ORIGINAL ARTICLE

Journal Section

Sequential machine learning for data assimilation with parameter estimation via hierarchical Bayesian optimization

Yang Sun[1*] | **Song Xi Chen**[2†] | **Yumou Qiu**[1*]

[1]School of Mathematical Sciences and and Center for Statistical Science, Peking University, Beijing, 100871, China

[2]Department of Statistics and Data Sciences, Tsinghua University, Beijing, 100084, China

Correspondence
Song Xi Chen, PhD, Department of Statistics and Data Sciences, Tsinghua University, Beijing, 100084, China;
Yumou Qiu, PhD, School of Mathematical Sciences and and Center for Statistical Science, Peking University, Beijing, 100871, China
Email: sxchen@tsinghua.edu.cn; qiuyumou@math.pku.edu.cn

Funding information

The performance of data assimilation critically depends on well-specified physical models, particularly accurate assignment of key parameters. Misspecified parameters can introduce model bias and uncertainty, degrading both forecasting and assimilation quality. This paper proposes a sequential data assimilation procedure with a hierarchical Bayesian machine learning approach for sequentially calibrating parameters in the physical model that integrates the ensemble Kalman filter. We define an objective function for parameter estimation by an optimally weighted distance between the observations and forecast states over a time segment. Based on this criterion, we develop a Bayesian learning algorithm to efficiently estimate the unknown parameters throughout the assimilation window. Statistical properties of the proposed method are developed. Numerical experiments on the Lorenz'96 system and a two-dimensional shallow water model demonstrate that the proposed method outperforms conventional approaches in terms of estimation accuracy and stability.

KEYWORDS
Data assimilation, ensemble Kalman filter, hierarchical Bayesian optimization, nonlinear system, machine learning, parameter estimation

1 | INTRODUCTION

In large-scale geophysical models such as those used in atmospheric and oceanic systems, key physical parameters exert significant influence on dynamic modeling and forecast accuracy. Proper specification of these parameters is essential to ensure realistic dynamical behavior that reflects the underlying physics, as misspecified parameters can introduce systematic errors, degrade the data assimilation performance, and ultimately reduce forecast accuracy. Therefore, estimating the key parameters of the models is a fundamental task in data assimilation, aiming to improve both the state and the underlying model representation.

Data assimilation integrates observations with a dynamic numerical model to provide an estimator of the unknown time-evolving state [1, 2, 3]. As a critical tool in diverse fields such as meteorology [3, 4, 5], oceanography [6, 7], and automatic control [8], data assimilation enhances the accuracy of forecasting systems by assimilating model forecasts with observations, leading to more accurate forecasts and imputation of the state variables. The Kalman Filter (KF) [9] was a pioneering method for linear system assimilation with Gaussian noise, by providing optimal recursive solutions to the state variables. It propagates the mean and covariance of the state distribution through time, updating the state estimate whenever new observations become available. For nonlinear physical models, the ensemble Kalman Filter (EnKF) [2, 10] uses the forecast ensembles to estimate the forecast error covariance matrix and updates the system state via the analysis formula of the KF. It has become a popular data assimilation method.

There are a set of methods for implementing the EnKF. The inflation method was proposed to increase the spread of the forecast error covariance and to counter the potential model error, thereby reducing filter divergence [11, 12, 13]. To overcome the challenge in high-dimension covariance estimation, researchers developed the localization method to counter the deterioration of the ensemble forecast error covariance estimation, a technique that imposes sparsity by tapering or banding the covariance between two far away locations to zero, see [14, 15, 16] for details of the techniques.

Despite progress in data assimilation algorithms, parameter uncertainty remains a major limitation. Many geophysical models, such as oceanic or atmospheric systems, contain empirical estimates of parameters (e.g., vertical diffusivity, mixing and drag coefficients, optical depths) that strongly affect model dynamics and forecasting accuracy [17]. These parameters are typically pre-chosen based on prior knowledge, but any misspecification would lead to biased forecasts and inaccurate analysis states, regardless of the assimilation method used. There have been growing efforts devoted to incorporating parameter estimation within data assimilation systems. Early strategies include state augmentation, where parameters are augmented as state variables and are updated as part of the EnKF. Augmented state assimilation treats uncertain model parameters as an additional state variable and estimates it simultaneously with the dynamic state [4, 18]. Recently, machine learning methods have been incorporated in augmented data assimilation procedures [19].

Although the augmented assimilation method has a relatively low computational cost and has been widely used in fields such as the atmosphere and the ocean, it is sometimes difficult to ensure algorithmic convergence due to the unstable results of parameter estimation. The quasi maximal likelihood estimation (QMLE) based on the expectation–maximization algorithm was used for parameter estimation in nonlinear state space models [20, 21]. However, it relies on the ensemble Kalman smoother (EnKS), and requires re-running of the entire sequence of the EnKS for every candidate value of parameters, which incurs a heavy computation burden and may not be suitable for high-dimensional nonlinear systems. Meanwhile, [22, 23] proposed a nonlinear least squares (NLS) estimation for unknown parameters based Kalman filters. Similar as the QMLE method, it also requires re-running the entire data assimilation procedure each time the model parameters are updated during the optimization process, which inevitably leads to a substantial computational burden. Both QMLE and NLS methods do not support sequential parameter updates, and hence, can

not handle time-varying parameters.

To overcome the limitations of the existing methods, this paper aims to develop an efficient sequential machine learning algorithm coupled with the EnKF for data assimilation and estimating the underlying model parameter with a statistical guarantee. The proposed machine learning-based data assimilation algorithm meets the following four objectives: (i) sequential updating of both state variables and model parameters, (ii) computationally efficient without repeatedly recalculation of the forecast error covariance matrix in the optimization procedure for the model parameter, (iii) a more statistically efficient estimator for parameter estimation than the existing approaches, (iv) a statistical guarantee for the consistency and asymptotic distribution of the parameter estimate.

To achieve the aforementioned objectives, our method combines modern hierarchical Bayesian machine learning algorithms [24] with the EnKF framework, which leverages the adaptability of data-driven learning while preserving the predictive strengths of ensemble-based filtering through a sequential learning mechanism. We develop a data assimilation framework that integrates parameter estimation with state updating. Specifically, first, the entire time interval is divided into several segments. Parameter estimation and state updating are performed sequentially within each segment. Second, the estimated parameters in the current segment are used for an initial data assimilation for the next segment to compute the forecast error covariance matrices. This approach avoids repeated computation of the Kalman gain matrices during parameter optimization, thereby reducing the computational burden. A statistically efficient weighting matrix is selected to construct the empirical loss function for parameter estimation, which achieves minimum asymptotic variance. Finally, a hierarchical expected improvement (HEI) algorithm is employed to search for the minimum of the loss function. The HEI is an iterative algorithm, which searches for the global optimum with faster convergence by strengthening the exploration of areas with high prediction uncertainty.

To evaluate the performance of our proposed method, we conduct numerical simulations based on two benchmark models: the Lorenz'96 model with a single parameter and the shallow water model with two parameters. Numerical experiments show that our approach consistently outperformed traditional EnKF-based methods without parameter estimation, augmented data assimilation methods and existing parameter estimation methods in terms of the accuracy of both data assimilation and parameter estimation.

This paper is organized as follows. Section 2 presents the problem setting, introduces the identification conditions for parameter estimation in state-space models, and discusses the challenges associated with existing methods. Section 3 constructs the proposed sequential machine learning method for data assimilation and parameter estimation and establishes its asymptotic properties. Section 4 presents numerical studies on the Lorenz-96 model and the two-dimensional shallow water equations. Section 5 concludes the paper with discussions on future topics. The proofs of the theorems and additional numerical results are relegated to the supplementary material (SM).

2 | PROBLEM SETTING AND BACKGROUND

We consider the data assimilation (DA) setting with unknown parameters in the state-space models. Suppose the underlying state variable $\boldsymbol{X}_t \in \mathcal{X} \subset \mathbb{R}^p$ and its observation $\boldsymbol{Y}_t \in \mathbb{R}^q$ follow

$$\boldsymbol{X}_t = \mathcal{M}_t(\boldsymbol{X}_{t-1}, \boldsymbol{\theta}^*) + \eta_t, \tag{1}$$

$$\boldsymbol{Y}_t = H_t \boldsymbol{X}_t + e_t \tag{2}$$

for $t = 1, \ldots, T$, where η_t and e_t are the model and observation errors with zero-mean and covariance matrices Q_t and R_t, respectively. This is the state-space model where $\{\boldsymbol{X}_t\}_{t=1}^T$ may be unobservable while $\{\boldsymbol{Y}_t\}_{t=1}^T$ are observable. It is

assumed that both η_t and e_t are independent of $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{t-1}\}$, respectively. In this study, we do not impose specific distributional assumptions (e.g., Gaussianity) on η_t and e_t, allowing them to follow unknown distributions. Moreover, $\mathcal{M}_t(\cdot,\cdot) : \mathbb{R}^p \times \Theta \to \mathbb{R}^p$ is a nonlinear physical model operator, H_t is a $q \times p$ known linear observational matrix, and $\boldsymbol{\theta}^*$ is an interior point of $\Theta \subset \mathbb{R}^d$, which represents the unknown true value of the key model parameters.

The purpose of data assimilation is to impute the state $\boldsymbol{X}_t$ from observation $\boldsymbol{Y}_t$ based on the physical model, which depends on choosing a quality estimate of $\boldsymbol{\theta}^*$. Misspecification of the parameter $\boldsymbol{\theta}$ can result in systematic bias in the physical model and the data assimilation procedure [25]. In particular, when the physical model $\mathcal{M}_t(\cdot, \boldsymbol{\theta})$ is highly sensitive to $\boldsymbol{\theta}$, such as the general circulation models (GCMs) [26], the resulting bias may degrade the performance of data assimilation and forecast, and in severe cases, lead to filter divergence or failure. The goal of this paper is to propose a joint estimation procedure for both states and parameters with a guarantee on statistical consistency, which can reduce model bias and maintain the quality of the analysis state.

Let $\boldsymbol{X}^a_{t-1,j}$ be the j-th perturbed analysis ensemble at $t-1$ for $j = 1, \ldots, n$. Let $\boldsymbol{X}^f_{t,j}(\boldsymbol{\theta}^*) = \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \boldsymbol{\theta}^*)$ be the one-step ensemble forecast of $\boldsymbol{X}_t$ under the true parameter. Denote the j-th forecast error at $\boldsymbol{\theta}^*$ as $\boldsymbol{\delta}_{t,j} = \boldsymbol{X}_t - \boldsymbol{X}^f_{t,j}(\boldsymbol{\theta}^*)$. Then, Equation (2) can be expressed as

$$\boldsymbol{Y}_t = H_t \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \boldsymbol{\theta}^*) + H_t \boldsymbol{\delta}_{t,j} + e_t. \tag{3}$$

Let $\|\cdot\|$ denote the Euclidean norm of vectors. We make the following assumptions for the statistical identification of $\boldsymbol{\theta}^*$ based on the observed data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t$ via the expression in (3).

(S1) The analysis ensemble randomness among $\{\boldsymbol{X}^a_{t-1,j}\}$ is independent with $\{e_t\}_{t=1}^T$.

(S2) The forecast error satisfies $\mathbb{E}\{\mathbb{E}_{\text{ens}}(\boldsymbol{\delta}_{t,j}) \mid \boldsymbol{Y}_{1:t-1}\} = 0$, where $\mathbb{E}_{\text{ens}}$ represents taking expectation over the ensembles.

(S3) For every $\epsilon > 0$, we have $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \geq \epsilon} q^{-1} \mathbb{E} \left\| H_t \mathbb{E}_{\text{ens}}(\mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \boldsymbol{\theta}) - \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \boldsymbol{\theta}^*)) \right\|^2 > 0$.

Assumption S1 is a natural condition as the analysis ensembles are determined by the observations $\boldsymbol{Y}_{1:t-1}$ before time t. Assumption S2 assumes the unbiasedness of the average ensemble forecast of $\boldsymbol{X}^f_t$ at the true parameter value $\boldsymbol{\theta}^*$. In applications, violations of S1 may arise from the misspecification of the model or observation operators, in which case bias modeling is required [18]. Assumption S3 imposes distinguishability of the parameter $\boldsymbol{\theta}$ after applying the observation operator H_t. It ensures sufficient curvature of the population loss function at the true value $\boldsymbol{\theta}^*$ so that the observed data can identify the target parameter.

For a $q \times q$ positive definite matrix W, let $\|\boldsymbol{a}\|_W = \boldsymbol{a}' W \boldsymbol{a}$ for $\boldsymbol{a} \in \mathbb{R}^q$. The following Lemma provides an identification criterion for $\boldsymbol{\theta}^*$.

Lemma 1 *Under the model in (1) and (2), if Assumptions S1–S3 hold at t, then for any $q \times q$ positive definite matrix W, the true parameter $\boldsymbol{\theta}^*$ is the unique minimum of the optimization problem:*

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ \mathbb{E} \ \left\| \boldsymbol{Y}_t - H_t \mathbb{E}_{\text{ens}} \{ \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \boldsymbol{\theta}) \} \right\|^2_W. \tag{4}$$

The lemma provides a theoretical basis for estimating $\boldsymbol{\theta}^*$ by minimizing the squared residual between the observation and forecast states at a single time point. To conduct parameter estimation along with data assimilation as time progresses, we introduce the main steps of EnKF in the following.

For a given parameter value $\boldsymbol{\theta}$ and the analysis ensembles $\{\boldsymbol{X}^a_{t-1,j}\}_{j=1}^n$ at time $t-1$, the forecast ensembles are obtained by $\boldsymbol{X}^f_{t,j}(\boldsymbol{\theta}) = \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \boldsymbol{\theta})$. Let $\Sigma^f_t = \Sigma^f_t(\boldsymbol{\theta}^*) = \text{Var}(\boldsymbol{X}^f_{t,j}(\boldsymbol{\theta}^*))$ be the forecast covariance matrix at the true

parameter value. Let $\hat{\Sigma}_t^f(\boldsymbol{\theta})$ be the sample covariance of $\{\boldsymbol{X}_{t,j}^f(\boldsymbol{\theta})\}_{j=1}^n$ and $\hat{K}_t(\boldsymbol{\theta}) = \hat{\Sigma}_t^f(\boldsymbol{\theta})H_t'(H_t\hat{\Sigma}_t^f(\boldsymbol{\theta})H_t' + R_t)^{-1}$ be the Kalman gain matrix, which highlights the role of $\boldsymbol{\theta}$ on these quantities. Let $e_{t,j}$ be an independent copy of the observation error e_t, with mean 0 and covariance R_t. The EnKF updates the analysis ensembles at t via

$$\boldsymbol{X}_{t,j}^a(\hat{K}_t(\boldsymbol{\theta}),\boldsymbol{\theta}) = \boldsymbol{X}_{t,j}^f(\boldsymbol{\theta}) + \hat{K}_t(\boldsymbol{\theta})(\boldsymbol{Y}_t - H_t\boldsymbol{X}_{t,j}^f(\boldsymbol{\theta}) + e_{t,j}) \tag{5}$$

for $j = 1, \cdots, n$, where $\boldsymbol{X}_{t,j}^a(\hat{K}_t(\boldsymbol{\theta}),\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ through both the Kalman gain $\hat{K}_t(\boldsymbol{\theta})$ and the forecast ensembles via $\mathcal{M}_t(\cdot,\boldsymbol{\theta})$. Therefore, for each change of $\boldsymbol{\theta}$, we need to re-run the ensembles and re-estimate the Kalman gain matrix to obtain an updated value of $\boldsymbol{X}_{t,j}^a(\hat{K}_t(\boldsymbol{\theta}),\boldsymbol{\theta})$. This would make parameter estimation very computationally expensive for dynamic state-space models.

To overcome the challenge, we consider a sequential updating approach by partitioning the entire time horizon evenly into K segments: $\{1,\ldots,T\} = \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_K = \cup_{k=1}^K \mathcal{T}_k$, where $L = T/K$ and the k-th segment $\mathcal{T}_k = \{(k-1)L+1,\ldots,kL\}$. Our strategy is to update estimates of $\boldsymbol{\theta}^*$ iteratively over the segments. Starting from an initial value, say $\boldsymbol{\theta}^{(0)}$, we run an initial data assimilation procedure to obtain the Kalman gain matrices $\{\hat{K}_t(\boldsymbol{\theta}^{(0)})\}$ using $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$ for $t \in \mathcal{T}_1$. Fix those Kalman gain matrices, we calculate the analysis and forecast ensembles $\boldsymbol{X}_{t,j}^a(\hat{K}_t(\boldsymbol{\theta}^{(0)}),\boldsymbol{\theta})$ and $\boldsymbol{X}_{t,j}^f(\boldsymbol{\theta})$ in this segment by varying the value of $\boldsymbol{\theta}$ in the model operator $\mathcal{M}_t(\cdot,\boldsymbol{\theta})$. We obtain an improved estimate $\hat{\boldsymbol{\theta}}^{(1)}$ from the first segment $\mathcal{T}_1$ by minimizing an empirical estimate of the loss function in (4). In general, given a $\hat{\boldsymbol{\theta}}^{(k-1)}$ on the $(k-1)$-th segment, a new estimate $\hat{\boldsymbol{\theta}}^{(k)}$ is obtained based on $\hat{\boldsymbol{\theta}}^{(k-1)}$ and the observations in the k-th segment $\mathcal{T}_k$. The rationale and detailed procedure of the proposed approach will be outlined in the next section.

Augmented-state EnKF methods (e.g., [19, 27]) incorporate unknown parameters into the state vector and apply artificial evolution models (such as random walks) to enable joint state-parameter estimation. These approaches are computationally straightforward and allow parameters to be updated sequentially in every time step. However, their performance may be limited in complex or nonlinear systems, since the imposed parameter dynamics do not necessarily reflect the true evolution, potentially affecting estimation stability and accuracy. On the other hand, optimization-based methods such as the NLS estimator [23, 22] provide statistically consistent estimators by minimizing a global loss function. However, they are computationally expensive, as evaluating each candidate parameter requires rerunning the full data assimilation cycle over all time steps, resulting in substantial computational burden, high memory consumption, and limited algorithmic scalability.

Our proposed framework bridges these two approaches by combining their respective strengths: it performs parameter updates sequentially over time segments, like the augmented EnKF, while ensuring statistical consistency through segment-wise optimization, as in the NLS. This hybrid framework enables scalable and accurate parameter estimation in both high-dimensional and nonlinear state-space models.

3 | METHODOLOGY

We elaborate the detailed method for updating $\hat{\boldsymbol{\theta}}^{(k-1)}$ to $\hat{\boldsymbol{\theta}}^{(k)}$ using the forecast states and the observations in the k-th segment $\mathcal{T}_k$. Specifically, for each segment $\mathcal{T}_k$ given the parameter estimate from the previous segment $\hat{\boldsymbol{\theta}}^{(k-1)}$, we build an objective function of $\boldsymbol{\theta}$ using the weighted residuals between the forecast states and the observations in this period by recycling a common Kalman gain matrix and the forecast error covariance over the segment. This avoids repeatedly estimating the forecast error covariance matrices and the Kalman gain matrices while the value of $\boldsymbol{\theta}$ changes. Motivated by this, we propose a sequential estimation framework for jointly estimating the states and parameters of the state-space model via hierarchical Bayesian optimization [24].

3.1 | Sequential objective function

Given the estimate $\hat{\boldsymbol{\theta}}^{(k-1)}$ from the $(k-1)$-th segment, we run the EnKF using $\hat{\boldsymbol{\theta}}^{(k-1)}$ to obtain the initial analysis ensembles for every time point t in the k-th segment $\mathcal{T}_k$ as in (5). We denote $\boldsymbol{X}^a_{t,j}(\hat{\boldsymbol{\theta}}^{(k-1)}|\hat{\boldsymbol{\theta}}^{(k-1)}) = \boldsymbol{X}^a_{t,j}(\hat{K}_t(\hat{\boldsymbol{\theta}}^{(k-1)}), \hat{\boldsymbol{\theta}}^{(k-1)})$ for notation simplicity, where the two arguments of $\boldsymbol{\theta}$ in $\boldsymbol{X}^a_{t,j}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ indicate that $\boldsymbol{\theta}_1$ is used to run the model operator $\mathcal{M}_{t-1}(\cdot, \boldsymbol{\theta}_1)$ at $t-1$ and $\boldsymbol{\theta}_2$ is used to estimate the Kalman gain matrix. The initial forecast ensembles are obtained by $\{\boldsymbol{X}^f_{t,j}(\hat{\boldsymbol{\theta}}^{(k-1)}) = \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}(\hat{\boldsymbol{\theta}}^{(k-1)}|\hat{\boldsymbol{\theta}}^{(k-1)}), \hat{\boldsymbol{\theta}}^{(k-1)})\}_{j=1}^n$, where $\hat{\boldsymbol{\theta}}^{(k-1)}$ is used in $\mathcal{M}_t(\cdot, \boldsymbol{\theta})$ for all $t \in \mathcal{T}_k$. The Kalman gain matrices $\hat{K}_t(\hat{\boldsymbol{\theta}}^{(k-1)}) = \hat{\Sigma}^f_t(\hat{\boldsymbol{\theta}}^{(k-1)})H_t'(H_t\hat{\Sigma}^f_t(\hat{\boldsymbol{\theta}}^{(k-1)})H_t' + R_t)^{-1}$ for $\mathcal{T}_k$, where $\hat{\Sigma}^f_t(\hat{\boldsymbol{\theta}}^{(k-1)})$ is the sample covariance of the forecast ensembles $\{\boldsymbol{X}^f_{t,j}(\hat{\boldsymbol{\theta}}^{(k-1)})\}_{j=1}^n$.

By keeping the Kalman gain matrices fixed at the parameter value $\hat{\boldsymbol{\theta}}^{(k-1)}$ estimated in the previous segment over the k-th segment, we find an updated estimate of $\boldsymbol{\theta}^*$ by evaluating forecast states in the k-th segment with changing values of $\boldsymbol{\theta}$. Specifically, given $\{\hat{K}_t(\boldsymbol{\theta}^{(k-1)})\}_{t\in\mathcal{T}_k}$ and using the analysis ensembles as the end of the $(k-1)$-th segment as the initial states for the k-th segment, we obtain the analysis ensembles at a $\boldsymbol{\theta}$ and $t \in \mathcal{T}_k$ as

$$\boldsymbol{X}^a_{t,j}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)}) = \boldsymbol{X}^f_{t,j}(\boldsymbol{\theta}) + \hat{K}_t(\hat{\boldsymbol{\theta}}^{(k-1)})(\boldsymbol{Y}_t - H_t\boldsymbol{X}^f_{t,j}(\boldsymbol{\theta}) + e_{t,j}) \tag{6}$$

for $j = 1, \cdots, n$, where $\boldsymbol{X}^f_{t,j}(\boldsymbol{\theta}) = \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)}), \boldsymbol{\theta})$, $e_{t,j} \sim \mathcal{N}(0, R_t)$ and $\mathcal{N}(0, R_t)$ denotes the normal distribution with mean zero and covariance R_t. Different from $\boldsymbol{X}^a_{t,j}(\hat{K}_t(\boldsymbol{\theta}), \boldsymbol{\theta})$ in (5) and the initial analysis ensembles $\{\boldsymbol{X}^a_{t,j}(\hat{\boldsymbol{\theta}}^{(k-1)}|\hat{\boldsymbol{\theta}}^{(k-1)})\}$, the sequentially updated analysis ensembles in (6) use different values of $\boldsymbol{\theta}$ for obtaining the Kalman gain matrices and computing the model forecasts, as reflected by the notation $\boldsymbol{X}^a_{t,j}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$. This allows us to find a $\boldsymbol{\theta}$ that minimizes weighted a distance between $\boldsymbol{Y}_t$ and $H_t\boldsymbol{X}^f_{t,j}(\boldsymbol{\theta})$ for $t \in \mathcal{T}_k$ without re-computing the Kalman gain matrix at each time point of the k-th segment. We consider the following objective function in the k-th segment $\mathcal{T}_k$ to minimize with respect to $\boldsymbol{\theta}$:

$$L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)}) := \sum_{t\in\mathcal{T}_k}(\boldsymbol{Y}_t - H_t\hat{\boldsymbol{X}}^f_t(\boldsymbol{\theta}))'W_t(\hat{\boldsymbol{\theta}}^{(k-1)})(\boldsymbol{Y}_t - H_t\hat{\boldsymbol{X}}^f_t(\boldsymbol{\theta})), \tag{7}$$

where $W_t(\hat{\boldsymbol{\theta}}^{(k-1)}) = (H_t\hat{\Sigma}^f_t(\hat{\boldsymbol{\theta}}^{(k-1)})H_t' + R_t)^{-1}$ and $\hat{\boldsymbol{X}}^f_t(\boldsymbol{\theta}) = n^{-1}\sum_{j=1}^n \boldsymbol{X}^f_{t,j}(\boldsymbol{\theta})$. The weighting matrix $W_t(\hat{\boldsymbol{\theta}}^{(k-1)})$ is introduced to improve the efficiency of parameter estimation in the presence of heteroscedastic forecast errors. See the detailed explanation in the following paragraphs and the theoretical comparison with the NLS estimator without weighting in Theorem 4. Thus, we can obtain an updated estimate $\hat{\boldsymbol{\theta}}^{(k)}$ of $\boldsymbol{\theta}$ in the k-th segment as

$$\hat{\boldsymbol{\theta}}^{(k)} = \underset{\boldsymbol{\theta}\in\boldsymbol{\theta}}{\arg\min}\, L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)}). \tag{8}$$

In evaluating the objective function, obtaining the inverse of the $q \times q$ matrix $(H_t\hat{\Sigma}^f_t(\hat{\boldsymbol{\theta}}^{(k-1)})H_t' + R_t)$ can be computationally realized by using the Sherman–Morrison–Woodbury formula if R_t is a spherical matrix [28]. Hence, the objective function, despite being weighted by $W_t(\hat{\boldsymbol{\theta}}^{(k-1)})$, does not incur additional computational cost compared to the standard EnKF. Figure 1 illustrates the main steps of the proposed sequential updating and data assimilation process in the k-th segment.

It is noted that the NLS estimator minimizes the squared Euclidean distance $\|\boldsymbol{Y}_t - H_t\hat{\boldsymbol{X}}^f_t(\boldsymbol{\theta})\|^2$ between $\boldsymbol{Y}_t$ and $\hat{\boldsymbol{X}}^f_t(\boldsymbol{\theta})$ without weighting, which is used in estimating the parameter for state-space models. However, the sequentially updating procedure was not developed. Evaluating the NLS loss function over the entire time period requires repeated re-runs of the whole data assimilation procedure, which is computationally heavy. Furthermore, even if the

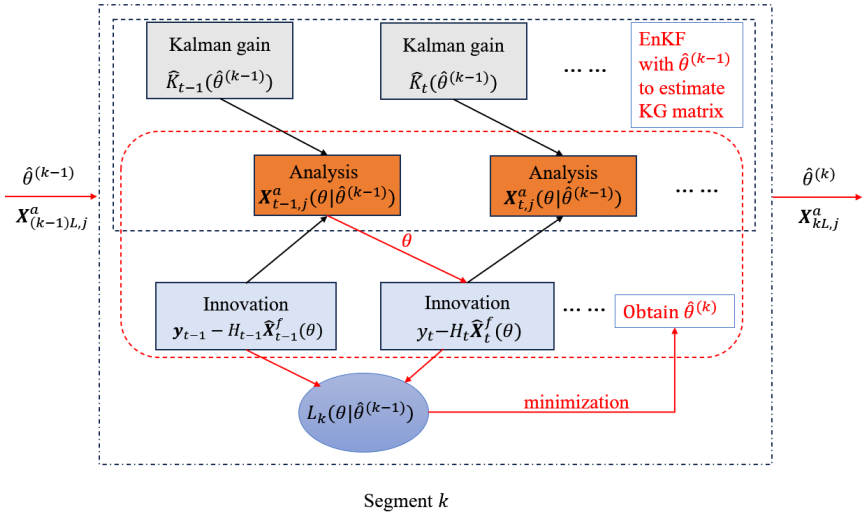

FIGURE 1 The entire time span is divided into L segments, and parameter estimation is performed within each segment using a machine learning-based optimization scheme introduced in Section 3.2. This plot shows the workflow for the k-th segment $\mathcal{T}_k$, where the parameter $\hat{\boldsymbol{\theta}}^{(k-1)}$ from the previous segment is used in an initial EnKF procedure to compute the Kalman gain $\hat{K}_t(\hat{\boldsymbol{\theta}}^{(k-1)})$ for $t \in \mathcal{T}_k$. Keeping those Kalman gain matrices unchanged, the analysis states are calculated with varying $\boldsymbol{\theta}$ in the model $\mathcal{M}_t(\cdot, \boldsymbol{\theta})$ to evaluate the forecast error at each time step. The objective function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$ of the weighted forecast errors is minimized to obtain the updated estimate $\hat{\boldsymbol{\theta}}^{(k)}$. The red arrow indicates that the unknown parameter $\boldsymbol{\theta}$ is propagated through the model, enabling a sequential update of parameters and states.

proposed sequential estimation procedure is applied, the NLS estimator

$$\hat{\boldsymbol{\theta}}^{(k)}_{\text{NLS}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \sum_{t\in\mathcal{T}_k} \|\boldsymbol{Y}_t - H_t\hat{\boldsymbol{X}}^f_t(\boldsymbol{\theta})\|^2 \tag{9}$$

without weighting on the k-th segment is not statistically efficient due to the heteroscedastic variances and correlations of the forecast errors $\boldsymbol{Y}_t - H_t\hat{\boldsymbol{X}}^f_t(\boldsymbol{\theta})$. Namely, $\hat{\boldsymbol{\theta}}^{(k)}_{\text{NLS}}$ does not achieve the lowest possible estimation variance as the sample size increases.

Note that $\text{Var}(\boldsymbol{Y}_t|\boldsymbol{Y}_{1:t-1}) = H_t\Sigma^f_t(\boldsymbol{\theta}^*)H'_t + R_t = (W^*_t)^{-1}$ if $\Sigma^f_t(\boldsymbol{\theta}^*) = \text{Var}(\boldsymbol{X}_t|\boldsymbol{Y}_{1:t-1})$. The application of the weighting matrix $W_t(\hat{\boldsymbol{\theta}}^{(k-1)})$ in our loss function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$ leads to whitened forecast errors. This is in the same spirit as the weighted least squares estimation in linear regression. Theorem 4 shows that the proposed estimator $\hat{\boldsymbol{\theta}}^{(k)}$ in (8) is more efficient than the NLS estimator $\hat{\boldsymbol{\theta}}^{(k)}_{\text{NLS}}$ under suitable conditions. In practical terms, $\hat{\boldsymbol{\theta}}^{(k)}_{\text{NLS}}$ tends to be more sensitive to the variability of forecast errors and thus may produce a less stable parameter estimate. The proposed weighted least squares approach enhances the stability of parameter estimation by weighting the forecast errors, and thus, reduces the asymptotic variance of the estimate.

3.2 | Hierarchical expected optimization algorithm

Because the physical model $\mathcal{M}_t(\cdot, \boldsymbol{\theta})$ is computationally expensive, particularly for high-resolution geophysical models, the optimization problem in (8) effectively becomes a black-box optimization task with substantial computational cost. To reduce the number of computationally expensive forward model evaluations, we adopt the hierarchical Bayesian optimization framework in [24], which uses a hierarchical Bayesian surrogate model to approximate the loss function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$. At each iteration, instead of numerically computing the derivative of $\mathcal{M}_t(\cdot, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, our method selects the most informative parameter value by maximizing an expected improvement function, called hierarchical expected improvement (HEI). Compared to the Bayesian optimization procedure using expected improvement in [23], the proposed HEI procedure provides a more efficient and stable search for $\hat{\boldsymbol{\theta}}^{(k)}$ in (8) by balancing the exploitation of the surrogate model and exploration of the areas with high prediction uncertainty.

Let $\{g(\boldsymbol{\theta})\} \sim \mathcal{GP}(0, \sigma^2 V_\tau)$ for $\boldsymbol{\theta} \in \Theta$ denote a zero–mean Gaussian process (GP) on $\boldsymbol{\theta}$ with marginal variance σ^2 and correlation function $V_\tau(\cdot, \cdot)$, where $V_\tau(\boldsymbol{\theta}, \boldsymbol{\theta}) = 1$ for any value of $\boldsymbol{\theta}$ and τ is the parameter. The correlation function V_τ encodes the dependence of this process over $\boldsymbol{\theta} \in \Theta$. For this GP, $(g(\boldsymbol{\theta}_1), \ldots, g(\boldsymbol{\theta}_m))^\top$ follows a multivariate normal distribution with mean zero and covariance $\sigma^2(V_\tau(\boldsymbol{\theta}_{i_1}, \boldsymbol{\theta}_{i_2}))_{m\times m}$ for any $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \in \Theta$. To construct a surrogate model for the objective function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$, we introduce a hierarchical GP process model $\tilde{L}_k(\boldsymbol{\theta}; \boldsymbol{\beta}_k, \sigma^2, \tau)$ as

$$\tilde{L}_k(\boldsymbol{\theta}; \boldsymbol{\beta}_k, \sigma^2, \tau) \sim \mathcal{GP}(\boldsymbol{b}(\boldsymbol{\theta})'\boldsymbol{\beta}_k, \sigma^2 V_\tau), \tag{10}$$

$$\boldsymbol{\beta}_k \propto \mathbf{1} \text{ and } \sigma^2 \sim \mathrm{IG}(\alpha_{k,1}, \alpha_{k,2}), \tag{11}$$

where $\boldsymbol{b}(\boldsymbol{\theta}) = (\boldsymbol{b}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{b}_r(\boldsymbol{\theta}))'$ are r basis functions, $\boldsymbol{\beta}_k \in \mathbb{R}^r$ are their coefficients, $\boldsymbol{\beta}_k \propto 1$ denotes a flat (non-informative) prior, meaning that all values of $\boldsymbol{\beta}_k$ are assigned equal prior probability, and $\mathrm{IG}(\alpha_{k,1}, \alpha_{k,2})$ represents the inverse Gamma distribution with parameters $\alpha_{k,1}, \alpha_{k,2}$. Under this hierarchical Bayesian model in (10)–(11), the mean $\boldsymbol{b}(\boldsymbol{\theta})'\boldsymbol{\beta}_k$ of the GP fits the overall trend of the objective function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$, while its variance part $\sigma^2 V_\tau$ fits the local structures and provides a credible interval for the fitted curve. Here, the prior distributions specified in (11) are designed to increase the error variance σ^2 of the model, and hence, to widen the length of the credible intervals. This helps faster exploration of the objective function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$ over the range of $\boldsymbol{\theta}$.

The GP model in (10) with basis function regression $\boldsymbol{\phi}(\boldsymbol{\theta})'\beta$ is a common statistical model used in atmospheric sciences [29, 30, 31]. The covariance function V_τ encodes spatial–temporal dependence, analogous to error covariance structures in numerical weather prediction and data assimilation [3]. The flat prior $\beta_k \propto 1$ indicates that no strong prior knowledge is imposed on the coefficients, allowing the data to determine the deterministic trend. Similar approaches have been successfully applied in climate and weather prediction [32, 33].

The HEI optimization procedure is conducted in an iterative manner. Let $\boldsymbol{\theta}_{k,1}, \ldots, \boldsymbol{\theta}_{k,m}$ be m pre-specified design points, where we evaluate the objective function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$. To build a surrogate model, we assume the GP $\tilde{L}_k(\boldsymbol{\theta}; \boldsymbol{\beta}_k, \sigma^2, \tau)$ takes the same value as $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$ on those design points $\{\boldsymbol{\theta}_{k,i}\}_{i=1}^m$. That is, we use the set

$$\mathcal{D}_{k,0} = \left\{\boldsymbol{\theta}_{k,i}, L_k(\boldsymbol{\theta}_{k,i}|\hat{\boldsymbol{\theta}}^{(k-1)})\right\}_{i=1}^m,$$

as the training data for the GP, where $\tilde{L}_k(\boldsymbol{\theta}_{k,i}; \boldsymbol{\beta}_k, \sigma^2, \tau) = L_k(\boldsymbol{\theta}_{k,i}|\hat{\boldsymbol{\theta}}^{(k-1)})$ on $\mathcal{D}_{k,0}$. The posterior distribution of $\tilde{L}_k(\boldsymbol{\theta}; \boldsymbol{\beta}_k, \sigma^2, \tau)$ given $\mathcal{D}_{k,0}$ provides the predictive mean and variance, which are considered as a surrogate function of $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$ and its uncertainty quantification, respectively. This posterior distribution is then utilized to construct the HEI updating function for Bayesian optimization.

Let $T_\nu(a, b^2)$ be the location-scale t distribution with the location parameter a, scale parameter b and degree of

freedom ν, where $T_\nu(a, b^2) = a + bt_\nu$ and t_ν is the Student's t distribution with degree of freedom ν. Given the training set $\mathcal{D}_{k,0}$ and the parameter τ in the correlation function V_τ, let $\hat{\boldsymbol{\beta}}_k$ and $\hat{\sigma}_k^2$ be the maximum likelihood estimators (MLEs) of $\boldsymbol{\beta}_k$ and σ^2 for the model in (10), respectively. Let $\tilde{\sigma}_k^2 = \tilde{\alpha}_{k,2}/\tilde{\alpha}_{k,1}$, where $\tilde{\alpha}_{k,1} = \alpha_{k,1} + (m-r)/2$ and $\tilde{\alpha}_{k,2} = \alpha_{k,2} + m\hat{\sigma}_k^2/2$. Note that all those estimates depend on the training set $\mathcal{D}_{k,0}$ in the updating procedure of HEI. Here, we ignore this dependence in those notations for simplicity of the presentation. Under the hierarchical model in (10)-(11), the marginal posterior distribution $\tilde{L}_k(\boldsymbol{\theta})|\mathcal{D}_{k,0}$ conditioned on the training set $\mathcal{D}_{k,0}$ that integrates over the prior distributions of $\boldsymbol{\beta}_k$ and σ^2 follows a location-scale t-distribution as

$$\tilde{L}_k(\boldsymbol{\theta})|\mathcal{D}_{k,0} \sim T_{2\tilde{\alpha}_{k,1}}(\hat{L}_k(\boldsymbol{\theta}), \tilde{\sigma}_k^2 s_k^2(\boldsymbol{\theta})), \tag{12}$$

where $\hat{L}_k(\boldsymbol{\theta}) = \boldsymbol{b}(\boldsymbol{\theta})'\hat{\boldsymbol{\beta}}_k + \boldsymbol{v}_k(\boldsymbol{\theta})'\boldsymbol{V}_k^{-1}(\boldsymbol{L}_k - \boldsymbol{P}_k\hat{\boldsymbol{\beta}}_k)$ and $s_k^2(\boldsymbol{\theta}) = 1 - \boldsymbol{v}_k(\boldsymbol{\theta})'\boldsymbol{V}_k^{-1}\boldsymbol{v}_k(\boldsymbol{\theta}) + \boldsymbol{h}_k(\boldsymbol{\theta})'(\boldsymbol{P}_k'\boldsymbol{V}_k^{-1}\boldsymbol{P}_k)^{-1}\boldsymbol{h}_k(\boldsymbol{\theta})$ with $\boldsymbol{v}_k(\boldsymbol{\theta}) = (V_\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_{k,1}), \ldots, V_\tau(\boldsymbol{\theta}, \boldsymbol{\theta}_{k,m}))'$, $\boldsymbol{V}_k = (V_\tau(\boldsymbol{\theta}_{k,i_1}, \boldsymbol{\theta}_{k,i_2}))_{i_1,i_2=1}^m$, $\boldsymbol{L}_k = (L_k(\boldsymbol{\theta}_{k,1}|\hat{\boldsymbol{\theta}}^{(k-1)}), \ldots, L_k(\boldsymbol{\theta}_{k,m}|\hat{\boldsymbol{\theta}}^{(k-1)}))'$, $\boldsymbol{P}_k = (\boldsymbol{b}(\boldsymbol{\theta}_{k,1}), \ldots, \boldsymbol{b}(\boldsymbol{\theta}_{k,m}))'$ and $\boldsymbol{h}_k(\boldsymbol{\theta}) = \boldsymbol{b}(\boldsymbol{\theta}) - \boldsymbol{P}_k'\boldsymbol{V}_k^{-1}\boldsymbol{v}_k(\boldsymbol{\theta})$.

The posterior distribution $\tilde{L}_k(\boldsymbol{\theta})|\mathcal{D}_{k,0}$ serves as a probabilistic surrogate representation of the objective function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$, allowing uncertainty quantification through the GP covariance function and the prior distribution on σ^2. In the posterior distribution, $\hat{L}_k(\boldsymbol{\theta})$ is the estimated conditional mean of $\tilde{L}_k(\boldsymbol{\theta})$ given the set $\mathcal{D}_{k,0}$ of training data. Its first term, $\boldsymbol{b}(\boldsymbol{\theta})'\hat{\boldsymbol{\beta}}_k$, gives the overall trend prediction by the regression basis functions, and its the second term provides a covariance-based correction, which exactly reproduces the same values of $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$ at the evaluation points $\{\boldsymbol{\theta}_{k,i}\}_{i=1}^m$, while smoothly interpolates in between. Figure 2 evaluates the fitted curve of $\hat{L}_k(\boldsymbol{\theta})$ in a Lorenz–96 system. The detailed simulation settings are introduced in Section 4.1. From this figure, the fitted curve closely aligned with the true objective function and the leave one out cross validation errors were small over all simulation repetitions, demonstrating that the GP model offers a good surrogate for our loss function of $\boldsymbol{\theta}$.

Note that the expression of the estimated conditional mean $\hat{L}_k(\boldsymbol{\theta})$ is the same as that of the universal kriging or Bayesian optimization based on a Gaussian process with a fixed covariance function [34]. Universal kriging and Bayesian optimization are widely used in geoscience to develop statistical emulators for physical model operators. See, for example [23] and [31]. However, our posterior variance is different due to adding a prior on the error variance σ^2 in (11). This results in a t-distribution which has a larger variance than the normal distributed posterior in the universal kriging and Bayesian optimization. A larger posterior variance helps for faster exploring outside the training points $\{\boldsymbol{\theta}_{k,i}\}_{i=1}^m$ in the optimization procedure introduced in the following.

Using the posterior distribution $\tilde{L}_k(\boldsymbol{\theta})|\mathcal{D}_{k,0}$ in (12), we now introduce an iterative procedure to minimize the objective function $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$. Let $L_{k,\min} = \min\{L_k(\boldsymbol{\theta}_{k,i}|\hat{\boldsymbol{\theta}}^{(k-1)}) : i = 1, \ldots, m\}$. Based on the current training set $\mathcal{D}_{k,0}$, we aim to find a new evaluation point of $\boldsymbol{\theta}$ that balances the exploitation and exploration. Here, exploitation means finding the minimum value of the current fitted curve $\hat{L}_k(\boldsymbol{\theta})$, and exploration means finding the value of $\boldsymbol{\theta}$ outside $\mathcal{D}_{k,0}$ where the fitted curve is most uncertain, namely, the locations with large posterior variances. This can be achieved by maximizing the expected improvement function, defined as

$$\mathrm{HEI}_0(\boldsymbol{\theta}) = \mathbb{E}_{\tilde{L}_k(\boldsymbol{\theta})|\mathcal{D}_{k,0}}\{L_{k,\min} - \tilde{L}_k(\boldsymbol{\theta})\}_+,$$

where $\{x\}_+ = \max\{x, 0\}$ is the positive truncation function. The improvement function $\{L_{k,\min} - \tilde{L}_k(\boldsymbol{\theta})\}_+$ measures the potential gain at $\boldsymbol{\theta}$ if $\tilde{L}_k(\boldsymbol{\theta})$ is smaller than the current minimum value $L_{k,\min}$ in the set $\mathcal{D}_{k,0}$. Based on the result in (12), the HEI updating function can be calculated as

$$\mathrm{HEI}_0(\boldsymbol{\theta}) = I_k(\boldsymbol{\theta})\Phi_{2\tilde{\alpha}_{k,1}}\left(\frac{I_k(\boldsymbol{\theta})}{\tilde{\sigma}s_k(\boldsymbol{\theta})}\right) + \nu_k\tilde{\sigma}s_k(\boldsymbol{\theta})\phi_{2(\tilde{\alpha}_{k,1}-1)}\left(\frac{I_k(\boldsymbol{\theta})}{\nu_k\tilde{\sigma}s_k(\boldsymbol{\theta})}\right), \tag{13}$$

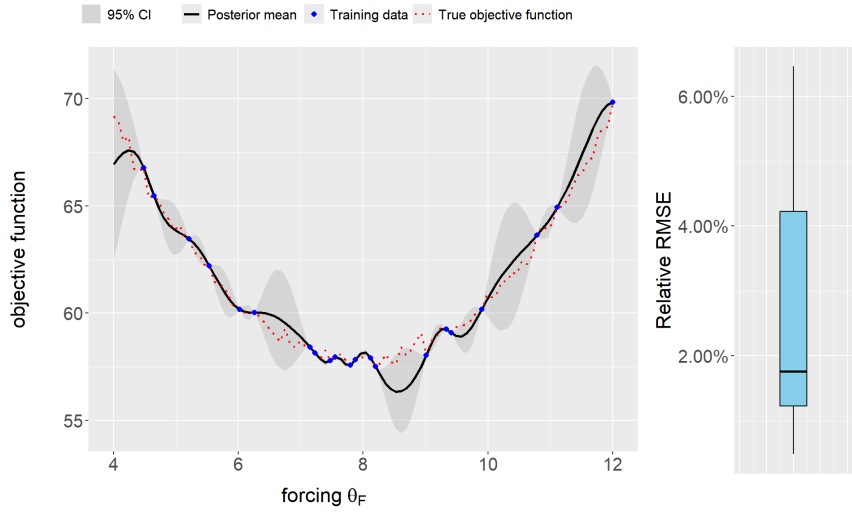

FIGURE 2 Fitted curve (left) of the loss function for the Lorenz–96 system with $T = 1000$, $p = 50$ and $q = 40$, where the parameter of interest is the forcing θ_F of this system and its true value is $\theta_F^* = 8$. For demonstration of the Gaussian process fitting, only one segment of data assimilation is considered ($K = 1$). The objective function (red dots) $L_1(\theta_F \mid \theta_F^{(0)})$ is derived using the standard ensemble Kalman filter with ensemble size $n = 50$ and initial value $\theta_F^{(0)} = 11$. The posterior mean (black line) $\hat{L}_k(\theta)$ is computed under the model in (10) using the training data (blue points) with size $m = 20$, where the design points $\{\theta_{1,i}\}_{i=1}^m$ are uniformly selected from $(4, 12)$. The shaded area represents the 95% credible interval. The boxplot (right) shows the distribution of the relative RMSE (RMSE normalized by the true values of objective function) based on 200 replications.

where $I_k(\boldsymbol{\theta}) = L_{k,\min} - \hat{L}_k(\boldsymbol{\theta})$, $v_k = \{\tilde{\alpha}_{k,1}/(\tilde{\alpha}_{k,1} - 1)\}^{1/2}$, and $\Phi_a(\cdot)$ and $\phi_a(\cdot)$ represent the cumulative distribution function and the probability density function of a t-distribution with degree of freedom a, respectively. The hyperparameters τ, $\alpha_{k,1}$ and $\alpha_{k,2}$ in the model (10)–(11) can be estimated empirically, and their estimates are plugged-in to evaluate the HEI function $\mathrm{HEI}_0(\boldsymbol{\theta})$.

Given the expression of $\mathrm{HEI}_0(\boldsymbol{\theta})$ in (13), we obtain the next evaluation point $\boldsymbol{\theta}_{k,m+1}$ by maximizing $\mathrm{HEI}_0(\boldsymbol{\theta})$, namely,

$$\boldsymbol{\theta}_{k,m+1} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\ \mathrm{HEI}_0(\boldsymbol{\theta}). \tag{14}$$

The training set is then updated as $\mathcal{D}_{k,1} = \mathcal{D}_{k,0} \cup \{(\boldsymbol{\theta}_{k,m+1}, L_k(\boldsymbol{\theta}_{k,m+1} | \hat{\boldsymbol{\theta}}^{(k-1)})\}$. Based on this updated training dataset $\mathcal{D}_{k,1}$, we compute a new fitted curve and the posterior distribution $\tilde{L}_k(\boldsymbol{\theta}) | \mathcal{D}_{k,1}$ as (12). Then, we evaluate the HEI function $\mathrm{HEI}_1(\boldsymbol{\theta}) = \mathbb{E}_{\tilde{L}_k(\boldsymbol{\theta})|\mathcal{D}_{k,1}}\{L_{k,\min} - \tilde{L}_k(\boldsymbol{\theta})\}_+$ as (13) and calculate its maximum to obtain the subsequent evaluation point $\boldsymbol{\theta}_{k,m+2}$. This iterative HEI procedure continues until the number of iterations reaches a prespecified threshold $s_{\max}$, which produces a sequence of evaluation points $\{\boldsymbol{\theta}_{k,i} : i = 1, \ldots, m + s_{\max}\}$ for the objective function in the k-th segment. The estimate $\hat{\boldsymbol{\theta}}^{(k)}$ in (8) is computed as

$$\hat{\boldsymbol{\theta}}^{(k)} = \underset{\boldsymbol{\theta}\in\{\boldsymbol{\theta}_{k,i}: i=1,\ldots,m+s_{\max}\}}{\operatorname{argmin}} L_k(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(k-1)}), \tag{15}$$

which is the minimum value of the objective function in the set of evaluation points selected by the HEI procedure.

Remark The HEI algorithm is a computationally efficient global optimization method in machine learning which avoids intensive computation of the derivatives of the model operator $\mathcal{M}_t(\cdot, \boldsymbol{\theta})$. The HEI function in (13) is a combination of two parts, where the first part $L_{k,\min} - \hat{L}_k(\boldsymbol{\theta})$ reaches the maximum at the minimum value of the fitted curve $L_k(\boldsymbol{\theta})$, and the second part is determined by the posterior standard deviation $\tilde{\sigma} s_k(\boldsymbol{\theta})$, which penalizes the fitted value by its uncertainty quantification. Specifically, in the expression of $s_k(\boldsymbol{\theta})$, a large value of $\boldsymbol{h}_k(\boldsymbol{\theta})$ implies deviation of $\boldsymbol{\theta}$ from the training data, which leads to a large posterior variance. This means that $s_k(\boldsymbol{\theta})$ could be large for the points that are far away from the training data, as demonstrated in Figure 2. Therefore, searching a new evaluation point by maximizing the HEI function in (14) intrinsically considers an exploitation-exploration trade-off: minimizing the fitted curve based on the training data while exploring the regions outside the training set with large prediction uncertainty. The posterior distribution of $\tilde{L}_k(\boldsymbol{\theta})$ integrates over the prior uncertainty of σ^2 in (11), resulting a heavier-tailed t-distribution rather than the normal distribution used in universal kriging. This heavier tail allows HEI to put more weight on exploration, so that it may find the global optimizer using fewer steps than Bayesian optimization.

3.3 | Sequential algorithm for parameter estimation

Based on the sequential objective function introduced in Section 3.1 and the HEI algorithm in Section 3.2, we propose an efficient sequential machine learning framework for joint data assimilation and parameter estimation under the models in (1) and (2). The time domain is divided into K segments $\{\mathcal{T}_k\}_{k=1}^K$, and the parameter estimation is performed once per segment based on the accumulated observations therein. This strategy retains the adaptivity of the augmented EnKF which updates the parameter at every assimilation step, while also enjoying the statistical properties of minimizing a global loss function across the entire domain of data assimilation. Given the observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ and the initial condition $\boldsymbol{X}_0$ and the covariance Σ_0 of the background field, we develop the following algorithm for data assimilation with parameter estimation.

- 1. (**Initialization at** $t = 0$) Give the initial condition $\boldsymbol{X}_0$ and the background error covariance Σ_0, generate n ensemble members $\{X_{0,1}, \ldots, X_{0,n}\}$ from the distribution $\mathcal{N}(\boldsymbol{X}_0, \Sigma_0)$ of the background field; give an initial value $\boldsymbol{\theta}^{(0)}$ of $\boldsymbol{\theta}$.
- 2. (**Initial EnKF with** $\hat{\boldsymbol{\theta}}^{(k-1)}$) In the k-th time segment, carry forward the analysis ensemble at the end of $\mathcal{T}_{k-1}$ as the initial state and fix the parameter at $\hat{\boldsymbol{\theta}}^{(k-1)}$. Run EnKF over $t \in \mathcal{T}_k$ using the physical model $\mathcal{M}_t(\cdot, \hat{\boldsymbol{\theta}}^{(k-1)})$ and obtain the initial forecast ensembles $\{\boldsymbol{X}_{t,j}^f(\hat{\boldsymbol{\theta}}^{(k-1)})\}$, then calculate their sample covariance $\hat{\Sigma}_t^f(\hat{\boldsymbol{\theta}}^{(k-1)})$. Using $\hat{\Sigma}_t^f(\hat{\boldsymbol{\theta}}^{(k-1)})$, compute the weighting matrix $W_t(\hat{\boldsymbol{\theta}}^{(k-1)})$ and Kalman gain $K_t(\hat{\boldsymbol{\theta}}^{(k-1)})$ as introduced in Section 3.1.
- 3. (**Loss function in** $\mathcal{T}_k$) With $\{W_t(\hat{\theta}^{(k-1)})\}$ and $\{\hat{K}_t(\hat{\theta}^{(k-1)})\}$ held fixed for $t \in \mathcal{T}_k$, we evaluate the forecast and analysis ensembles for a candidate value θ over $\mathcal{T}_k$ and construct the objective function $L_k(\theta \mid \hat{\theta}^{(k-1)})$ defined in (7). Generate m design points $\boldsymbol{\theta}_{k,1}, \ldots, \boldsymbol{\theta}_{k,m}$ uniformly sampled from Θ by the space-filling design (Chapter 5 of [35]). Obtain m evaluations of $L_k(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})$ at the design points $\boldsymbol{\theta}_{k,1}, \ldots, \boldsymbol{\theta}_{k,m}$.
- 4. (**HEI optimization**) Minimize the objective function $L_k(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k-1)})$ and obtain an update $\hat{\boldsymbol{\theta}}^k$ via the HEI algorithm as follows:
  - i) construct the HEI updating funcition $\mathrm{HEI}(\theta)$ by (13);
  - ii) perform $s_{\max}$ iterations of HEI, yielding $s_{\max}$ new evaluation points $\{\boldsymbol{\theta}_{k,m+1}, \ldots, \boldsymbol{\theta}_{k,m+s_{\max}}\}$.
  - iii) obtain the updated parameter $\hat{\boldsymbol{\theta}}^{(k)}$ from all evaluation points, including the initial m design points, by (15).
- 5. (**EnKF with** $\hat{\boldsymbol{\theta}}^{(k)}$) Based on the updated parameter $\hat{\boldsymbol{\theta}}^{(k)}$, compute the updated forecast states $\boldsymbol{X}_{t,j}^f(\hat{\boldsymbol{\theta}}^{(k)})$ and

analysis states $\boldsymbol{X}^{a}_{t,j}(\hat{\boldsymbol{\theta}}^{(k)}|\hat{\boldsymbol{\theta}}^{(k-1)})$ for $j = 1, \ldots, n$.

- 6. Iterate Steps 2–5 for all the time segments $\mathcal{T}_1, \ldots, \mathcal{T}_K$.

Figure 3 presents a flowchart to illustrate this sequential data assimilation algorithm with parameter estimation by Bayesian optimization. Due to the rapid convergence of the hierarchical Bayesian optimization algorithm, the overall computational cost can be kept low.

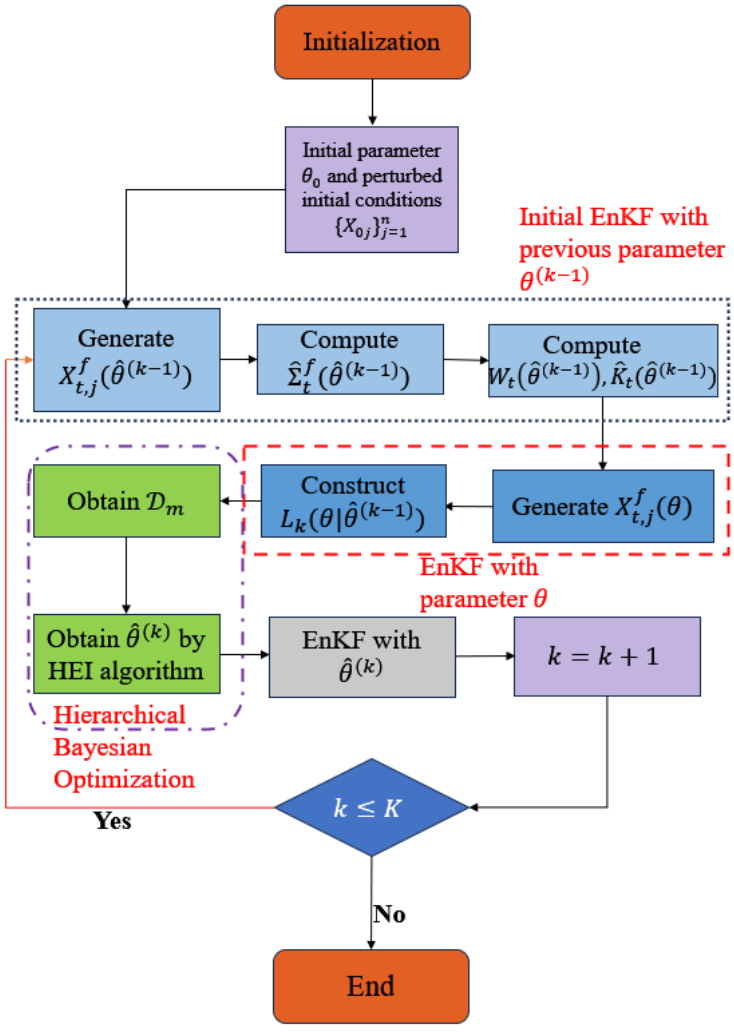

FIGURE 3 Flowchart illustrating the sequential machine learning algorithm for data assimilation and parameter estimation by the hierarchical Bayesian optimization.

3.4 | Theoretical properties

In this subsection, we investigate the theoretical properties of the proposed estimator, which establishes its consistency, asymptotic normality, and efficiency under suitable regularity conditions. The proofs of all theorems can be found in the SM.

Theorem 2 *Under Assumptions S1'–S3' and A1–A5 in the SM, $\hat{\boldsymbol{\theta}}^{(k)}$ converges to $\boldsymbol{\theta}^*$ in probability for $k = 1, \ldots, K$ as $n, L \to \infty$.*

Theorem 2 establishes the consistency of the proposed estimator of $\boldsymbol{\theta}^*$ in each segment. Recall that $\mathcal{N}(0, \Sigma)$ denotes the multivariate normal distribution with zero mean and covariance Σ. The following theorem shows the asymptotic normality of $\hat{\boldsymbol{\theta}}^{(k)}$.

Theorem 3 *Under Assumptions S1'–S3' and A1–A8 in the SM, $\sqrt{L}(\hat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^*)$ converges to $\mathcal{N}_d(0, U_1^{-1} U_{01} U_1^{-1})$ in distribution, and for $k = 2, \ldots, K$, $\sqrt{L}(\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*)$ converges to $\mathcal{N}_d(0, U_k^{-1})$ in distribution, as $n, L \to \infty$, where*

$$U_{01} = \lim_{L\to\infty} \frac{1}{L} \sum_{t=1}^{L} \mathbb{E}\left\{ \mathbb{E}_{ens}\{J_t(\boldsymbol{\theta}^*)\}' H_t' (H_t \Sigma_t^f(\boldsymbol{\theta}^{(0)}) H_t' + R_t)^{-1} (H_t \Sigma_t^f H_t' + R_t)(H_t \Sigma_t^f(\boldsymbol{\theta}^{(0)}) H_t' + R_t)^{-1} H_t \mathbb{E}_{ens}\{J_t(\boldsymbol{\theta}^*)\} \right\}$$

and

$$U_k = \lim_{L\to\infty} \frac{1}{L} \sum_{t \in \mathcal{T}_k} \mathbb{E}\left\{ \{\mathbb{E}_{\text{ens}} J_t(\boldsymbol{\theta}^*)\}' H_t' W_t^* H_t \{\mathbb{E}_{\text{ens}} J_t(\boldsymbol{\theta}^*)\} \right\}, k = 1, \ldots, K$$

with $J_t(\boldsymbol{\theta}^*) = \frac{\partial}{\partial \theta} \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \boldsymbol{\theta}^*)$

Theorem 3 derives the asymptotic variance of the proposed estimator $\hat{\boldsymbol{\theta}}^{(k)}$ in the k-th segment and shows its asymptotic normality property. From this theorem, the convergence rate of $\hat{\boldsymbol{\theta}}^{(k)}$ is at the order $L^{-1/2}$, meaning a larger L (longer segment) leads to a more accurate estimate of $\boldsymbol{\theta}^*$. However, this would increase the computation burden in that segment. The $(1-\alpha) \times 100\%$ confidence interval of the ℓ-th component of $\boldsymbol{\theta}$ can be constructed as

$$\hat{\theta}_\ell^{(k)} \pm z_{1-\alpha/2} \sqrt{L^{-1} (\hat{U}_k^{-1})_{\ell\ell}}$$

for $\ell = 1, \ldots, d$, where

$$\hat{U}_k = \lim_{L\to\infty} \frac{1}{L} \sum_{t \in \mathcal{T}_k} \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \theta} \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \hat{\boldsymbol{\theta}}^{(k)}) \right\}' H_t' W_t(\hat{\boldsymbol{\theta}}^{(k)}) H_t \left\{ \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \theta} \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \hat{\boldsymbol{\theta}}^{(k)}) \right\}$$

is an estimator of U_k, $\frac{\partial}{\partial \theta} \mathcal{M}_t(\boldsymbol{X}^a_{t-1,j}, \hat{\boldsymbol{\theta}}^{(k)})$ can be obtained by numerical differentiation, $(\hat{U}_k^{-1})_{\ell\ell}$ denotes the ℓ-th diagonal element of the inverse matrix of $\hat{U}_k$, and $z_{1-\alpha/2}$ refers to the $(1-\alpha/2)$ quantile of the standard normal distribution.

To demonstrate the advantage of the proposed estimator $\hat{\boldsymbol{\theta}}^{(k)}$ using a weighted loss function in (7) over the nonlinear least squares estimator $\hat{\boldsymbol{\theta}}^{(k)}_{\text{NLS}}$ in (9) without weighting, we compare their asymptotic variances and theoretically show the proposed estimator is statistically more efficient.

Theorem 4 *Under the same Assumptions as Theorem 3, we have* $\sqrt{L}(\hat{\boldsymbol{\theta}}_{\mathrm{NLS}}^{(k)} - \boldsymbol{\theta}^*)$ *converges to* $\mathcal{N}_d(0, \tilde{U}_k^{-1}\tilde{U}_{0k}\tilde{U}_k^{-1})$ *in distribution, where the expressions of* $\tilde{U}_{0k}$ *and* $\tilde{U}_k$ *are given in formulae (S3) and (S5) in the SM. Furthermore,* $\tilde{U}_k^{-1}\tilde{U}_{0k}\tilde{U}_k^{-1} - U_k^{-1}$ *is positive semidefinite for* $k = 2, \ldots, K$ *as* $n, L \to \infty$.

Since $\tilde{U}_k^{-1}\tilde{U}_{0k}\tilde{U}_k^{-1} - U_k^{-1}$ is positive semidefinite, the variances of the proposed estimator are smaller than those of the NLS estimator as the ensemble size n and the length L of each segment grow large. This demonstrates that the proposed method yields more precise parameter estimates with smaller asymptotic variance.

Remark Both the proposed estimator and the NLS estimator are constructed from the forecast ensembles. In contrast, one may also consider an alternative estimator

$$\hat{\boldsymbol{\theta}}_a^{(k)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{t\in\mathcal{T}_k} \|\boldsymbol{Y}_t - H_t\hat{\boldsymbol{X}}_t^a(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k-1)})\|^2. \tag{16}$$

based on the analysis ensembles, as the mean of the analysis ensembles is generally closer to the true state than that of the forecast ensembles. However, our theoretical analysis in the SM shows that $\hat{\boldsymbol{\theta}}_a^{(k)}$ is not necessarily superior to $\hat{\boldsymbol{\theta}}_{\mathrm{NLS}}^{(k)}$. In fact, the estimator $\hat{\boldsymbol{\theta}}_a^{(k)}$ can in fact be viewed as a special case of the weighted loss framework, where the weighting matrix is specified as $\tilde{W}_t = (I_q - \hat{K}_t(\hat{\boldsymbol{\theta}}^{(k-1)})H_t)'(I_q - \hat{K}_t(\hat{\boldsymbol{\theta}}^{(k-1)})H_t)$. In the SM, we have shown that the proposed estimator $\hat{\boldsymbol{\theta}}^{(k)}$ is optimal within the class of estimators that minimize a weighted loss as (7). This means that the proposed estimator also has higher precision than $\hat{\boldsymbol{\theta}}_a^{(k)}$.

4 | NUMERICAL STUDIES

To evaluate the performance of the proposed method, we conducted a series of numerical experiments using the Lorenz'96 model and the two-dimensional shallow water equations. All simulations were implemented in `R` and were based on 50 repeated runs. The proposed sequential data assimilation method (Seq_HEI) was compared against five existing approaches: the standard EnKF, EnKF with inflation (EnKF_infla), augmented assimilation (Aug), augmented assimilation with inflation (Aug_infla), and the standard EnKF using the NLS method for parameter estimation (EnKF_NLS) without sequential data assimilation and updating, as proposed in [23].

4.1 | Simulations on Lorenz'96 model

We considered the Lorenz'96 model [36], a widely used nonlinear dynamical system that has been extensively studied in the context of data assimilation. Let the true state variable at time t be denoted by $\boldsymbol{X} = (X_1, \ldots, X_p)'$, which evolved according to the following equation:

$$\frac{dX_i}{dt} = (X_{i+1} - X_{i-2})X_{i-1} - X_i + \theta_F, \quad i = 1, \ldots, p, \tag{17}$$

where $X_{-1} = X_{p-1}$, $X_{-2} = X_{p-2}$ and $\theta_F^* = 8$ was the true parameter. We solved the equation (17) by the fourth-order Runge-Kutta time integration scheme with a time step of 0.05 non-dimensional time unit. We set the initial state $\boldsymbol{X}_0 = (X_{01}, \ldots, X_{0p})'$ with the j-th element $X_{0i} = \theta_F^*$ for $i \neq \lfloor p/2 \rfloor$, $X_{0i} = \theta_F^* + 0.001$ for $i = \lfloor p/2 \rfloor$, $i = 1, \ldots, p$, and the initial parameter $\theta_F^{(0)} = 11$. Recall that $Q = (Q_{i_1 i_2})$ and $R = (R_{i_1 i_2})$ are the the model error covariance and observation error covariance in (1), respectively. We set $Q_{i_1 i_2} = \sigma_1 \times 0.3^{|i_1 - i_2|}$ and $R_{i_1 i_2} = \sigma_2 \times 0.1^{|i_1 - i_2|}$ for $i_1, i_2 = 1, \ldots, p$. We considered two settings for the simulation experiment. In the first setting, we fixed $\sigma_1 = 0.1$ and let $\sigma_2 = 0.2, 0.4, 0.8$.

In the second setting, we fixed $\sigma_2 = 0.2$ and allowed σ_1 to take values of 0.2, 0.4, and 0.8.

For all methods, we set $T = 2000$, $p = 200, 400$, $q = 0.8p$, and the ensemble size $n = 30, 60$. For the proposed method, we divided the time period $\{1, \ldots, T\}$ into $K = 20$ segments. For the standard EnKF and EnKF_infla that don't involve parameter estimation, the initial value $\theta_F^{(0)} = 11$ was set to obtain their prediction states. To compare the performance between different methods, we used the average RMSE over spatial locations for the differences between the observations and the analysis states, that is,

$$\text{RMSE}_t = \sqrt{\frac{1}{q}\left\| \boldsymbol{Y}_t - \frac{1}{n}\sum_{j=1}^{n} H_t \boldsymbol{X}_{t,j}^a(\hat{\boldsymbol{\theta}}^{(k)} | \hat{\boldsymbol{\theta}}^{(k-1)}) \right\|^2}, \quad \text{for } t \in \mathcal{T}_k,\ k = 1, \ldots, K. \tag{18}$$

To carry out the proposed sequential algorithm introduced in Section 3.3 based on Lorenz'96 model, we used orthogonal polynomial basis functions to fit the loss function under model (10), and the polynomial order was determined according to the Bayesian Information Criterion rule. Due to space limits, some of the simulation results were provided in the SM.

Figure 4 together with Figures S1–S2 in the SM report the RMSEs under observation variances $\sigma_2 = 0.2, 0.4, 0.8$, respectively. From those figures, Seq_HEI consistently achieved the lowest RMSE. At $\sigma_2 = 0.2$ with $n = 30$, its RMSE remained below 1.5, while standard EnKF and EnKF_infla exceeded 2.5. Increasing n to 60 reduced Seq_HEI's RMSE to around 1.0, whereas other methods improved only marginally. As σ_2 grew, all methods' errors increased, yet Seq_HEI retained its lead, especially at small ensemble sizes, demonstrating superior performance under limited ensembles.

Figure 5 and Figures S3–S4 in the SM report the RMSEs under model variance $\sigma_1 = 0.2, 0.4, 0.8$, respectively. At $\sigma_1 = 0.2$ and $n = 30$, Seq_HEI stabilized around 2.0, outperforming EnKF_NLS (2.5 to 3.0), Aug (around 3.0), Aug_infla (2.5 to 3.0), standard EnKF and EnKF_infla (both around 3.5). When n increased to 60, its RMSE fell to around 1.5 while others remained above 2.5. Although all methods' RMSEs rose with the increase of the model variance σ_1, Seq_HEI consistently yielded the most accurate data assimilation results.

We also summarize the mean and standard deviation of the RMSEs for the analysis state over simulation repetitions in Tables 1 and 2, which also demonstrated that Seq_HEI consistently attained the lowest RMSE with the smallest variability, and its advantage became more pronounced as p grew. Table 3 presents the average and standard deviation of the estimates of θ_F^* by four methods, the proposed Seq_HEI, EnKF_NLS, Aug and Aug_infla, over simulation repetitions. For the proposed method, we reported the estimate $\hat{\theta}^{(K)}$ from the last segment $\mathcal{T}_K$. Consistent with previous results, Seq_HEI provided the most accurate and stable estimates as shown in Table 3. These results confirmed that our method provided both superior data assimilation accuracy and reliable parameter estimation.

4.2 | Simulations on shallow water equations

The shallow water equations (SWE) are simplified PDEs for a thin fluid layer (e.g., oceans, rivers, lakes) where horizontal scales greatly exceed the vertical depth, so vertical accelerations are negligible relative to horizontal accelerations. In atmospheric and oceanic studies, SWE are also widely used to assess the performance of data assimilation methods

TABLE 1 The average (standard derivation) of RMSEs of the data assimilation results by six different methods, Seq_HEI (proposed), EnKF_NLS, EnKF, EnKF_infla, Aug, Aug_infla, for the the Lorenz'96 model with $p = 200, 400$, $q = 160, 320$ and $T = 2000$, $n = 30$.

$n = 30, p = 200, q = 160$						
Method	$\sigma_1 = 0.1$			$\sigma_2 = 0.2$		
	$\sigma_2 = 0.2$	$\sigma_2 = 0.4$	$\sigma_2 = 0.8$	$\sigma_1 = 0.2$	$\sigma_1 = 0.4$	$\sigma_1 = 0.8$
EnKF	2.84(0.170)	2.99(0.231)	3.30(0.250)	2.68(0.203)	2.83(0.278)	3.27(0.282)
EnKF_infla	2.80(0.215)	2.97(0.232)	3.27(0.218)	2.64(0.203)	2.84(0.260)	3.26(0.223)
Seq_HEI	1.71(0.160)	1.84(0.151)	1.99(0.133)	1.71(0.182)	2.05(0.186)	2.02(0.174)
EnKF_NLS	2.22(0.215)	2.43(0.142)	2.52(0.215)	2.11(0.209)	2.38(0.203)	2.72(0.342)
Aug	2.43(0.521)	2.50(0.384)	2.69(0.447)	2.45(0.474)	2.52(0.385)	3.32(0.705)
Aug_infla	2.23(0.377)	2.47(0.279)	2.65(0.358)	2.20(0.444)	2.51(0.336)	2.82(0.543)
$n = 30, p = 400, q = 320$						
Method	$\sigma_1 = 0.1$			$\sigma_2 = 0.2$		
	$\sigma_2 = 0.2$	$\sigma_2 = 0.4$	$\sigma_2 = 0.8$	$\sigma_1 = 0.2$	$\sigma_1 = 0.4$	$\sigma_1 = 0.8$
EnKF	3.69(0.126)	3.87(0.139)	4.06(0.163)	3.63(0.168)	3.69(0.189)	3.74(0.207)
EnKF_infla	3.68(0.118)	3.83(0.093)	4.01(0.143)	3.62(0.189)	3.60(0.148)	3.68(0.182)
Seq_HEI	1.94(0.089)	2.05(0.111)	2.24(0.120)	2.01(0.116)	2.05(0.116)	2.21(0.116)
EnKF_NLS	2.65(0.352)	2.69(0.449)	2.97(0.350)	2.59(0.394)	2.60(0.344)	2.79(0.307)
Aug	2.84(1.224)	3.13(0.871)	2.89(0.487)	2.93(0.700)	2.96(0.660)	3.21(0.643)
Aug_infla	2.51(0.442)	2.73(0.545)	2.88(0.361)	2.62(0.659)	2.83(0.664)	3.02(0.444)

TABLE 2 The average (standard derivation) of RMSEs of the data assimilation results by six different methods, Seq_HEI (proposed), EnKF_NLS, EnKF, EnKF_infla, Aug, Aug_infla, for the the Lorenz'96 model with $p = 200, 400$, $q = 160, 320$ and $T = 2000$, $n = 60$.

$n = 60, p = 200, q = 160$						
Method	$\sigma_1 = 0.1$			$\sigma_2 = 0.2$		
	$\sigma_2 = 0.2$	$\sigma_2 = 0.4$	$\sigma_2 = 0.8$	$\sigma_1 = 0.2$	$\sigma_1 = 0.4$	$\sigma_1 = 0.8$
EnKF	1.72(0.280)	2.18(0.252)	2.65(0.396)	1.58(0.193)	1.82(0.247)	1.98(0.240)
EnKF_infla	1.68(0.226)	2.17(0.253)	2.62(0.425)	1.57(0.218)	1.71(0.224)	1.98(0.237)
Seq_HEI	0.68(0.069)	0.87(0.076)	1.22(0.143)	0.85(0.094)	1.08(0.124)	1.36(0.146)
EnKF_NLS	1.08(0.181)	1.37(0.169)	1.97(0.391)	1.21(0.418)	1.48(0.175)	1.78(0.224)
Aug	1.16(0.259)	1.51(0.318)	1.94(0.375)	1.53(0.442)	1.60(0.165)	1.90(0.310)
Aug_infla	1.16(0.242)	1.54(0.308)	1.95(0.315)	1.32(0.132)	1.60(0.094)	1.89(0.306)
$n = 60, p = 400, q = 320$						
Method	$\sigma_1 = 0.1$			$\sigma_2 = 0.2$		
	$\sigma_2 = 0.2$	$\sigma_2 = 0.4$	$\sigma_2 = 0.8$	$\sigma_1 = 0.2$	$\sigma_1 = 0.4$	$\sigma_1 = 0.8$
EnKF	2.76(0.122)	2.92(0.186)	3.26(0.199)	3.16(0.246)	3.27(0.215)	3.50(0.241)
EnKF_infla	2.74(0.140)	2.91(0.161)	3.19(0.164)	3.13(0.249)	3.24(0.213)	3.34(0.129)
Seq_HEI	1.64(0.096)	1.75(0.097)	1.98(0.123)	1.76(0.185)	1.97(0.162)	2.32(0.172)
EnKF_NLS	2.10(0.178)	2.19(0.152)	2.50(0.141)	2.65(0.165)	2.80(0.232)	2.96(0.223)
Aug	2.27(0.416)	2.45(0.266)	2.57(0.292)	3.08(0.715)	3.12(0.546)	3.28(0.800)
Aug_infla	2.24(0.372)	2.29(0.198)	2.52(0.185)	2.96(0.476)	3.11(0.529)	3.27(0.458)

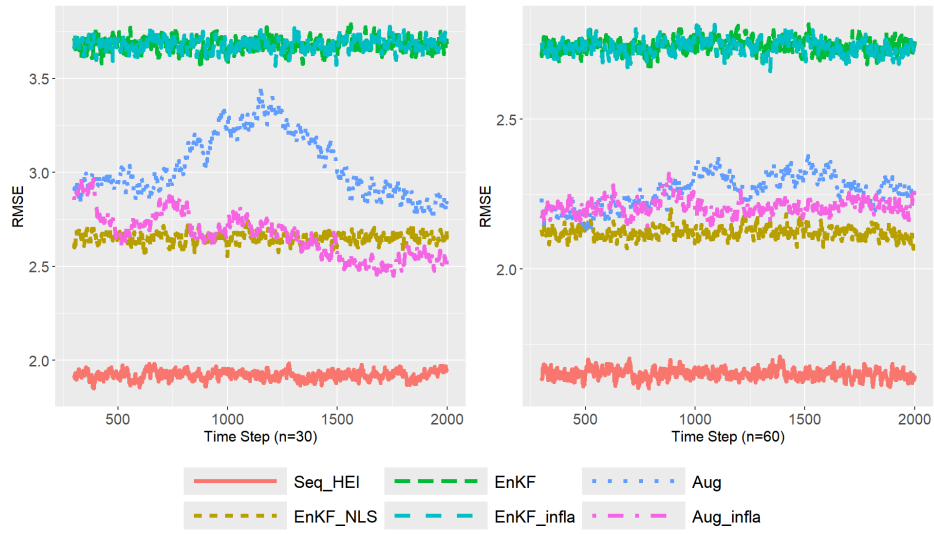

FIGURE 4 RMSE of the data assimilation results by six different methods, Seq_HEI (proposed), EnKF_NLS, EnKF, EnKF_infla, Aug, Aug_infla, for the the Lorenz'96 model with $(p, q, T) = (400, 320, 2000)$ and $(\sigma_1, \sigma_2) = (0.1, 0.2)$.

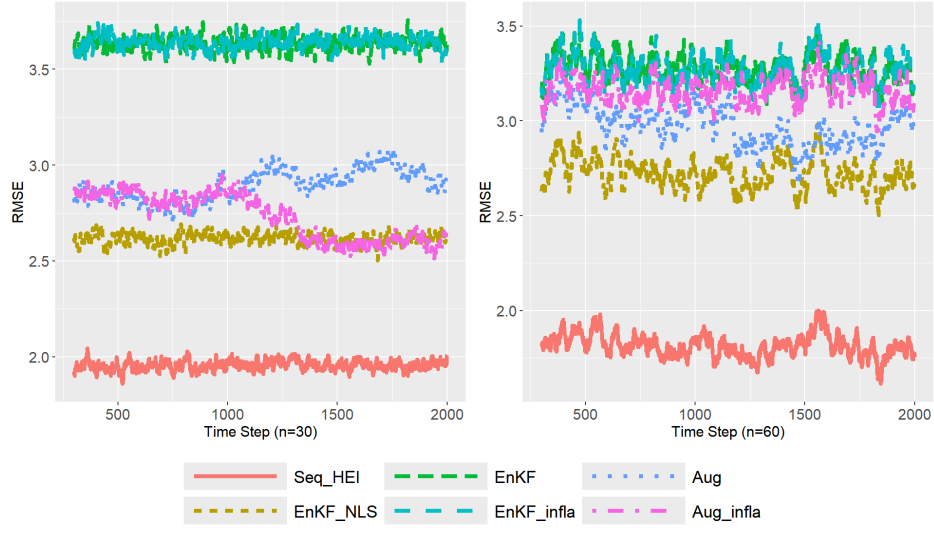

FIGURE 5 RMSE of the data assimilation results by six different methods, Seq_HEI (proposed), EnKF_NLS, EnKF, EnKF_infla, Aug, Aug_infla, for the the Lorenz'96 model with $(p, q, T) = (400, 320, 2000)$ and $(\sigma_1, \sigma_2) = (0.2, 0.2)$.

TABLE 3 The average (standard derivation) of estimated values of the unknown parameter in the Lorenz'96 model by four different methods, Seq_HEI (proposed), EnKF_NLS, Aug, Aug_infla, with $n = 60$, $p = 200, 400$, $q = 160, 320$ and $\sigma_1, \sigma_2 = 0.2, 0.4, 0.8$.

$n = 60, p = 200, q = 160$						
Method	$\sigma_1 = 0.1$			$\sigma_2 = 0.2$		
	$\sigma_2 = 0.2$	$\sigma_2 = 0.4$	$\sigma_2 = 0.8$	$\sigma_1 = 0.2$	$\sigma_1 = 0.4$	$\sigma_1 = 0.8$
Seq_HEI	8.00(0.039)	8.00(0.048)	7.97(0.043)	8.00(0.045)	8.04(0.054)	8.05(0.303)
EnKF_NLS	7.90(0.193)	7.91(0.172)	7.88(0.243)	7.87(0.216)	7.88(0.128)	7.82(0.399)
Aug	8.11(1.744)	8.04(1.692)	7.66(1.778)	8.95(3.549)	6.99(1.673)	7.79(4.031)
Aug_infla	7.85(1.353)	8.06(1.692)	7.23(1.926)	7.62(2.028)	5.95(4.109)	7.45(3.897)
$n = 60, p = 400, q = 320$						
Method	$\sigma_1 = 0.1$			$\sigma_2 = 0.2$		
	$\sigma_2 = 0.2$	$\sigma_2 = 0.4$	$\sigma_2 = 0.8$	$\sigma_1 = 0.2$	$\sigma_1 = 0.4$	$\sigma_1 = 0.8$
Seq_HEI	7.72(0.454)	7.69(0.406)	7.53(0.216)	8.03(0.354)	7.95(0.224)	7.95(0.065)
EnKF_NLS	7.69(0.620)	7.10(1.304)	6.59(0.215)	7.76(0.442)	7.65(0.526)	7.48(0.674)
Aug	7.71(2.578)	6.99(1.376)	6.82(1.432)	7.41(4.579)	8.21(4.454)	9.18(3.332)
Aug_infla	7.37(2.221)	8.66(1.502)	8.75(1.312)	8.59(4.877)	8.35(4.641)	9.15(5.034)

[25, 37, 38]. Referring to [25], the continuous fields $u(z_1, z_2, t)$, $v(z_1, z_2, t)$ and $h(z_1, z_2, t)$ satisfy

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial z_1} + v\,\frac{\partial u}{\partial z_2} - f\,v = -\,g\,\frac{\partial h}{\partial z_1} + k\,\nabla^2 u,$$
$$\frac{\partial v}{\partial t} + u\,\frac{\partial v}{\partial z_1} + v\,\frac{\partial v}{\partial z_2} + f\,u = -\,g\,\frac{\partial h}{\partial z_2} + k\,\nabla^2 v,$$
$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial z_1} + \frac{\partial (vh)}{\partial z_2} = k\,\nabla^2 h,$$

where u, v and h are simplified notations for $u(z_1, z_2, t)$, $v(z_1, z_2, t)$ and $h(z_1, z_2, t)$, respectively, $z_1 \in [0, L]$, $z_2 \in [0, D]$, $g = 9.8\,\mathrm{m\,s^{-2}}$, $k = 5\theta_1 \times 10^4\,\mathrm{m^2\,s^{-1}}$ is the diffusion coefficient, and $f = \theta_2 \times 10^{-4}\,\mathrm{s^{-1}}$ is the Coriolis parameter. We uniformly discretized $[0, L] \times [0, D]$ to a $p_1 \times p_2$ grid. Let $\boldsymbol{u}_t = \{u(z_1, z_2, t)\}$, $\boldsymbol{v}_t = \{v(z_1, z_2, t)\}$ and $\boldsymbol{h}_t = \{h(z_1, z_2, t)\}$ be the grid-stacked vectors of state variables. We applied the proposed method to estimate the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ in the Coriolis and diffusion coefficients, and recovery the true state vector $\boldsymbol{X}_t = (\boldsymbol{u}_t', \boldsymbol{v}_t', \boldsymbol{h}_t')'$. Note that the dimension of $\boldsymbol{X}_t$ is $p = 3\tilde{p}$, where $\tilde{p} = p_1 p_2$.

We set $L = 500\,\mathrm{km}$ and $D = 300\,\mathrm{km}$, took the true parameter $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)' = (1, 1)'$, and set the observation dimension to $\tilde{q} = 0.2\,\tilde{p}$. For the proposed method, we divided the time period $\{1, \ldots, T\}$ into $K = 12$ segments. The observation error covariance was taken constant over time as

$$R_t \equiv R = \mathrm{diag}(\sigma_u^2 I_{\tilde{q}},\ \sigma_v^2 I_{\tilde{q}},\ \sigma_h^2 I_{\tilde{q}}),$$

where $\sigma_u^2 = \sigma_v^2 = 0.5$ and $\sigma_h^2 = 1.0$. The initial condition $\boldsymbol{X}_0 = (\boldsymbol{u}_0', \boldsymbol{v}_0', \boldsymbol{h}_0')'$ was specified as follows. First, the initial height (depth) field was prescribed by

$$h(z_1, z_2) = H_0 + H_1 \tanh\left(\frac{9(D/2 - z_2)}{2D}\right) + H_2\ \mathrm{sech}^2\left(\frac{9(D/2 - z_2)}{D}\right) \sin\left(\frac{2\pi z_1}{L}\right),$$

with $H_0 = 50\,\mathrm{m}$, $H_1 = 5.5\,\mathrm{m}$, and $H_2 = 3.325\,\mathrm{m}$. Then, the initial velocity fields (u_0, v_0) were then obtained from h_0

using the geostrophic balance. We integrated the SWE for 24.5 hours using 30-second time steps, for a total of 2940

FIGURE 6 RMSE of the data assimilation results by six different methods, Seq_HEI (proposed), EnKF_NLS, EnKF, EnKF_infla, Aug, and Aug_infla, for the shallow water model with $(T, n) = (3000, 25)$.

steps, after a 60-step spin-up. Observations were assimilated every 60 steps (i.e., every 30 minutes). The initial value of $\boldsymbol{\theta}$ was set as $\boldsymbol{\theta}^{(0)} = (0.8, 0.8)'$. For the standard EnKF and EnKF_infla that don't involve parameter estimation, we used this initial value to run the SWE.

Figure 6 reports the data assimilation RMSEs between the true state and analysis state by the proposed method (Seq_HEI) and five existing methods considered in Section 4.1 under two settings of model dimensions (p, q). In both settings, the proposed method Seq_HEI achieved the lowest RMSE around 0.4, demonstrating its estimation accuracy and robustness to the increase of dimensions. The EnKF_NLS method performed slightly worse, but still outperformed the other methods. The augmented data assimilation approaches, Aug and Aug_infla, showed moderate performance, achieving lower RMSE than the standard EnKF and EnKF_infla, which suffered from large initial spikes due to the misspecification of $\boldsymbol{\theta}$.

Overall, the results from both simulation studies confirm that the proposed method can achieve better data assimilation accuracy by a sequentially updated estimation of the model parameters. The improvement is particularly evident when compared to traditional EnKF methods and their inflation-based variants that don't consider model calibration or parameter estimation.

5 | CONCLUSION AND DISCUSSION

The unknown parameter in state-space models can introduce significant bias that even advanced data assimilation techniques cannot fully correct. Existing methods for parameter estimation often break down under strong non-

linearity, high dimensionality, or large error variances, and they incur heavy computational costs. To address these challenges, this paper proposes a sequential machine learning algorithm that couples the ensemble Kalman filter with hierarchical Bayesian optimization. This method not only provides effective sequential estimates of the parameters but also achieves an efficient "estimating-while-assimilating" computational paradigm. The asymptotic properties including the consistency and the asymptotic normality of the proposed estimator have been established in this paper. Numerical experiments on the Lorenz-96 and the shallow water models demonstrate the accuracy and efficiency of the proposed method in high-dimensional settings with large variance in both observation and model errors.

Since this paper focuses primarily on parameter estimation, it overlooks certain data assimilation techniques, for example, the localization methods in high-dimensional data assimilation. In atmospheric or oceanic data assimilation, the dimension of state variables can be much higher than the time step and the ensemble size. In such cases, localization can be incorporated into our sequential data assimilation procedure to improve the accuracy of forecast error covariance estimation. The loss function in (7) and parameter estimation by the proposed HEI algorithm can be similarly constructed with localized data assimilation methods. Our study also assumes that the forecast error is unbiased conditional on the historical observations (Assumption S2). If the data assimilation and prediction results are biased, a debiasing step [4] is necessary before constructing the loss function for parameter estimation. We will leave those two extensions of the proposed method in future works.

Conflict of interest

The authors declare that they have no conflict of interest.

references

[1] Lorenc AC. Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society 1986;112(474):1177–1194.

[2] Evensen G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans 1994;99(C5):10143–10162.

[3] Carrassi A, Bocquet M, Bertino L, Evensen G. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. Wiley Interdisciplinary Reviews: Climate Change 2018;9:e535.

[4] Dee DP, Da Silva AM. Data assimilation in the presence of forecast bias. Quarterly Journal of the Royal Meteorological Society 1998;124(545):269–295.

[5] Zheng X. An adaptive estimation of forecast error covariance parameters for Kalman filtering data assimilation. Advances in Atmospheric Sciences 2009;26:154–160.

[6] Oke PR, Brassington GB, Griffin DA, Schiller A. The Bluelink ocean data assimilation system (BODAS). Ocean Modelling 2008;21(1):46–70.

[7] Sakov P, Counillon F, Bertino L, Lisæter KA, Oke PR, Korablev A. TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic. Ocean Science 2012;8(4):633–656.

[8] Friedland B. Treatment of bias in recursive filtering. IEEE Transactions on Automatic Control 1969;14(4):359–367.

[9] Kalman RE. A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering 1960;82(1):35–45.

[10] Evensen G. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean dynamics 2003;53:343–367.

[11] Anderson JL, Anderson SL. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly Weather Review 1999;127(12):2741–2758.

[12] Liang X, Zheng XG, Zhang SP, Wu GC, Dai YJ, Li Y. Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. Quarterly Journal of the Royal Meteorological Society 2012;138:263–273.

[13] Raanes PN, Bocquet M, Carrassi A. Adaptive covariance inflation in the ensemble Kalman filter by Gausssian scale mixtures. Quarterly Journal of the Royal Meteorological Society 2018;145:53–75.

[14] Greybush SJ, Kalnay E, Miyoshi T, Ide K, Hunt BR. Balance and ensemble Kalman filter localization techniques. Monthly Weather Review 2011;139(2):511–522.

[15] Perianez A, Reich H, Potthast R. Optimal localization for ensemble Kalman filter systems. Journal of the Meteorological Society of Japan Ser II 2014;92(6):585–597.

[16] Sun HX, Wang S, Zheng X, Chen SX. High-dimensional Ensemble Kalman Filter with Localization, Inflation and Iterative Updates. Quarterly Journal of the Royal Meteorological Society 2024;150:4870–4884.

[17] Mellor GL, Yamada T. Development of a turbulence closure model for geophysical fluid problems. Review of Geophysics 1982;20(4):851–875.

[18] Dee DP. Bias and data assimilation. Quarterly Journal of the Royal Meteorological Socity 2005;131(613):3323–3343.

[19] Sawada Y. An efficient estimation of time-varying parameters of dynamic models by combining offline batch optimization and online data assimilation. Journal of Advances in Modeling Earth Systems 2022;14:e2021MS002882.

[20] Yang Y, Mémin E. Estimation of physical parameters under location uncertainty using an ensemble^2 expectation–maximization algorithm. Quarterly Journal of the Royal Meteorological Society 2019;145:418–433.

[21] Lucini MM, Van Leeuwen PJ, Pulido M. Model error estimation using the expectation maximization algorithm and a particle flow filter. SIAM/ASA Journal on Uncertainty Quantification 2021;9(2):681–707.

[22] Ahn KW, Chan KS. Approximate conditional least squares estimation of a nonlinear state-space model via an unscented Kalman filter. Computational Statistics and Data Analysis 2014;69.

[23] Lunderman S, Morzfeld M, Posselt DJ. Using global Bayesian optimization in ensemble data assimilation: parameter estimation, tuning localization and inflation, or all of the above. Tellus A: Dynamic Meteorology and Oceanography 2021;73(1):1–16.

[24] Chen ZH, Mak S, Wu CFJ. A Hierarchical Expected Improvement Method for Bayesian Optimization. Journal of the American Statistical Association 2024;119(546):1619–1632.

[25] Wang S, Chen SX, Sun HX. High Dimensional Ensemble Kalman Filter. arXiv preprint arXiv:250500283v3 2025;.

[26] Zhang GJ, McFarlane NA. Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model. Atmosphere-ocean 1995;33(3):407–446.

[27] Sawada Y, Duc L. An efficient and robust estimation of spatio-temporally distributed parameters in dynamic models by an ensemble Kalman fliter. Journal of Adavance in Modeling Earth Systems 2024;16:e2023MS003821.

[28] Hager WW. Updating the inverse of a matrix, vol. 31. SIAM; 1989.

[29] Castruccio S, McInerney DJ, Stein ML, Liu Crouch F, Jacob RL, Moyer EJ. Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs. Journal of Climate 2014;27(5):1829–1844.

[30] Girard S, Mallet V, Korsakissok I, Mathieu A. Emulation and Sobol' Sensitivity Analysis of an Atmospheric Dispersion Model Applied to the Fukushima Nuclear Accident. Journal of Geophysical Research: Atmospheres 2016;121(7):3484–3496.

[31] Song Y, Khalid Z, Genton MG. Efficient stochastic generators with spherical harmonic transformation for high-resolution global climate simulations from CESM2-LENS2. Journal of the American Statistical Association 2024;119(548):2493–2507.

[32] Rougier J, Goldstein M, House L. Second-order exchangeability analysis for multimodel ensembles. Journal of the American Statistical Association 2013;108(503):852–863.

[33] Sansom PG, Ferro CA, Stephenson DB, Goddard L, Mason SJ. Best practices for postprocessing ensemble climate forecasts. Part I: Selecting appropriate recalibration methods. Journal of Climate 2016;29(20):7247–7264.

[34] Cressie NAC. Statistics for spatial data. John Wiley & Sons; 1993.

[35] Santner T, Williams B, Notz W. The Design and Analysis of Computer Experiments. 2nd ed. New York: Springer; 2018.

[36] Papamichail C, Bouzebda S, Limnios N. Predictability: a problem partly solved. Proceedings on Seminar on Predictability 1996;1:1–18.

[37] Greybush SJ, Kalnay E, Miyoshi T, Ide K, Hunt BR. Balance and ensemble Kalman filter localization techniques. Monthly Weather Review 2011;139:511–522.

[38] Lei L, Stauffer DR, Deng A. A hybrid nudging-ensemble Kalman filter approach to data assimilation. Part II: application in a shallow-water model. Tellus A: Dynamic Meteorology and Oceanography 2012;64:18485.