are similar to the current forecast will produce an ensemble of perturbations that is more appropriate for use in EnOI than an ensemble that is representative of the climatology of the model. Ensemble perturbations drawn from the climatology represent the correlations in the climatology, which can be a poor proxy for correlations in the forecast error. Analog ensemble perturbations come from the part of the dynamical system's attractor (or pullback attractor for non-autonomous systems) that is close to the actual forecast, and therefore represent correlations on a specific part of the model attractor rather than over the whole climatology. As a result they are expected to provide a more realistic representation of forecast error, since the forecast error distribution should be expected to cover a neighborhood of the attractor close to the forecast mean.

Model states that are similar to the current forecast are called 'analogs' (Lorenz, 1969) and have a long history in weather forecasting and forecast downscaling (Delle Monache et al., 2013; Eckel and Delle Monache, 2016; Zhao and Giannakis, 2016). Van den Dool (1994) considered finding analogs from a large historical catalog of model states, and showed that to make an effective analog global weather forecast would require an impossibly large catalog - on the order of 10^{30} years of data. Nevertheless, in the current setting one may still expect some degree of success with analogs drawn from a practically-sized catalog since the analogs are not being used for forecasting, but only to improve the background covariance within the data assimilation framework.

One way of avoiding the impossibly large size requirements of a catalog for analog forecasting is to use a reasonably-sized catalog to construct analogs, and there are many ways of doing this (Van den Dool et al., 2003; Hidalgo et al., 2008; Maurer et al., 2010; Abatzoglou and Brown, 2012; Tippett and DelSole, 2013; Pierce et al., 2014). This investigation explores a new way of constructing analogs using variational autoencoders (Kingma and Welling, 2019). A standard autoencoder consists of two functions: an encoder $\boldsymbol{e}(\boldsymbol{x})$ that maps the model state $\boldsymbol{x} \in \mathbb{R}^d$ to a latent space $\boldsymbol{z} \in \mathbb{R}^l$ where $l \ll d$, and a decoder $\boldsymbol{d}(\boldsymbol{z})$ that maps a vector in the latent space to a model state. Both $\boldsymbol{e}$ and $\boldsymbol{d}$ are usually specified as deep artificial neural networks. Given a catalog of model states $\{\boldsymbol{x}_i\}_{i=1}^N$, the parameters of $\boldsymbol{e}$ and $\boldsymbol{d}$ are chosen to minimize

$$\sum_i \|\boldsymbol{x}_i - \boldsymbol{d}(\boldsymbol{e}(\boldsymbol{x}_i))\|_2^2$$

or some similar loss function. A standard autoencoder does not impose any particular structure on the latent space. For example, a sufficiently powerful autoencoder might simply learn the map $i = \boldsymbol{e}(\boldsymbol{x}_i)$, $\boldsymbol{d}(i) = \boldsymbol{x}_i$. As a result, standard autoencoders are not always useful as generative models: If $\boldsymbol{z}_i = \boldsymbol{e}(\boldsymbol{x}_i)$, then $\boldsymbol{d}(\boldsymbol{z}_i + \boldsymbol{\epsilon})$ need not be very similar to $\boldsymbol{x}_i$ for small $\boldsymbol{\epsilon}$. Variational autoencoders attempt to impose structure in the latent space; specifically, they aim to choose the parameters of $\boldsymbol{e}$ and $\boldsymbol{d}$ so that the structure of the data in latent space is approximately Gaussian. This is accomplished by two devices. First, the latent space is divided in two so that $\boldsymbol{e}(\boldsymbol{x}) = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ where $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^l$. Then, a latent space vector is constructed as $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a standard normal Gaussian random vector and $\circ$ denotes the elementwise product (also known as Hadamard product,

or Schur product). Second, the loss function is altered by the addition of a term that penalizes deviations of the latent space distribution from a standard normal. (For details on the form of this additional penalty term see Kingma and Welling (2019).) This investigation uses a variational autoencoder (VAE) to generate analogs that are then used to construct the ensemble of perturbations for use in data assimilation.

Data assimilation algorithms using analogs have been proposed in the context of geophysical data assimilation by Lguensat et al. (2017, 2019). Two key differences between this approach and the current approach are (i) the current approach relies on a model simulation to make the forecast, whereas in Lguensat et al. (2017, 2019) the dynamics are data-driven in a manner similar to analog forecasting, and (ii) the current approach investigates the use of VAEs to construct analogs. The data-driven approach of Lguensat et al. (2017, 2019) is expected to be clearly superior in cases where there is no reliable dynamical model for the system in question, or where such a model would be prohibitively expensive.

The investigation is carried out in the context of a multiscale Lorenz-'96 model, which is described in section 2. The configuration and training of the VAE is described in section 3. The data assimilation system setup is described in section 4, and the results of data assimilation experiments are described in section 5. Conclusions are offered in section 6.

2 | MULTISCALE LORENZ-'96 MODEL CONFIGURATION

Many data assimilation methods have been initially explored in the context of the Lorenz-'96 model (Lorenz, 1996, 2006). Higher dimensionality can be obtained in this model by simply retaining the model form while increasing the dimension; alternatively there is a two-scale version also described by Lorenz (1996). This latter two-scale model has two sets of variables, K variables X_i describing the large, slow scales and JK variables Y_j describing the small, fast scales. Grooms and Lee (2015) introduced a multiscale Lorenz-'96 model with a single set of variables x_i with distinct large-scale and small-scale parts. The model is governed by the following system of ordinary differential equations

$$\dot{\boldsymbol{x}} = h\boldsymbol{N}_S(\boldsymbol{x}) + J\mathbf{T}^T\boldsymbol{N}_L(\mathbf{T}\boldsymbol{x}) - \boldsymbol{x} + F\mathbf{1} \tag{1}$$

where the state vector $\boldsymbol{x}$ has length JK with $K = 41$, where $h, F \in \mathbb{R}$, $J \in \mathbb{N}$, $\mathbf{1}$ is a vector of ones, and the nonlinearities have the form

$$(\boldsymbol{N}_S(\boldsymbol{x}))_i = -x_{i+1}(x_{i+2} - x_{i-1}) \tag{2}$$

$$(\boldsymbol{N}_L(\boldsymbol{X}))_k = -X_{k-1}(X_{k-2} - X_{k+1}). \tag{3}$$

The experiments presented here use $h = 0.5$ and $F = 8$. The matrix $\mathbf{T}$ projects onto the 41 largest-scale discrete Fourier modes and then evaluates that projection at 41 equally-spaced points. The vector $\boldsymbol{X} = \mathbf{T}\boldsymbol{x}$ has length $K = 41$. The matrix $J\mathbf{T}^T$ spectrally interpolates a vector of length 41 back to the full dimension of $\boldsymbol{x}$, so that for example $J\mathbf{T}^T\mathbf{T}\boldsymbol{x}$ is the large-scale part of $\boldsymbol{x}$. The number of state variables in $\boldsymbol{x}$ is $41J$; here $J = 64$ for a total system dimension of 2624. In the definition of the nonlinear terms the indices are assumed to extend periodically, as in the Lorenz-'96 model.

The large-scale part of the model dynamics, which can be extracted by applying $\mathbf{T}$ to $\boldsymbol{x}$, is identical to the dynamics of the standard Lorenz-'96 model, except that the large scales are coupled to small scales via the term $h\mathbf{T}\boldsymbol{N}_S(\boldsymbol{x})$. While the Lorenz-'96 model is often configured with $K = 40$ large-scale variables (e.g. Lorenz and Emanuel, 1998), this multiscale model uses 41 variables so that the real and imaginary parts of the 20^{th} Fourier mode are not split between large and small scales. At small scales, the dynamics are the same as those of original Lorenz-'96 model but with the direction of indexing reversed, and with coupling to the large scales. Coupling to the large scales drives small-scale instabilities, which then grow and cause feedback

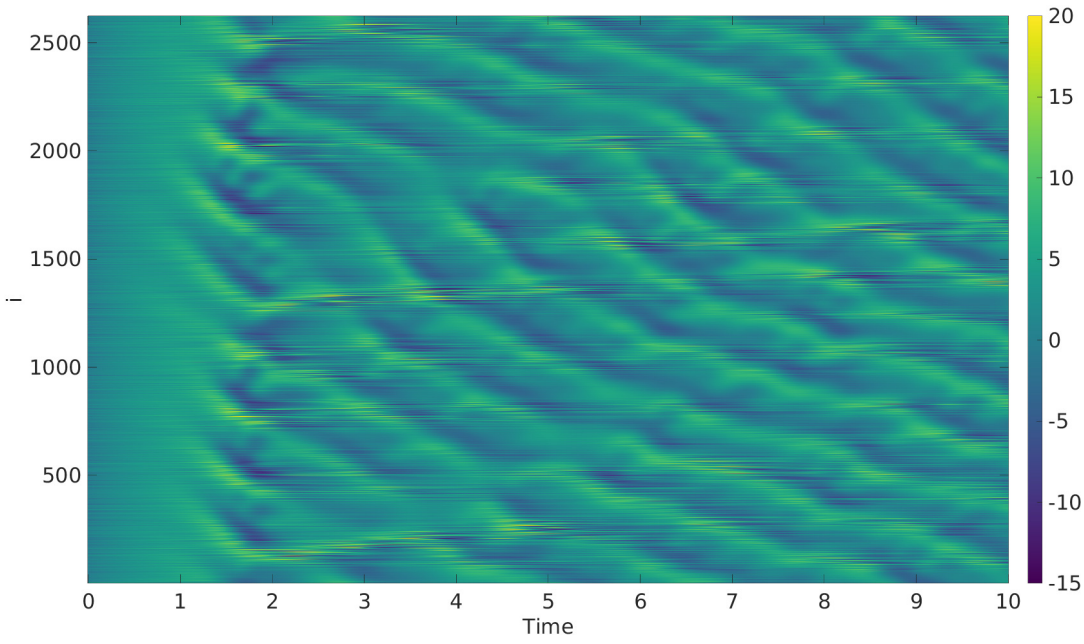

FIGURE 1 A simulation of the multiscale Lorenz-'96 model initialized at $t = 0$ with a sample from a standard normal distribution.

onto the large-scale flow. Figure 1 shows the result of a simulation of this model initialized at $t = 0$ with a sample from a standard normal distribution. After a short transient the dynamics settle onto an attractor, with large-scale nonlinear waves propagating eastward and small-scale instabilities transiently excited by the large-scale waves. All simulations are performed using an adaptive fourth-order Runge-Kutta with relative error tolerance 10^{-3} and absolute error tolerance 10^{-6}. This model has recently been used to study hybrid particle/ensemble Kalman filter performance (Robinson and Grooms, 2020).

3 | VARIATIONAL AUTOENCODER

Variational autoencoders (VAEs) were described generally in the introduction; the architecture of the VAE used here is described in this section. Higham and Higham (2019) provide an introduction to machine learning with artificial neural networks and the associated terminology. The architecture of the autoencoder is summarized in Fig. 2. The encoder $\boldsymbol{e}(\boldsymbol{x})$ is constructed as follows:

1. A convolutional layer with three filters of size 3×1
2. A convolutional layer with nine filters of size 3×1
3. A convolutional layer with 27 filters of size 3×1
4. A max pooling layer with 2×1 pool size
5. Two convolutional layers with 27 filters each of size 3×1
6. A max pooling layer with 2×1 pool size
7. Two convolutional layers with 27 filters each of size 3×1
8. A max pooling layer with 2×1 pool size

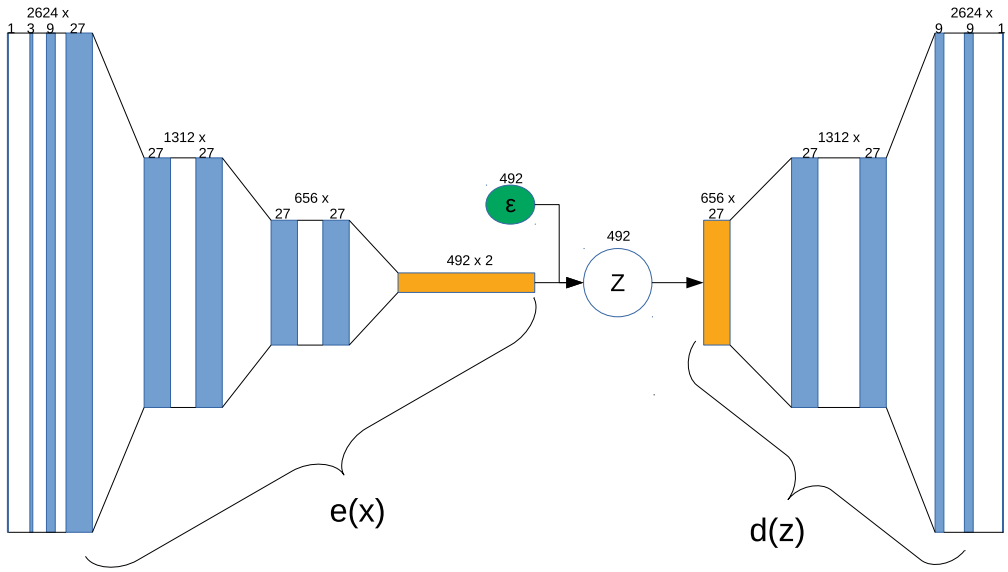

FIGURE 2 Architecture of the variational autoencoder. The leftmost vertical line indicates the data $\boldsymbol{x}$. The blue rectangles in the left half indicate convolutional layers, and the blue rectangles in the right half indicate transposed convolutional layers. The yellow rectangles in the middle indicate fully-connected layers. The green oval indicates the random noise $\boldsymbol{\epsilon}$. The rightmost vertical line indicates the output. Labels at the top of each layer indicate the size of the layer.

9. A fully connected layer with two outputs, each of size 492.

The decoder $\boldsymbol{d}(\boldsymbol{z})$ is constructed as follows:

1. A fully connected layer whose output is reshaped to 27 channels, each of size 656.
2. A transposed convolutional layer with 27 filters of size 3×1 and stride of 2
3. A convolutional layer with 27 filters of size 3×1 and stride of 1
4. A transposed convolutional layer with nine filters of size 3×1 and stride of 2
5. A convolutional layer with nine filters of size 3×1 and stride of 1
6. A convolutional layer with one filter of size 3×1 and a stride of 1.

The convolutional layers, transposed convolutional layers, and fully connected layers all use the exponential linear unit activation function, with the form

$$\mathrm{eLU}(s) = \begin{cases} s & s \geq 0 \\ e^s - 1 & s < 0 \end{cases} \tag{4}$$

The eLU function has a continuous first derivative, which implies that the the decoder also has a continuous first derivative, since it is a composition of continuously-differentiable functions. By contrast, the use of max pooling layers in the encoder implies that the encoder is continuous, but its first derivative is only piecewise continuous.

The VAE architecture used here has a latent space dimension approximately five times smaller than the state dimension, which is not a significant reduction. However, it is important to note that the point of the VAE in this application is not to reduce the dimension, but to act as a generative model. In fact, the dimension reduction (in this application) serves only to keep the number of parameters manageable: a deep network of fully-connected layers with a latent space the same size as the state space would probably perform far better than the dimension-reducing convolutional VAE used here, but would be orders of magnitude more difficult to train, both in terms of data requirements and in terms of computational cost.

The data used to train the VAE consists of 70,000 snapshots of the state of the multiscale Lorenz-'96 model described in the previous section. These snapshots are generated by initializing the model from a standard normal, then running the simulation until it reaches a statistical equilibrium, then taking data every 1 time unit, which corresponds to 5 days in the standard dimensionalization of the Lorenz-'96 model. Using training data from the model's attractor means that the variational autoencoder is attempting to learn a map that transforms the stationary invariant measure on the system attractor to a Gaussian distribution in latent space. The model is trained (i.e. the parameters of $\boldsymbol{e}$ and $\boldsymbol{d}$ are estimated) using stochastic gradient descent. The batch size is 3500 snapshots, and the optimization was trained for 272 epochs, at which point the objective function had saturated. The relatively large batch size and number of epochs are enabled by the small size of the problem and the relative simplicity of the VAE architecture. In more realistic applications than the multiscale Lorenz-'96 model it would be of interest to explore multiple architectures and training regimes to investigate whether the VAE strikes a balance between being sufficiently expressive and being simple enough to train with limited data. For the purpose here of demonstrating the proof of concept in a simple model, a single architecture suffices. Training machine learning methods to generate synthetic realizations of three-dimensional turbulent flows is an active area of research (e.g. Mohan et al., 2019; Rodriguez et al., 2020).

Figure 3 shows the energy spectrum from a test data set consisting of 10,000 snapshots. A dashed black line at wavenumber $k = 20$ marks the dividing line between large and small scales. A solid black line at $k = 328$ marks the Nyquist wavenumber of the observing system; the bump in energy to the right of $k = 328$ results from a small-scale instability, and is not resolvable by the observing system. The trained VAE is used to reconstruct the test data, and Fig. 3 shows both the energy spectrum of the

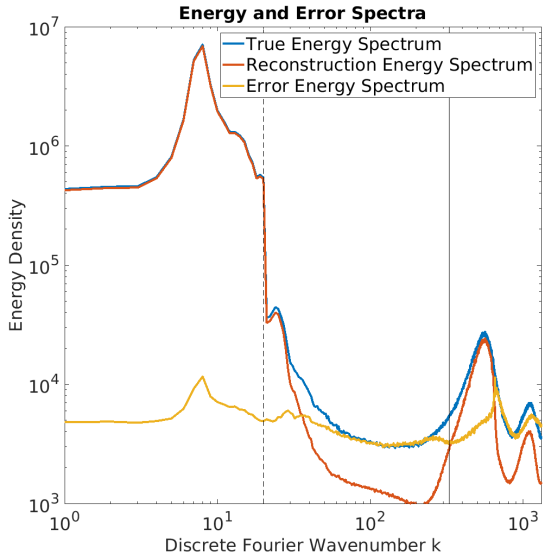

FIGURE 3 Energy spectra from (i) the test data, (ii) the test data reconstructed using the VAE, and (iii) the reconstruction error. The dashed black line is at wavenumber $k = 20$, which is the dividing line between large and small scales. The solid black line is at wavenumber $k = 328$, which is the Nyquist wavenumber of the observing system.

reconstructed data and the spectrum of the reconstruction error. The reconstruction does an excellent job at large scales. The peak of the energy spectrum at small scales is captured reasonably well by the VAE, but the other aspects of the small scales generally have too little energy. In the application to data assimilation the VAE is used to generate a synthetic forecast ensemble. These results on reconstruction accuracy suggest that the forecast error covariance associated with the synthetic forecast ensemble will generally have unrealistically small variance at small scales, as well as at scales intermediate between the main small-scale instability and the large scales at $k \leq 20$.

4 | DATA ASSIMILATION: METHODS

The observations are taken at every fourth point in space and at every 0.2 time units (which corresponds to 1 day in the standard dimensionalization of the Lorenz-'96 model). At every assimilation cycle there are effectively 16 observations for each of the 41 large-scale Lorenz-'96 modes. The observation errors are Gaussian with zero mean and variance 1/2. Experiments with the IEnKF in a similar model suggest that nonlinearity begins to assert itself over an observation window of this length (Bocquet and Sakov, 2012; Sakov et al., 2012; Bocquet, 2016).

Each of the various data assimilation methods described below has several tunable parameters, e.g. localization radius and inflation factor. To optimize the performance of each method, a range of parameters is explored. For each parameter combination that is tested, at least 8 experiments are run. For each experiment a reference simulation is initialized from standard normal noise and run for 9 time units, by which time it has reached a statistical equilibrium. Observations are taken starting at time $t = 9$, every 0.2 time units (1 day) for 73 time units (one year), which corresponds to 365 assimilation cycles. This time window is somewhat longer than the standard 0.05 time units for the Lorenz-96 model; the density of the observing system means that the large-scale part of the field is very well observed, and the uncertainty in the forecast does not grow much over a forecast of length 0.05. The longer window of 0.2 time units is used here because it results in a more difficult test, enhancing the performance

disparity between the methods. The first 73 assimilation cycles of each experiment are considered a burn-in period, and are discarded when computing performance statistics. Once optimal parameters are found, eight longer experiments are run, each having 1,000 assimilation cycles, to verify that the performance does not change over a longer time window; the results of the longer experiments are in every case not statistically-significantly different from the shorter experiments.

At each assimilation cycle the performance of the filter is measured using the root mean square error (RMSE), defined as the 2-norm of the error between the reference simulation and the filter analysis mean. At each parameter value this procedure results in at least $8 \times (365 - 73) = 2336$ values of RMSE. The mean of these values is used to summarize the performance of the method for that specific combination of parameters. RMSE based on the forecast is available, but does not behave qualitatively differently from analysis RMSE and is therefore not shown.

The following subsections detail the different assimilation methods to be compared.

4.1 | Serial Ensemble Square Root Filter

The point of this investigation is to consider methods that improve on EnOI but that are less computationally costly than an EnKF. As such, it is useful to run an EnKF as a baseline for comparison, giving an upper bound on the expected performance of the other methods. The baseline method used here is the serial ensemble square root assimilation of Whitaker and Hamill (2002), with Schur-product localization in observation space and multiplicative inflation. This method is referred to as ESRF in the results.

The initial ensemble is constructed by initializing each ensemble member using an independent draw from a standard normal distribution, then forecasting this initial condition for 9 time units (45 days), by which time they have reached the system attractor. The final condition of each simulation at $t = 9$ is used to initialize the ensemble, so the initial ensemble is completely independent of the reference simulation used to generate the observations. The localization function has the form

$$\ell_i = e^{-\frac{1}{2}\left(\frac{i}{L}\right)^2} \tag{5}$$

where L is the localization radius. For reference, the large-scale Lorenz-'96 modes in this model are effectively 64 units apart, so to convert L to a comparable localization radius for the standard Lorenz-'96 model it suffices to divide L by 64. The multiplicative inflation is applied to the analysis ensemble, since El Gharamti et al. (2019) recently found that posterior inflation is more appropriate and more effective in situations without model error. Inflation is applied by multiplying the analysis ensemble perturbations by an inflation factor of $r \geq 1$.

The three tunable parameters for the ESRF are the ensemble size N_e, the localization radius L, and the inflation factor r. Some limited exploration of ensemble size N_e was performed. First a range of L and r were explored at $N_e = 100$. Then, a range was explored at $N_e = 200$. The optimal RMSE obtained at $N_e = 200$ was not significantly better than at $N_e = 100$, so all results reported here for all methods described below (EnOI, AnEnOI, CAnEnOI, and ESRF) use an ensemble size of $N_e = 100$.

4.2 | Ensemble OI

EnOI can also be considered as a baseline for comparison of the analog methods, providing a lower bound on performance to complement the upper bound provided by the ESRF. The EnOI used here is configured exactly the same as the ESRF, except that no inflation needs to be applied. A different ensemble of perturbations is drawn randomly for each experiment from a catalog of 41,000 model states (once drawn, the ensemble perturbations remain time-independent for all assimilation cycles within a single experiment). This catalog is different from the one used to train the VAE, but is constructed in the same way. The climatological spread represented by this ensemble is too large to be an accurate representation of the forecast error, so the ensemble of perturbations is scaled to a pre-defined forecast spread, which forms the second tunable parameter (together with

localization radius) for the EnOI method. Ensemble size is $N_e = 100$.

4.3 | Analog Ensemble OI

The analog ensemble OI (AnEnOI) method is exactly the same as the EnOI method except for the following: At each assimilation cycle the ensemble is chosen to be the $N_e = 100$ members of the catalog that are closest to the current forecast. Future work will investigate using a weighted analog ensemble to construct the forecast error covariance matrix, with weights depending on distance to the forecast mean. The impact of the size of the catalog is briefly explored by performing experiments using (i) a catalog of only 1,000 members, and (ii) the full catalog of 41,000 members. Results reported below are for the smaller catalog, unless noted otherwise. The similarity of analogs to the forecast is defined using the 2-norm, i.e. $\boldsymbol{x}_1$ is considered to be similar to $\boldsymbol{x}_2$ when $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$ is small. The impact of using other, more dynamically motivated measures of similarity is not explored. Ensemble size is $N_e = 100$.

4.4 | Constructed Analog Ensemble OI

The constructed analog ensemble OI (cAnEnOI) is exactly the same as the AnEnOI except for the construction of the analogs. To construct analogs, the forecast mean is first encoded using the encoder $\boldsymbol{e}(\boldsymbol{x})$. Recall that the encoder produces two vectors, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, and during training the encoded state is $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a standard normal random variable. For the purposes of constructing analogs, an ensemble in latent space is constructed as follows:

$$\boldsymbol{z}_i = \boldsymbol{\mu} + r_z \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N_e \tag{6}$$

where $\boldsymbol{\epsilon}_i$ are independent draws from a standard normal distribution and r_z is a tunable parameter governing the spread of the ensemble in the latent space. The $\boldsymbol{\sigma}$ vector produced by the encoder is not used here. The analog ensemble is then constructed using the decoder as $\boldsymbol{x}_i = \boldsymbol{d}(\boldsymbol{z}_i)$. Ensemble size is $N_e = 100$.

As noted above, the decoder is a continuously-differentiable function. The ensemble in latent space is Gaussian, so for small enough r_z the analog ensemble will also be approximately Gaussian distributed, with a covariance matrix approximately

$$r_z^2 \mathbf{D}\mathbf{D}^T \tag{7}$$

where $\mathbf{D}$ is the Jacobian derivative of $\boldsymbol{d}$ evaluated at $\boldsymbol{\mu}$. The rank of this covariance matrix is less than or equal to the dimension of the latent space, since $\mathbf{D}$ is a $d \times l$ matrix. Of course the analog ensemble covariance matrix will also have rank less than or equal to $N_e - 1$.

For small r_z the correlation structure depends only on the forecast mean, and not on r_z. For larger r_z the nonlinearity of the decoder comes into play with two important consequences. First, the analog ensemble becomes increasingly non-Gaussian, which allows the rank of the covariance matrix to exceed the dimension of the latent space (though the ensemble covariance matrix still must have rank bounded by $N_e - 1$). Second, the correlation structure of the analog ensemble begins to depend on r_z as well as on the forecast mean.

It is desirable to decouple the forecast spread of the analog ensemble from the correlation structure of the analog ensemble covariance matrix. This can be achieved by first constructing the analog ensemble as described above, and then rescaling the ensemble perturbations to have the desired spread. As a result, the cAnEnOI method has three main tunable parameters: (i) localization radius, (ii) r_z which controls the correlation structure of the analog ensemble perturbation covariance matrix, and (iii) the forecast spread.

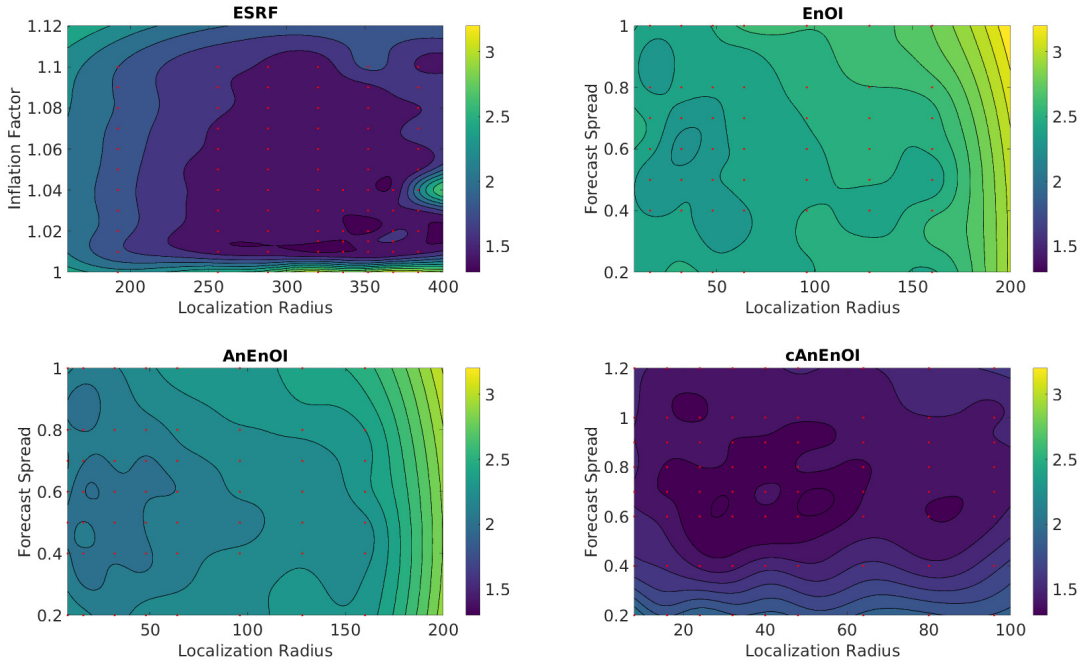

FIGURE 4 Mean RMSE for the four methods, as a function of the governing parameters. The experimental results are shown as red dots; values in between are interpolated. Values for cAnEnOI are at $r_z = 0.7$. The colorbar is the same for all plots. Note that all methods include localization radius as a tunable parameter, but ESRF has inflation as a tunable parameter while the other methods have forecast spread. The axis limits on each panel are different.

5 | DATA ASSIMILATION: RESULTS

Figure 4 shows analysis RMSE as a function of tunable parameters for the four methods (ESRF, EnOI, AnEnOI, and cAnEnOI). Results for the cAnEnOI method are shown for $r_z = 0.7$; the dependence on r_z is discussed below. The ESRF has a fairly broad well in parameter space where the analysis RMSE is around 1.5. The optimal observed RMSE is 1.35, which occurs at inflation factor $r = 1.02$ and localization radius $L = 320$. This is about a factor of two larger than the raw observation error $1/\sqrt{2}$, which is not bad given that only one quarter of the state variables are observed. At localization radii smaller than 192 or larger than 384 the performance begins to degrade. For larger localization radii the ESRF performance becomes erratic, being limited by the deleterious effects of rare spurious long-range correlations: some experiments perform well, while others diverge. For smaller localization radii the ESRF performance also degrades, for reasons that are not entirely clear and could be related to dynamical imbalance of the analysis or to the fact that localization is suppressing non-spurious long distance correlations.

The EnOI (with a catalog of 1,000 model states) has an optimal analysis RMSE of 2.27, which occurs at a localization radius of 32 and a forecast spread of 0.6. Though significantly worse than ESRF, the EnOI still produces reasonably-accurate analyses; for comparison, a random draw from the climatological distribution would produce an RMSE of 4.97. The optimal localization radius for EnOI is a factor of 10 smaller than for ESRF. This is presumably because the correlations encoded in the ESRF ensemble are far more meaningful (i.e. representative of forecast error correlations) at long range than the climatological correlations associated with the EnOI ensemble.

The use of analogs significantly improves the EnOI method: the optimal analysis RMSE for AnEnOI is 2.01, which occurs

at a localization radius of 16 and a forecast spread of 0.6. It is not clear why the optimal localization radius decreases, but on the other hand as seen in Fig. 4 the analysis RMSE of AnEnOI is not too strongly sensitive to changes in localization radius or forecast spread. This is very encouraging, since a catalog of only 1,000 states would presumably be far too small to produce accurate analog forecasts for this system. Increasing the catalog size to 41,000 further improves the analysis RMSE to 1.90: a very modest improvement for a very large increase in catalog size. To put a positive spin on this, it suggests that the bulk of the benefits that can be obtained by moving from EnOI to AnEnOI do not require an unrealistically large catalog. On the other hand, in a real application with a much larger dimension AnEnOI might require an unrealistically large library to produce even marginal improvement compared to EnOI.

The real success comes from using constructed analogs. The optimal analysis RMSE obtained using cAnEnOI at $r_z = 0.6$ is 1.30: slightly better than obtained using ESRF! Each of the 8 experiments per parameter setting produces an independent estimate of the mean analysis RMSE at that parameter setting; from these 8 estimates of the mean one can calculate the 'standard error in the mean' (SEM) as a way to quantify how close the estimated mean is to the true mean that would be obtained over an infinite number of experiments. The SEM for the computed analysis RMSE is 0.05 both for the optimal cAnEnOI method and for the optimal ESRF method, which means that the differences are not statistically significant: on the basis of the results presented here one cannot conclude with certainty that cAnEnOI is better than ESRF, or vice versa. Furthermore, the standard deviation of analysis RMSE values is 0.33 for optimal ESRF and 0.29 for optimal cAnEnOI, so the performance of the methods is quite similar overall. The optimal localization radius and forecast spread are 40 and 0.7, respectively, but as shown in Fig. 4, the performance of cAnEnOI is not strongly sensitive to changes in these parameters: performance comparable to ESRF can be obtained over a wide range of localization radii and forecast spread.

Though neither method is strongly sensitive to deviations from the optimal localization radius, cAnEnOI does seem to be even less sensitive in this regard than ESRF: good results for cAnEnOI can be obtained from localization radii of at least 10 to 100 (a factor of 10), whereas good results for ESRF can be obtained from localization radii of about 250 to 400 (a factor of only 1.6). These differences may stem from the fact that cAnEnOI does not forecast the full ensemble (only the ensemble mean). If one or a few members of the analysis ensemble are dynamically unstable it will have a detrimental impact on ESRF performance, while having no impact on cAnEnOI.

Figure 5 shows the cAnEnOI analysis RMSE as a function of latent space spread r_z for a fixed localization radius of 24 (left panel) and a fixed forecast spread of 0.7 (right panel). The method is extremely robust to varying all three parameters (forecast spread, localization radius, and latent space spread), and is able to produce RMSE comparable to ESRF over a wide range of parameters. As noted above, the correlation structure of the constructed analog ensemble is independent of r_z for small r_z. Consistent with this, for r_z between 0.05 and 0.2, cAnEnOI produces RMSE of 1.38 (at optimal values of forecast spread and localization radius), which is comparable to ESRF. As r_z increases the performance improves, with excellent results in the range $.2 \leq r_z \leq 1$. As r_z increases further the performance slowly degrades, but even at $r_z = 2$ the performance is better than the optimal results using the AnEnOI method with 'found' analogs.

Differences between cAnEnOI and AnEnOI are conjectured to stem primarily from the size of the library in AnEnOI. Even for a relatively small model like the one used here, it would presumably take an astronomically large library to achieve a dense coverage of the model's attractor. The cAnEnOI method apparently circumvents this limitation in a manner analogous to the way that constructed analogs circumvent the limitations on library size in analog forecasting applications. The AnEnOI method could also potentially be improved by applying unequal weights, related to the distance between the analog and the forecast, when computing the forecast error covariance matrix.

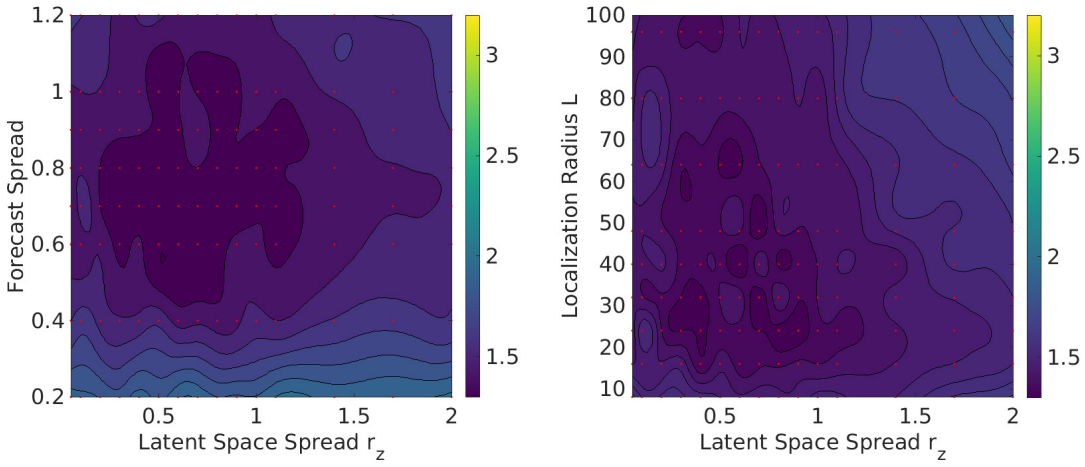

FIGURE 5 Mean RMSE for the cAnEnOI method, as a function of latent space spread r_z and forecast spread (left panel), and as a function of r_z and localization radius L (right panel). The experimental results are shown as red dots; values in between are interpolated. The colorbar is the same for both panels, and is the same as in Fig. 4.

6 | CONCLUSIONS

This work introduces a new use for analogs, besides forecasting and downscaling: to construct an ensemble background covariance matrix for use in data assimilation, as in the EnOI or EnVar frameworks. The research was carried out in the context of a multiscale Lorenz-'96 model invented by Grooms and Lee (2015). Two methods were formulated: one based on finding analogs within a catalog of historical states (AnEnOI), the other based on constructing analogs using a variational autoencoder (VAE; Kingma and Welling, 2019) trained on a catalog of historical states (cAnEnOI). It was found that AnEnOI outperforms a basic EnOI method even with a relatively small catalog of 1,000 members, and further improvements were marginal when the catalog size was increased to 41,000. The cAnEnOI method was able to perform as well as an optimized ensemble square root filter (ESRF), and was quite robust to variations in the tuning parameters of the method. Several alternate methods exist for constructing analogs (Van den Dool et al., 2003; Hidalgo et al., 2008; Maurer et al., 2010; Abatzoglou and Brown, 2012; Tippett and DelSole, 2013; Pierce et al., 2014); these could also be adapted for use in a cAnEnOI method.

Analogs have previously been used in data assimilation by Lguensat et al. (2017, 2019). The key conceptual difference between the methods is that the method proposed here uses a model integration to generate the forecast, while the methods of Lguensat et al. (2017, 2019) use analog forecasting. A key procedural innovation of the present work is the use of machine learning to generate analogs.

In real geophysical applications the model states are much larger than in the simple model considered here. This leads to two difficulties in implementing the analog ensemble data assimilation methods proposed here: (i) finding analogs within the catalog is expensive, and (ii) training a VAE to reproduce an entire model state is likely far more difficult and may be practically impossible. Fortunately, given the long history of analogs, there is already research on efficient ways to find analogs within a large catalog of large model states; see, e.g., Raoult et al. (2018) and Yang and Alessandrini (2019). Since the method using constructed analogs is far more successful, the second difficulty of real geophysical models is more pertinent. To overcome this limitation it is suggested to use a local analysis in the vein of the Local Ensemble Kalman Filter (LEnKF; Brusdal et al., 2003; Evensen, 2003; Ott et al., 2004). This framework uses many local ensembles: for each model grid point a local ensemble analysis is performed using observations near that grid point. The cAnEnOI method developed here could easily be used in this

local framework: For each grid point an analog ensemble is constructed for use in the local assimilation. The benefit of such a local analysis is that the VAE would only have to be trained to generate local subsets of the model state, rather than, e.g., the full state of a global coupled climate model. A similar localization procedure could be leveraged in the case of 'found' analogs rather than constructed ones.

Overall, the results are quite promising. EnOI is a widely used method (Xie et al., 2011; Backeberg et al., 2014; Mignac et al., 2015; Deng et al., 2018; Wu et al., 2018) because of its acceptable performance and significantly reduced cost compared to EnKF, and ensemble background covariances are widely used in EnVar and hybrid data assimilation methods (Gharamti et al., 2014; Bannister, 2017). The results here suggest that improvements could be obtained using either found analogs or constructed analogs; the increased cost of using analogs will be situation-dependent, but if the costs can be made lower than the cost of forecasting an ensemble, then the analog EnOI or EnVar methods may be an attractive alternative. The methods used here to construct analogs could also be used to study predictability (Anderson and Hubeny, 1997), which was the original context for analogs (Lorenz, 1996).

ACKNOWLEDGMENTS

The author is grateful to Jeff Anderson for a discussion on the history of analog weather forecasting, and to Marc Bocquet and two anonymous reviewers for their constructive criticisms. This work used the Extreme Science and Engineering Discovery Environment (XSEDE; Towns et al., 2014) Bridges (Nystrom et al., 2015) at the Pittsburgh Supercomputing Center through allocation TG-DMS190025. This work was funded by the US National Science Foundation under grant number DMS 1821074. Code defining the VAE architecture is available at (Grooms, 2020b) and the trained VAE is available at (Grooms, 2020a).

REFERENCES

John T Abatzoglou and Timothy J Brown. A comparison of statistical downscaling methods suited for wildfire applications. *International Journal of Climatology*, 32(5):772–780, 2012.

JL Anderson and V Hubeny. A reexamination of methods for evaluating the predictability of the atmosphere. *Nonlinear Proc. Geoph.*, 4:157–165, 1997.

Björn C Backeberg, François Counillon, Johnny A Johannessen, and Marie-Isabelle Pujol. Assimilating along-track sla data using the enoi in an eddy resolving model of the agulhas system. *Ocean Dynamics*, 64(8):1121–1136, 2014.

RN Bannister. A review of operational methods of variational and ensemble-variational data assimilation. *Quart. J. Roy. Meteor. Soc.*, 143(703):607–633, 2017.

M. Bocquet. Localization and the iterative ensemble Kalman smoother. *Quart. J. Roy. Meteor. Soc.*, 142:1075–1089, 2016. doi: 10.1002/qj.2711.

M. Bocquet and P. Sakov. Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Proc. Geoph.*, 19:383–399, 2012. doi: 10.5194/npg-19-383-2012.

K Brusdal, Jean-Michel Brankart, G Halberstadt, Geir Evensen, Pierre Brasseur, Peter Jan van Leeuwen, Eric Dombrowsky, and Jacques Verron. A demonstration of ensemble-based assimilation methods with a layered ogcm from the perspective of operational ocean forecasting systems. *Journal of Marine Systems*, 40:253–289, 2003.

Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble kalman filter. *Mon. Wea. Rev.*, 126(6): 1719–1724, 1998.

Luca Delle Monache, F Anthony Eckel, Daran L Rife, Badrinath Nagarajan, and Keith Searight. Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, 141(10):3498–3516, 2013.

Ziwang Deng, Jinliang Liu, Xin Qiu, Xiaolan Zhou, and Huaiping Zhu. Downscaling RCP8. 5 daily temperatures and precipitation in Ontario using localized ensemble optimal interpolation (EnOI) and bias correction. *Climate Dynamics*, 51(1-2):411–431, 2018.

F Anthony Eckel and Luca Delle Monache. A hybrid nwp–analog ensemble. *Mon. Wea. Rev.*, 144(3):897–911, 2016.

Mohamad El Gharamti, Kevin Raeder, Jeffrey Anderson, and Xuguang Wang. Comparing adaptive prior and posterior inflation for ensemble filters using an atmospheric general circulation model. *Mon. Wea. Rev.*, 147(7):2535–2553, 2019.

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res. Oceans*, 99(C5):10143–10162, 1994.

Geir Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

ME Gharamti, Johan Valstar, and Ibrahim Hoteit. An adaptive hybrid EnKF-OI scheme for efficient state-parameter estimation of reactive contaminant transport models. *Advances in water resources*, 71:1–15, 2014.

I Grooms and Y Lee. A framework for variational data assimilation with superparameterization. *Nonlinear Proc. Geoph.*, 22(5): 601–611, 2015.

Ian Grooms. Trained VAE for Grooms-QJRMS (submitted 2020). 9 2020a. doi: 10.6084/m9.figshare.12912275.v1. URL `https://figshare.com/articles/dataset/Trained_VAE_for_Grooms-QJRMS_submitted_2020_/12912275`.

Ian Grooms. iangrooms/canenoi-qjrms: Zenodo release, September 2020b. URL `https://doi.org/10.5281/zenodo.4013825`.

HG Hidalgo, MD Dettinger, and DR Cayan. Downscaling with constructed analogues: Daily precipitation and temperature fields over the United States, 2008. California Energy Commission, PIER Energy-Related Environmental Research. CEC-500-2007-123.

Catherine F Higham and Desmond J Higham. Deep learning: An introduction for applied mathematicians. *SIAM Review*, 61(4): 860–891, 2019.

Peter L Houtekamer and Herschel L Mitchell. Data assimilation using an ensemble kalman filter technique. *Mon. Wea. Rev.*, 126(3): 796–811, 1998.

DP Kingma and M Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019. doi: 10.1516/2200000056.

Redouane Lguensat, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet. The analog data assimilation. *Mon. Wea. Rev.*, 145(10):4093–4107, 2017.

Redouane Lguensat, Phi Huynh Viet, Miao Sun, Ge Chen, Tian Fenglin, Bertrand Chapron, and Ronan Fablet. Data-driven interpolation of sea level anomalies using analog data assimilation. *Remote Sensing*, 11(7):858, 2019.

Edward N Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, 26(4):636–646, 1969.

Edward N Lorenz and Kerry A Emanuel. Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, 55(3):399–414, 1998.

EN Lorenz. Predictability: A problem partly solved. In *Proceedings of Seminar on Predicability*, volume 1, pages 1–18. ECMWF, Reading, UK, 1996.

EN Lorenz. Predictability: A problem partly solved. In T Palmer and R Hagedorn, editors, *Predictability of Weather and Climate*, pages 40–58. Cambridge University Press, 2006.

E. P. Maurer, H. G. Hidalgo, T. Das, M. D. Dettinger, and D. R. Cayan. The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in california. *Hydrology and Earth System Sciences*, 14(6):1125–1138, 2010. doi: 10.5194/hess-14-1125-2010. URL `https://www.hydrol-earth-syst-sci.net/14/1125/2010/`.

D Mignac, CAS Tanajura, AN Santana, LN Lima, and J Xie. Argo data assimilation into hycom with an enoi method in the atlantic ocean. *Ocean Science*, 11(1), 2015.

Arvind Mohan, Don Daniel, Michael Chertkov, and Daniel Livescu. "compressed convolutional lstm: An efficient deep learning framework to model high fidelity 3d turbulence", 2019.

Nicholas A. Nystrom, Michael J. Levine, Ralph Z. Roskies, and J. Ray Scott. Bridges: A uniquely flexible hpc resource for new communities and data analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE '15, pages 30:1–30:8, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3720-5. doi: 10.1145/2792745.2792775. URL http://doi.acm.org/10.1145/2792745.2792775.

Peter R Oke, John S Allen, Robert N Miller, Gary D Egbert, and P Michael Kosro. Assimilation of surface velocity data into a primitive equation coastal ocean model. *J. Geophys. Res. Oceans*, 107(C9):5–1, 2002.

Edward Ott, Brian R Hunt, Istvan Szunyogh, Aleksey V Zimin, Eric J Kostelich, Matteo Corazza, Eugenia Kalnay, DJ Patil, and James A Yorke. A local ensemble kalman filter for atmospheric data assimilation. *Tellus A*, 56(5):415–428, 2004.

David W Pierce, Daniel R Cayan, and Bridget L Thrasher. Statistical downscaling using localized constructed analogs (loca). *Journal of Hydrometeorology*, 15(6):2558–2585, 2014.

Baudouin Raoult, Giuseppe Di Fatta, Florian Pappenberger, and Bryan Lawrence. Fast retrieval of weather analogues in a multi-petabytes archive using wavelet-based fingerprints. In *International Conference on Computational Science*, pages 697–710. Springer, 2018.

Gregor Robinson and Ian Grooms. A hybrid particle-ensemble Kalman filter for problems with medium nonlinearity. *arXiv e-prints*, art. arXiv:2006.04699, June 2020.

Arturo Rodriguez, Carlos Cuellar, Luis Rodriguez, Armando Garcia, Jose Terrazas, VM Kotteda, Rao Gudimetla, Vinod Kumar, and Jorge Munoz. Simulation of atmospheric turbulence with generative machine learning models. *Bull. Amer. Phys. Soc.*, 65, 2020.

P. Sakov, D. S. Oliver, and L. Bertino. An iterative EnKF for strongly nonlinear systems. *Mon. Wea. Rev.*, 140:1988–2004, 2012. doi: 10.1175/MWR-D-11-00176.1.

O Talagrand. Variational assimilation. In W Lahoz, B Khattatov, and R Menard, editors, *Data Assimilation Making Sense of Observations*, pages 41–67. Springer, 2010.

Michael K Tippett and Timothy DelSole. Constructed analogs and linear regression. *Mon. Wea. Rev.*, 141(7):2519–2525, 2013.

J Towns, T Cockerill, M Dahan, I Foster, K Gaither, A Grimshaw, V Hazlewood, S Lathrop, D Lifka, GD Peterson, R Roskies, JR Scott, and N Wilkins-Diehr. XSEDE: Accelerating Scientific Discovery. 16:62–74, 2014.

HM Van den Dool. Searching for analogues, how long must we wait? *Tellus A*, 46(3):314–324, 1994.

Huug Van den Dool, Jin Huang, and Yun Fan. Performance and analysis of the constructed analogue method applied to US soil moisture over 1981–2001. *J. Geophys. Res. Atmospheres*, 108(D16), 2003.

Jeffrey S Whitaker and Thomas M Hamill. Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, 130(7):1913–1924, 2002.

Bo Wu, Tianjun Zhou, and Fei Zheng. EnOI-IAU Initialization Scheme Designed for Decadal Climate Prediction System IAP-DecPreS. *Journal of Advances in Modeling Earth Systems*, 10(2):342–356, 2018.

J Xie, F Counillon, J Zhu, L Bertino, and A Schiller. An eddy resolving tidal-driven model of the South China Sea assimilating along-track SLA data using the EnOI. *Ocean Science*, 7(5), 2011.

Dazhi Yang and Stefano Alessandrini. An ultra-fast way of searching weather analogs for renewable energy forecasting. *Solar Energy*, 185:255–261, 2019.

Zhizhen Zhao and Dimitrios Giannakis. Analog forecasting with dynamics-adapted kernels. *Nonlinearity*, 29(9):2888, 2016.