Research Article

Manhong Fan, Yulong Bai*, Lili Wang, Lihong Tang, and Lin Ding

Coupling the K -nearest neighbors and locally weighted linear regression with ensemble Kalman filter for data-driven data assimilation

<https://doi.org/10.1515/geo-2020-0312>
received June 29, 2021; accepted October 27, 2021

Abstract: Machine learning-based data-driven methods are increasingly being used to extract structures and essences from the ever-increasing pool of geoscience-related big data, which are often used in relation to the atmosphere, oceans, and land surfaces. This study focuses on applying a data-driven forecast model to the classical ensemble Kalman filter process to reconstruct, analyze, and elucidate the model. In this study, a nonparametric sampler from a catalog of historical datasets, namely, a nearest neighbor or analog sampler, is given by numerical simulations. Based on this catalog (sampler), the dynamics physics model is reconstructed using the K -nearest neighbors algorithm. The optimal values of the surrogate model are found, and the forecast step is performed using locally weighted linear regression. Several numerical experiments carried out using the Lorenz-63 and Lorenz-96 models demonstrate that the proposed approach performs as good as the ensemble Kalman filter for larger catalog sizes. This approach is restricted to the ensemble Kalman filter form. However, the basic strategy is not restricted to any particular version of the Kalman filter. It is found that this combined approach can outperform the generally used sequential data

Graphical abstract

assimilation approach when the size of the catalog is substantially large.

Keywords: data assimilation, data-driven, machine learning, Lorenz model, K -nearest neighbor, locally weighted linear regression

1 Introduction

Data assimilation (DA) is a fundamental approach for combining numerical simulations and observational and experimental big data [1]. Sequential DA is also called filtering algorithms [2,3]; it includes the forecast and analysis steps [4,5]. Some examples of such filtering algorithms are the Kalman filter (KF) [6], ensemble KF [7,8], particle filter (PF) [9], robust filter [10], and evolutionary computation [11]. The choice of a DA algorithm depends on the dynamic physical model of the system, the error of the model, and the uncertainty of the problems that directly affect the assimilation results. Data from remote sensing observations, archives of *in situ* measurements, and numerical simulations provide novel ideas for

* **Corresponding author: Yulong Bai**, College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, China, e-mail: baiyulong@nwnu.edu.cn, tel: +86-931-7971503

Manhong Fan: College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, China, e-mail: fanmanhong@nwnu.edu.cn

Lili Wang: College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, China, e-mail: wanglili@nwnu.edu.cn

Lihong Tang: College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, China, e-mail: tanglihong@sim.ac.cn

Lin Ding: College of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, China, e-mail: 2020117107@nwnu.edu.cn

improving the understanding, modeling, and reconstruction of geophysical dynamics [12,13]. Indeed, in the last two decades, *in situ* measurements and satellites have provided significant amounts of data in high temporal and spatial resolutions. However, noteworthy difficulties remain, such as the complexity of physical models, the high dimensionality of states, and the uncertainty of chaotic behaviors.

Although the abundance of data, physical laws, or governing equations remain elusive, this is true for problems in climate science, finance, and neuroscience [14]. Making full use of large amounts of data to solve relevant problems has become a growing issue in geophysics [15–19]. The classical approach reconstructs geophysical systems from observational data. However, it frequently depends on known nonlinear dynamic physical models. Consequently, there are serious limitations with regard to computational costs, model reliability, and error size. Especially, when the model is uncertain or utterly ignorant, classical DA schemes may not deal with assimilation problems.

Only during the last few years, the DA strategies are starting to approach machine learning (ML) models to improve the efficiency of DA models. To achieve the accuracy of the DA solution and reduce the execution time, various ML approaches have been applied to generate simulation models of fully observed low-order chaotic systems and then used for forecasting purposes. Examples of successful approaches in analog forecast strategies are as follows: the analog method to replace the numerical forecast model [20–22], a model-free filter based on the ensemble Kalman filter (EnKF) equations [23,24], a model-free prediction of spatiotemporally chaotic systems of arbitrarily large spatial extents and attractor dimensions [25], and an analog data assimilation [26]. Some scholars have also tried to apply ML to deal with errors, such as embedded residual neural network architecture and bilinear layers in nonlinearity [27], combined DA and ML to emulate a dynamical model from sparse and noisy observations in Lorenz 96 model [28], dealing with model errors in DA [29,30], and employing neural networks (NNs) to simulate the local ensemble transform KF (LETKF) [31]. In contrast, there are also some meaningful applications in remote sensing, industry, wind speed prediction, and atmosphere aspects, such as ensemble transform Kalman filter used in satellite data assimilation [32], forecasting the production of gas from mature gas wells [33], the prediction of severe weather [34], the recurrent network used in time-series data prediction scheme [35,36], the global atmospheric forecast model [37], and the combined power of NNs and high-performance computing to assimilate meteorological data [38]. Essentially,

the development of ML, deep learning, and big data will improve the forecast abilities of DA and complex time series. Most analog models are trained on complete and noise-free observations of a system. Many numerical results have proven this perspective to be accurate. For simple problems, straightforward ML models may provide accurate forecasts. However, owing to the nonstationary property of the data, uncertainty in the timing and values of the observation data, and more nonlinear dynamics that often come with increased resolution, the quality of ML forecasts can be degraded. Such issues require increasing complexity in ML model design. These hybrid approaches are also what the Universal Approximation Theorem claims [39].

Combining sequential DA algorithms with ML has notably been explored along with two potential directions we outline now. First, DA approaches can be seen as an optimizer to train neural networks [27,39]. Second, this combinative approach is used to emulate a dynamical model from small ensembles or sparse and noisy observations [26,28,40]. The latter direction is definitely closer to our study. The key features of this study are threefold: First, a nonparametric sampler is obtained, which directly reflects the dynamic model, from a representative catalog of historical datasets. Second, the model forecast approach is replaced with the catalog and ML, and then, we apply the new forecast values to the analysis step of the EnKF. Finally, numerical experiments are performed using the Lorenz-63 and Lorenz-96 models. Herein, this approach is called data-driven data assimilation (DD-DA).

The structure of this article is as follows: In Section 2, the principle of the ensemble KF and the core ideas of DD-DA are introduced briefly. This section includes the generation methods of data catalog sets, in addition to the application and implementation of the ML methods. In Section 3, we present the experimental setup using the Lorenz-63 and Lorenz-96 models chosen to illustrate the approach. Section 4 summarizes the advantages and disadvantages of this approach and puts forward a new perspective and challenges for future work.

2 Methods

2.1 Ensemble Kalman filter

The EnKF is a sequential data assimilation algorithm based on the theory of stochastic dynamic forecasting. It was proposed by Evensen in 1994 [8]. This approach combines ensemble forecasting with KF and incorporates the ideas of Monte Carlo methods. It normally uses a

group of random variables coming from a Gaussian distribution to represent the probability density function in random dynamic forecasting. Then, it calculates the population probability density function of the next state using forward integration and obtains the statistical characteristics, such as mean and covariance. In the forecast step, we give a group of random variables that conforms to the Gaussian distribution ($i = 1, \dots, N$) at time t in initial conditions. The forecasting values $\mathbf{X}_{i,t+1}^f$ of each random variable and the observation values $\mathbf{y}_{i,t+1}$ at time $t + 1$ are expressed in equations (1–2):

$$\mathbf{X}_{i,t+1}^f = \mathbf{M}_t(\mathbf{X}_{i,t}^a) + \mathbf{w}_{i,t}, \mathbf{w}_{i,t} \sim N(0, \mathbf{Q}_t), \quad (1)$$

$$\mathbf{y}_{i,t+1} = \mathbf{H}_t(\mathbf{X}_{i,t}^f) + \varepsilon_{i,t}, \varepsilon_{i,t} \sim N(0, \mathbf{R}_t), \quad (2)$$

where $\mathbf{X}_{i,t}^a$ stands for the analysis values of the random variable at time t , $\mathbf{M}_t(\cdot)$ is the nonlinear model operator, $\mathbf{Q}_t$ denotes the model error covariance matrix, and $\mathbf{w}_{i,t}$ is a random perturbation added to represent model uncertainty. We choose it to be the Gaussian white noise with the expectation values and variances equal to 0 and $\mathbf{Q}_t$, respectively. $\mathbf{R}_t$ is the covariance matrix of the observation errors. $\varepsilon_{i,t}$ is the Gaussian white noise with the expectation values and variances equal to 0 and $\mathbf{R}_t$, respectively. The Kalman gain matrix $\mathbf{K}_{t+1}$ at time $t + 1$ is calculated using equations (3–6):

$$\mathbf{K}_{t+1} = \mathbf{P}_{t+1}^f \mathbf{H}^T (\mathbf{H} \mathbf{P}_{t+1}^f \mathbf{H}^T + \mathbf{R}_{t+1})^{-1}, \quad (3)$$

$$\mathbf{P}_{t+1}^f \mathbf{H}^T = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{X}_{i,t+1}^f - \overline{\mathbf{X}_{t+1}^f}) \cdot (\mathbf{H}(\mathbf{X}_{i,t+1}^f) - \mathbf{H}(\overline{\mathbf{X}_{t+1}^f}))^T, \quad (4)$$

$$\mathbf{H} \mathbf{P}_{t+1}^f \mathbf{H}^T = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{H}(\mathbf{X}_{i,t+1}^f) - \mathbf{H}(\overline{\mathbf{X}_{t+1}^f})) \cdot (\mathbf{H}(\mathbf{X}_{i,t+1}^f) - \mathbf{H}(\overline{\mathbf{X}_{t+1}^f}))^T, \quad (5)$$

$$\overline{\mathbf{X}_{t+1}^f} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_{i,t+1}^f, \quad (6)$$

where N is the size of the ensemble, $\overline{\mathbf{X}_{t+1}^f}$ is the mean of the forecast values at time $t + 1$, and $\mathbf{H}$ is the observation operator.

In the analysis step, the observation and forecast values are weighted to obtain the best estimation values at time $t + 1$. The model is reinitialized according to the current state, repeating the aforementioned two steps until all observation values are updated. The mean $\overline{\mathbf{X}_{t+1}^a}$ of the analysis values and the background field error covariance matrix $\mathbf{P}_{t+1}^a$ at time $t + 1$ are calculated by using equations (7–8):

$$\mathbf{X}_{i,t+1}^a = \mathbf{X}_{i,t+1}^f + \mathbf{K}_{t+1} [\mathbf{y}_{i,t+1} - \mathbf{H}(\mathbf{X}_{i,t+1}^f) + \mathbf{v}_{i,t+1}], \mathbf{v}_{i,t+1} \sim N(0, \mathbf{R}_t), \quad (7)$$

$$\overline{\mathbf{X}_{t+1}^a} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_{i,t+1}^a, \quad (8)$$

where $\mathbf{X}_{i,t+1}^a$ is the analysis value at time $t + 1$, and $\mathbf{v}_{i,t+1}$ is the Gaussian white noise with the expectation values and variances equal to 0 and $\mathbf{R}_t$, respectively. This is the principle used by the EnKF in one cycle.

2.2 Data-driven model and DD-DA approach

In EnKF, the forecast values are obtained from a physical model (Lorenz-63 or Lorenz-96). The general idea of the DD-DA approach is to replace the dynamical model applied to each ensemble member by a data-driven model applied to each ensemble member. A diagram of two assimilation procedures is shown in Figure 1, and it includes one assimilation cycle. The difference between EnKF and DD-DA is that the former uses a dynamics physical model, whereas the latter uses a data-driven model that comprises a dataset. This data-driven model is composed of representative data catalog sets [41]. For each continuous state variable, the catalog contains analogs and successors, which are formed by pairs of consecutive state vectors and separated by the same time interval. The second component of each pair is called the successor of its analogs. The catalog can be obtained from observations or numerical simulations. We chose the latter for this study.

(1) In the forecast step of DD-DA, we assume that there is a catalog that represents the simulation dynamics model. The size of the catalog is M , the size of the dimension is n , and the variance of the model error to generate the catalog is Ca . The catalog data are presented in Table 1. For instance, Table 2 shows the Lorenz-63 data catalog sets to intuitively understand the dataset values. It contains three dimensions ($n = 3$): $x(t)$, $y(t)$, and $z(t)$. $x(t)$ is related to convection, $y(t)$ represents horizontal temperature variation, and $z(t)$ represents vertical temperature variation. Each dimension has M continuous state variables. We can see that the successor values at time $t + 1$ correspond to the analog values at time t . This relationship between analogs and successors also corresponds to inputs and outputs in training sets.

Following the idea from the EnKF method in Section 2.1, the analysis value $\mathbf{X}_{i,t}^a$ means that using

Figure 1: EnKF and DD-DA, including one assimilation cycle. (Top) EnKF: during the forecast step, the model (Lorenz-63 or Lorenz-96) is integrated into the time of the next observation. During the analysis step, the forecast value is updated using the input observation to compute the analysis. (Bottom) DD-DA: The model is replaced by a data-driven model, which is composed of representative data catalog sets. The forecast approach is coupled K -NN with LWLR.

the state variables over time, we can find K samples that are closest to $\mathbf{X}_{i,t}^a$ at time t using the K -nearest neighbors (K -NN) algorithm [42–44]. Kailei proposed coupling of the K -NN procedure with KF for real-time updating of the hydraulic model [45]. This algorithm is one of the most basic methods in data mining technology and can be used for classification as well as regression. Compared with the general naive Bayesian

algorithm (mainly include Polynomial Naive Bayesian, Gaussian Naive Bayesian, and Bernoulli naive Bayesian) [46,47], this approach has a high degree of accuracy. In this case, the samples can be expressed as $(\mathbf{X}_{j,a}, \mathbf{X}_{j,s})$, $j = 1, 2, \dots, K$ [48–50]. We can assign different weights to the different neighbors at different distances and use K neighbors to represent the

Table 1: Samples of data catalog sets (analog and successors): it contains n dimensions; each dimension has M continuous state variables; A_{nM} represents one of the catalog data

Analog				Successors			
A_{00}	A_{10}	$\dots$	A_{n0}	A_{01}	A_{11}	$\dots$	A_{n1}
A_{01}	A_{11}	$\dots$	A_{n1}	A_{02}	A_{12}	$\dots$	A_{n2}
A_{02}	A_{12}	$\dots$	A_{n2}	A_{03}	A_{13}	$\dots$	A_{n3}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
A_{0M-2}	A_{1M-2}	$\dots$	A_{nM-2}	A_{0M-1}	A_{1M-1}	$\dots$	A_{nM-1}
A_{0M-1}	A_{1M-1}	$\dots$	A_{nM-1}	A_{0M}	A_{1M}	$\dots$	A_{nM}

Table 2: Lorenz-63 data catalog sets (analog and successors) containing three dimensions; each dimension has M continuous state variables ($M = 1,000$)

Analog (t)			Successors ($t + 1$)		
$x(t)$	$y(t)$	$z(t)$	$x(t + 1)$	$y(t + 1)$	$z(t + 1)$
−6.7456	−11.710	14.0628	−7.2584	−12.548	14.5310
−7.2584	−12.548	14.5310	−7.8034	−13.411	15.1135
−7.8034	−13.411	15.1135	−8.3794	−14.2873	15.8220
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
13.732	20.0323	26.2375	14.329	19.9331	28.3183
14.329	19.9331	28.3183	14.8456	19.5347	30.4189

Figure 2: Operation of K -NN in sample space. (a) State variable $X_{i,t}^a$ at time t (red dot) and a sample of data catalog sets (blue dots). (b) The K samples that are closest to $X_{i,t}^a$ ($K = 3$), in terms of Euclidean distance in sample space, are selected to form its neighborhood. (c) Three nearest neighbors can be found ($X_{1,a}$, $X_{2,a}$, and $X_{3,a}$).

samples. Euclidean distance is normally calculated using equation (9).

$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^K (\mathbf{x}_i - \mathbf{y}_i)^2}, \quad (9)$$

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i\}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_i\}$, $i = 1, 2, \dots, K$. In this study, the ensemble of the EnKF is chosen as N . There are two main strategies for selecting K . One predefined the number of the nearest neighbors and the other selected the analogs that are closer than the predefined threshold on distance. For simplicity, the first strategy is used in this study. To ensure the consistency of the ensemble of the two approaches, we defined the number of the nearest neighbors $K = N$. Figure 2 shows the operation of K -NN in sample space when $K = 3$.

For each variable at time t , we can use the nearest neighbor samples to calculate the possible forecast values at time $t + 1$. Traditional linear regression (LR) was used for the forecast step of the DA, which was called the analog ensemble KF (AnEnKF) [26]. Locally weighted linear regression (LWLR) can determine the parameters each time when a new sample is predicted to achieve better predictions. It makes up for the problems of underfitting and overfitting in the traditional LR models [48,51,52] and needs to relearn and keep the training samples all the time. LWLR builds a regression model for each ensemble member, so we adopted LWLR to calculate the forecast values in this study. Figure 3 compares the forecast results of the two methods. We can see that the forecast result of the LWLR is superior to LR. LWLR uses a “Gaussian kernel” to give higher weights to nearby neighbors.

Its weights are normally calculated using equation (10). A weight matrix with only diagonal elements $\mathbf{W}(j, i)$ is also called the exponential decay function.

$$\mathbf{W}(j, i) = \exp\left(-\frac{\|\mathbf{X}_{j,a} - \mathbf{X}_{i,t}^a\|^2}{2\tau^2}\right), \quad (i = 1 \dots N, j = 1 \dots K), \quad (10)$$

where $\|\cdot\|$ is the norm of $\mathbf{X}_{j,a} - \mathbf{X}_{i,t}^a$ and τ is the wavelength parameter controlling the speed of the weight values falling with distance. When τ is too large, the result is prone to underfitting. In contrast, when τ is too small, the result is prone to overfitting. In general, regularization is an effective method to determine the τ value. It is typically introduced to prevent the estimated function from overfitting the training

Figure 3: Comparison of the forecast results of two methods. Blue solid point circles represent training sets. The red and green lines represent the forecast results of LWLR and LR, respectively.

data, as well as avoid potential numerical issues in the course of solving the minimization problem [53]. In Figure 3, the green dotted line is the LR forecast result with $\tau = 1$, and the red solid line is the LWLR forecast result with $\tau = 0.02$. For general training sets $(\mathbf{X}_{1,a}, \mathbf{X}_{1,s}), \dots, (\mathbf{X}_{j,a}, \mathbf{X}_{j,s}), \dots, (\mathbf{X}_{K,a}, \mathbf{X}_{K,s})$, the regression coefficient parameter is expressed as follows:

$$\boldsymbol{\theta} = [(\theta_1(\mathbf{X}_{1,a}), \theta_1(\mathbf{X}_{1,s})), \dots, (\theta_j(\mathbf{X}_{j,a}), \theta_j(\mathbf{X}_{j,s})), \dots, (\theta_K(\mathbf{X}_{K,a}), \theta_K(\mathbf{X}_{K,s}))]^T$$

The LWLR loss function is expressed as follows:

$$L(\boldsymbol{\theta}) = \sum_{j=1}^K \mathbf{W}(j, i) (\mathbf{X}_{j,s} - \boldsymbol{\theta} \cdot (\mathbf{X}_{j,a}))^2. \quad (11)$$

One can solve the optimization problem of equation (11) as follows: $\nabla L(\boldsymbol{\theta}) = 0$. $\boldsymbol{\theta}$ is given by equation (12):

$$\boldsymbol{\theta} = (\mathbf{X}_{j,a}^T \mathbf{W}(j, i) \mathbf{X}_{j,a})^{-1} \mathbf{X}_{j,a}^T \mathbf{W}(j, i) \mathbf{X}_{j,s}. \quad (12)$$

We assign the weights to all neighbors near the forecast values. The weights are introduced to the loss function, and the size of the loss function is determined. Here, the regression coefficient parameter $\theta_j(\mathbf{X}_{j,a})$ represents the slope, $\theta_j(\mathbf{X}_{j,s})$ represents the intercept, and $\mathbf{Y}_j(\mathbf{X}) = \mathbf{X}_{j,s} - (\theta_j(\mathbf{X}_{j,a})\mathbf{X}_{j,a} + \theta_j(\mathbf{X}_{j,s}))$ represents the residual. The analog forecast values $\mathbf{X}_{i,t+1}^{af}$ by Gaussian sampling are shown in equation (13):

$$\mathbf{X}_{i,t+1}^{af} \sim N(\theta_j(\mathbf{X}_{j,a})\mathbf{X}_{i,a} + \theta_j(\mathbf{X}_{j,s}), \text{cov}(\mathbf{X}_{j,s} - (\theta_j(\mathbf{X}_{j,a})\mathbf{X}_{j,a} + \theta_j(\mathbf{X}_{j,s}))), \quad (13)$$

where the expectation values are $\theta_j(\mathbf{X}_{j,a})\mathbf{X}_{i,a} + \theta_j(\mathbf{X}_{j,s})$, and the variances are $\mathbf{P}_{t+1}^{af} = \text{cov}(\mathbf{X}_{j,s} - (\theta_j(\mathbf{X}_{j,a})\mathbf{X}_{j,a} + \theta_j(\mathbf{X}_{j,s})))$.

- (2) In the analysis step of DD-DA, the observation and forecast values are weighted to obtain the best estimation values at time $t + 1$. The mean of the analog analysis values, $\overline{\mathbf{X}}_{t+1}^{aa}$, and the analog background field error covariance matrix, $\mathbf{P}_{t+1}^{aa}$, at time $t + 1$ are calculated using equations (14–17):

$$\mathbf{K}_{t+1}^a = \mathbf{P}_{t+1}^{af} \mathbf{H}^T (\mathbf{H} \mathbf{P}_{t+1}^{af} \mathbf{H}^T + \mathbf{R}_t)^{-1}, \quad (14)$$

$$\mathbf{X}_{i,t+1}^{aa} = \mathbf{X}_{i,t+1}^{af} + \mathbf{K}_{t+1}^a [\mathbf{y}_{i,t+1} - \mathbf{H}(\mathbf{X}_{i,t+1}^{af}) + \mathbf{v}_{i,t+1}], \mathbf{v}_{i,t+1} \sim N(0, \mathbf{R}_t), \quad (15)$$

$$\overline{\mathbf{X}}_{t+1}^{aa} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_{i,t+1}^{aa}, \quad (16)$$

$$\mathbf{P}_{t+1}^{aa} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{X}_{i,t+1}^{aa} - \overline{\mathbf{X}}_{t+1}^{aa})(\mathbf{X}_{i,t+1}^{aa} - \overline{\mathbf{X}}_{t+1}^{aa})^T, \quad (17)$$

where $\mathbf{X}_{i,t+1}^{aa}$ is the analog analysis value at time $t + 1$, $\mathbf{K}_{t+1}^a$ is the analog Kalman filter gain, and $\mathbf{v}_{i,t+1}$ is the Gaussian white noise with the expectation values and variances equal to 0 and $\mathbf{R}_t$, respectively. This is the principle used by the DD-DA approach in one cycle.

2.3 Metrics for assimilation evaluation

We use the root mean square error (RMSE), as shown in equation (18), to quantify the overall assimilation performance in the total assimilation times. Because the analysis step remains unchanged in EnKF and DD-DA, the assimilation evaluation also reflects the performance of forecast results, and it is equivalent to the testing of the training process in the DD-DA approach.

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{t=0}^{m-1} (\overline{\mathbf{X}}_{t+1}^a - \mathbf{X}(t))^2}, \quad (18)$$

where m is the total assimilation times and $\mathbf{X}(t)$ is the variables of the true states (Lorenz-63 or Lorenz-96) at time t .

We also use the RMSE- f to reflect the ensemble mean of the forecast values at time t :

$$\text{RMSE-}f = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (\mathbf{X}_{i,t}^f - \mathbf{X}(t))^2}. \quad (19)$$

3 Evaluation of the assimilation approaches

3.1 Numerical experiments in the Lorenz-63 model

The Lorenz-63 model is a nonlinear spectral model found in the study of the finite-amplitude convection of a fluid and was proposed by Lorenz and Saltzman in 1963 [54]. Because of its nonlinear chaotic characteristics and low dimension, it is often used to verify the performance of data assimilation systems. The Lorenz-63 model is defined as shown in equation (20):

$$\begin{aligned} \frac{dx(t)}{dt} &= \sigma(y(t) - x(t)) \\ \frac{dy(t)}{dt} &= rx(t) - y(t) - x(t)z(t) \\ \frac{dz(t)}{dt} &= x(t)y(t) - bz(t), \end{aligned} \quad (20)$$

Figure 4: RMSE for two approaches using different M (DD-DA and AnEnKF). The horizontal axis represents catalog sizes, and the vertical axis represents RMSE. The black dotted line represents $\text{RMSE}_{\text{EnKF}}$, the red dotted line represents $\text{RMSE}_{\text{AnEnKF}}$, and the red solid line represents $\text{RMSE}_{\text{DD-DA}}$.

In this configuration, $\sigma = 10$, $r = 28$, $b = 8/3$, and the integration of the system is based on fourth-order Runge-Kutta explicit iterative methods [51]. The integral time step is $t = 0.01$ units, the total assimilation times is $m = 1,000$, and the size of the dimension is $n = 3$. The observation time step is $\text{Obs} = 0.08$ units (every 8 integration time steps or 6-hour time step in the atmosphere). The observed error covariance is 2. These configurations have been used in previous studies. We choose variable $x(t)$ to verify the performance of prediction and assimilation. The performance of the DD-DA approach is examined in comparison to that of EnKF.

3.1.1 Effectiveness of the DD-DA approach

To verify the correctness of the new approach, we compared the performance of our proposed approach to that of the former approach (AnEnKF) [26]. To avoid divergence of the filtering methods, we choose $N = 50$. The catalog noise variance $\text{Ca} = 0$. As shown in Figure 4, choosing the size of the catalog $M = \{10^3, 10^4, 10^5, \text{ and } 10^6\}$, we can see that (1) The RMSE of both approaches decreases when the size of the catalog increases. (2) The results of AnEnKF are closer to those of EnKF when $M = 10^6$. However, DD-DA performs slightly better than AnEnKF when $M \geq 10^5$. These results were similar to those obtained by Lguensat using the Lorenz-63 model. In conclusion, the effectiveness of the new DD-DA approach was demonstrated for this low-dimensional system.

3.1.2 Performance of DD-DA with different catalog sizes

To investigate the performance of our approach with the catalog sizes, we experimented with two approaches. Figure 5 compares the performance of DD-DA with EnKF when $M = \{10^3, 10^5\}$. Through the experiments, we found

that $\text{RMSE}_{\text{EnKF}} = 0.927$; furthermore, $\text{RMSE}_{\text{DD-DA}} = 1.104$ when $M = 10^3$ (top), and $\text{RMSE}_{\text{DD-DA}} = 0.921$ when $M = 10^5$ (bottom). This shows that the performance of both approaches is significantly satisfactory in terms of an overall low RMSE. The DD-DA approach performs almost as good as that of EnKF, especially when catalog sets become larger. The partial magnification also reflects this performance. The ensemble means of the forecast values in two approaches are compared in Figure 6. It also reflects that RMSE_f decreases with the increase of datasets. These results demonstrate that a data-driven algorithm could be used here as a model-free approach to search for optimal forecast values within the best range of catalog sets.

3.1.3 Performance of DD-DA with different catalog sizes and noise variances

Another group of experiments with $M = \{10^3, 10^5\}$ was conducted using a catalog noise variance of $\text{Ca} = 0.1$, which is different from the one used in the numerical simulation. Other configurations are the same as in the previous experiments. Figure 7 compares the performance of the DD-DA approach with that of EnKF. Through the experiments, we found that $\text{RMSE}_{\text{DD-DA}} = 1.373$ when $M = 10^3$ (top), and $\text{RMSE}_{\text{DD-DA}} = 1.099$ when $M = 10^5$ (bottom). Although the noise variances are added in the catalog, the RMSE of the DD-DA also decreases when M is larger. We also compared more configurations where the size of the catalog $M = \{10^3, 10^4, 10^5, 10^6\}$, and the noise variances $\text{Ca} = \{0, 0.1, 0.2\}$. Figure 8 shows the RMSE of the two approaches in more situations. From the aforementioned comparisons, the assimilation performance is not intuitively obvious when the catalog noises become increasingly severe. However, with a larger catalog size, the performance significantly improves especially when $M > 10^5$.

Figure 5: Performance comparison of two approaches ($Ca = 0$, $M = \{10^3, 10^5\}$). Left column represents the size of the catalog: the phase diagram between analogs and successors. Right column represents the assimilation results of the two approaches: (top) $M = 10^3$, (bottom) $M = 10^5$. The black solid lines, the blue solid lines, the green solid points, and the red solid lines represent the true states, the EnKF assimilation results, the observations, and the DD-DA results, respectively.

Figure 6: The ensemble mean of the forecast values in two approaches ($M = \{10^3, 10^5\}$). Red and blue lines represent the RMSE- f of the EnKF and DD-DA, respectively.

3.1.4 Performance of DD-DA with different catalog sizes and observation steps

We further compared the performance of the DD-DA approach with that of EnKF built using different observation steps Obs to demonstrate the effectiveness of using sparse observations. Figure 9 demonstrates the scatter plots between two consecutive values of the variable $x(t)$ and the assimilation results. We also calculated the

RMSE of the two approaches (DD-DA and EnKF) in the different M and Obs to evaluate the adaptability of the different models. The comparison of RMSE values of different approaches is presented in Table 3. The results demonstrate that (1) the two approaches are outperformed when the Obs is smaller ($= 0.01$), and it is difficult to determine which is better; (2) the RMSE of the DD-DA approach decreases when the size of the catalog is larger; and (3) $RMSE_{DD-DA}$ is clearly lower than $RMSE_{EnKF}$ when

Figure 7: Performance comparison of two approaches ($Ca = 0.1$, $M = \{10^3, 10^5\}$). (Top) $M = 10^3$, (bottom) $M = 10^5$. The black solid lines, the blue solid lines, the green solid points, and the red solid lines represent the true states, the EnKF assimilation results, the observations, and the DD-DA results, respectively.

M increases ($>10^5$), especially when $Obs \geq 0.04$. This conclusion is shown in Figure 10. Under the perfect-model assumption, the choice of the size of the catalog, M , does not seem very critical to the DD-DA performance, as long as the observation step Obs is short ($Obs \leq 0.01$). However, when the observations step Obs is longer,

the performance of DD-DA is superior to that of EnKF, and the advantage of DD-DA over EnKF becomes quite obvious, especially when the size of the catalog is $M \geq 10^5$. Therefore, the DD-DA approach can make up for insufficient observations. It also shows that the choice of the size of the catalog plays a significant

Figure 8: RMSE for two approaches using different M and Ca values (DD-DA and EnKF). The horizontal axis represents catalog size, and the vertical axis represents RMSE. The black dotted line represents $RMSE_{EnKF}$, and solid lines represent $RMSE_{DD-DA}$ ($Ca = 0, 0.1$ and 0.2 , respectively).

Figure 9: Performance comparison of two approaches ($\text{Obs} = \{0.01, 0.04, 0.16\}$, $M = \{10^3, 10^4, 10^5\}$, and $\text{Ca} = 0$). The left column represents the scatter plots between two consecutive values of the variable $x(t)$. The right column displays the assimilation results. (Top) $M = 10^3$, $\text{Obs} = 0.01$; (middle) $M = 10^4$, $\text{Obs} = 0.04$, and (bottom) $M = 10^5$, $\text{Obs} = 0.16$.

role in determining whether DD-DA could be successfully implemented.

3.1.5 Compared with other approaches

To verify the forecasting performance of DD-DA approach, we replace EnKF with ensemble Kalman smoother (EnKS)

and PF. EnKS is an extension of EnKF, which can update state variables using all observation information in the entire assimilation windows. PF is another nonparametric method, which would try to reproduce the real posterior distribution. The configuration is $M = \{10^3, 10^5\}$, the size of the ensemble (particles) is 50, and $\text{Ca} = 0$. Figure 11 compares the ensemble mean of the forecast values in these approaches. The results also reflect that $\text{RMSE-}f$ decreases with the increase of datasets.

Table 3: RMSE of two approaches (EnKF and DD-DA) with different M and Obs

M	Obs	$RMSE_{EnKF}$	$RMSE_{DD-DA}$
10^3	0.01	0.176	0.238
	0.04	0.431	0.668
	0.16	1.242	1.754
10^4	0.01	0.176	0.186
	0.04	0.431	0.523
	0.16	1.242	1.421
10^5	0.01	0.176	0.175
	0.04	0.431	0.457
	0.16	1.242	1.275
10^6	0.01	0.176	0.172
	0.04	0.431	0.404
	0.16	1.242	1.201

3.2 Numerical experiments in the Lorenz-96 model

The Lorenz-96 system is a simplified dynamic model derived from the dynamic model by Lorenz and Emanuel in 1996 [55]. It is also widely used to verify various assimilation algorithms. The Lorenz-96 model expression is as follows:

$$\frac{dx_j(t)}{dt} = -x_{j-2}(t)x_{j-1}(t) + x_{j-1}(t)x_{j+1}(t) - x_j(t) + F, \quad j = 0, 1, 2, \dots, \quad (21)$$

The model variables (the number is j) are uniformly distributed on a latitude circle. They embody several basic features of the atmosphere, such as nonlinear advection-like terms, a damping term, and an external forcing. In this experiment, we set $j = \{0, 1, \dots, 39\}$ and $F = 8$. For computational stability, the time step was 0.05 units,

corresponding to a 6-hour time step in the atmosphere. A fourth-order Runge-Kutta scheme was used for temporal integration in this study. These configurations have also been used in previous studies.

First, the other parameter are observation step $Obs = 0.2$, observed error covariance is 2, size of the catalog $M = 10^4$, size of the ensemble $N = 100$, and the catalog noise variance $Ca = 0$. RMSE was used to reflect the performance of data assimilation. Figure 12 compares the performance of DD-DA with that of EnKF. The absolute errors between the true values and the assimilation results were also calculated in these experiments. In these configurations, $RMSE_{EnKF}$ and $RMSE_{DD-DA}$ were equal to 1.012 and 1.282, respectively. These results show that DD-DA performs worse than EnKF, especially when j is close to 2 and 17. Another group of experiments, with $M = 10^6$ and $N = 200$, was also conducted in the numerical simulation. Other configurations are the same as in the previous experiments. Figure 13 shows the performance of the two approaches in these configurations. $RMSE_{EnKF}$ and $RMSE_{DD-DA}$ were equal to 0.971 and 0.949, respectively. In conclusion, the absolute errors and RMSEs demonstrate that with an increase in catalog size, the performance of the DD-DA approach improves significantly, almost as good as that of EnKF.

For a more comprehensive verification of the DD-DA performance in the Lorenz-96 model, we also designed another group of experiments using the size of the catalog $M = \{10^3, 10^4, 10^5, 10^6, 10^7\}$, size of the ensemble $N = \{50, 100, 200, 500\}$, and the observation step $Obs = \{0.05, 0.2, 0.5\}$. The RMSEs for the two approaches with these configurations are shown in Figure 14. With the increase in the size of the ensemble N , EnKF performs better than before. DD-DA can also perform better than EnKF if M is increased to a certain degree ($M \geq 10^6$). At the

Figure 10: RMSE for two approaches using different M and Obs. The horizontal axis represents the size of the catalog, and the vertical axis represents the RMSE. The dotted lines represent $RMSE_{EnKF}$ and the solid lines represent $RMSE_{DD-DA}$ ($Obs = 0.01, 0.04, 0.08$, and 0.16 , respectively).

Figure 11: Comparison of the ensemble mean of the forecast values in EnKS and PF.

same time, similar superior characteristics occur if the observation step *Obs* becomes longer. Therefore, the DD-DA approach makes up for the defect of the longer observation step.

4 Discussion

Several numerical experiments performed using the Lorenz-63 and Lorenz-96 models show that the proposed DD-DA approach performs better than EnKF when the size of the catalog is larger. This approach restricts the numerical experiments to the EnKF form. However, the basic strategy is not restricted to any other version of KF. Several key characteristics were analyzed and proposed in this study.

4.1 Data-driven approaches

The data-driven model also realizes the model's reconstructions, analyses, and understanding. We believe that this technology opens new research avenues for future nonlinear dynamic analysis. In this study, the data-driven model comprises representative data catalog sets. It contains analogs and successors that are formed by pairs of consecutive state vectors and separated by the same time interval. Successors are used for forecasting the subsequent time values. The catalog is obtained from the numerical simulations. Therefore, these prediction approaches are also called model-free filters.

The data-driven approach will be also combined with evolutionary computation to solve optimization problems, such as genetic algorithms, particle swarm optimization,

and differential algorithms. The purpose of data-driven models is to approximate real problems. Data-driven models are used to evaluate the solution instead of the optimization problem. A suitable solution to a real problem is preselected to reduce the evaluation time of the real problem.

Compared with the study by Brajard *et al.* [28], there are two differences between the data-driven models: (1) Model training comes from different sources. Brajard *et al.* used the forecast values and analysis values as input and output of the neural network, respectively. The aim of the proposed approach is to correct the model from sparse and noisy observations. Our approach used analogs and successors as the input and output for the training sets, respectively; it implemented an analog model. (2) These two approaches used different machine learning algorithms. Brajard *et al.* used a convolutional neural network (CNN), while we used *K*-NN and LWLR. Because of the different effects in the data-driven model, it is difficult to identify the better approach; however, both approaches provide useful ideas using a hybrid numerical correction model in small-ensemble and high-dimension data assimilation.

4.2 Combining *K*-nearest neighbors machine learning and locally weighted linear regression

The advantage of the *K*-NN algorithm is that it is easy to understand and implement. Because of its high accuracy rates and simple regression principle, the *K*-NN algorithm is currently one of the most widely used search algorithms. Furthermore, independent learning

Figure 12: Performance comparison of the two approaches ($M = 10^4$, $N = 100$). The horizontal axis represents $x_j(t)$ $j = \{0, 1, \dots, 39\}$, and the vertical axis represents the time of the Lorenz-96 model: (top left) the true trajectory, (top right) the observed trajectory, (middle) the assimilation results of the two approaches, and (bottom) the absolute errors of two approaches.

processes can be decomposed; therefore, K -NN algorithms are particularly suitable for realizing parallel computation

in nonlinear models. It is important to choose the size of K optimally in K -NN algorithms. If the value of K is too

Figure 13: Performance comparison of the two approaches ($M = 10^6$, $N = 200$). The horizontal axis represents $x_j(t)$ $j = \{0, 1, \dots, 39\}$, and the vertical axis represents the time of the Lorenz-96 model: (top) the assimilation results of two approaches, and (bottom) the absolute errors of two approaches.

small, the presence of noise will have a significant impact on the prediction. For instance, there will be deviations when the nearest value is noise when $K = 1$, and a decrease in the value of K means that the overall model becomes complex and prone to overfitting. Similarly, it is necessary to use the training sets in a large neighborhood when the value of K is too large. Here, the approximate errors of learning increase, and the values that are far away from the input also affect the prediction. When the size of K is too large, the overall model becomes simple. For instance, we choose the size of the ensemble $N = 50$ and the size of the catalog $M = 10^5$ in Lorenz-63. We compared the RMSE of the DD-DA when $K = \{3, 20, 50, 100, 200\}$. From Table 4, we can see that when K is too small or too large, the assimilation performance becomes increasingly severe.

The traditional LR algorithm is a learning method based on parameters with fixed values. However, in several cases, there is a significant deviation between numerical simulations and LR models. Therefore, inaccurate and wrong data often affect the forecast results. LWLR is a nonparametric learning algorithm that aims to address some of the inevitable drawbacks in traditional LR models and overcome the problems of underfitting and overfitting [52,56]. Owing to the weighting factor of each data point, the impact of useful data increases, whereas that of the deviated data decreases.

We propose a new analysis approach, which is based on K -NN using LWLR, combined with the characteristics of the two algorithms. We then use the data-driven model to realize sequential data assimilation. The results indicate that the DD-DA approach helps ameliorate the final

Figure 14: RMSE for the two approaches using different M , Obs , and N . Dashed lines represent $\text{RMSE}_{\text{EnKF}}$, and solid lines represent $\text{RMSE}_{\text{DD-DA}}$. $M = \{10^3, 10^4, 10^5, 10^6, 10^7\}$, $N = \{50, 100, 200, 500\}$, and $\text{Obs} = \{0.05, 0.2, 0.5\}$.

Table 4: RMSE of the DD-DA approach in Lorenz-63 using $K = \{3, 20, 50, 100, 200\}$

K	3	20	50	100	200
$\text{RMSE}_{\text{DD-DA}}$	2.153	1.151	0.921	0.962	1.917

assimilation results significantly, especially when the observation step is long or the size of the ensemble is small. The results also demonstrated that ML approaches perform well at solving complex nonlinear problems and big data assimilation.

4.3 Uncertainties

The fairly good performance of DD-DA that we observed in the experiments conducted using Lorenz-63 and Lorenz-96 indicates its potential application in real numerical models (atmosphere, ocean, or land surface). However, as is the case with all data-driven models, there are still some uncertainties and limitations to the current study and its

modeling approaches. For instance, (1) although the DD-DA approach works well in Lorenz models, choosing an appropriate size for the catalog and ensemble still needs to be addressed. (2) This approach has not been verified in the high-dimensional model (quasi-geostrophic model, Noah-MP land surface model, and weather research and forecasting model).

The application of the DD-DA approach to larger systems is another important challenge for the future. In this study, the amount of data in a catalog should increase exponentially with the intrinsic dimension of the state to guarantee the retrieval of analogs at a given precision [57]. This forecast method is probably inappropriate for high-dimensional systems. Coupling the DD-DA strategies with multiscale decompositions is a very promising research direction [58]. Katzfuss used EnKF methods for dimension reduction [59]. By using real data and simulated data, it was shown that this approach outperforms those existing methods. For real high-dimensional application systems, the dimension reduction algorithm becomes part of the data preprocessing. The principal component analysis is the most commonly used linear

Table 5: Assimilation time comparison of the two approaches, EnKF and DD-DA, in Lorenz-63 (Unit: seconds)

M	10^3	10^4	10^5	10^6
DD-DA	19.57	22.65	113.52	1650.21
EnKF	25.23			

dimensionality reduction method. In addition, other methods are linear discriminant analysis, locally linear embedding and Laplacian eigenmaps.

4.4 Evaluation of assimilation times

A full cost–benefit analysis would be useful to assess the tradeoffs between data lookup and search and integrate an ensemble of models for EnKF. Under the same parameters and hardware conditions, the assimilation times of the two approaches in Lorenz-63 are compared in Table 5. We can see that the assimilation time become longer when the size of the catalog $\geq 10^5$. Therefore, how to reduce the assimilation times when the catalog is larger still needs to be addressed.

In the DD-DA approach, the computational cost of the forecast step is related to the cost of the K -NN step. The most popular K -NN algorithm is K -d trees [60] and Ball trees [61]. These algorithms expedite the K -NN search. In this study, we chose K -d trees to search nearest neighbors. Here, D denotes the dimension of the system, M denotes the size of the catalog, K denotes the number of nearest neighbors to be retrieved, and v denotes the size of the neighborhood used for the search for analogs. The amount of data on average is as follows:

$$M \geq K \frac{\ln(0.05)}{\ln(1 - \sigma^{2v+1})} \approx \frac{3K}{\sigma^{2v+1}}, \quad (22)$$

where σ is the integral of the standard Gaussian probability density function. In each forecast step, the computational cost contains two parts: the K -NN search and the creation of the K -d trees. As an example, take the Lorenz-63 model: $D = 3$, $v = 2$, $K = 50$, and $\sigma = 0.15$. We would only need about $M \approx 1 \times 10^5$ samples. Combining the data-driven approaches with dimension reduction algorithms would be a very meaningful study.

5 Conclusions and expectations

In this study, the data-driven model approach was combined with EnKF to exploit the complementary strengths

of the two methodologies. Although the update phase of KF is proposed in the known kinetic evolution equations, we proved there was little or no knowledge of the form of the population distribution. Therefore, it is necessary to use statistical inference methods. The most basic method to solve nonparametric problems is to use ML or ML models, which can complement traditional physical modeling. So, a nonparametric sampler from a catalog of historical datasets, namely, a nearest neighbor or analog sampler, was given by numerical simulations. Based on this catalog, the dynamics physics model was reconstructed using the K -NN algorithm. The optimal values of the surrogate model were found, and the forecast step was performed using LWLR.

In recent years, our ability to collect data has improved significantly. However, there is still insufficient data regarding effective techniques of data mining. It is mainly reflected in two aspects: extracting useful information from data and obtaining better models from data [62,63]. Currently, ML methods have significantly developed with regard to studying the atmosphere, biology, land, space, and oceans. They have become a universal method for the classification and detection of changes and anomalies in Earth science [64,65]. Particularly, deep learning neural networks [66] are being increasingly used in geosciences to make better use of spatial and temporal structures in data [40]. Therefore, ML methods undoubtedly bring significant improvements to classification and forecasting problems. The data-driven ML methods can also mine new information from data that are yet to be known and promote the generation of new mechanisms and new understandings [67–69]. It is useful in the big data era of geosciences. However, there are many problems, such as identifying the particularity of the data, the interpretability of inference, the estimation of uncertainty, and the verification of complex physical patterns. Physical models and ML will be further combined in the future. They will play a complementary and crucial role in the final implementation of hybrid modeling. ML and artificial intelligence technology also represent new opportunities and challenges for the analysis of geoscience data.

Acknowledgements: The authors would like to thank the valuable comments and suggestions of the anonymous reviewers.

Funding information: This research was funded by the NSFC (National Natural Science Foundation of China) project (grant number: 41861047, 41461078), the Northwest Normal University Scientific research capability upgrading program (No. NWNNU-LKQN-17-06), the Foundation of Northwest Normal University of China (No. NWNNU-LKQN2020-21), the

Scientific Research Program of the Higher Education Institutions of Gansu Province (2020A-016), and the Foundation of Northwest Normal University of China (No. NWNNU-LKQN2019-18).

Author contributions: Conceptualization: M.F. and Y.B.; formal analysis: L.W. and L.T.; methodology: Y.B., L.D., and M.F.; resources: M.F.; supervision: Y.B.; validation: Y.B. and M.F.; writing – original draft preparation: M.F.; writing – review and editing: Y.B.; funding acquisition: Y.B. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: Authors state no conflict of interest.

Citing: <https://github.com/ptandeo/AnDA>. This Python library is attached to the following publication: Lguensat R, Tandeo P, Ailliot P, Pulido M & Fablet R. The Analog Data Assimilation. *Mon Wea Rev.* 2017; 145(10): 4093–4107.

References

- [1] Ren D. Adjoint retrieval of prognostic land surface model variables for an NWP model: assimilation of ground surface temperature. *Open Geosci.* 2010;2:83–102.
- [2] Asch M, Bocquet M, Nodet M. Data assimilation: Methods, algorithms, and applications. Series: Fundamentals of Algorithms. Philadelphia, USA: SIAM; 2016.
- [3] Bannister RN. A review of operational methods of variational and ensemble-variational data assimilation. *Q J Roy Meteor Soc.* 2017;143:607–33.
- [4] Han YQ, Zhang YC, Wang YF, Ye S, Fang HX. A new sequential data assimilation method. *Sci China Ser E-Tech Sci.* 2009;52:1027–38.
- [5] Bai YL, Li X, Han XJ. A review of error problems for land data assimilation systems. *Adv Earth Sci.* 2011;26:795–804.
- [6] Reichle RH. Data assimilation methods in the Earth sciences. *Adv Water Resour.* 2008;31:1411–8.
- [7] Evensen G. Data assimilation: the ensemble Kalman filter. 2nd edn. Springer-Verlag Berlin Heidelberg; 2009.
- [8] Hoteit I, Luo XD, Bocquet M, Köhl A, Ait-El-Fquih B. Data assimilation in oceanography: Current status and new directions. In: Chassignet E, Pascual A, Tintoré J, Verron J, editors. *New frontiers in operational oceanography*. GODAE Ocean View. 2018. p. 465–512. doi: 10.17125/gov2018.ch17.
- [9] Carrassi A, Bocquet M, Bertino L, Evensen G. Data assimilation in the geosciences – an overview on methods, issues and perspectives. *WIREs Clim Change*; 2018.
- [10] Luo X, Hoteit I. Robust ensemble filtering and its relation to covariance inflation in the ensemble Kalman filter. *Mon Wea Rev.* 2011;139(12):3938–53. doi: 10.1175/MWR-D-10-05068.1.
- [11] Bai YL, Li X. Evolutionary algorithm-based error parameterization methods for data assimilation. *Mon Wea Rev.* 2011;139:2668–85.
- [12] Marciniak A, Stan-Kłeczek I, Idziak A, Majdański M. Uncertainty based multi-step seismic analysis for near-surface imaging. *Open Geosci.* 2019;11:727–37.
- [13] Zawadzki J, Kędzior M. Statistical analysis of soil moisture content changes in Central Europe using GLDAS database over three past decades. *Open Geosci.* 2014;6:344–53.
- [14] Fablet R, Huynh Viet P, Lguensat R, Horrein PH, Chapron B. Spatio-temporal interpolation of cloudy SST fields using conditional analog data assimilation. *Remote Sens.* 2018;10:310.
- [15] Ruiz J, Saulo AC, Nogués-Paegle J. WRF model sensitivity to choice of parameterization over South America: validation against surface variables. *Mon Wea Rev.* 2010;138:3342–55.
- [16] Lott F, Miller M. A new subgrid-scale orographic drag parameterization: its formulation and testing. *Q J Roy Meteor Soc.* 1997;123:101–27.
- [17] Luo X, Bhakta T, Jakobsen M, Nævdal G. Efficient big data assimilation through sparse representation: a 3D benchmark case study in petroleum engineering. *PLoS One.* 2018;13(7):e0198586. doi: 10.1371/journal.pone.0198586.
- [18] Miyoshi T, Kunii M, Ruiz J, Lien GY, Satoh S, Ushio T, et al. “Big data assimilation” revolutionizing severe weather prediction. *B Am Meteorol Soc.* 2016;97(8):1347–54. doi: 10.1175/BAMS-D-15-00144.
- [19] Soares RV, Luo X, Evensen G, Tuhin B. Handling big models and big data sets in history-matching problems through an adaptive local analysis scheme. *SPE J.* 2021;26(2):973–92. doi: 10.2118/204221-PA.
- [20] Hamill TM, Whitaker JS. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon Wea Rev.* 2006;134(11):3209–29.
- [21] Delle Monache L, Nipen T, Liu Y, Roux G, Stull R. Kalman filter and analog schemes to post-process numerical weather predictions. *Mon Wea Rev.* 2011;139:3554–70.
- [22] Delle Monache L, Eckel T, Rife D, Nagarajan B, Searight K. Probabilistic weather predictions with an analog ensemble. *Mon Wea Rev.* 2013;131:3498–516.
- [23] Hamilton F, Berry T, Sauer T. Ensemble Kalman filtering without a model. *Phys Rev X.* 2016;6:011021.1–12.
- [24] Hamilton F, Berry T, Sauer T. Predicting chaotic time series with a partial model. *Phys Rev E.* 2015;92:010902.1–5.
- [25] Pathak J, Hunt B, Girvan M, Lu Z, Ott E. Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach. *Phys Rev L.* 2018;120:024102.1–15.
- [26] Lguensat R, Tandeo P, Ailliot P, Pulido M, Fablet R. The analog data assimilation. *Mon Wea Rev.* 2017;145:4093–107.
- [27] Fablet R, Ouala S, Herzet C. Bilinear residual neural network for the identification and forecasting of dynamical systems. *European Signal Processing Conference: Rome, Italy*; 2017. p. 1–5.
- [28] Brajard J, Carrassi A, Bocquet M, Bertino L. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *J Comput Sci.* 2020;44:1877–7503.
- [29] Luo X, Lorentzen RJ, Bhakta T. Accounting for model errors of rock physics models in 4D seismic history matching problems: A perspective of machine learning. *J Pet Sci Eng.* 2021;196:107961. doi: 10.1016/j.petrol.2020.107961.
- [30] Luo X. Ensemble-based kernel learning for a class of data assimilation problems with imperfect forward simulators. *PLoS One.* 2019;14(7):e0219247. doi: 10.1371/journal.pone.0219247.

- [31] Cintra RS, de Campos, Velho HF. Data assimilation by artificial neural networks for an atmospheric general circulation model. arXiv preprint; 2018.
- [32] Bishop CH, Whitaker JS, Lei L. Gain form of the ensemble transform Kalman filter and its relevance to satellite data assimilation with model space ensemble covariance localization. *Mon Wea Rev.* 2017;145:4575–92.
- [33] Loh K, Omrani PS, van der Linden R. Deep learning and data assimilation for real-time production prediction in natural gas wells. arXiv preprint, 2018.
- [34] Miyoshi BT, Lien GY, Satoh S, Ushio T, Bessho K. “Big data assimilation” toward post-petascale severe weather prediction: An overview and progress. *Proc IEEE.* 2016;104:2155–79.
- [35] Park D, Zhu Y. Bilinear recurrent neural network. *IEEE ICNN’94.* 2002;3:1459–64.
- [36] Park DC. A time series data prediction scheme using bilinear recurrent neural network. Seoul, Korea (South): ICISA; 2010. p. 1–7. doi: 10.1109/ICISA.2010.5480383.
- [37] Arcomano T, Szunyogh I, Pathak J, Wikner A, Hunt BR, Ott E. A machine learning based global atmospheric forecast model. *Geophys Res Lett.* 2020;47:e2020GL087776.
- [38] De Campos Velho H, Stephany S, Preto A, Vijaykumar N, Nowosad A. A neural network implementation for data assimilation using MPI. In: Brebbia CA, Melli P, Zanasi A, editors. *Applications of high performance computing in engineering VII.* 2002;27:211–20.
- [39] Schäfer AM, Zimmermann HG. Recurrent neural networks are universal approximators. *ICANN2006.* 2006;4131:632–40.
- [40] Bocquet M, Brajard J, Carrassi A, Bertino L. Data assimilation as a deep learning tool to infer ODE representations of dynamical models. *Nonlin Process Geophys.* 2019;26:143–62.
- [41] Tandeo P, Ailliot P, Ruiz J, Hannart A, Chapron B. Combining analog method and ensemble data assimilation: Application to the Lorenz-63 chaotic system. In: Lakshmanan V, Gilleland E, McGovern A, Tingley M, editors. *Machine learning and data mining approaches to climate science.* Cham: Springer; 2015. p. 3–12. doi: 10.1007/978-3-319-17220-0_1.
- [42] Ugur D, Cyrus S, Farnoush B. Efficient K -nearest neighbor search in time-dependent spatial networks. Los Angeles, CA, US: University of Southern California; 2013.
- [43] Liu Y, Jing N, Chen L, Xiong W. Algorithm for processing k -nearest join based on R-tree in MapReduce. *J Softw.* 2013;24:1836–51.
- [44] Xue T, Li TT, Sun B. Research on parallelization of KNN locally weighted linear regression algorithm based on MapReduce. *J Commun Technol.* 2015;10:864–9.
- [45] Kailei L. Coupling the k -nearest neighbor procedure with the Kalman filter for real-time updating of the hydraulic model in flood forecasting. *Int J Sediment Res.* 2016;31:149–58.
- [46] He Y, Xie J, Xu C. An improved Naive Bayesian algorithm for Web page text classification. *FSKD 2011.* Shanghai, China; 2011. p. 1765–8. doi: 10.1109/FSKD.2011.6019801.
- [47] Langley P, Wayne I, Thompson K. An analysis of Bayesian classifiers. *Proceedings of the 10th National Conference on AI.* San Jose, California; 1992. p. 223–28.
- [48] Denoeux T. A k -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans Syst Man Cybern.* 1995;219:804–13.
- [49] Peterson L. K -nearest neighbor. *Scholarpedia.* 2009;4(2):1883. doi: 10.4249/scholarpedia.1883.
- [50] Park J, Bhuiyan MZA, Kang M, Son J, Kang K. Nearest neighbor search with locally weighted linear regression for heartbeat classification. *Soft Comput.* 2018;22:1225–36.
- [51] Dormand JR, Prince PJ. A family of embedded Runge–Kutta formulae. *J Comput Appl Math.* 1980;6:19–26.
- [52] Jhun M. Bootstrap choice of smoothing parameter of locally weighted linear regression. *J Jpn Soc Comp Stat.* 1993;6:25–32.
- [53] Luo X, Stordal AS, Lorentzen RJ, Geir N. Iterative ensemble smoother as an approximate solution to a regularized minimum-average-cost problem: theory and applications. *SPE J.* 2015;20:962–82. doi: 10.2118/176023-PA.
- [54] Anderson JL, Anderson SL. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon Wea Rev.* 1999;127:2741–58.
- [55] Lorenz E. Predictability: a problem partly solved. *Proc Seminar on Predictability ECMWF.* U Kingd. 1996;1:1–19.
- [56] Alrasheedi M. Parametric and non-parametric bootstrap: A simulation study for a linear regression with residuals from a mixture of Laplace distributions. *Eur Sci J.* 2013;9:120–31.
- [57] Van den Dool HM. Searching for analogues, how long must we wait? *Tellus A.* 1994;46:314–24.
- [58] Fablet R, Viet PH, Lguensat R, Chapron B. Data-driven assimilation of irregularly-sampled image time series. *IEEE ICIP 2017; 2017.* p. WQ-PB.2.
- [59] Katzfuss M, Stroud JR, Wikle CK. Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. *J Am Stat Assoc.* 2018;115:866–85.
- [60] Bentley JL. Multidimensional binary search trees used for associative searching. *Commun ACM.* 1975;18:509–17.
- [61] Omohundro SM. Five balltree construction algorithms. Berkeley, California, USA: International Computer Science Institute; 1989.
- [62] Kalnay E. *Atmospheric modeling, data assimilation and predictability.* Cambridge: Cambridge University Press; 2002. doi: 10.1017/CBO9780511802270.
- [63] Lguensat R, Viet PH, Sun M, Chen G, Fenglin T, Chapron B, et al. Data-driven interpolation of sea level anomalies using analog data assimilation. *Remote Sens.* 2019;11:858.
- [64] Hong SY, Dudhia J. Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and large eddies. *Bull Am Meteorol Soc.* 2012;93:ES6–9.
- [65] Nagao H. What is required for data assimilation that is applicable to big data in the solid Earth science? Importance of simulation-/data-driven data assimilation. 17th International Conference on Information Fusion, Salamanca, Spain; 2014.
- [66] Maimaitijiang M, Sagan V, Sidike P, Hartling S, Esposito F, Fritschi FB. Soybean yield prediction from UAV using multi-modal data fusion and deep learning. *Remote Sens Env.* 2020;237:111599.
- [67] Higuchi T. Embedding reality in a numerical simulation with data assimilation. *Proceedings of 14th International Conference on Information Fusion (FUSION);* 2011. p. 1–7.
- [68] Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N. Deep learning and process understanding for data-driven Earth system science. *Nature.* 2019;566:195–204.
- [69] Teweldebrhan AT, Burkhart JF, Schuler TV, Xu CY. Improving the informational value of MODIS fractional snow cover area using fuzzy logic based ensemble smoother data assimilation frameworks. *Remote Sens.* 2019;11:28.

Appendix A

EnKF and DD-DA algorithms

EnKF

```

-initialization (generate true state, observations)
for  $t$  in 1: $t$  do
  for  $i$  in 1: $N$  do (forecast step)
    -compute forecast values  $\mathbf{X}_{i,t+1}^f$  using
      equation (1)
    -compute observation values  $\mathbf{y}_{i,t+1}$  using
      equation (2)

  end

  -compute error covariance matrix  $\mathbf{P}_{t+1}^f$ 

  -compute The Kalman gain matrix  $\mathbf{K}_{t+1}$  using
    equations (3–5)
  for  $i$  in 1: $N$  do
    -compute the analysis values  $\mathbf{X}_{i,t+1}^a$  using
      equation (7)
  end
  -compute the mean of the analysis values
     $\overline{\mathbf{X}}_{t+1}^a$  using equation (8) (analysis step)
  -update the background field error
    covariance matrix  $\mathbf{P}_{t+1}^a$ 
end

```

DD-DA

```

-initialization (generate true state, observations, catalog)
for  $t$  in 1: $t$  do
  for  $i$  in 1: $N$  do (forecast step)
    -find the  $K$  neighbors in the catalog using
      equation (9) ( $K$ -NN step)

    -compute the weight matrix  $\mathbf{W}(j, i)$  using equation
      (10) (LWLR step)
    -compute the regression coefficient parameter  $\theta$ 
      using equation (12) (LWLR step)
    -compute the analog forecast values  $\mathbf{X}_{i,t+1}^{af}$  using
      equation (13) (LWLR step)

  end

  -compute the analog error covariance matrix  $\mathbf{P}_{t+1}^{af}$  of
    the  $\mathbf{X}_{i,t+1}^{af}$ 

  -compute The analog Kalman gain matrix  $\mathbf{K}_{t+1}^a$  using
    equation (14)
  for  $i$  in 1: $N$  do
    -compute the analog analysis values  $\mathbf{X}_{i,t+1}^{aa}$  using
      equation (15)
  end

  -compute the mean of the analog analysis values  $\overline{\mathbf{X}}_{t+1}^{aa}$ 
    using equation (16) (analysis step)
  -update the analog background field error covariance
    matrix  $\mathbf{P}_{t+1}^{aa}$  using equation (17)
end

```