Online state and time-varying parameter estimation using the implicit equal-weights particle filter

Received: 9 October 2023 | Revised: 10 January 2024 | Accepted: 10 February 2024

DOI: 10.1002/qj.4698

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Online state and time-varying parameter estimation using the implicit equal-weights particle filter

Mineto Satoh[1,2] | **Peter Jan van Leeuwen**[3,4] | **Shin'ya Nakano**[1,5]

[1]Graduate Institute for Advanced Studies, SOKENDAI, Tachikawa, Japan

[2]Data Science Laboratories, NEC Corporation, Kawasaki, Japan

[3]Department of Meteorology and National Centre for Earth Observation, University of Reading, Reading, UK

[4]Department of Atmospheric Sciences, Colorado State University, Fort Collins, Colorado USA

[5]The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa, Japan

Correspondence
Mineto Satoh, Graduate Institute for Advanced Studies, SOKENDAI, Tachikawa, 190-8562, Japan.
Email: mineto-s@ism.ac.jp

Funding information
Grants-in-Aid for Scientific Research (KAKENHI) by the Japan Society for the Promotion of Science (JSPS), Grant/Award Number: 22H03553

Abstract

A method is proposed for resilient and efficient estimation of the states and time-varying parameters in nonlinear high-dimensional systems through a sequential data assimilation process. The importance of estimating time-varying parameters lies not only in improving prediction accuracy but also in determining when model characteristics change. We propose a particle-filter-based method that incorporates nudging techniques inspired by optimization algorithms in machine learning by taking advantage of the flexibility of the proposal density in particle filtering. However, as the model resolution and number of observations increase, filter degeneracy tends to be the obstacle to implementing the particle filter. Therefore, this proposed method is combined with the implicit equal-weights particle filter (IEWPF), in which all particle weights are equal. The method is validated using the 1000-dimensional linear model with an additive parameter and the 1000-dimensional Lorenz-96 model, where the forcing term is parameterized. The method is shown to be capable of resilient and efficient parameter estimation for parameter changes over time in our application with a linear observation operator. This leads to the conjecture that it applies to realistic geophysical, climate, and other problems.

KEYWORDS

data assimilation, nondegeneracy, parameter estimation, particle filter

1 | INTRODUCTION

Online parameter estimation is the process of inferring values that are often included in numerical models as unobservable quantities using sequentially collected observations. Since such parameters in numerical models are simplified representations of the modeled characteristics, parameter estimation plays an important role in obtaining accurate and reliable predictions. There are several approaches to parameter estimation, such as using an optimization algorithm under given state variables in the model and using data assimilation (DA) techniques (Evensen et al., 2022).

DA is known as the procedure to incorporate observations into numerical models and obtain posteriors of the state variables, especially in high-dimensional dynamical systems. Although DA usually focuses on generating an optimal initial state and forecasting the temporal

evolution of millions of time-varying state variables (Clayton et al., 2013), parameter estimation is often combined to calibrate the models (i.e., estimate the appropriate model characteristics). Therefore, parameter estimation is key to improving the prediction accuracy and is as complex as state estimation due to nonlinearities, even for linear dynamical models (Evensen et al., 1998).

Further, parameters can be considered not only as static but also as time-variant. For example, in hydrological modeling, parameters are usually assumed to be constant and calibrated using a particular data record to obtain an optimal parameter set or stationary parameter distributions. Still, it is necessary to use time-variant parameters to accurately simulate state variables wherein the calibration period may contain different climate conditions and hydrological regimes compared with the simulation period (Deng et al., 2016). As another example, according to Zhu et al. (2017), state and parameter estimation plays an important role in the application of process monitoring, online optimization, and process control. The difficulty of these applications is in identifying changes in model parameters when the operating conditions of the processing system have changed, or some faults have occurred in the processing system. From the above examples, it can be seen that estimating time-varying parameters plays an important role not only in improving prediction accuracy but also in determining when model characteristics change abruptly. However, the challenging issue is to distinguish whether the cause of the inaccuracy is incorrectly estimated state variables or a change in the model characteristics (i.e., parameters).

A typical method for time-varying state and parameter estimation in high-dimensional dynamical systems is the state augmentation technique, in which the parameter vector is incorporated into the state vector. This technique is also called joint estimation. Generally, the Kalman filter-based method is used for linear Gaussian systems, whilst the particle filter (PF) based method can be applied to nonlinear non-Gaussian systems. Santitissadeekorn and Jones (2015) indicate that the state augmentation method may become ineffective when the impact of parameters on the state is weak, and they propose a two-stage filter that combines a PF and an ensemble Kalman filter. This method estimates the static parameters and the tracking of the dynamic variables alternatively. Although similar approaches using an independent dual PF (Cooper & Perez, 2018) and a nested hybrid filter (Pérez-Vieites et al., 2018) have been proposed, they are only applicable to the estimation of static parameters. Extension to time-varying parameters requires identifying whether the change in observed states originates from state variables or parameters, but the amenability in practical contexts depends on the cross-covariance between states and parameters. In particular, detecting abrupt changes in characteristics in high-dimensional and partially observed nonlinear systems may be problematic because of the relatively low correlation between the observed state and parameters.

Another issue concerns nonlinearities due to the temporal evolution of the system and augmented state vector. As in the example using PF above, the parameter estimation method combined with PF can deal with nonlinearities, but filter degeneracy might be a critical obstacle for high-dimensional systems such as geophysical and climate systems. To overcome this problem, several approaches have been proposed, including the PF method by hybridizing with the ensemble Kalman filter (EnKF: Santitissadeekorn & Jones, 2015), as mentioned above. The approach of the equivalent-weights particle filter (EWPF: e.g., Van Leeuwen, 2010; Ades & Van Leeuwen, 2015) allows the proposal density to depend on all particles at the previous time step and assigns equivalent weights to most particles to avoid filter degeneracy. Zhu et al. (2016) proposed the implicit equal-weights particle filter (IEWPF), which combines the method of EWPF and implicit sampling (Chorin & Tu, 2009) to eliminate the need for parameter tuning. Skauvold et al. (2019) proposed a two-stage IEWPF method to correct the systematic bias in predictions caused by a gap in the proposal distribution in IEWPF (Zhu et al., 2016). Other approaches to eliminate filter degeneracy are also reviewed in Van Leeuwen et al. (2019). However, the above methods focus on estimating state variables or constant parameters.

In this article, we focus on a nonlinear time-varying system where the dimension of the state vector is large, while that of the model parameters is comparatively small, with a view to application in geophysical, climate, and other high-dimensional contexts. Then, we propose a new PF-based parameter estimation method and assess the capability of detecting abrupt changes in characteristics by applying it to the above system. We provide a methodology and results based on the IEWPF of Zhu et al. (2016) as an example of avoiding filter degeneracy. In our application, we assume a linear observation operator and require partial derivatives with respect to the parameters depending on the dimension of the parameters, although the methodology does apply to nonlinear observation operators and can work with approximate derivatives.

The remainder of the article is organized as follows. Section 2 describes the methodology for estimating time-varying parameters. First, to estimate states and parameters simultaneously, we extend IEWPF to an augmented state-space model with a correlated covariance matrix. We then propose the IEWPF-based method that

incorporates an optimization algorithm from machine learning into the parameter time evolution model by taking advantage of the flexibility of the proposal density in particle filtering. In Section 3, the effectiveness and advantages of the proposed method are evaluated through comparison with a method without incorporation of an optimization technique by using the linear model and the Lorenz-96 model (Lorenz, 1996). A summary and conclusions are put forward in Section 4.

2 | METHODOLOGY

2.1 | Correlated perturbation in augmented state-space model

A typical state-space model for a nonlinear system containing model parameters is described as

$$\begin{aligned} x^n &= f\left(x^{n-1}, \theta^{n-1}\right) + \beta^n, \\ y^n &= H_x\left(x^n\right) + \epsilon^n, \end{aligned} \tag{1}$$

where x^n is the state variable at time step n and y^n is the observation vector at time step n. f is the known possibly nonlinear function that maps the state from time t^{n-1} to t^n, and H_x is the known nonlinear observation operator. θ is the vector of model parameters, the true values of which are unknown and possibly time-varying. β is a random model perturbation drawn from the model-error probability density function (pdf) $\mathcal{N}(0, Q_\beta)$, while the observation error ϵ is drawn from the observation-error pdf $\mathcal{N}(0, R)$. To estimate time-varying parameters sequentially, the state vector is updated according to the following dynamical system by augmenting parameters as artificial states:

$$\begin{pmatrix} x^n \\ \theta^n \end{pmatrix} = \begin{pmatrix} f\left(x^{n-1}, \theta^{n-1}\right) \\ \theta^{n-1} \end{pmatrix} + \begin{pmatrix} \beta^n \\ \eta^n \end{pmatrix}. \tag{2}$$

Here, η^n is a random parameter perturbation drawn from the pdf $\mathcal{N}(0, Q_\eta)$, and we require that f is a differentiable function with respect to the parameter. Then, the above state updating function f can be approximately expressed by a first-order Taylor series expansion at the previous parameter θ^{n-2}:

$$f\left(x^{n-1}, \theta^{n-1}\right) \simeq f\left(x^{n-1}, \theta^{n-2}\right) + \left.\frac{\partial f}{\partial \theta}\right|_{\theta^{n-2}} \left(\theta^{n-1} - \theta^{n-2}\right). \tag{3}$$

Then, by using the time evolution model in the previous time step $n-1$:

$$\theta^{n-1} = \theta^{n-2} + \eta^{n-1}, \tag{4}$$

we can rewrite Equation 2 as

$$\begin{aligned} z^n &\equiv \begin{pmatrix} x^n \\ \theta^{n-1} \end{pmatrix} \\ &= \begin{pmatrix} f\left(x^{n-1}, \theta^{n-2}\right) \\ \theta^{n-2} \end{pmatrix} + \left(\left.\frac{\partial f}{\partial \theta}\right|_{n-2} \eta^{n-1} + \beta^n \eta^{n-1} \right) \\ &\equiv \tilde{f}\left(z^{n-1}\right) + \tilde{\rho}^n, \end{aligned} \tag{5}$$

where we introduce the augmented vector $z^n = [x^{n\mathrm{T}}, \theta^{n-1\mathrm{T}}]^\mathrm{T}$, model $\tilde{f}$, and perturbation $\tilde{\rho}$ representation. We also rewrite the observation operator H_x in Equation 1 as follows:

$$y^n = H_z\left(z^n\right) + \epsilon^n. \tag{6}$$

The augmented perturbation $\tilde{\rho}$ can drawn from the error pdf $\mathcal{N}(0, \tilde{Q}^n)$, which is expressed as

$$\tilde{Q}^n = \begin{pmatrix} \mathrm{cov}\left[\beta'^n, \beta'^n\right] & \mathrm{cov}\left[\beta'^n, \eta^{n-1}\right] \\ \left(\mathrm{cov}\left[\beta'^n, \eta^{n-1}\right]\right)^\mathrm{T} & \mathrm{cov}\left[\eta^{n-1}, \eta^{n-1}\right] \end{pmatrix}, \tag{7}$$

where $\beta'^n = (\partial f / \partial \theta)\eta^{n-1} + \beta^n$. Since model perturbation β and parameter perturbation η are independent of each other and both have zero means, each matrix element in Equation 7 can be calculated as follows:

$$\begin{aligned} \mathrm{cov}\left[\beta'^n, \beta'^n\right] &= E\left[\left(\frac{\partial f}{\partial \theta}\eta^{n-1} + \beta^n \right) \left(\frac{\partial f}{\partial \theta}\eta^{n-1} + \beta^n \right)^\mathrm{T} \right] \\ &= E\left[\frac{\partial f}{\partial \theta}\eta^{n-1}\left(\eta^{n-1}\right)^\mathrm{T} \left(\frac{\partial f}{\partial \theta}\right)^\mathrm{T} + \beta^n \left(\beta^n\right)^\mathrm{T} \right] \\ &= \frac{\partial f}{\partial \theta} Q_\eta^{n-1} \frac{\partial f}{\partial \theta}^\mathrm{T} + Q_\beta^n, \end{aligned} \tag{8}$$

$$\begin{aligned} \mathrm{cov}\left[\beta'^n, \eta^{n-1}\right] &= E\left[\left(\frac{\partial f}{\partial \theta}\eta^{n-1} + \beta^n \right) \left(\eta^{n-1}\right)^\mathrm{T} \right] \\ &= E\left[\frac{\partial f}{\partial \theta}\eta^{n-1}\left(\eta^{n-1}\right)^\mathrm{T} \right] \\ &= \frac{\partial f}{\partial \theta} Q_\eta^{n-1}, \end{aligned} \tag{9}$$

$$\begin{aligned} \mathrm{cov}\left[\eta^{n-1}, \eta^{n-1}\right] &= E\left[\eta^{n-1}\left(\eta^{n-1}\right)^\mathrm{T}\right] \\ &= Q_\eta^{n-1}. \end{aligned} \tag{10}$$

Then, Equation 7 can be expressed as

$$\tilde{Q}^n = \begin{pmatrix} \frac{\partial f}{\partial \theta} Q_\eta^{n-1} \frac{\partial f}{\partial \theta}^\mathrm{T} + Q_\beta^n & \frac{\partial f}{\partial \theta} Q_\eta^{n-1} \\ \left(\frac{\partial f}{\partial \theta} Q_\eta^{n-1}\right)^\mathrm{T} & Q_\eta^{n-1} \end{pmatrix}. \tag{11}$$

Note that the Taylor expansion in Equation 3 is used up to the first-order term, so the augmented perturbation $\tilde{\rho}$ from $\tilde{Q}$ includes the linear impact of the parameters on the model evolution over one time step.

2.2 | State and parameter update with IEWPF

In this section, we explain how to apply the IEWPF to the update equation Equation 5 and how to avoid filter degeneracy. Considering a Markovian system with observational errors that are independent from one time to another, the prior pdf can be written as

$$p(z^n) = \int p(z^n|z^{n-1})p(z^{n-1})\, dz^{n-1}. \tag{12}$$

Then, plugging Equation 12 into Bayes Theorem as a prior pdf, the posterior pdf of the model state given observations can be written as

$$p(z^n|y^{1:n}) = \frac{p(y^n|z^n)}{p(y^n)} \int p(z^n|z^{n-1})p(z^{n-1}|y^{1:n-1})\, dz^{n-1}. \tag{13}$$

Suppose we run a particle filter, and the particle weight for the ensemble at the previous time step $n-1$ is given by

$$p(z^{n-1}|y^{1:n-1}) = \frac{1}{N}\sum_{i=1}^{N}\delta(z^{n-1} - z_i^{n-1}). \tag{14}$$

Then plugging Equation 14 into Equation 13, we can obtain

$$p(z^n|y^{1:n}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(y^n|z^n)p(z^n|z_i^{n-1})}{p(y^n)}. \tag{15}$$

Introducing the proposal density $q(z^n|\boldsymbol{Z^{n-1}}, y^n)$, which is conditioned on all particles at time $n-1$, which indicated by the $\boldsymbol{Z^{n-1}}$, Equation 15 can be expressed as

$$p(z^n|y^{1:n}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(y^n|z^n)p(z^n|z_i^{n-1})}{p(y^n)q(z^n|\mathbf{Z}^{n-1}, y^n)}q(z^n|\mathbf{Z}^{n-1}, y^n). \tag{16}$$

The well-known problem of filter degeneracy means the weight will concentrate on only some particles, and most particles will have a negligible weight after a few propagations. Snyder et al. (2015) described that the particle filter using the optimal proposal yields minimal degeneracy and provides performance bounds. This could be a serious obstacle to implementing the particle filter when the number of states and observations increases, that is, a high-dimensional system. Therefore, we use the IEWPF (Zhu et al., 2016), which can avoid this filter degeneracy problem. From Equation 14, Equation 16 can be expressed as

$$p(z^n|y^{1:n}) = \frac{1}{N}\sum_{i=1}^{N} w_i\delta(z^{n-1} - z_i^{n-1}), \tag{17}$$

where w_i is the weight for particle i and is expressed as follows using the proposal density expressed in Equation 16:

$$w_i = \frac{p(y^n|z_i^n)}{p(y^n)}\frac{p(z_i^n|z_i^{n-1})}{q(z_i^n|\mathbf{Z}^{n-1}, y^n)}. \tag{18}$$

Instead of drawing directly from proposal density q, we can draw a standard Gaussian distributed proposal density $q(\xi)$, which is related by

$$q(\xi) = q(z^n|\boldsymbol{Z^{n-1}}, y^n)\left\|\frac{dz}{d\xi}\right\|, \tag{19}$$

where $\|dz/d\xi\|$ denotes the absolute value of the determinant of the Jacobian matrix, which expresses the following transformation:

$$z_i^n = \zeta_i^n + \alpha_i^{1/2}P^{1/2}\xi_i^n, \tag{20}$$

where ζ_i^n express the mode of $q(z^n|\mathbf{Z}^{n-1}, y^n)$, P is a measure of the width of that pdf, and α_i is a scalar factor. Note that this expression is similar to the original IEWPF (Zhu et al., 2016), but z_i^n denotes the augmented vector $z^n = [x^{n\mathrm{T}}, \theta^{n-1\mathrm{T}}]^{\mathrm{T}}$. This means that transformed variable ξ also has the dimension of the augmented vector. Then, Equation 18 can be expressed as follows:

$$w_i = \frac{p(y^n|z_i^n)}{p(y^n)}\frac{p(z_i^n|z_i^{n-1})}{q(\xi)}\left\|\frac{dz}{d\xi}\right\|. \tag{21}$$

In general, the ζ_i^n can be obtained via a minimization of $-\log q(z^n|\mathbf{Z}^{n-1}, y^n)$, similar to for example, a 3DVar, and also the equal weights can be obtained numerically. In this article, we will follow Zhu et al. (2016) and assume a linear observation operator, which will allow for an analytical solution for the equal weights.

2.3 | Linear observation model and Gaussian error

Assuming the linear observation model $\tilde{H}$ and Gaussian model and observation error as shown in Equations 5 and 6, ζ_i^n in Equation 20 can be expressed as explained in

Zhu et al. (2016):

$$\zeta_i^n = \tilde{f}\left(z_i^{n-1}\right) + K\left(y^n - \tilde{H}\tilde{f}\left(z_i^{n-1}\right)\right), \tag{22}$$

where

$$K = \tilde{Q}\tilde{H}^{\mathrm{T}}\left(\tilde{H}\tilde{Q}\tilde{H}^{\mathrm{T}} + R\right)^{-1} \tag{23}$$

and P in Equation 20 is

$$P = \left(\tilde{Q}^{-1} + \tilde{H}^{\mathrm{T}}R^{-1}\tilde{H}\right)^{-1}. \tag{24}$$

Note that $\tilde{Q}$ is the model-error covariance matrix described in Equation 11 and R is the observation-error covariance matrix. Therefore, from Equations 20–22, equal-weight particle z_i sampled from posterior pdf Equation 16 can be constructed using the scalar factor α_i.

The factor α_i needs to be determined so that the weight of each particle i represented by Equation 21 is the same target weight for all particles. Introducing w_i^{prev}, which denotes the weight from previous time steps, we can express Equation 21 as

$$w_i = \frac{p(y^n|z_i^n)p(z_i^n|z_i^{n-1})}{q(\xi)}\left\|\frac{dz}{d\xi}\right\| \cdot w_i^{\mathrm{prev}}. \tag{25}$$

With the above Gaussian assumption, we can write

$$\begin{aligned} &p(y^n|z^n)p(z^n|z_i^{n-1}) \\ &\propto \exp\left[-\frac{1}{2}\left(y^n - \tilde{H}z^n\right)^{\mathrm{T}}R^{-1}\left(y^n - \tilde{H}z^n\right)\right. \\ &\left.-\frac{1}{2}\left(z^n - \tilde{f}\left(z_i^{n-1}\right)\right)^{\mathrm{T}}\tilde{Q}^{-1}\left(z^n - \tilde{f}\left(z_i^{n-1}\right)\right)\right] \\ &= \exp\left[-\frac{1}{2}\left(z^n - z_i^a\right)^{\mathrm{T}}P^{-1}\left(z^n - z_i^a\right)\right]\exp\left(-\frac{1}{2}\phi_i\right), \end{aligned} \tag{26}$$

where

$$\phi_i = \left(y^n - \tilde{H}\tilde{f}\left(z_i^{n-1}\right)\right)^{\mathrm{T}}\left(\tilde{H}\tilde{Q}\tilde{H}^{\mathrm{T}} + R\right)^{-1}\left(y^n - \tilde{H}\tilde{f}\left(z_i^{n-1}\right)\right). \tag{27}$$

Taking the logarithm of Equation 25 leads to

$$\begin{aligned} -2\log w_i = &-2\log w_i^{\mathrm{prev}} \\ &+\left[-2\log\left(\frac{p(y^n|z_i^n)p(z_i^n|z_i^{n-1})}{q(\xi)}\left\|\frac{dz}{d\xi}\right\|\right)\right]. \end{aligned} \tag{28}$$

Substituting Equations 26 and 20 in Equation 28, we find

$$\begin{aligned} -2\log w_i = &-2\log w_i^{\mathrm{prev}} + \alpha_i\xi_i^{n\mathrm{T}}P^{1/2}P^{-1}P^{1/2}\xi_i^n \\ &+\phi_i - \xi_i^{n\mathrm{T}}\xi_i^n - 2\log\left(\left\|\frac{dz}{d\xi}\right\|\right). \end{aligned} \tag{29}$$

Using Equation 20 and the simplified expression for the Jacobian in Zhu et al. (2016), we can rewrite

$$\begin{aligned} -2\log w_i &= -2\log w_i^{\mathrm{prev}} + (\alpha_i - 1)\xi_i^{n\mathrm{T}}\xi_i^n \\ &\quad + \phi_i - 2\log\left(\alpha_i^{N_x/2}\left\|P^{1/2}\right\|\left|1 + \frac{\partial\alpha_i^{1/2}}{\partial\xi_i^n}\frac{\xi_i^n}{\alpha_i^{1/2}}\right|\right) \\ &= -2\log w_i^{\mathrm{prev}} + (\alpha_i - 1)\xi_i^{n\mathrm{T}}\xi_i^n + \phi_i \\ &\quad - 2N_x\log\alpha_i^{1/2}d - 2\log\left(\left\|P^{1/2}\right\|\right) \\ &\quad - 2\log\left(\left|1 + \frac{\partial\alpha_i^{1/2}}{\partial\xi_i^n}\frac{\xi_i^n}{\alpha_i^{1/2}}\right|\right), \end{aligned} \tag{30}$$

where N_x is the dimension of the model state. Setting the weights of all particles to the target weight is equivalent to setting all $\log w_i$ equal to the constant C, which leads to the following equation for α_i:

$$\begin{aligned} &(\alpha_i - 1)\xi_i^{n\mathrm{T}}\xi_i^n - 2N_x\log\alpha_i^{1/2} \\ &\quad - 2\log\left(\left|1 + \frac{\partial\alpha_i^{1/2}}{\partial\xi_i^n}\frac{\xi_i^n}{\alpha_i^{1/2}}\right|\right) = C - \left(\phi_i - 2\log w_i^{\mathrm{prev}}\right), \end{aligned} \tag{31}$$

in which constant value $2\log\left(\left\|P^{1/2}\right\|\right)$ is included in C. Here, let c_i denote the log-weight offsets for each particle i from the target weight C as

$$c_i = C - \left(\phi_i - 2\log w_i^{\mathrm{prev}}\right). \tag{32}$$

In practice, this c_i can be determined using the values of ϕ for all particles as

$$c_i = \max_j\{\phi_j\} - \phi_i. \tag{33}$$

Therefore, α_i is obtained as a solution satisfying Equation 31 with c_i determined by Equation 33.

Further assuming that the factor α_i depends on ξ_i^n only through $g_i = \xi_i^{n\mathrm{T}}\xi_i^n$, Equation 31 can be simplified to

$$\begin{aligned} &\exp\left(-\frac{\alpha_i g_i}{2}\right)(\alpha_i g_i)^{N_x/2-1}\left\|\frac{d(\alpha_i g_i)}{dg_i}\right\| \\ &= \exp\left(-\frac{g_i}{2}\right)g_i^{N_x/2-1}\exp\left(-\frac{c_i}{2}\right) \end{aligned} \tag{34}$$

(see Appendix in Zhu et al., 2016). For every particle to reach the target weight, $c_i \geq 0$ should be satisfied, therefore $0 < \exp(-c_i/2) \leq 1$ in Equation 34. Furthermore, since the function of the left-hand side $\exp(-\alpha g_i/2)(\alpha g_i)^{N_x/2-1}$ has an extremum at $\alpha_i = (N_x - 2)/g_i$, it is suggested that the solution α_i of Equation 34 allows two values. According to Zhu

et al. (2016), Equation 34 can be integrated from $N/2$ to ∞, then yields the following equation:

$$\Gamma\left(\frac{N_x}{2}, \frac{\alpha_i g_i}{2}\right) = \begin{cases} \exp\left(-\frac{c_i}{2}\right)\Gamma\left(\frac{N_x}{2}, \frac{g_i}{2}\right) & \text{if } \frac{d(\alpha_i g_i)}{dg_i} > 0, \\ \exp\left(\frac{c_i}{2}\right)\Gamma\left(\frac{N_x}{2}, \frac{g_i}{2}\right) & \text{if } \frac{d(\alpha_i g_i)}{dg_i} < 0, \end{cases} \tag{35}$$

where Γ is the monotonically decreasing upper incomplete gamma function. Therefore the solution α_i for every particle i that satisfies Equation 35 is allowed both $\alpha \leq 1$ and $\alpha \geq 1$ theoretically. Although $\alpha \geq 1$ solutions are known to lead to systematic bias (Zhu et al., 2016), the bias decreases when the state-space dimension N_x increases, that is, the high-dimensional case. As another solution, Skauvold et al. (2019) proposed the two-stage IEWPF that can eliminate this bias.

In practice, the following should be considered when generating the posterior distribution by calculating α_i that satisfies Equation 35. The first point is the computational cost of finding α_i numerically for each particle. To avoid this calculation, Zhu et al. (2016) proposed an approximation under the limiting case of $N_x \to \infty$. Then, the solution α can be expressed analytically using the Lambert W function (Corless et al., 1996), which has two branches: $\alpha > 1$, which gives a large ensemble spread, and $\alpha < 1$, which gives the opposite effect. The authors proposed adjusting the ratio of sampling α_i for each particle i from either branch in order to bring the shape of the distribution closer to the ideal one. The results of this α dependence will be shown later. The second point is the guarantee of convergence to the posterior distribution. IEWPF can equalize the weights of all particles, but the convergence of the filter distribution to the posterior distribution was only confirmed experimentally by Zhu et al. (2016) and not shown theoretically.

2.4 | Parameter nudging with proposal density

The effectiveness of the method proposed in the previous section, which augments parameters as artificial states, depends on the cross-covariance between states and parameters. To improve the accuracy and resilience of time-varying parameters, we introduce an optimization algorithm from machine learning into the parameter time evolution model using the flexibility of the proposal density in particle filtering. According to Equation 11, the model transition density is expressed as

$$p\left(z^n|z^{n-1}\right) = \mathcal{N}\left(\tilde{f}\left(z^{n-1}\right), \tilde{Q}^n\right). \tag{36}$$

The prior pdf expressed in Equation 12 is allowed to both divide and multiply the model transition density by a proposal transition density q, leading to

$$p\left(z^n\right) = \int \frac{p\left(z^n|z^{n-1}\right)}{q\left(z^n|\mathbf{Z}^{n-1}, y^n\right)} q\left(z^n|\mathbf{Z}^{n-1}, y^n\right) p\left(z^{n-1}\right)\, dz^{n-1}. \tag{37}$$

Drawing from $p\left(z^n|z^{n-1}\right)$ corresponds to using the original model transition density Equation 36. Still, we could instead draw from $q\left(z^n|\mathbf{Z}^{n-1}, y^n\right)$, which would correspond to any other model transition that we choose. This allows us to control the transition of both state and parameters by choosing proposal density q.

Sequential observation data can be considered as samples for the stochastic gradient descent (SGD) algorithm based on the similarity between sequential DA and online learning or stochastic optimization, in that the data are given sequentially. The ideas in stochastic optimization have advanced in recent years in machine learning and deep learning with large-scale data. The basic problem structure classification and associated solutions are summarized in Hannah (2015). The effectiveness of SGD for large-scale learning problems, that is, cases with large-scale data, is also described in Bottou (2010). The optimization algorithm used in the proposed method is described in the next section. Assume an objective function $L_i^n(\theta)$ and consider the problem of minimizing this function, where the parameter θ minimizes $L_i^n(\theta)$. The parameter θ^n can be updated by the following iteration:

$$\theta^n \leftarrow \theta^{n-1} - \lambda g^n, \quad \left(g^n \in \nabla L_i^n(\theta)\right), \tag{38}$$

where λ is the step size, sometimes called the learning rate in machine learning contexts. The function g^n expresses the update rule for the parameter.

Here, we consider introducing the above parameter update analogy to the transition density modification. In the next step of the last observation n, that is, $n+1$, let us assume that instead of original transition density Equation 12, the proposal density q at time step $n+1$ for augmented state z can described as

$$\begin{aligned} q\left(z_i^{n+1}|z_i^n, y^n\right) &= \mathcal{N}\left(\tilde{f}\left(z_i^n\right) + \begin{pmatrix} 0 \\ -\lambda g\left(\theta_i^{n-1}, y^n\right) \end{pmatrix}, \tilde{Q}^{n+1}\right) \\ &\equiv \mathcal{N}\left(\tilde{f}\left(z_i^n\right) + \tilde{g}^n,\ \tilde{Q}^{n+1}\right), \end{aligned} \tag{39}$$

where the augmented nudging term is denoted as $\tilde{g}^n$. Therefore, the step size λ and the function $g(\theta_i^{n-1}, y^n)$ have the same role as Equation 38 and together express the nudging term forcing estimated model parameters towards true values, and y^n is the last observed data vector. $\tilde{Q}^n$ is the same augmented model-error covariance matrix as

described in Equation 11 with correlated perturbation. Then updating of the augmented state vector after the last observation step n is given as follows, instead of the original updating expressed in Equation 5:

$$z_i^{n+1} = \tilde{f}\left(z_i^n\right) + \hat{\rho}_i^{n+1}, \tag{40}$$

where

$$p\left(\hat{\rho}^{n+1}\right) = \mathcal{N}\left(\tilde{g}^n,\ \tilde{Q}^{n+1}\right). \tag{41}$$

This corresponds to only the modification of augmented perturbation $\hat{\rho}^{n+1}$, which shifts the mean value of parameters. Note that sampling from this proposal transition density instead of the original model is compensated by an extra weight as described in Ades and Van Leeuwen (2015):

$$\begin{aligned}
&\frac{p\left(z_i^{n+1}|z_i^n\right)}{q\left(z_i^{n+1}|z_i^n, y^n\right)} \\
&\propto \exp\left[-\frac{1}{2}\left(z_i^{n+1} - \tilde{f}(z_i^n)\right)^{\mathrm{T}}\tilde{Q}^{-1}\left(z_i^{n+1} - \tilde{f}(z_i^n)\right)\right. \\
&\left.+\frac{1}{2}\left(z_i^{n+1} - \left(\tilde{f}(z_i^n) + \tilde{g}^n\right)\right)^{\mathrm{T}}\tilde{Q}^{-1}\left(z_i^{n+1} - \left(\tilde{f}(z_i^n) + \tilde{g}^n\right)\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(\hat{\rho}_i^{n+1}\right)^{\mathrm{T}}\tilde{Q}^{-1}\ \hat{\rho}_i^{n+1}\right. \\
&\left.+\frac{1}{2}\left(\hat{\rho}_i^{n+1} - \tilde{g}^n\right)^{\mathrm{T}}\tilde{Q}^{-1}\left(\hat{\rho}_i^{n+1} - \tilde{g}^n\right)\right].
\end{aligned} \tag{42}$$

2.5 | Adam-method-based parameter nudging

As mentioned above, we introduced a nudging term for the parameters by taking advantage of the flexibility of the proposal density in particle filtering. One of the main points in this article is that we can choose any term that forces the parameters toward the true value. Therefore, our scheme is combined with a well-known gradient descent optimization algorithm that has evolved in recent years as deep learning progresses (Alom et al., 2018). In general, a task in machine learning and deep learning is often expressed as the problem of finding parameters that minimize (or maximize) the objective function, and the key is how quickly the optimal parameters can be found. Typical optimization formulations and algorithms are summarized in Sun et al. (2019).

Regarding gradient-based optimization algorithms, Ruder (2016) showed a classification of algorithms and a description of typical examples. Momentum-based algorithms accumulate a decaying sum of the previous gradients into a momentum vector and use that instead of the true gradients. This method has the advantage of accelerating optimization along dimensions where the gradient remains relatively consistent and slowing it along turbulent dimensions where the gradient is significantly oscillating. Another approach is norm-based algorithms, which divide a portion of the gradient by the L_2 norm of all previous gradients. This has the advantage of slowing down along dimensions that have already changed and accelerating along dimensions that have only changed slightly. In our method, we use the adaptive moment estimation (Adam) proposed by Kingma and Ba (2014), which combines the above two approaches.

Our proposed formulation of the function $g(\theta_i^{n-1}, y^n)$ for the parameter nudging term in Equation 39 is as follows. First, $\tilde{f}(z_i^{n-1})$ can be regarded as the expected value of z_i^n given z_i^{n-1} and is defined by

$$\bar{z}_i^n = E[z_i^n|z_i^{n-1}] = \tilde{f}(z_i^{n-1}). \tag{43}$$

Next, we chose the log-likelihood of $p\left(y^n|\bar{z}_i^n\right)$ as the aforementioned objective function L_i^n in Equation 38 as follows:

$$L_i^n \equiv -2\log\left[p\left(y^n|\bar{z}_i^n\right)\right]. \tag{44}$$

Here, Equation 44 can be calculated from the likelihood with respect to the observed value y^n at observation step n and ensemble member i, given $\bar{z}_i^n$, as follows:

$$p\left(y^n|\bar{z}_i^n\right) \propto \exp\left[-\frac{1}{2}\left(y^n - \tilde{H}\bar{z}_i^n\right)^{\mathrm{T}}R^{-1}\left(y^n - \tilde{H}\bar{z}_i^n\right)\right]. \tag{45}$$

Then, we define the function $g(\theta_i^{n-1}, y^n)$ in Equation 39 by using the gradient of the objective function L_i^n as follows. Following Kingma and Ba (2014), we introduce the moving averages of the gradient and the squared gradient, and denote them as m_i^n and v_i^n, respectively. Their update equations are expressed using the gradient of L_i^n as follows:

$$\begin{aligned}
m_i^n &= \mu_m m_i^{n-1} + (1-\mu_m)\nabla_\theta L_i^n, \\
v_i^n &= \mu_v v_i^{n-1} + (1-\mu_v)\left(\nabla_\theta L_i^n\right)^2,
\end{aligned} \tag{46}$$

where the hyperparameters μ_m and μ_v control the decay rate of these moving averages. Note that the gradient $\nabla_\theta L_i^n$ requires computing the partial derivatives of the likelihood with respect to the parameters in Equation 45 or an approximation thereof. Since these moving averages are initialized (as a vector of zeros), the moment estimates are biased toward zero, especially during the initial time step and especially when the decay rates are low (i.e., μ_m and μ_v are chosen to be close to 1). Therefore, m_i^n and v_i^n in Equation 46 are modified as follows to cancel these biases:

$$\hat{m}_i^n = \frac{m_i^n}{1-\mu_m},\ \hat{v}_i^n = \frac{v_i^n}{1-\mu_v}. \tag{47}$$

Finally, the function $g(\theta_i^{n-1}, y^n)$ expressed in Equation 39 is yielded as follows:

$$g(\theta_i^{n-1}, y^n) = \frac{\hat{m}_i^n}{\sqrt{\hat{v}_i^n} + \delta}. \tag{48}$$

Here, the factor $\sqrt{\hat{v}_i^n}$ represents the L_2 norm of the past gradients via the v_i^{n-1} term and current gradient in Equation 46, and scales the gradient. Note that δ is a factor to avoid dividing by zero and set to 1.0×10^{-8} in the following experiment.

The proposed method contains two procedures dependent on the observation: (1) state and parameter update by IEWPF and computation of likelihood gradient at the observation step, and (2) parameter nudging with proposal density between observations. The algorithm is summarized as follows:

(1) State and parameter update at the observation step

- Sample initial particle for state x_i^0 and parameter $\theta_i^0, i = 1, \ldots, N$.
- For every model time step k:
  - Perform forecast based on model transition and error covariance Q_β^k.
  - Generate parameter perturbation from parameter error covariance Q_η^k.
  - Compute parameter differentiation using model and parameter perturbation, then update augmented covariance matrix $\tilde{Q}^k$.

- When the model reaches the observation time t, for each particle i:
  - Compute ϕ_i for all particles by Equation 27, then determine c_i from Equation 33.
  - Calculate α_i that satisfies Equation 35 using the analytical solution of the Lambert W function.
  - Update the state and parameter using Equations 20 and 22.
  - Normalize and update the weight.

- In preparation for the next forecast step:
  - Compute likelihood L_i^t from observation y^t and observation-error covariance R by Equation 45, then obtain likelihood gradient $\nabla_\theta L_i^t$ from Equation 44.
  - Compute parameter nudging term $\lambda g(\theta_i^{t-1}, y^t)$ from Equation 48, by using hyperparameters μ_m, μ_v, and step-size factor λ.

(2) Parameter nudging at the forecast step

- The time step $t+1$ in the next step after observation, for each particle i:
  - Generate parameter perturbation using the computed parameter nudging term $\lambda g(\theta_i^{t-1}, y^t)$ from Equation 41.
  - Compute extra weight in Equation 42.
  - Perform forecast using Equation 40.

3 | NUMERICAL EXPERIMENTS

The effectiveness of the proposed method is demonstrated through two synthetic test cases as follows. The first case is the linear model with additive parameters, where all model states are observed directly at every time step. Although this article focuses on a nonlinear system, we use a linear model to verify that the shape of the posterior pdf is close to the true one. The second case is the Lorenz-96 model (Lorenz, 1996) with parameterized forcing, where only the model states are observed directly at every fourth step.

3.1 | Linear model with an unknown parameter

In order to compare the estimates of the proposed method with the analytically calculated true values, we use the following linear model as the time evolution expressed in Equation 2:

$$\begin{pmatrix} x^n \\ \theta^n \end{pmatrix} = \begin{pmatrix} F_x & F_{x\theta} \\ O & I \end{pmatrix} \begin{pmatrix} x^{n-1} \\ \theta^{n-1} \end{pmatrix} + \begin{pmatrix} \beta^n \\ \eta^n \end{pmatrix}, \tag{49}$$

where $x \in \mathbb{R}^{N_x}$ is the model state vector with dimension N_x and $\theta \in \mathbb{R}^{N_\theta}$ is the parameter vector with dimension N_θ. β and η are random perturbations drawn from the model-error pdf $\mathcal{N}(0, Q_\beta)$ and parameter error pdf $\mathcal{N}(0, Q_\eta)$, respectively. The matrix $F_x \in \mathbb{R}^{N_x \times N_x}$ and $F_{x\theta} \in \mathbb{R}^{N_x \times N_\theta}$ represent the linear model. Here, we define the matrices $\tilde{F}$ and $\tilde{G}$ as follows:

$$\tilde{F} = \begin{pmatrix} F_x & F_{x\theta} \\ O & I \end{pmatrix}, \quad \tilde{G} = \begin{pmatrix} I & F_{x\theta} \\ O & I \end{pmatrix}. \tag{50}$$

Then, Equation 49 can be rewritten by using Equation 4 as follows:

$$\begin{pmatrix} x^n \\ \theta^{n-1} \end{pmatrix} = \tilde{F} \begin{pmatrix} x^{n-1} \\ \theta^{n-2} \end{pmatrix} + \tilde{G} \begin{pmatrix} \beta^n \\ \eta^{n-1} \end{pmatrix}, \tag{51}$$

When the initial prior pdf is Gaussian, the true posterior pdf should also be Gaussian. Assuming that the posterior pdf at time $n-1$ is Gaussian with covariance

matrix $P_{n-1|n-1}$, the predicted covariance matrix $P_{n|n-1}$ of the prior pdf expressed in Equation 51 can be calculated as follows:

$$P_{n|n-1} = \tilde{F} P_{n-1|n-1} \tilde{F}^{\mathrm{T}} + \tilde{G} \tilde{Q} \tilde{G}^{\mathrm{T}}, \tag{52}$$

where

$$\tilde{G} \tilde{Q} \tilde{G}^{\mathrm{T}} = \begin{pmatrix} F_{x\theta} Q_\eta F_{x\theta}^{\mathrm{T}} + Q_\beta & F_{x\theta} Q_\eta \\ \left(F_{x\theta} Q_\eta\right)^{\mathrm{T}} & Q_\eta \end{pmatrix}, \tag{53}$$

and this term is equivalent to Equation 11 when using the linear model $\tilde{F}$ defined in Equations 50 and 51.

In the following experiments, we choose the dimension of the model state $N_x = 1000$ and the parameter $N_\theta = 1$, in order to consider a simple high-dimensional system with a parameter. Setting the model $F_x = I$, $F_{x\theta} = 0.1$, the time evolution model described in Equation 51 and observation model are expressed as

$$\begin{aligned} z^n &= \begin{pmatrix} x_j^n \\ \theta^{n-1} \end{pmatrix} = \begin{pmatrix} x_j^{n-1} + 0.1\, \theta^{n-2} \\ \theta^{n-2} \end{pmatrix} + \begin{pmatrix} \beta^n + 0.1 \eta^{n-1} \\ \eta^{n-1} \end{pmatrix}, \\ y^n &= \tilde{H} z^n + \epsilon^n, \end{aligned} \tag{54}$$

where index $j = 1, \ldots, N_x$ indicates the elements of the model states x. Here, the observation model $\tilde{H} = (I \;\; 0)$, assuming that all variables are observed, and ϵ is the observation error drawn from the observation-error pdf $\mathcal{N}(0, R)$. Since we assume a time-independent state transition matrix $\tilde{F}$, the covariance matrix satisfying the linear system defined by Equation 54 converges to the steady-state matrix P such that $P_{n|n-1} = P_{n-1|n-2} \equiv P$, and satisfies the discrete-time Riccati equation (Wonham, 1968) as follows:

$$P = \tilde{F} P \tilde{F}^{\mathrm{T}} - \tilde{F} P \tilde{H}^{\mathrm{T}} \left(\tilde{H} P \tilde{H}^{\mathrm{T}} + R\right)^{-1} \tilde{H} P \tilde{F}^{\mathrm{T}} + \tilde{G} \tilde{Q} \tilde{G}^{\mathrm{T}}. \tag{55}$$

Therefore, the shape of the true posterior pdf of Equation 54 can be obtained by solving Equation 55 numerically and compared with the distribution obtained from the proposed IEWPF.

The procedure of the comparison using synthetic data is as follows. Let us assume the initial ensemble members z_i^0 are sampled from the background error $\mathcal{N}(0, B)$. First, one member from the ensemble generated under the model-error covariance matrix Q and the background-error covariance matrix B is used as the "truth". Observations are then created from this "truth" and the observation error defined by covariance matrix R. In the following experiments, the true value of the parameter is 0, and the true model-error covariance matrix Q is chosen as a diagonal matrix with the main diagonal value 0.04 for states and 0 for the parameter. The background-error covariance matrix B is a diagonal matrix with the main diagonal values of 1 for states and 0 for the parameter. The observation-error matrix R is diagonal, and the main diagonal value is set to 0.01.

Next, for the assimilation, we choose the same matrix Q_β, B for states, and R as when the observation was generated. The matrix Q_η and B for parameters are set to be the same as those of the states. The number of particles is set to $N = 20$ to demonstrate the validity of the estimation with few particles. Regarding observations, consider the condition that all model state variables x are observed at every step. Note that the step size λ in Equation 39 is set to 0 to evaluate the parameter augmentation method of IEWPF described in Section 2.2. In order to investigate the dependence of the aforementioned α_i on the shape of the posterior pdf, we compare the variance of pdfs estimated with the values sampled from the $\alpha_i \geq 1$ branch at three sampling percentages: 0%, 50%, and 100%. Note that 50% means sampling from both branches of $\alpha_i \geq 1$ and $\alpha_i \leq 1$, which is the closest to the true pdf according to Zhu et al. (2016). Thus, 0% and 100% mean sampling only from $\alpha_i \leq 1$ branch and $\alpha_i \geq 1$ branch, respectively.

Figure 1 shows histograms of variance accumulated from the 20th to 1000th steps for comparing the two sampling cases of α with the diagonal value of $R = 0.01$. The variances of both (a) states $Var(x)$ and (b) parameter $Var(\theta)$ are averaged over the dimension, that is, $N_x = 1000$ and $N_\theta = 1$ for the variables and parameter, respectively, and the number of particles N_p for each dimension, as follows:

$$\begin{aligned} \overline{Var(x_j^n)} &= \frac{1}{N_x} \sum_{j=1}^{N_x} \frac{1}{N_p} \sum_{i=1}^{N_p} \left(x_i^n - \overline{x^n}\right)_j^2, \\ \overline{Var(\theta^n)} &= \frac{1}{N_p} \sum_{i=1}^{N_p} \left(\theta_i^n - \overline{\theta^n}\right)^2, \end{aligned} \tag{56}$$

where the index j denote the elements of the states x, and $\overline{x^n}$ and $\overline{\theta^n}$ are the ensemble mean. Note that the dimension of the parameter θ is one. The true variances based on the solution of Equation 55 are shown as "True". From these comparisons, both the states and parameter variances are close to the "True" value when sampling 50% from the $\alpha_i \leq 1$ branch. On the other hand, when sampling only from the $\alpha \leq 1$ branch and the $\alpha \geq 1$ branch, we see that the variance becomes smaller and larger with the same trend as for Zhu et al. (2016), respectively.

Figure 2 compares the posterior pdf obtained in the 50% sampling case with the true pdf for the diagonal value of R of 0.01. Since the ensemble size is too small compared with the number of model dimensions, both of the estimated pdfs are shown as the histogram accumulated over the time evolution from 20th to 1000th steps for the state and parameter, respectively. From Figure 2a,b, we see that

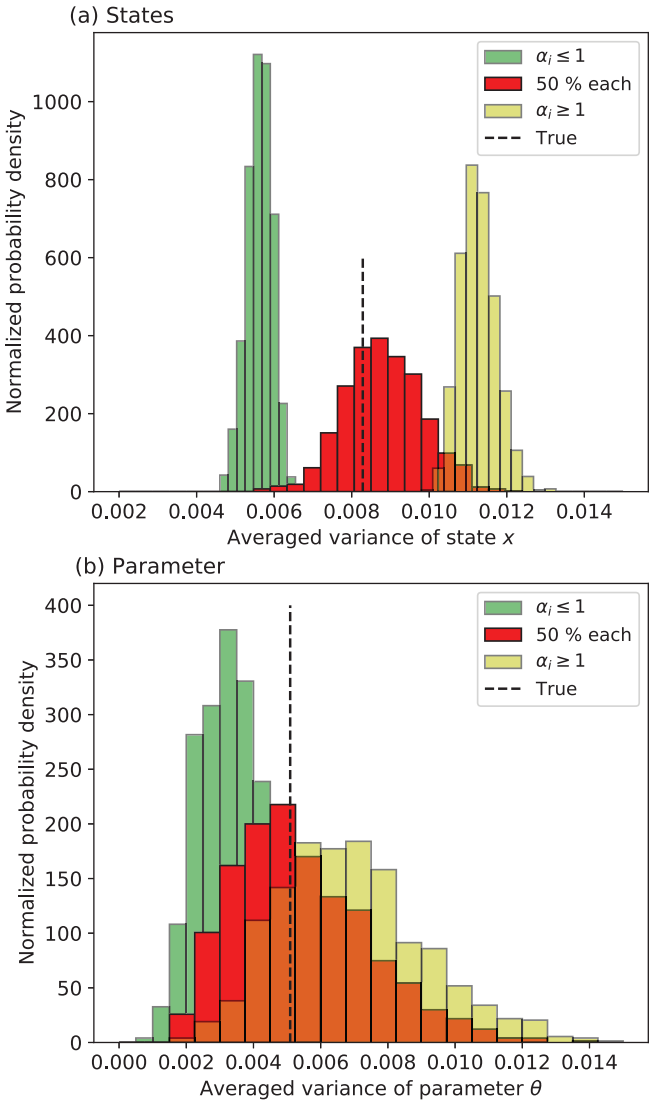

FIGURE 1 Histogram of cumulative variance comparing the diagonal values of $R = 0.01$ for (a) states and (b) parameter, respectively. Three sampling percentages from the $\alpha \leq 1$ branch: 100%, 50%, and 0% are compared with the true variance (dashed line). [Colour figure can be viewed at wileyonlinelibrary.com]

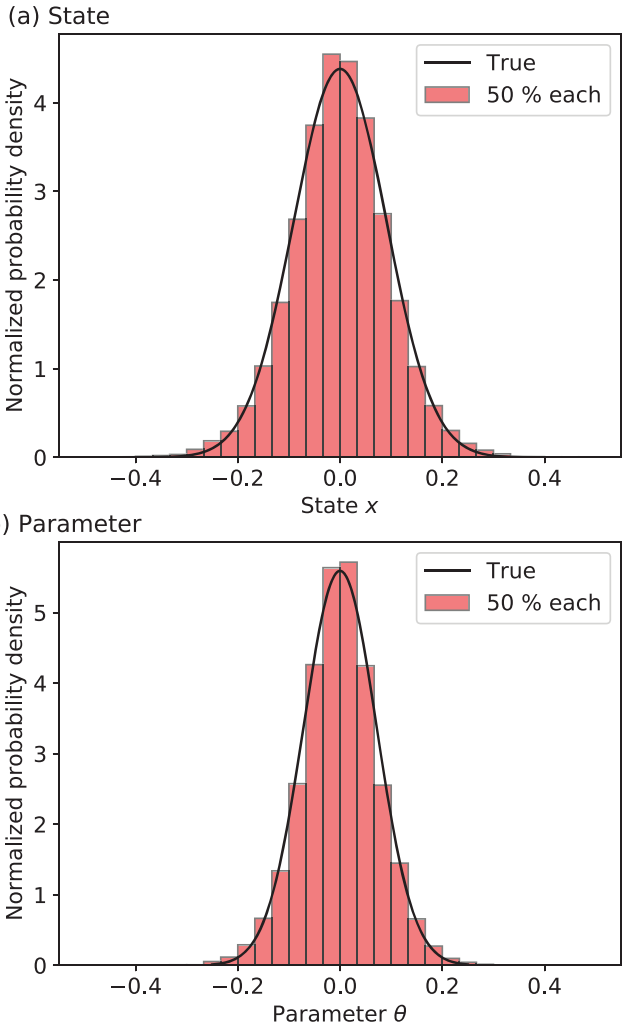

FIGURE 2 Posterior pdf represented by the particles using the 50% sampling case compared with true pdf (full line) for (a) state x_1 of element one and (b) parameter θ, respectively. [Colour figure can be viewed at wileyonlinelibrary.com]

the obtained pdf of the state x_1 and parameter θ is close to the true pdf.

These results indicate that the method of extending IEWPF to the proposed augmented state-space model is valid, and the variance and shape of the posterior pdf for the parameter are also close to those of true pdf under the condition that the variance and shape of the posterior pdf for the state are close to those of true pdf.

3.2 | Lorenz-96 model with parameterized forcing

The Lorenz 1996 model with parameterized forcing is used as the time evolution expressed in Equation 1 to explore the validity of the proposed method in a nonlinear high-dimensional system. The original Lorenz-96 model (Lorenz, 1996) is the dynamical nonlinear model given by

$$\frac{d}{dt}x_j = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F_j, \tag{57}$$

where index $j = 1, \ldots, N_j$ with cyclic indices, x_j is the state variable of the model at position j, N_j is total dimension, and F_j is the forcing function parameterized by

$$F_j(\theta_0, \theta_1, \theta_2) = c_0\theta_0 + c_1\theta_1 \sin\left(\frac{2\pi}{c_2\theta_2}j\right), \tag{58}$$

for which c_0, c_1, c_2 are true values, and $\theta_0, \theta_1, \theta_2$ are their scale parameters that have to be estimated. For the evaluation of nonlinearity, this value of F_j, which is typically chosen to be 8 or more to generate chaotic behavior, is set as follows. The values of c_0, c_1 are set to 8, 4 respectively,

and c_2 is set to the same value as the dimension of the model state: N_j. Then, the scale parameters $\theta_0, \theta_1, \theta_2$ are estimated, and their true values are 1 each. By introducing this parameterized forcing term $F_j(\theta_0, \theta_1, \theta_2)$, each state variable x_j contains a parameter-dependent chaotic behavior. This model is numerically solved by the fourth-order Runge–Kutta scheme with a time step of $\Delta_t = 0.05$.

The procedure for the following experiment is the same as for the previous linear model. The true model-error covariance matrix Q_β for states is chosen as a tridiagonal matrix, the main diagonal value being 0.10 and both sub- and superdiagonal values being 0.025. The background-error covariance matrix B is a diagonal matrix with the main diagonal value 1 for states. In the experiments below, the true observation-error matrix R is diagonal, with main diagonal values of 0.02. For the assimilation, we choose the same matrix Q_β, B for states and R as when the observation was generated, that is, the true one. The matrices Q_η, B for parameters are diagonal matrices with main values $5.0 \times 10^{-6}, 0.001$, respectively. The step size λ for the Adam method is set to 0.001. The number of particles is set to $N = 20$ to demonstrate the validity of the estimation with few particles. To consider high-dimensional cases, N_j is chosen as 1000, the same as in the linear-model experiment.

In contrast to the previous evaluation using the linear model and a static parameter, this experiment investigates the ability of the proposed methods for estimating time-varying (i.e., dynamic) parameters in nonlinear high-dimensional systems. Regarding observations, consider the condition that all of the model states are observed every fourth step (i.e., the assimilation interval is 4). Moreover, this 1000-dimensional evaluation with only 20 particles can validate its feasibility to apply to realistic geophysical, climate, and other problems. First, we compare the methods outlined in Section 2 in terms of the RMSE and the ensemble spread (Spread). Next, we compare the impact of the parameter error covariance Q_η and the step size factor λ on the ensemble. The performance indicator of parameter estimation is not only the RMSE but also the ratio of the RMSE to the spread in the ensemble, and it is preferable that their ratio becomes one for Gaussian variables. Note that, for non-Gaussian variables, this is only true for the forecast ensemble (Fortin et al., 2014).

3.2.1 | Comparison of the methods

Figure 3 compares the true values and particle trajectories in the three methods mentioned above for the state x_1 and the three scale parameters $\theta_0, \theta_1, \theta_2$. All variables are observed every four steps, setting the main diagonal value of matrix R to 0.02. Each true parameter is increased by 30% at the 200th step, as the dashed red line shows. The figure shows the difference in tracking performance of the three methods for abrupt parameter changes and the advantage of the proposed method. The method shown in Figure 3a MH1 is the conventional augmented method expressed as Equation 2. There are some steps where the trajectories of each ensemble deviate from the true trajectory in the state, and the ensemble spreads out greatly and cannot track abrupt changes in all three parameters. Then, both of the methods shown in Figure 3b MH2 and Figure 3c MH3 are based on the proposed state-space model expressed as Equation 5 with the covariance matrix $\tilde{Q}$. The method shown in Figure 3c MH3 further applies the Adam-method-based nudging described in Section 2.5 with step-size factor $\lambda = 0.001$. The results for the state show that the trajectories of each ensemble are close to the true trajectory. Although both methods tend to approach the true values for θ_0 and θ_2, the Adam-method-based nudging is more accurate and responsive to abrupt changes, especially for θ_1.

Figure 4a,b shows the comparisons of time series RMSE for the states and parameters, respectively. The horizontal axis indicates the time steps in the 100th–600th steps, where the difference between methods is significant in Figure 3. For the state, since the assimilation interval is four, each value represents the average of all elements (i.e., 1000) for the third step, which has the largest prediction error after filtering, while for the parameter, the average values of all elements (i.e., 3) for all steps are shown. The results show that the estimation error of the parameters after the parameter abrupt change (200th step) increases the error in the forecast step of the model states, and the estimation error of the proposed method (MH3) decreases the fastest for both states and parameters.

Figure 5a,b shows the RMSE and spread comparisons for the states and parameters, respectively. Each box plot shows the time-averaged RMSE and spread at the forecast and filtering steps in the 100th–1500th steps shown in Figure 3, including the abrupt change (at 200th steps). Therefore, the interquartile range (IQR) of the box plot indicates the dispersion across the dimensions of the model states (1000) and parameters (3). Note that outliers are not plotted, to exclude estimation errors immediately after abrupt changes in the 200th step. From the result for the states shown in Figure 5a, the proposed methods (i.e., MH2 and MH3) have smaller RMSE values and dispersion than the conventional methods (i.e., MH1), especially in the forecast step. The result for the parameters shown in Figure 5b clearly shows that both the RMSE values and dispersion of MH3 (i.e., with nudging) are smaller than the others, and the spread is also smaller. The fact that the RMSE dispersion of MH3 is smaller than that of MH2 means that the difference in RMSE in the three parameters

is small. Thus, the proposed nudging method reduces differences in estimation accuracy for each parameter, which is the effectiveness of combining IEWPF with Adam.

3.2.2 | Dependence of parameter error covariance and step-size factor

In the following, we investigate the impact of the parameter error covariance Q_η and the step-size factor λ on estimation accuracy (RMSE) and ensemble spread (spread). Figure 6 shows the true values and the particle trajectories of the scale parameter θ_0 under the combination of different values of Q_η and λ, respectively. Note that Q_η is chosen as a diagonal matrix and we denote it as $Q_\eta = \sigma_\eta^2 I$. The graph shown in Figure 6 as exp2 is the reference condition with $\sigma_\eta^2 = 5.0 \times 10^{-6}$, $\lambda = 0.001$, and is the same graph shown for scale parameter θ_0 in Figure 3c. The other graphs exp1, exp3, and exp4 in Figure 6 show the cases where σ_η^2 is 1.0×10^{-6}, 1.0×10^{-5}, and 5.0×10^{-5}, respectively, under the same value of $\lambda = 0.001$. These graphs show that the larger the parameter covariance, the larger the ensemble spread and the less overshoot after the parameter abrupt change.

Next, we quantitatively evaluate the impact of the parameter error covariance Q_η on the ensemble. Figure 7 shows the dependence of the parameter error covariance Q_η on RMSE and spread for (a) states and (b) parameters, respectively. Each box plot shows the time-averaged RMSE and spread at the forecast and filtering steps in the 100th–1500th steps. The forecast RMSE and spread include three cycles of forecast steps, since the filtering interval is four. The four values of σ_η^2 shown on the horizontal axis are for exp1, exp2, exp3, and exp4 in Figure 6. Note that outliers are not plotted to exclude estimation errors immediately after abrupt changes in the 200th step. For the states, we can see from Figure 7a that neither the value of RMSE nor the value of spread depends on the diagonal value of the parameter error covariance Q_η. In addition, the values of forecast RMSE and spread are close, that is, their ratio is close to one. On the other hand, for the parameters, Figure 7b shows that as the diagonal values σ_η^2 increase, the values of spread also increase, and

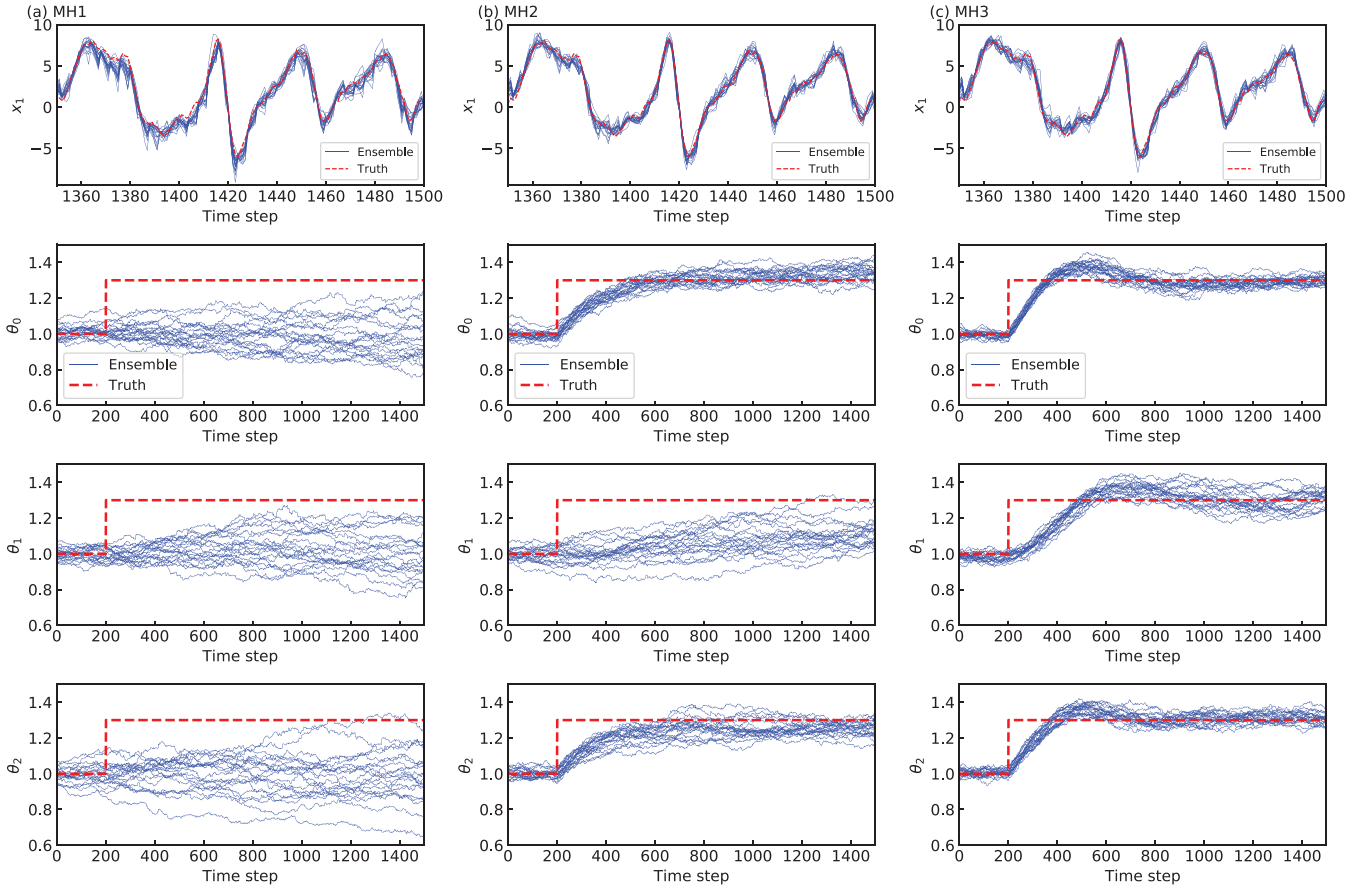

FIGURE 3 Comparison of estimated state and parameter trajectories between (a) conventional augmented method (MH1), (b) without nudging: $\lambda = 0$ (MH2), and (c) with nudging: $\lambda = 0.001$ (MH3). The solid lines show each of the 20 ensemble members, and the dashed lines show the true parameter value. Only the 1350–1500th steps are shown for the state, and each true parameter is increased by 30% at the 200th step. [Colour figure can be viewed at wileyonlinelibrary.com]

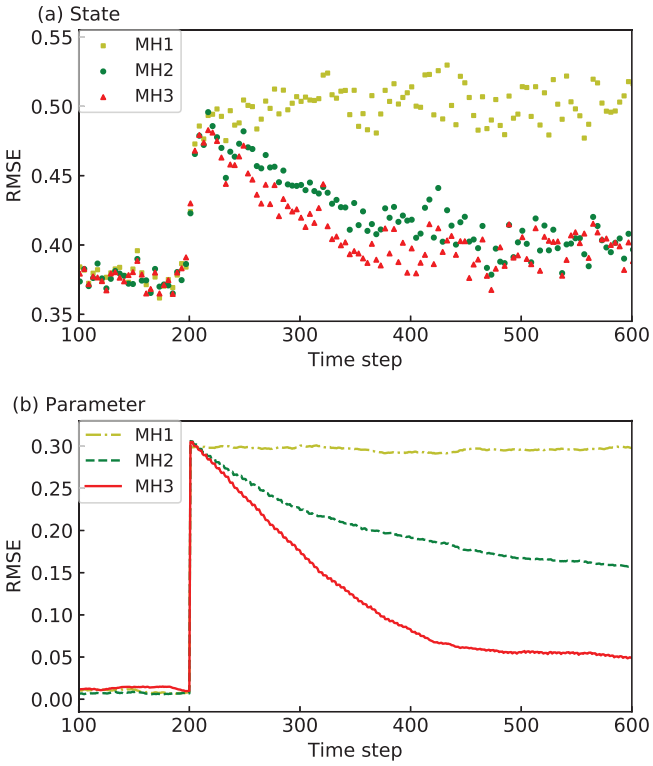

FIGURE 4 Comparison of time series RMSE after parameter abrupt change (200th step) between augmented method (MH1), without nudging: $\lambda = 0$ (MH2) and with nudging: $\lambda = 0.001$ (MH3) as per Figure 3. The third step after the filtering for the (a) state and all steps for the (b) parameter are shown. Each value is averaged over all elements. [Colour figure can be viewed at wileyonlinelibrary.com]

the values of RMSE decrease. Especially in the case of $\sigma_\eta^2 = 5.0 \times 10^{-5}$, the values of forecast RMSE and spread are close, that is, their ratio is close to one.

Figure 8 shows the true values and the particle trajectories, as in Figure 6. The graph of exp2 is the same as in Figure 6 exp2 of the reference condition with $\sigma_\eta^2 = 5.0 \times 10^{-6}$, $\lambda = 0.001$. The exp5, exp6, and exp7 in Figure 8 show the cases where λ is 0.0005, 0.002, and 0.004, respectively, under the same value of $\sigma_\eta^2 = 5.0 \times 10^{-6}$. These graphs show that the larger the step-size factor, the faster the value approaches the true value after the abrupt change, but the more likely it is to overshoot.

Figure 9 shows the dependence of the step-size factor λ on RMSE and spread for (a) states and (b) parameters, respectively. Each box plot shows the time-averaged RMSE and spread at the forecast and filtering steps during the 100th–1500th steps, and the forecast RMSE and spread include three cycles of forecast steps, as in Figure 7. The four values of λ shown on the horizontal axis are for exp5, exp2, exp6, and exp7 in Figure 8. Note that outliers are not plotted as in Figure 7. Similarly to the trend shown in Figure 7, there is almost no dependence of the step-size factor λ on the RMSE and spread for states. For parameters,

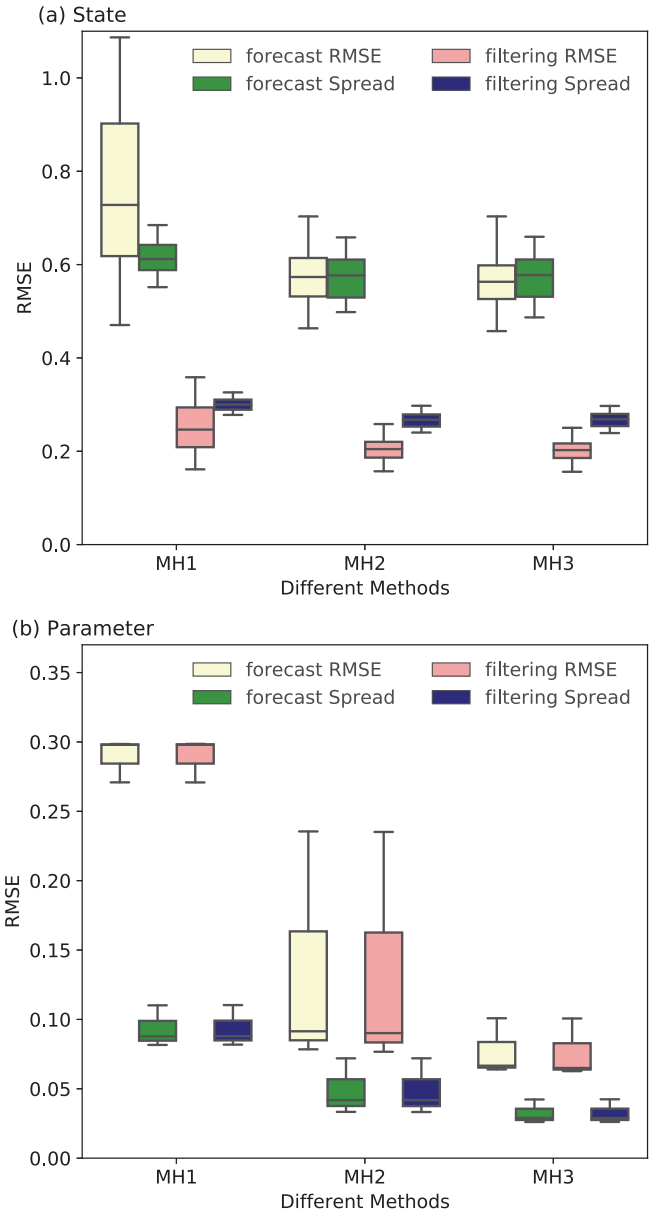

FIGURE 5 Box plot showing the comparisons of RMSE and spread for forecast and filtered ensembles between augmented method (MH1), without nudging: $\lambda = 0$ (MH2) and with nudging: $\lambda = 0.001$ (MH3) as per Figure 3. Each IQR indicates the dispersion of the (a) state and (b) parameter elements averaged over the forecast and filtering steps in 100–1500, respectively. Outliers are not plotted. [Colour figure can be viewed at wileyonlinelibrary.com]

the spread does not increase even as the step-size factor λ increases, but the RMSE decreases, that is, the ratio of the forecast RMSE to spread approaches one.

3.2.3 | Dependence of observation error and number of observations

In order to evaluate the dependence of the observation error and number of observations, we compare the large step-size condition: $\lambda = 0.004$ (exp7) with two additional

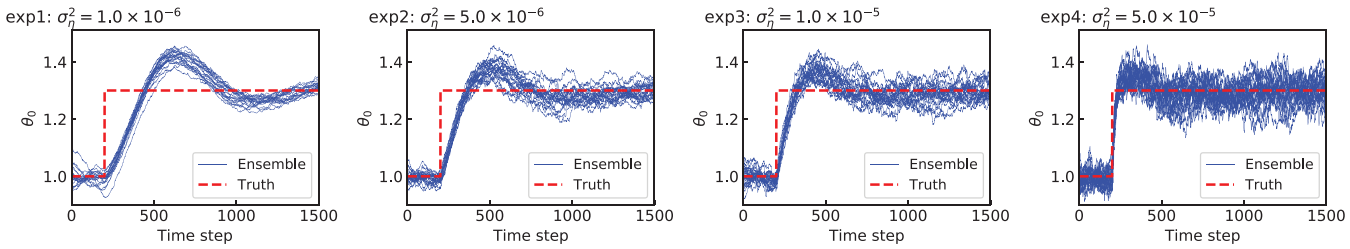

FIGURE 6 Comparison of estimated parameter trajectories between different values of σ_η^2: 1.0×10^{-6} (exp1), 5.0×10^{-6} (exp2), 1.0×10^{-5} (exp3), and 5.0×10^{-5} (exp4) under the same value of $\lambda = 0.001$. The solid lines show each of the 20 ensemble members, and the dashed lines show the true parameter value. Each true parameter is increased by 30% at the 200th step. [Colour figure can be viewed at wileyonlinelibrary.com]

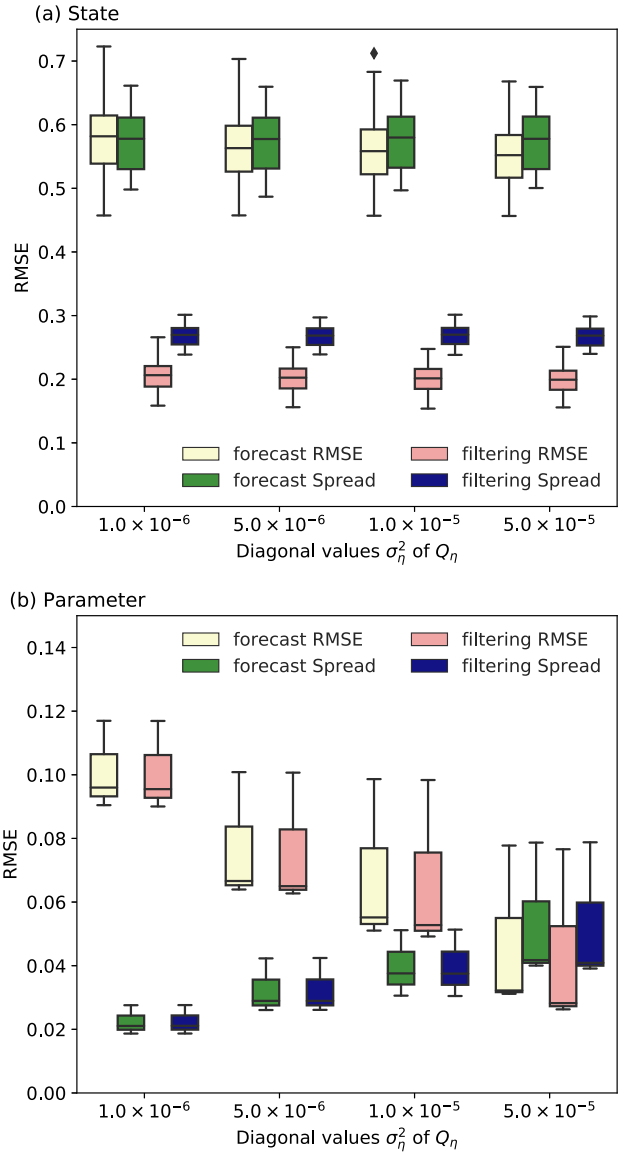

FIGURE 7 Box plot showing the comparison of RMSE and spread for each of the forecast and filtered ensembles between different values of $\sigma_\eta^2 = 1.0 \times 10^{-6}, 5.0 \times 10^{-6}, 1.0 \times 10^{-5}$, and 5.0×10^{-5} as per Figure 6. Each IQR indicates the dispersion of the (a) state and (b) parameter elements averaged over the forecast and filtering steps in 100–1500, respectively. Outliers are not plotted. [Colour figure can be viewed at wileyonlinelibrary.com]

experiments (exp8 and exp9). The first (exp8) is the case where the main diagonal value of the matrix R is large, and in the following, the value is set to 0.08. Note that this experiment (exp8) uses observation data generated at $R = 0.08$. Hence, R for data generation and assimilation are the same value. The second (exp9) is when the state is observed at every other grid point, so that $H_x x^n = \left(x_1^n, x_3^n, \ldots\right)^{\mathrm{T}}$. In both additional experiments, the conditions of the step size and the diagonal value of the parameter error covariance are the same as for exp7, that is, $\lambda = 0.004$, $\sigma_\eta^2 = 5.0 \times 10^{-6}$.

Figure 10 shows a comparison of RMSE and spread for different observation conditions for (a) state and (b) parameter. The description of the box plot is the same as in Figure 9. Figure 10 exp7 shows the results of the reference condition, that is, $R = 0.02$, and all model states are observed. From the comparison of the state in Figure 10a exp7 and exp8, the change in R from 0.02–0.08 increases both RMSE and spread, but spread is somewhat more pronounced. For the parameter in Figure 10b, RMSE values and dispersion tend to increase compared with spread. From comparison of the state in Figure 10a exp7 and exp9, because the number of observed variables was reduced to half, both RMSE and spread are increasing except for the filtering value of the observed variable. As for the parameters, both RMSE and spread show a small increase in median values, but an increase in dispersion. The results indicate that increasing observation error and decreasing observation density increase differences in estimation accuracy between parameters. In other words, the decrease in observed information has reduced the estimation accuracy of parameters with little impact (i.e., low sensitivity) on the model state. This could potentially be mitigated by adjusting the step size and the parameter error covariance.

4 | CONCLUSION

This article proposed a resilient and efficient state and time-varying parameter estimation method for nonlinear high-dimensional systems through a sequential DA

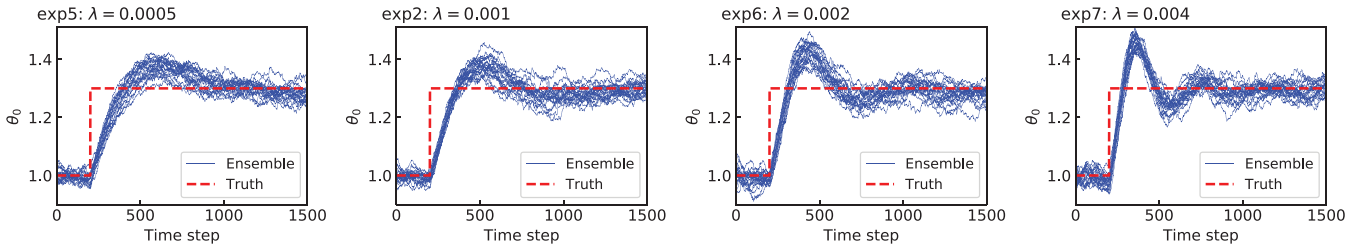

FIGURE 8 Comparison of estimated parameter trajectories between different values of λ: 0.0005 (exp5), 0.001 (exp2), 0.002 (exp6), and 0.004 (exp7) under the same value of $\sigma_\eta^2 = 5.0 \times 10^{-6}$. The solid lines show each of the 20 ensemble members, and the dashed lines show the true parameter value. Each true parameter is increased by 30% at the 200th step. [Colour figure can be viewed at wileyonlinelibrary.com]

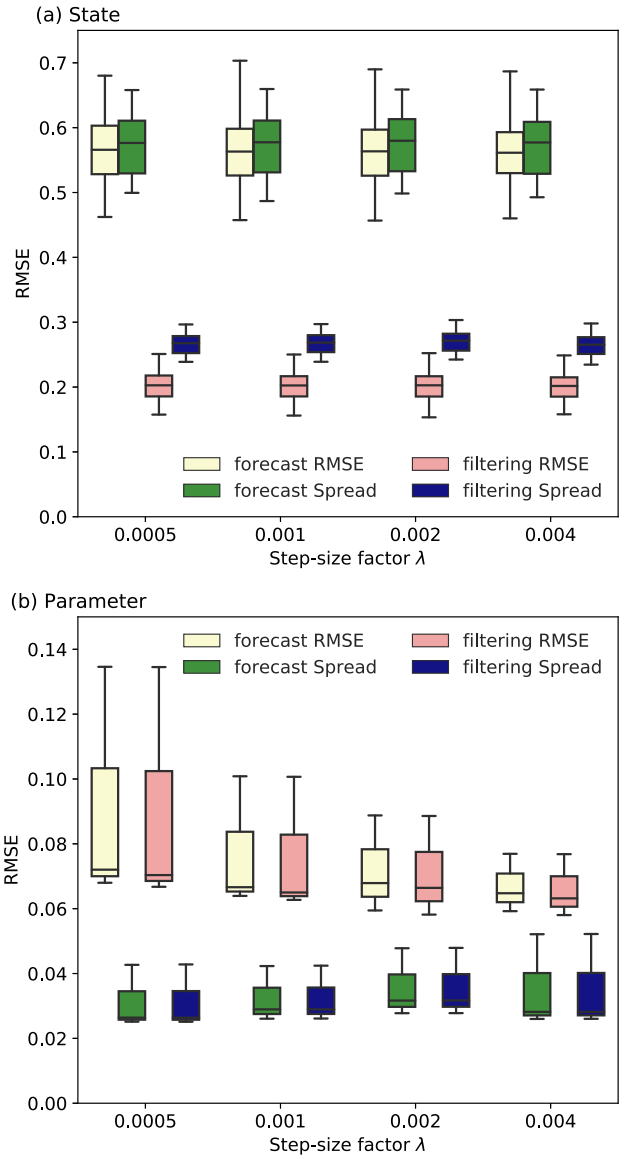

FIGURE 9 Box plot showing the comparison of RMSE and spread for each of the forecast and filtered ensembles between different values of $\lambda = 0.0005$, 0.001, 0.002, and 0.004 as per Figure 8. Each IQR indicates the dispersion of the (a) state and (b) parameter elements averaged over the forecast and filtering steps in 100–1500, respectively. Outliers are not plotted. [Colour figure can be viewed at wileyonlinelibrary.com]

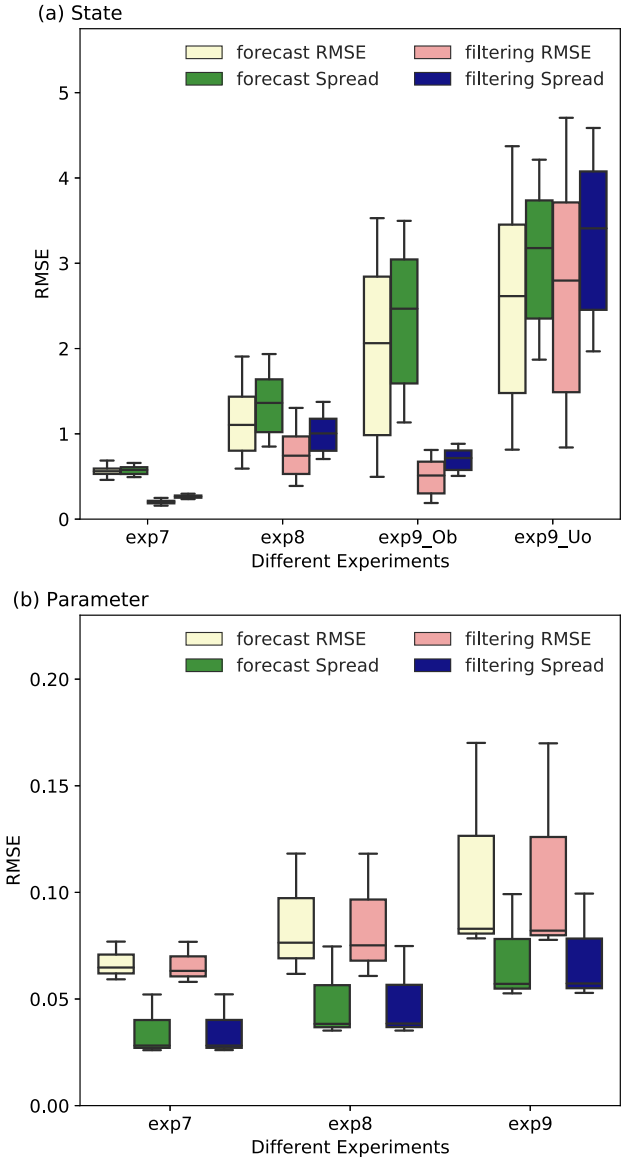

FIGURE 10 Box plot showing the comparisons of RMSE and spread for forecast and filtered ensembles between the large step-size condition (exp7), large observation error: $R = 0.08$ (exp8), and partially observed (exp9). Each IQR indicates the dispersion of the (a) state and (b) parameter elements averaged over the forecast and filtering steps in 100–1500, respectively. Outliers are not plotted. "Ob" and "Uo" represent observed and unobserved states. [Colour figure can be viewed at wileyonlinelibrary.com]

process. First, we introduced an extension of IEWPF to an augmented state-space model with a correlated covariance matrix. We then proposed the IEWPF-based method that incorporates the nudging technique inspired by optimization algorithms in machine learning into the parameter time evolution model by using the flexibility of the proposal density in particle filtering.

The performance of the method is examined in the 1000-dimensional linear model and nonlinear Lorenz-96 model. Experiments using the linear model with the static parameter indicate that the impact of the scalar factor α on the variance of the parameter is similar to that on the variance of the state. Numerically, under the condition that the variance and shape of the posterior pdf for the states are close to the true ones, those for the parameter are also close to the true ones.

The experimental results of the nonlinear Lorenz-96 model with the time-varying parameters show the following points. First, the proposed state augmentation method successfully estimates states and parameters simultaneously, even when the number of ensemble members is much smaller than the model dimension. This result indicates that filter degeneracy is avoided when extending to an augmented state-space model. Second, the proposed parameter nudging method inspired by optimization algorithms accelerates the tracking for abrupt parameter changes and reduces the difference in estimation accuracy for each parameter. This result suggests the effectiveness of combining IEWPF with Adam, one of the optimization algorithms. Thirdly, from evaluating the impact of the parameter error covariance and the step-size factor on the time-averaged RMSE and the ensemble spread (spread), the former increases the spread and decreases the RMSE, while the latter decreases the RMSE. Properly determining these values so that the ratio of the RMSE to the spread approaches one will allow for good ensemble generation. However, its systematic method will be a subject of future research. Finally, from evaluating the dependence of the observation error and number of observations, the decrease in observed information has reduced the estimation accuracy of parameters with little impact (i.e., low sensitivity) on the model state. This could potentially be mitigated by adjusting the step-size factor and the parameter error covariance. Alternatively, it may be beneficial to narrow the parameters to be estimated to those with high sensitivity through a preliminary sensitivity analysis.

In the numerical experiments in this article, the Lorenz-96 model with parameterized forcing was used mainly to evaluate the nonlinearity of time evolution of the model states, but further investigation of the nonlinearity of the parameters is needed. Adam optimization is a first-order gradient-based method, and it is widely used to learn the weights in deep neural networks, that is, nonlinear functions. Thus, our Adam-based nudging term can work theoretically in nonlinear problems. However, even for nonlinear convex problems, there are conditions and limits to convergence, and new methods have been proposed (Reddi et al., 2018). Furthermore, convergence for nonconvex problems is still an open question, though Chen et al. (2019) developed an analysis framework and a set of sufficient conditions that guarantee convergence. Therefore, the applicability of the proposed method to various nonlinear problems in data assimilation needs to be investigated and is a topic for future research.

In this article, we applied the proposed online parameter estimation scheme to IEWPF as an example of a PF that can avoid filter degeneracy. The method is shown to be capable of resilient and efficient parameter estimation for time-varying parameters. The results lead to the conjecture that the proposed method is applicable to realistic geophysical, climate, and other problems. Since several approaches have been proposed to avoid filter degeneracy (e.g., Skauvold et al., 2019), the evaluation of another combination will be a subject of future research.

AUTHOR CONTRIBUTIONS

Mineto Satoh: conceptualization; investigation; methodology; software; validation; writing – original draft. **Peter Jan van Leeuwen:** methodology; software; writing – review and editing. **Shin'ya Nakano:** funding acquisition; supervision; writing – review and editing.

ACKNOWLEDGEMENTS

We thanks the reviewers for their help in improving the article.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest, financial or otherwise.

DATA AVAILABILITY STATEMENT

Data sharing not applicable—no new data generated, or the article describes entirely theoretical research.

ORCID

Mineto Satoh https://orcid.org/0000-0002-2312-0051
Peter Jan van Leeuwen https://orcid.org/0000-0003-2325-5340
Shin'ya Nakano https://orcid.org/0000-0003-0772-4610

REFERENCES

Ades, M. & Van Leeuwen, P.J. (2015) The equivalent-weights particle filter in a high-dimensional system. *Quarterly Journal of the Royal Meteorological Society*, 141, 484–503.

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S. et al. (2018) The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164.

Bottou, L. (2010) Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT'2010: 19th international conference on computational statistics Paris France, august 22-27, 2010 keynote, invited and contributed papers, 177–186. Springer.

Chen, X., Liu, S., Sun, R. & Hong, M. (2019) On the convergence of a class of adam-type algorithms for non-convex optimization. Paper presented at: 7th international conference on learning representations, ICLR 2019.

Chorin, A.J. & Tu, X. (2009) Implicit sampling for particle filters. *Proceedings of the National Academy of Sciences*, 106, 17249–17254.

Clayton, A.M., Lorenc, A.C. & Barker, D.M. (2013) Operational implementation of a hybrid ensemble/4d-var global data assimilation system at the met office. *Quarterly Journal of the Royal Meteorological Society*, 139, 1445–1461.

Cooper, M. & Perez, T. (2018) Dual-particle-filtering for recursive estimation of agricultural-machinery dynamics. *IFAC-PapersOnLine*, 51, 658–663.

Corless, R.M., Gonnet, G.H., Hare, D.E., Jeffrey, D.J. & Knuth, D.E. (1996) On the lambert w function. *Advances in Computational Mathematics*, 5, 329–359.

Deng, C., Liu, P., Guo, S., Li, Z. & Wang, D. (2016) Identification of hydrological model parameter variation using ensemble kalman filter. *Hydrology and Earth System Sciences*, 20, 4949.

Evensen, G., Dee, D.P. & Schröter, J. (1998) Parameter estimation in dynamical models. In: *Ocean Modeling and parameterization*. Dordrecht, Netherlands: Springer, pp. 373–398.

Evensen, G., Vossepoel, F.C. & van Leeuwen, P.J. (2022) *Data assimilation fundamentals: a unified formulation of the state and parameter estimation problem*. Cham, Switzerland: Springer Nature.

Fortin, V., Abaza, M., Anctil, F. & Turcotte, R. (2014) Why should ensemble spread match the rmse of the ensemble mean? *Journal of Hydrometeorology*, 15, 1708–1713.

Hannah, L.A. (2015) Stochastic optimization. *International Encyclopedia of the Social & Behavioral Sciences*, 2, 473–481.

Kingma, D.P. & Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lorenz, E.N. (1996) Predictability: a problem partly solved. Proc. seminar on predictability, vol. 1.

Pérez-Vieites, S., Mariño, I.P. & Míguez, J. (2018) Probabilistic scheme for joint parameter estimation and state prediction in complex dynamical systems. *Physical Review E*, 98, 063305.

Reddi, S.J., Kale, S. & Kumar, S. (2018) On the convergence of adam and beyond. Paper presented at: International conference on learning representations.

Ruder, S. (2016) An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Santitissadeekorn, N. & Jones, C. (2015) Two-stage filtering for joint state-parameter estimation. *Monthly Weather Review*, 143, 2028–2042.

Skauvold, J., Eidsvik, J., Van Leeuwen, P.J. & Amezcua, J. (2019) A revised implicit equal-weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, 145, 1490–1502.

Snyder, C., Bengtsson, T. & Morzfeld, M. (2015) Performance bounds for particle filters using the optimal proposal. *Monthly Weather Review*, 143, 4750–4761.

Sun, S., Cao, Z., Zhu, H. & Zhao, J. (2019) A survey of optimization methods from a machine learning perspective. *IEEE Transactions on Cybernetics*, 50, 3668–3681.

Van Leeuwen, P.J. (2010) Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136, 1991–1999.

Van Leeuwen, P.J., Künsch, H.R., Nerger, L., Potthast, R. & Reich, S. (2019) Particle filters for high-dimensional geoscience applications: a review. *Quarterly Journal of the Royal Meteorological Society*, 145, 2335–2365.

Wonham, W.M. (1968) On a matrix riccati equation of stochastic control. *SIAM Journal on Control*, 6, 681–697.

Zhu, M., Van Leeuwen, P.J. & Amezcua, J. (2016) Implicit equal-weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, 142, 1904–1919.

Zhu, Z., Meng, Z., Cao, T., Zhang, Z. & Dai, Y. (2017) Particle filter-based robust state and parameter estimation for nonlinear process systems with variable parameters. *Measurement Science and Technology*, 28, 065003.

How to cite this article: Satoh, M., van Leeuwen, P.J. & Nakano, S. (2024) Online state and time-varying parameter estimation using the implicit equal-weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, 1–17. Available from: https://doi.org/10.1002/qj.4698