

Uniform error bounds of the ensemble transform Kalman filter for chaotic dynamics with multiplicative covariance inflation*

Kota Takeda[†] and Takashi Sakajo[‡]

Abstract. Data assimilation is a method of uncertainty quantification to estimate the hidden true state by updating the prediction owing to model dynamics with observation data. As a prediction model, we consider a class of nonlinear dynamical systems on Hilbert spaces including the two-dimensional Navier-Stokes equations and the Lorenz 63 and 96 equations. For nonlinear model dynamics, the ensemble Kalman filter (EnKF) is often used to approximate the mean and covariance of the probability distribution with a set of particles called an ensemble. In this paper, we consider a deterministic version of the EnKF known as the ensemble transform Kalman filter (ETKF), performing well even with limited ensemble sizes compared to other stochastic implementations of the EnKF. When the ETKF is applied to large-scale systems, an ad-hoc numerical technique called a covariance inflation is often employed to reduce approximation errors. Despite the practical effectiveness of the ETKF, little is theoretically known. The present study aims to establish the theoretical analysis of the ETKF. We obtain that the state estimation error of the ETKF with and without the covariance inflation is bounded for any finite time. Furthermore, when the system is finite-dimensional, we show that the uniform-in-time error bound is obtained when an inflation parameter is chosen appropriately, justifying the effectiveness of the covariance inflation in the ETKF.

Key words. data assimilation, dissipative dynamics, ensemble Kalman filter, accuracy, filtering problem

MSC codes. 65C05, 35R30, 35Q93, 62M20, 62F15

1. Introduction. Data assimilation is a method of uncertainty quantification to estimate the hidden true state by combining model dynamics with observation data. This approach encompasses various problems such as the interpolation of missing data by model dynamics. In this paper, we consider the filtering problem [24, 32], where we estimate the time series of the true state using observations up to the current time. We suppose that observation data is obtained by applying a linear operator to the true state and adding noise, which is the standard setting adopted in many filtering problems. We then solve the problem with a Bayesian approach, in which the uncertainty in the state estimation is represented by a probability distribution. One assimilation cycle consists of prediction and analysis steps. In the prediction step, the dynamical model provides the prior distribution for the current state. It is then updated to the posterior distribution by Bayes' formula from the current observation in the analysis step. The time series of the true state is estimated by repeating this assimilation cycle iteratively. This approach is widely used in many applications. See the comprehensive

*Submitted to the editors August 24, 2024.

Funding: The first author is supported by RIKEN Junior Research Associate Program and JST SPRING JP-MJSP2110. The second author is supported by JST MIRAI JPMJMI22G1.

[†]Department of Mathematics, Kyoto University, Kitashirakawa Oiwakecho, Sakyo-ku, Kyoto, Kyoto, 606-8502, Japan, and RIKEN Center for Computational Science, 7-1-26, Minatojima Minamimachi, Chuo-ku, Kobe, Hyogo, 650-0047, Japan (takeda.kota.53r@st.kyoto-u.ac.jp).

[‡]Department of Mathematics, Kyoto University, Kitashirakawa Oiwakecho, Sakyo-ku, Kyoto, Kyoto, 606-8502, Japan (sakajo@math.kyoto-u.ac.jp).

textbooks [19, 32].

When the model dynamics is a linear system, the standard Bayesian assimilation method is the Kalman filter (KF) [18], in which the probability distribution of the state is represented by the Gaussian distribution. A remarkable property of linear dynamical systems is that they transform one Gaussian distribution to another. Hence, we directly obtain the prior Gaussian distribution for the current time from that for the previous time in the prediction step. Then, in the analysis step, the prior Gaussian distribution is updated to the posterior one using Bayes' formula with the current observation through the linear observation operator. The KF is theoretically well-understood, and it works practically as well under the assumption that the additive observation noises and the probability distribution for the initial state follow Gaussian distributions [24].

In the meantime, for nonlinear model dynamics, the ensemble Kalman filter (EnKF) [12] is employed as an extension of the KF. The EnKF approximates probability distributions using a set of particles referred to as an ensemble. Variants of the EnKF use the same prediction step, where each ensemble member evolves according to the nonlinear dynamical model. In the analysis step, each variant uses a different way to generate the analysis ensemble from the prediction ensemble and observation data. As a simple and stochastic method, the perturbed observation (PO) method [7] is proposed, which adds Gaussian noises into the observation data to create artificial observations for each ensemble member. Then, each member of the analysis ensemble is obtained by the weighted average of a replicated observation and each member of the prediction ensemble. As noted in [36], the artificial noises added in the analysis step of the PO method introduce additional errors in approximating the analysis distribution. To address this issue, deterministic versions of the EnKF, known as the ensemble square root filters (ESRF), have been proposed [3, 6, 36]. In particular, we focus on a specific implementation of the ESRF known as the ensemble transform Kalman filter (ETKF) [6]. In the ETKF method, the mean of the analysis ensemble is calculated by the weighted average of the observation and the prediction ensemble mean. The deviations of the analysis ensemble are determined by the linear transformation of the prediction ensemble so that the covariance coincides with that obtained in the KF. It is noteworthy that both the PO method and the ETKF are consistent with the KF [21, 30]. That is to say, the ensemble mean and covariance converge to those of the KF in the limit as $N \rightarrow \infty$ when considering a linear system with Gaussian noises. In contrast, it has been shown that the ETKF produces less error than the PO method in approximating the mean and covariance of the analysis distribution with finite N [1]. Additionally, numerical evidence indicates that the ETKF performs well with small N while the PO method requires large N to achieve comparable performance [29]. Another practical advantage of the ETKF is that it does not evaluate the ensemble covariance, avoiding redundant memory allocations in numerical computation [16].

In the EnKF, the approximation of the covariance is often underestimated due to limited ensemble sizes, which causes an inaccurate estimation of the true state. To handle the problem, ad-hoc covariance inflation techniques [5, 36] are introduced to modify the prediction ensemble covariance $\hat{C}$ in an additive or multiplicative way. In an additive inflation, one adds a scaled identity matrix $\alpha^2 I$ to $\hat{C}$ for $\alpha \geq 0$. On the other hand, in a multiplicative inflation, one multiplies $\hat{C}$ by a scaling factor α^2 for $\alpha \geq 1$. Numerical results have shown the improvement in accuracy of the EnKF with covariance inflations [5, 29, 36]. For a large-scale atmospheric

model, large α is required, and manual tuning of α is expensive [15]. To avoid the expensive manual tuning, adaptive tuning algorithms are developed [2, 4, 31]. Another study focuses on the residual, which is the difference between the measured observation and the simulated (or predicted) observation, and derives upper and lower bounds of α for the multiplicative inflation to ensure that the residual remains within a prescribed interval [27]. These results depend on realizations of observation noises, giving rise to a lack of a priori guarantee for the accuracy of filters. On the other hand, in [20], the authors establish a lower bound of α in the additive inflation to obtain an error bound for the PO method uniformly in time. Another technique to address covariance underestimation is localization, which leverages the local low dimensionality of atmospheric dynamics [13, 16, 36]. The central idea is to ignore observations in regions far from each state variable in the analysis. The effectiveness of localization has been demonstrated in many large-scale geoscience applications [16, 17, 36]. Localization is not covered in this paper although it is essential in the EnKF algorithms.

Recent theoretical studies reveal the basic properties of the EnKFs. In addition to the consistency with the KF [21, 30] and sampling errors [1] as we mentioned above, stability and ergodicity are studied for both the PO method and the ESRF [34, 35], in which the coupled system consisting of the true state, the observation, and the filter is considered. As a result, they have proved that the probability measure with respect to the coupled state converges exponentially to a unique invariant measure if the model dynamics possess a certain dissipation property. While these basic results have been established in general settings, the long-time accuracy of the EnKF is studied in limited situations. For the PO method, sufficiently large additive inflation provides a uniform-in-time error bound [20]. For the continuous time formulation of the deterministic EnKF, a uniform-in-time error bound is obtained with continuous time observation [11]. On the other hand, the error bound of the ETKF remains unexplored due to the complexity of its algorithm.

While many nonlinear filtering problems are considered on finite-dimensional model dynamics [11, 24, 34, 35], most dynamics in geoscience are modeled by nonlinear and chaotic partial differential equations based on the Navier-Stokes equations [19, 32], which are infinite-dimensional dynamical systems. In [20], the authors formulate the PO method for dynamical systems on infinite-dimensional Hilbert spaces. The purpose of this paper is to establish a uniform-in-time error bound for the ETKF and to compare it with the PO method. Hence, we start with the ETKF for chaotic dynamical systems on Hilbert spaces based on the formulation in [20]. We apply multiplicative covariance inflation and derive a sufficient condition on the inflation parameter α to obtain the error bound.

This paper is constructed as follows. In Section 2, we introduce some notations and the concept of nonlinear filtering problems and explain the algorithm of the ETKF and the covariance inflation technique. In Section 3, we show the two main results for the error analysis of the ETKF. Section 4 provides numerical examples demonstrating the scenario in the main theorem. Section 5 is a summary of the results and a discussion of future directions.

2. Nonlinear data assimilation problem and algorithm.

2.1. Notations. Let $\mathcal{U}$ be a separable Hilbert space endowed with the norm $|\cdot|$ and the inner product $\langle \cdot, \cdot \rangle$, and $\mathcal{L}(\mathcal{U}, \mathcal{V})$ denote the space of all bounded linear operators between two Hilbert spaces $\mathcal{U}$ and $\mathcal{V} \subset \mathcal{U}$. For $N \in \mathbb{N}$, we express the identity maps on $\mathcal{U}$ and $\mathbb{R}^N$ by

$I_{\mathcal{U}}$ and I_N respectively. For $A \in \mathcal{L}(\mathcal{U}) := \mathcal{L}(\mathcal{U}, \mathcal{U})$, $|A|_{\mathcal{L}}$ represents the operator norm of A , $\text{Ran}(A)$ denotes the range of A , and A^* is the adjoint of A . We also use $u^* \in \mathcal{U}^*$ as a dual of $u \in \mathcal{U}$. For $u, v \in \mathcal{U}$, we define their product $u \otimes v \in \mathcal{L}(\mathcal{U})$ by $u \otimes v : \mathcal{U} \ni w \mapsto u \langle v, w \rangle \in \mathcal{U}$, which is equivalent to $uv^* = u \otimes v$.

Let $\mathcal{L}_{sa}(\mathcal{U})$ denote the space of all self-adjoint operators in $\mathcal{L}(\mathcal{U})$ (i.e., $A^* = A$). For $A \in \mathcal{L}_{sa}(\mathcal{U})$, by $A \succeq 0$, we mean that A is positive semidefinite, i.e., $\langle u, Au \rangle \geq 0$ for all $u \in \mathcal{U}$. On the other hand, when there exists $c > 0$ such that $\langle u, Au \rangle \geq c|u|^2$ for all $u \in \mathcal{U}$, we say A is positive definite, denoted by $A \succ 0$. For $A, B \in \mathcal{L}_{sa}(\mathcal{U})$, the order $A \succ$ (resp. $\succeq$) B means $A - B \succ$ (resp. $\succeq$) 0 .

Let $N \in \mathbb{N}$, for $U = [u^{(n)}]_{n=1}^N$ and $V = [v^{(n)}]_{n=1}^N \in \mathcal{U}^N$, the ℓ_2 -norm $|U|_2$ and the products $UV^* \in \mathcal{L}(\mathcal{U})$ and $U^*V \in \mathbb{R}^{N \times N}$ are given by

$$|U|_2 = \left(\frac{1}{N} \sum_{n=1}^N |u^{(n)}|^2 \right)^{\frac{1}{2}}, \quad UV^* = \sum_{n=1}^N u^{(n)} \otimes v^{(n)}, \quad U^*V = \left[\langle u^{(i)}, v^{(j)} \rangle \right]_{i,j=1}^N.$$

When we write $u\mathbf{1} = [u, \dots, u] \in \mathcal{U}^N$ by $\mathbf{1} = (1, \dots, 1) \in (\mathbb{R}^N)^*$, it holds that $|u\mathbf{1}|_2^2 = |u|^2$ for all $u \in \mathcal{U}$. Moreover, for $m \in \mathcal{U}$ and $T \in \mathbb{R}^{N \times N}$, we define

$$m + U = m\mathbf{1} + U = [m + u^{(n)}]_{n=1}^N \in \mathcal{U}^N, \quad UT = \left[\sum_{k=1}^N u^{(k)} T_{k,n} \right]_{n=1}^N \in \mathcal{U}^N.$$

For an ensemble $V = [v^{(n)}]_{n=1}^N \in \mathcal{U}^N$, $\bar{v} = \frac{1}{N} \sum_{n=1}^N v^{(n)}$ is the ensemble mean and $dV = [v^{(n)} - \bar{v}]_{n=1}^N \in \mathcal{U}^N$ is the ensemble deviation. The ensemble V is then decomposed into the mean and the deviations, $V = \bar{v}\mathbf{1} + dV$. The (unbiased) ensemble covariance $\text{Cov}(V) \in \mathcal{L}_{sa}(\mathcal{U})$ is defined by

$$\text{Cov}(V) = \frac{1}{N-1} dV dV^*.$$

Note that it is easy to see $\text{Cov}(V) = \text{Cov}(dV)$ and $\text{Cov}(V) \succeq 0$.

The following lemma provides equivalent representations of the ℓ_2 -norm in terms of an ensemble V .

Lemma 2.1. *The ℓ_2 -norm satisfies*

$$(2.1) \quad |V|_2^2 = \frac{1}{N} \text{Tr}(V^*V) = \frac{1}{N} \text{Tr}(VV^*) = |\bar{v}|^2 + |dV|_2^2.$$

Proof. The first equality is derived from the definition of $|V|_2$.

$$|V|_2^2 = \frac{1}{N} \sum_{n=1}^N |v^{(n)}|^2 = \frac{1}{N} \sum_{n=1}^N \langle v^{(n)}, v^{(n)} \rangle = \frac{1}{N} \text{Tr}(V^*V).$$

Let $(\phi_i)_{i \in \mathbb{N}}$ be a complete orthonormal basis of $\mathcal{U}$, we have

$$\begin{aligned} |V|_2^2 &= \frac{1}{N} \sum_{n=1}^N |v^{(n)}|^2 = \frac{1}{N} \sum_{n=1}^N \sum_{i \in \mathbb{N}} \langle v^{(n)}, \phi_i \rangle^2 = \frac{1}{N} \sum_{i \in \mathbb{N}} \sum_{n=1}^N \langle v^{(n)}, \phi_i \rangle^2 \\ &= \frac{1}{N} \sum_{i \in \mathbb{N}} \sum_{n=1}^N \langle \phi_i, (v^{(n)} \otimes v^{(n)}) \phi_i \rangle = \frac{1}{N} \text{Tr}(VV^*). \end{aligned}$$

Owing to the relation $dV\mathbf{1}^* = 0$, we have $VV^* = \bar{v}\mathbf{1}\mathbf{1}^*\bar{v}^* + dVdV^* = N\bar{v}\bar{v}^* + dVdV^*$. Hence, we obtain $\frac{1}{N} \text{Tr}(VV^*) = |\bar{v}|^2 + |dV|_2^2$. ■

2.2. State space model. We consider the following dynamical system on $\mathcal{U}$.

$$(2.2) \quad \frac{du}{dt} = F(u), \quad u \in \mathcal{U}.$$

We suppose that a unique solution exists for any $u_0 \in \mathcal{U}$, and it generates an analytic semi-group $\Psi_t : \mathcal{U} \rightarrow \mathcal{U}$ for $t \geq 0$. For a fixed interval $h > 0$, we denote $\Psi = \Psi_h$ and consider a discrete dynamical system

$$(2.3) \quad u_j = \Psi(u_{j-1}), \quad j \in \mathbb{N}.$$

The Hilbert space $\mathcal{Y} \subset \mathcal{U}$ is the observation space. Then, the noisy observation $y_j \in \mathcal{Y}$ is obtained from u_j for each $j \in \mathbb{N}$,

$$(2.4) \quad y_j = Hu_j + \xi_j, \quad \xi_j \sim N(0, \Gamma),$$

where $H \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$ is the observation operator and $\Gamma \in \mathcal{L}_{sa}(\mathcal{Y})$ is the covariance operator with $\Gamma \succ 0$. Remark that the assumption $\Gamma \succ 0$ is necessary to ensure the bounded inverse Γ^{-1} , which makes the algorithm well-defined as discussed later.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\mathbb{E}$ denote expectation with respect to $\mathbb{P}$. We assume that the i.i.d. noise sequence $(\xi_j)_{j \in \mathbb{N}}$ is independent of u_0 . The filtering problem aims to construct an approximation of $(u_j)_{j \in \mathbb{N}}$ by a stochastic process $(v_j)_{j \in \mathbb{N}}$ adapted to the filtration $(\mathcal{F}_j)_{j \in \mathbb{N}} \subset \mathcal{F}$ where each σ -algebra $\mathcal{F}_j$ is generated by initial uncertainties u_0, v_0 , and the noise sequence $(\xi_k)_{k=1}^j$ for $j \in \mathbb{N}$. The conditional probability distribution $\mathbb{P}^{v_j}(\cdot | Y_j)$ is called by the filtering distribution for provided observations $Y_j = (y_k)_{k=1}^j$. Here, we focus on an algorithm computing $\mathbb{P}^{v_j}(\cdot | Y_j)$ from given $\mathbb{P}^{v_{j-1}}(\cdot | Y_{j-1})$ and y_j . The algorithm consists of two steps, which we refer to as the prediction and analysis steps. In the prediction step, the estimate of the current state $\hat{v}_j$ is provided by the time evolution of the dynamics $\hat{v}_j = \Psi(v_{j-1})$. Accordingly, the filtering distribution $\mathbb{P}^{v_{j-1}}(\cdot | Y_{j-1})$ is propagated into the predicted distribution $\mathbb{P}^{\hat{v}_j}(\cdot | Y_{j-1})$. In the analysis step, the current observation y_j is assimilated into the prediction $\hat{v}_j$ to obtain v_j , and the predicted distribution $\mathbb{P}^{\hat{v}_j}(\cdot | Y_{j-1})$ is updated to $\mathbb{P}^{v_j}(\cdot | Y_j)$.

2.3. The ensemble transform Kalman filter. For $N \in \mathbb{N}$, the ensemble transform Kalman filter (ETKF) approximates the filtering distribution by the empirical distribution of an ensemble of particles $V_j = [v_j^{(1)}, \dots, v_j^{(N)}] \in \mathcal{U}^N$, $\mathbb{P}^{v_j}(\cdot | Y_j) \approx \frac{1}{N} \sum_{n=1}^N \delta_{v_j^{(n)}}(\cdot)$, where $\delta_a(\cdot)$

denotes the Dirac measure. In the prediction step, each particle is driven by the dynamics (2.3). In the analysis step, the ensemble deviation is transformed by a matrix so that the posterior ensemble covariance corresponds to that in the minimum variance posterior ensemble. A detailed description is given as follows.

Definition 2.2. *The algorithm of the ETKF is as follows.*

- (I) Draw N independent samples $V_0 = [v_0^{(n)}]_{n=1}^N$ from some probability distribution μ_0 .
- (II) (the prediction step, input: $V_{j-1} \rightarrow$ output: $\hat{V}_j$) Compute the time evolution $\hat{v}_j^{(n)} = \Psi(v_{j-1}^{(n)})$, $n = 1, \dots, N$ and set $\hat{V}_j = [\hat{v}_j^{(n)}]_{n=1}^N \in \mathcal{U}^N$.
- (III) (the analysis step, input: $\hat{V}_j, y_j \rightarrow$ output: V_j) Decompose $\hat{V}_j = \bar{v}_j \mathbf{1} + d\hat{V}_j$, and put $\hat{C}_j = \text{Cov}(\hat{V}_j)$. Update the mean

$$(2.5) \quad \bar{v}_j = \bar{v}_j + K_j(y_j - H\bar{v}_j)$$

with the Kalman gain,

$$(2.6) \quad K_j = \hat{C}_j H^* (H \hat{C}_j H^* + \Gamma)^{-1}.$$

Take a symmetric transform matrix $T_j \in \mathbb{R}^{N \times N}$ satisfying

$$(2.7) \quad \frac{1}{N-1} d\hat{V}_j T_j (d\hat{V}_j T_j)^* = (I - K_j H) \hat{C}_j,$$

and transform the ensemble deviation $dV_j = d\hat{V}_j T_j$. Finally, set the analysis ensemble $V_j = \bar{v}_j \mathbf{1} + dV_j$.

Practically, we can use the following analysis step instead of (III). This avoids redundant memory allocations in practical numerical computation since we do not need to evaluate the covariance $\hat{C}_j$ explicitly.

- (III') (Input: $\hat{V}_j, y_j \rightarrow$ Output: V_j) Decompose $\hat{V}_j = \bar{v}_j \mathbf{1} + d\hat{V}_j$. Define the transform matrix

$$(2.8) \quad T_j = \left(I_N + \frac{1}{N-1} d\hat{V}_j^* H^* \Gamma^{-1} H d\hat{V}_j \right)^{-\frac{1}{2}},$$

and

$$\tilde{T}_j = \frac{1}{N-1} T_j^2 d\hat{V}_j^* H^* \Gamma^{-1} (y_j - H\bar{v}_j) \mathbf{1} + T_j.$$

Transform

$$V_j = \bar{v}_j \mathbf{1} + d\hat{V}_j \tilde{T}_j.$$

Note that (III') is equivalent to (III). The following theorem shows the existence of the transform matrix T_j , which indicates that the ETKF is well-defined.

Theorem 1. *For any $\hat{V}_j \in \mathcal{U}^N$, there exists a unique symmetric transform matrix $T_j \in \mathbb{R}^{N \times N}$ satisfying (2.7).*

To prove [Theorem 1](#), let us summarize the key property of the Kalman gain K_j .

Lemma 2.3. *Let $\hat{C}_j \succeq 0$, $H \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$, and $\Gamma \in \mathcal{L}(\mathcal{Y})$ with $\Gamma \succ 0$. The Kalman gain satisfies*

$$(2.9) \quad K_j = (I_{\mathcal{U}} - K_j H) \hat{C}_j H^* \Gamma^{-1},$$

and (2.5) is equivalent to

$$(2.10) \quad (I_{\mathcal{U}} + \hat{C}_j H^* \Gamma^{-1} H) \bar{v}_j = \bar{v}_j + \hat{C}_j H^* \Gamma^{-1} y_j.$$

Proof. For simplicity, we omit the time index j in the following proofs since $j \in \mathbb{N}$ is fixed. Owing to $\hat{C} \succeq 0$ and $\Gamma \succ 0$, we have $\Gamma + H\hat{C}H^* \succ 0$ and it is thus invertible. Then, $I_{\mathcal{Y}} + \Gamma^{-1}H\hat{C}H^* = \Gamma^{-1}(\Gamma + H\hat{C}H^*)$ is also invertible since a product of two positive definite operators has positive spectrum [14]. From (A.1) and the fact that $(AB)^{-1} = B^{-1}A^{-1}$ for invertible A, B , we have

$$\begin{aligned} (\Gamma + H\hat{C}H^*)^{-1} &= (I_{\mathcal{Y}} + \Gamma^{-1}H\hat{C}H^*)^{-1}\Gamma^{-1} = [I_{\mathcal{Y}} - (I_{\mathcal{Y}} + \Gamma^{-1}H\hat{C}H^*)^{-1}\Gamma^{-1}H\hat{C}H^*]\Gamma^{-1} \\ &= [I_{\mathcal{Y}} - (\Gamma + H\hat{C}H^*)^{-1}H\hat{C}H^*]\Gamma^{-1}. \end{aligned}$$

Hence, we have.

$$\begin{aligned} K &= \hat{C}H^*(\Gamma + H\hat{C}H^*)^{-1} = \hat{C}H^*[I_{\mathcal{Y}} - (\Gamma + H\hat{C}H^*)^{-1}H\hat{C}H^*]\Gamma^{-1} \\ &= (I_{\mathcal{U}} - KH)\hat{C}H^*\Gamma^{-1}, \end{aligned}$$

which is (2.9). On the other hand, from (2.9), we have

$$\bar{v} = \bar{v} + K(y - H\bar{v}) = (I_{\mathcal{U}} - KH)\bar{v} + Ky = (I_{\mathcal{U}} - KH)\bar{v} + (I_{\mathcal{U}} - KH)\hat{C}H^*\Gamma^{-1}y.$$

To show (2.10), it is sufficient to check $(I_{\mathcal{U}} + \hat{C}H^*\Gamma^{-1}H)(I_{\mathcal{U}} - KH) = I_{\mathcal{U}}$ with (2.6). This is confirmed straightforwardly as follows.

$$\begin{aligned} &(I_{\mathcal{U}} + \hat{C}H^*\Gamma^{-1}H)(I_{\mathcal{U}} - KH) \\ &= I_{\mathcal{U}} + \hat{C}H^*\Gamma^{-1}H - \hat{C}H^*(\Gamma + H\hat{C}H^*)^{-1}H - \hat{C}H^*\Gamma^{-1}H\hat{C}H^*(\Gamma + H\hat{C}H^*)^{-1}H \\ &= I_{\mathcal{U}} + \hat{C}H^*\Gamma^{-1}H - \hat{C}H^*\Gamma^{-1}\Gamma(\Gamma + H\hat{C}H^*)^{-1}H - \hat{C}H^*\Gamma^{-1}H\hat{C}H^*(\Gamma + H\hat{C}H^*)^{-1}H \\ &= I_{\mathcal{U}} + \hat{C}H^*\Gamma^{-1}H - \hat{C}H^*\Gamma^{-1}(\Gamma + H\hat{C}H^*)(\Gamma + H\hat{C}H^*)^{-1}H = I_{\mathcal{U}}. \end{aligned} \quad \blacksquare$$

Proof of Theorem 1. First, we prove the existence of T_j satisfying (2.7). Let $dY = Hd\hat{V}$. Then, the operator $\Gamma + \frac{1}{N-1}dYdY^* = \Gamma + H\hat{C}H^*$ is invertible, and we consider the symmetric matrix

$$S = I_N - \frac{1}{N-1}dY^* \left(\Gamma + H\hat{C}H^* \right)^{-1} dY \in \mathbb{R}^{N \times N}.$$

Then, we have

$$\begin{aligned} \frac{1}{N-1} d\hat{V} S d\hat{V}^* &= \frac{1}{N-1} d\hat{V} d\hat{V}^* - \frac{1}{N-1} d\hat{V} dY^* \left(\Gamma + H\hat{C}H^* \right)^{-1} \frac{1}{N-1} dY d\hat{V}^* \\ &= \hat{C} - \hat{C}H^*(\Gamma + H\hat{C}H^*)^{-1}H\hat{C} = (I - KH)\hat{C}, \end{aligned}$$

in which, we use $\frac{1}{N-1} dY d\hat{V}^* = H\hat{C}$. From (A.4), we have

$$(2.11) \quad S = (I_N + \frac{1}{N-1} dY^* \Gamma^{-1} dY)^{-1}$$

and $S \succ 0$. We finally define the transform matrix $T = S^{\frac{1}{2}}$, which is nothing but (2.8). Then T becomes symmetric by definition. $\blacksquare$

Remark 2.4. *Theorem 1 is an extension of the well-definedness of the ETKF algorithm [22] to Hilbert spaces which can be infinite-dimensional.*

We finally show another representation of the ensemble update, which is a variant of that in the PO method (Proposition 3.2 in [20]).

Lemma 2.5. *The following hold for the ensemble deviation and the transform matrix*

$$(2.12) \quad dV_j \mathbf{1}^* = d\hat{V}_j \mathbf{1}^* = 0 \in \mathcal{U},$$

$$(2.13) \quad T_j \mathbf{1}^* = \mathbf{1}^*.$$

Moreover, the ensembles satisfy the relation

$$(2.14) \quad (I_{\mathcal{U}} + \hat{C}_j H^* \Gamma^{-1} H) V_j = \hat{V}_j T_j^{-1} + \hat{C}_j H^* \Gamma^{-1} y_j \mathbf{1}.$$

Proof. The definition of the ensemble mean yields the first property (2.12).

$$dV \mathbf{1}^* = \sum_{n=1}^N (v^{(n)} - \bar{v}) = 0,$$

yielding

$$S^{-1} \mathbf{1}^* = \left(I_N + \frac{1}{N-1} dV^* H^* \Gamma^{-1} H dV \right) \mathbf{1}^* = \mathbf{1}^*.$$

Hence, we have

$$(2.15) \quad S \mathbf{1}^* = \mathbf{1}^*.$$

Then, we prove that $\mathbf{1}^*$ is also an eigenvector of $T = S^{\frac{1}{2}}$ with an eigenvalue 1. Since S is symmetric, it is diagonalized as $S = UDU^*$ for a unitary $U \in \mathbb{R}^{N \times N}$ and a diagonal $D \in \mathbb{R}^{N \times N}$. Then, (2.15) is equivalent to

$$S \mathbf{1}^* = \mathbf{1}^* \Leftrightarrow UDU^* \mathbf{1}^* = \mathbf{1}^* \Leftrightarrow DU^* \mathbf{1}^* = U^* \mathbf{1}^*.$$

Putting $\mathbf{u} = U^* \mathbf{1}^* = (u_1, \dots, u_N)^* \in \mathbb{R}^N$ and $d_n > 0$ as n -diagonal element of D for $n = 1, \dots, N$, the last equality is rewritten for each component

$$d_n u_n = u_n, \quad n = 1, \dots, N.$$

This implies that $d_n = 1$ or $u_n = 0$ for each $n = 1, \dots, N$. Hence, the following also holds

$$d_n^{\frac{1}{2}} u_n = u_n, \quad n = 1, \dots, N,$$

and we have $D^{\frac{1}{2}} U^* \mathbf{1}^* = U^* \mathbf{1}^*$. By definition, T is written as $T = U D^{\frac{1}{2}} U^*$, and this yields $T \mathbf{1}^* = \mathbf{1}^*$, which is (2.13).

The last equality (2.14) is shown as follows. From (2.10), we have

$$(I_{\mathcal{U}} + \widehat{C} H^* \Gamma^{-1} H) \bar{v} = \bar{v} + \widehat{C} H^* \Gamma^{-1} y \in \mathcal{U}.$$

By using $\widehat{C} = \text{Cov}(\widehat{V})$, (2.11) and $S = T^2$, we obtain

$$\begin{aligned} (I_{\mathcal{U}} + \widehat{C} H^* \Gamma^{-1} H) dV &= (I_{\mathcal{U}} + \widehat{C} H^* \Gamma^{-1} H) d\widehat{V} T = d\widehat{V} \left[I_{\mathcal{U}} + \frac{1}{N-1} d\widehat{V}^* H^* \Gamma^{-1} H d\widehat{V} \right] T \\ &= d\widehat{V} S^{-1} T = d\widehat{V} T^{-1} \in \mathcal{U}^N. \end{aligned}$$

Finally, owing to (2.10) and $\bar{v} \mathbf{1} T^{-1} = \bar{v} \mathbf{1}$,

$$\begin{aligned} (I_{\mathcal{U}} + \widehat{C} H^* \Gamma^{-1} H) V &= (I_{\mathcal{U}} + \widehat{C} H^* \Gamma^{-1} H) (\bar{v} \mathbf{1} + dV) = \bar{v} \mathbf{1} + \widehat{C} H^* \Gamma^{-1} y \mathbf{1} + d\widehat{V} T^{-1} \\ &= \bar{v} \mathbf{1} T^{-1} + d\widehat{V} T^{-1} + \widehat{C} H^* \Gamma^{-1} y \mathbf{1} = \widehat{V} T^{-1} + \widehat{C} H^* \Gamma^{-1} y \mathbf{1}. \end{aligned}$$

This finishes the proof. ■

3. Bounds for the filtering error of the ensemble transform Kalman filter. For each $j \in \mathbb{N} \cup \{0\}$, we define the filtering ensemble error by

$$E_j = [e_j^{(n)}]_{n=1}^N \quad \text{with} \quad e_j^{(n)} = v_j^{(n)} - u_j, \quad n = 1, \dots, N.$$

Note that $dE_j = dV_j$, and hence, $E_j = e_j \mathbf{1} + dV_j$ where $e_j = \bar{v}_j - u_j \in \mathcal{U}$. Furthermore, owing to (2.1), the decomposition of the norm of the ensemble error is given by

$$(3.1) \quad |E_j|_2^2 = |e_j|^2 + |dV|_2^2.$$

3.1. Well-posedness. We show the well-posedness of the ETKF, i.e., the ensemble error does not blow up faster than an exponential function. To do this, we make some assumptions on the dynamics and the observation to analyze the properties of the ETKF as in [20].

Assumption 1. *There exists $\rho > 0$ such that Ψ_t has an absorbing ball $B(\rho) = \{v \in \mathcal{U} \mid |v| \leq \rho\}$, i.e., $\Psi_t(v) \in B(\rho)$ for any $v \in B(\rho)$ and $t \geq 0$.*

Assumption 2. *There exists $\beta > 0$ such that, for any $u \in B(\rho)$ and $v \in \mathcal{U}$,*

$$(3.2) \quad \langle F(u) - F(v), u - v \rangle \leq \beta |u - v|^2.$$

Note that [Assumptions 1](#) and [2](#) are satisfied by Lorenz 63, Lorenz 96 equations and the incompressible Navier-Stokes equations on a two-dimensional torus [[23](#), [28](#), [29](#), [33](#)]. The following lemma was established in [[20](#)], providing the upper bound of the error growth rate due to the model dynamics.

Lemma 3.1. *Suppose that [Assumptions 1](#) and [2](#) hold. Then, for any $u \in B(\rho)$, $v \in \mathcal{U}$ and $t > 0$,*

$$(3.3) \quad |\Psi_t(u) - \Psi_t(v)| \leq e^{\beta t} |u - v|.$$

Proof. See Lemma 2.6 of [[20](#)]. ■

Regarding the observation operator and its covariance operator, we assume that they are trivial.

Assumption 3. *The state is fully observed, i.e., $H = I_{\mathcal{U}}$ and $\Gamma = \gamma^2 I_{\mathcal{U}}$ for some $\gamma > 0$.*

Theorem 2. *Under [Assumptions 1–3](#), we consider the ETKF provided in [Definition 2.2](#). Then, we have the following upper bound.*

$$(3.4) \quad \mathbb{E} [|E_j|_2^2] \leq e^{2\beta h j} \mathbb{E} [|E_0|_2^2] + (N-1)\gamma^2 \frac{e^{2\beta h j} - 1}{e^{2\beta h} - 1}, \quad j \in \mathbb{N}.$$

Proof. From [Assumption 3](#), the relation (2.14) becomes

$$(3.5) \quad (I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j) V_j = \hat{V}_j T_j^{-1} + \gamma^{-2} \hat{C}_j y_j \mathbf{1}.$$

Let $U_j = u_j \mathbf{1} \in \mathcal{U}^N$. From (2.13), we have $U_j = U_j T_j^{-1}$. Hence,

$$(3.6) \quad (I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j) U_j = U_j + \gamma^{-2} \hat{C}_j U_j = U_j T_j^{-1} + \gamma^{-2} \hat{C}_j U_j.$$

Setting $\hat{E}_j = \hat{V}_j - U_j$ and subtracting (3.6) from (3.5) yields

$$(I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j) E_j = \hat{E}_j T_j^{-1} + \gamma^{-2} \hat{C}_j (y_j - u_j) \mathbf{1} = \hat{E}_j T_j^{-1} + \gamma^{-2} \hat{C}_j \xi_j \mathbf{1}.$$

Owing to $\gamma^{-2} \hat{C}_j \succeq 0$, $I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j$ is invertible. Multiplying $(I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j)^{-1}$, we obtain

$$E_j = (I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j)^{-1} \hat{E}_j T_j^{-1} + (I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j)^{-1} \gamma^{-2} \hat{C}_j \xi_j \mathbf{1}.$$

Let us divide E_j into the following two terms and evaluate them separately.

$$(3.7) \quad R_1 = (I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j)^{-1} \hat{E}_j T_j^{-1},$$

$$(3.8) \quad R_2 = (I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j)^{-1} \gamma^{-2} \hat{C}_j \xi_j \mathbf{1}.$$

Here, the dimension of $\text{Ran}(C_j)$ is $N-1$ at most since $\hat{C}_j$ consists of N vectors with one constraint. Let Π_j be the projection to $\text{Ran}(C_j)$, and we have $R_2 = (I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j)^{-1} \gamma^{-2} \hat{C}_j \Pi_j \xi_j \mathbf{1}$. From (A.2), we have $(I_{\mathcal{U}} + \gamma^{-2} \hat{C}_j)^{-1} \gamma^{-2} \hat{C}_j \preceq I$ owing to $\gamma^{-2} \hat{C}_j \succeq 0$. This leads to

$$(3.9) \quad |R_2|_2^2 \leq |\Pi_j \xi_j \mathbf{1}|_2^2 = |\Pi_j \xi_j|^2.$$

Let $J = I_{\mathcal{U}} + \gamma^{-2}\hat{C}_j$. Then we have $J, J^{-1} \in \mathcal{L}_{sa}(\mathcal{U})$ and $|J^{-1}|_{\mathcal{L}} \leq 1$. We obtain

$$R_1 R_1^* = J^{-1} \hat{E}_j T_j^{-2} \hat{E}_j^* J^{-1}.$$

Considering the relations $d\hat{E}_j = d\hat{V}_j$ and $\hat{E}_j d\hat{V}_j^* = d\hat{V}_j \hat{E}_j^* = d\hat{V}_j d\hat{V}_j^*$, we have

$$\begin{aligned} \hat{E}_j T_j^{-2} \hat{E}_j^* &= \hat{E}_j \left[I_N + \frac{\gamma^{-2}}{N-1} d\hat{V}_j^* d\hat{V}_j \right] \hat{E}_j^* \\ &= \hat{e}_j \mathbf{1} (\hat{e}_j \mathbf{1})^* + d\hat{V}_j d\hat{V}_j^* + d\hat{V}_j d\hat{V}_j^* \gamma^{-2} \hat{C}_j \\ &= \hat{e}_j \mathbf{1} (\hat{e}_j \mathbf{1})^* + d\hat{V}_j d\hat{V}_j^* J, \end{aligned}$$

where $\hat{e}_j = \bar{v}_j - u_j$. Since J^{-1} is self-adjoint, we have

$$R_1 R_1^* = J^{-1} \hat{e}_j \mathbf{1} (\hat{e}_j \mathbf{1})^* J^{-1} + J^{-1} d\hat{V}_j d\hat{V}_j^* = J^{-1} \hat{e}_j \mathbf{1} (J^{-1} \hat{e}_j \mathbf{1})^* + J^{-1} d\hat{V}_j d\hat{V}_j^*.$$

Then, $|R_1|_2^2$ is bounded by

$$\begin{aligned} |R_1|_2^2 &= |J^{-1} \hat{e}_j|^2 + \frac{1}{N} \text{Tr} \left(J^{-1} d\hat{V}_j d\hat{V}_j^* \right) \leq |J^{-1}|_{\mathcal{L}}^2 |\hat{e}_j|^2 + |J^{-1}|_{\mathcal{L}} \frac{1}{N} \text{Tr} \left(d\hat{V}_j d\hat{V}_j^* \right) \\ &\leq |\hat{e}_j|^2 + |d\hat{V}_j|_2^2 = |\hat{E}_j|_2^2. \end{aligned}$$

The first inequality follows from the Hölder inequality about the trace norm and the operator norm [10], and the second inequality holds owing to $|J^{-1}|_{\mathcal{L}} \leq 1$. From this with Lemma 3.1, we obtain the upper bound of $|R_1|_2$ as follows.

$$(3.10) \quad |R_1|_2^2 \leq |\hat{E}_j|_2^2 \leq e^{2\beta h} |E_{j-1}|_2^2.$$

Since R_1 and R_2 are conditionally independent under $\mathcal{F}_{j-1}$, it follows from (3.9) and (3.10) that

$$\begin{aligned} \mathbb{E}_{j-1}[|E_j|_2^2] &= \mathbb{E}_{j-1}[|R_1|_2^2] + \mathbb{E}_{j-1}[|R_2|_2^2] \leq e^{2\beta h} \mathbb{E}_{j-1}[|E_{j-1}|_2^2] + \mathbb{E}_{j-1}[|\Pi_j \xi_j|^2] \\ &= e^{2\beta h} \mathbb{E}_{j-1}[|E_{j-1}|_2^2] + (N-1)\gamma^2, \end{aligned}$$

where the conditional expectation is denoted by $\mathbb{E}_{j-1}[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_{j-1}]$. Here, the conditional expectation satisfies $\mathbb{E}[\mathbb{E}_{j-1}[\cdot]] = \mathbb{E}[\cdot]$ in general. Therefore, taking the expectation yields

$$\mathbb{E}[|E_j|_2^2] \leq e^{2\beta h} \mathbb{E}[|E_{j-1}|_2^2] + (N-1)\gamma^2.$$

Applying this inequality repeatedly, we obtain (3.4). ■

Remark 3.2. *Theorem 2 ensures the short-term accuracy of the ETKF while the existing result [34] only shows the stability of the filter (boundedness of the ensemble). Additionally, Theorem 2 holds even when $\mathcal{U}$ is infinite-dimensional and is analogous to Theorem 4.2 for the PO method [20]. Compared to this, a constant coefficient 2 does not appear in the second term of the upper bound (3.4) in Theorem 2, indicating that the PO method is more influenced by the variance of observations due to its stochastic implementation.*

3.2. Error analysis with multiplicative covariance inflation. If the ensemble size N is smaller than the dimension of $\mathcal{U}$, the ensemble covariance $\widehat{C}_j$ degenerates. This implies that $|J^{-1}|_{\mathcal{L}} = |(I_{\mathcal{U}} + \gamma^{-2}\widehat{C}_j)^{-1}|_{\mathcal{L}} = 1$. In the PO method [20], a diagonal matrix to $\widehat{C}_j$ is added to avoid this degeneration. In the meantime, this additive inflation technique is not applicable to the ETKF because $\widehat{C}_j$ is not explicitly used in the computation of the transform matrix (2.8). Hence, multiplicative covariance inflation is used, in which we multiply an inflation factor $\alpha \geq 1$ by the deviations of the ensemble $d\widehat{V}_j$ after the prediction step (II) of Definition 2.2. To be specific, for the prediction ensemble $\widehat{V}_j = \bar{v}_j + d\widehat{V}_j$, we define the inflated prediction ensemble $\widehat{V}_j^\alpha$ by

$$\widehat{V}_j^\alpha = \bar{v}_j + \alpha d\widehat{V}_j.$$

In the analysis step (III) of Definition 2.2, we use $\widehat{V}_j^\alpha$ instead of $\widehat{V}_j$. First, it follows that $\widehat{C}_j^\alpha = \text{Cov}(\widehat{V}_j^\alpha) = \alpha^2 \widehat{C}_j$. Then, the mean update (2.5) is given by

$$\bar{v}_j^\alpha = \bar{v}_j + K_j^\alpha (y_j - H\bar{v}_j),$$

where the Kalman gain is $K_j^\alpha = \widehat{C}_j^\alpha H^* (H\widehat{C}_j^\alpha H^* + \Gamma)^{-1}$. Second, the transform matrix is given by

$$T_j^\alpha = \left(I_N + (d\widehat{V}_j^\alpha)^* H^* \Gamma^{-1} H d\widehat{V}_j^\alpha \right)^{-\frac{1}{2}} = \left(I_N + \alpha^2 d\widehat{V}_j^* H^* \Gamma^{-1} H d\widehat{V}_j \right)^{-\frac{1}{2}},$$

and the deviations are transformed by

$$dV_j^\alpha = d\widehat{V}_j^\alpha T_j^\alpha,$$

and we define $V_j^\alpha = \bar{v}_j^\alpha + dV_j^\alpha$. The relation between the prediction and analysis ensemble is summarized as follows. The mean update (2.10) is given by

$$(3.11) \quad (I_{\mathcal{U}} + \alpha^2 \widehat{C}_j H^* \Gamma^{-1} H) \bar{v}_j^\alpha = \bar{v}_j + \alpha^2 \widehat{C}_j H^* \Gamma^{-1} y_j,$$

and the analysis covariance satisfies

$$(3.12) \quad C_j^\alpha = \frac{\alpha^2}{N-1} d\widehat{V}_j (I_N + \alpha^2 d\widehat{V}_j^* H^* \Gamma^{-1} H d\widehat{V}_j)^{-1} d\widehat{V}_j^*.$$

In what follows, we omit the superscript α in the notations since the changes in terms of α explicitly appear as the multiplicative factor α^2 in (3.11) and (3.12), and no confusion occurs.

It is known that the multiplicative inflation is insufficient to resolve the issue of the covariance degeneration when the state space is infinite-dimensional since it does not affect the eigenvectors of the covariance. Hence, we need the following assumption that the actual dimension of the dynamics becomes finite and F is Lipschitz, which is stronger than Assumption 2.

Assumption 4. *The dimension of the state space $\mathcal{U}$ is finite $m \in \mathbb{N}$. For $\rho > 0$ in Assumption 1 and F in (2.2), there exists $\beta > 0$ such that for any $u, v \in B(\rho)$*

$$|F(u) - F(v)| \leq \beta |u - v|.$$

Let $u_0 \in B(\rho)$ and $v_0^{(n)} \in B(\rho)$ for $n = 1, \dots, N$. We write $\bar{v}_t = \frac{1}{N} \sum_{n=1}^N \Psi_t(v_0^{(n)})$ and $u_t = \Psi_t(u_0)$. Then, we have the following lemma.

Lemma 3.3. *Suppose [Assumptions 1](#) and [4](#). Then, for any $\epsilon > 0$, $t > 0$,*

$$(3.13) \quad |\bar{v}_t - u_t|^2 \leq e^{2(\beta+\epsilon)t}(|\bar{v}_0 - u_0|^2 + D) - D,$$

where $D = \frac{\beta^2 \rho^2}{(\beta+\epsilon)\epsilon}$.

Proof. We write $e_t = \bar{v}_t - u_t$ and $\bar{F}_t = \frac{1}{N} \sum_{n=1}^N F(v_t^{(n)})$. Then, we have $\frac{d}{dt} \bar{v}_t = \bar{F}_t$. Hence, we obtain

$$(3.14) \quad \begin{aligned} \frac{1}{2} \frac{d}{dt} |e_t|^2 &= \left\langle \frac{d}{dt} e_t, e_t \right\rangle = \langle \bar{F}_t - F(u_t), e_t \rangle \\ &\leq |\bar{F}_t - F(u_t)| |e_t| \\ &\leq (|F(\bar{v}_t) - F(u_t)| + |\bar{F}_t - F(\bar{v}_t)|) |e_t| \\ &\leq (\beta |e_t| + |\bar{F}_t - F(\bar{v}_t)|) |e_t|, \end{aligned}$$

where we can use [Assumption 4](#) owing to $\bar{v}_t, u_t \in B(\rho)$. The second term is estimated by

$$|\bar{F}_t - F(\bar{v}_t)|^2 \leq \frac{N}{N^2} \sum_{n=1}^N |F(v_t^{(n)}) - F(\bar{v}_t)|^2 \leq \frac{1}{N} \sum_{n=1}^N \beta^2 |v_t^{(n)} - \bar{v}_t|^2 \leq 4\beta^2 \rho^2.$$

Substituting it into (3.14) and using the Young inequality, we have

$$\begin{aligned} \frac{d}{dt} |e_t|^2 &\leq 2\beta |e_t|^2 + 4\beta \rho |e_t| = 2\beta |e_t|^2 + (2\beta \rho \epsilon^{-1/2})(2\epsilon^{1/2} |e_t|) \\ &\leq 2\beta |e_t|^2 + 2\beta^2 \rho^2 \epsilon^{-1} + 2\epsilon |e_t|^2 = 2(\beta + \epsilon)(|e_t|^2 + D), \end{aligned}$$

for any $\epsilon > 0$. Therefore, we obtain $|e_t|^2 \leq e^{2(\beta+\epsilon)t}(|e_0|^2 + D) - D$ from the Gronwall inequality. ■

We now consider the error between the ensemble mean and the true state.

$$e_j = \bar{v}_j - u_j \in \mathcal{U}.$$

Here, $\lambda_{\min}(C)$ denotes the minimum eigenvalue of a covariance operator C .

Theorem 3. *Under [Assumptions 1](#), [3](#), and [4](#), we consider the ETKF with the multiplicative inflation with $\alpha \geq 1$. Suppose also that the ensemble size $N \in \mathbb{N}$ is large enough to satisfy $\lambda_{\min}(C_0) \geq \lambda_0$ with $\lambda_0 > 0$, and that $v_j^{(n)} \in B(\rho)$ for all $n = 1, \dots, N$ and $j \in \mathbb{N}$. Then, for any $\epsilon > 0$, there exists $\alpha_0 = \alpha_0(\rho, \beta, N, \lambda_0, h, \gamma, \epsilon) \geq 1$ such that the following hold for any $\alpha \geq \alpha_0$.*

(i) *There exists $\lambda_* = \lambda_*(\rho, \beta, N, \lambda_0, h, \gamma, \alpha) > 0$ such that $\lambda_{\min}(\hat{C}_j) > \lambda_*$ for all $j \in \mathbb{N}$.*

(ii) *For $j \in \mathbb{N}$ and $\theta = (1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^{-2} e^{2(\beta+\epsilon)h}$,*

$$(3.15) \quad \mathbb{E}[|e_j|^2] \leq \theta^j (\mathbb{E}[|e_0|^2] + D) + m\gamma^2 \frac{1 - \theta^j}{1 - \theta} + \left(\frac{(1 - \theta^j)(1 - \Theta)}{1 - \theta} - 1 \right) D,$$

where $D = \frac{\beta^2 \rho^2}{(\beta + \epsilon)\epsilon}$ and $\Theta = (1 + \frac{\alpha_0^2}{\gamma^2} \lambda_*)^{-2}$. Moreover, if $\theta < 1$, we have

$$(3.16) \quad \limsup_{j \rightarrow \infty} \mathbb{E}[|e_j|^2] \leq \frac{m\gamma^2}{1-\theta} + \left(\frac{1-\Theta}{1-\theta} - 1 \right) D.$$

Proof. For convenience, we write $\hat{\lambda}_j^{min} = \lambda_{min}(\hat{C}_j)$ and $\lambda_j^{min} = \lambda_{min}(C_j)$. We first estimate the change from λ_{j-1}^{min} to $\hat{\lambda}_j^{min}$ in the prediction step. To this end, we write $\hat{v}_t^{(n)} = \Psi_t(v_{j-1}^{(n)})$, $\hat{C}_t = \frac{1}{N-1} \sum_{n=1}^N (\hat{v}_t^{(n)} - \bar{v}_t) \otimes (\hat{v}_t^{(n)} - \bar{v}_t)$, $\lambda_t = \lambda_{min}(\hat{C}_t)$ for $t \in [0, h]$. The differentiation of $\hat{C}_t$ with respect to t yields

$$(3.17) \quad \frac{d}{dt} \hat{C}_t = \frac{1}{N-1} \sum_{n=1}^N (F(\hat{v}_t^{(n)}) - \bar{F}_t) \otimes (\hat{v}_t^{(n)} - \bar{v}_t) + (\hat{v}_t^{(n)} - \bar{v}_t) \otimes (F(\hat{v}_t^{(n)}) - \bar{F}_t),$$

where $\bar{F}_t = \frac{1}{N} \sum_{n=1}^N F(\hat{v}_t^{(n)})$. Owing to $\frac{1}{N} \sum_{n=1}^N \hat{v}_t^{(n)} - \bar{v}_t = 0$, we have

$$(3.18) \quad \frac{d}{dt} \hat{C}_t = \frac{1}{N-1} \sum_{n=1}^N (F(\hat{v}_t^{(n)}) - F(\bar{v}_t)) \otimes (\hat{v}_t^{(n)} - \bar{v}_t) + (\hat{v}_t^{(n)} - \bar{v}_t) \otimes (F(\hat{v}_t^{(n)}) - F(\bar{v}_t)).$$

With a similar argument on the derivative of the eigenvalue of the covariance matrix in Section 3 of [11], there exists $w \in \mathcal{U}$ with $|w| = 1$ such that

$$\frac{d}{dt} \lambda_t = \left\langle w, \frac{d}{dt} \hat{C}_t w \right\rangle.$$

To derive the lower bound $\frac{d}{dt} \lambda_t$, we consider the absolute value of the right-hand side of (3.18). Owing to $|w| = 1$ and [Assumption 4](#), we have

$$\begin{aligned} \left| \left\langle w, \frac{d}{dt} \hat{C}_t w \right\rangle \right| &\leq \left| \frac{2}{N-1} \sum_{n=1}^N \left\langle F(\hat{v}_t^{(n)}) - F(\bar{v}_t), w \right\rangle \left\langle \hat{v}_t^{(n)} - \bar{v}_t, w \right\rangle \right| \\ &\leq 2 \left(\frac{1}{N-1} \sum_{n=1}^N \left\langle F(\hat{v}_t^{(n)}) - F(\bar{v}_t), w \right\rangle^2 \right)^{\frac{1}{2}} \left(\frac{1}{N-1} \sum_{n=1}^N \left\langle \hat{v}_t^{(n)} - \bar{v}_t, w \right\rangle^2 \right)^{\frac{1}{2}} \\ &\leq \beta \frac{1}{N-1} \sum_{n=1}^N |\hat{v}_t^{(n)} - \bar{v}_t|^2 \leq 8 \frac{N}{N-1} \beta \rho^2. \end{aligned}$$

The last inequality comes from owing to $\hat{v}_t^{(n)} \in B(\rho)$ by [Assumption 1](#) and the assumption of [Theorem 3](#). Hence, for $a = 8 \frac{N}{N-1} \beta \rho^2 > 0$,

$$\frac{d}{dt} \lambda_t \geq -a.$$

Integrating it from 0 to $t = h$, we have

$$(3.19) \quad \hat{\lambda}_j^{min} = \lambda_h \geq e^{-ah} \lambda_0 = e^{-ah} \lambda_{j-1}^{min}.$$

The next step is to address the change in the eigenvalue in the analysis step. From [Assumption 3](#), we have

$$C_{j-1} = \frac{\alpha^2}{N-1} d\widehat{V}_{j-1} (I_N + \alpha^2 \gamma^{-2} \widetilde{C}_{j-1})^{-1} d\widehat{V}_{j-1}^*,$$

where $\widetilde{C}_{j-1} = \frac{1}{N-1} d\widehat{V}_{j-1}^* d\widehat{V}_{j-1} \in \mathbb{R}^{N \times N}$. Next, for fixed $j \in \mathbb{N}$, we show that the eigenvectors of $\widehat{C}_{j-1}$ are also the eigenvectors of C_{j-1} . Indeed, if $\phi \in \mathcal{U}$ satisfies $\widehat{C}_{j-1}\phi = \lambda\phi$ with an eigenvalue $\lambda \geq 0$, we have

$$\widetilde{C}_{j-1} d\widehat{V}_{j-1}^* \phi = \frac{1}{N-1} d\widehat{V}_{j-1}^* d\widehat{V}_{j-1} d\widehat{V}_{j-1}^* \phi = d\widehat{V}_{j-1}^* \widehat{C}_{j-1} \phi = \lambda d\widehat{V}_{j-1}^* \phi.$$

Hence, it follows that

$$\begin{aligned} C_{j-1}\phi &= \frac{\alpha^2}{N-1} d\widehat{V}_{j-1} (I_N + \alpha^2 \gamma^{-2} \widetilde{C}_{j-1})^{-1} d\widehat{V}_{j-1}^* \phi = \frac{\alpha^2}{N-1} d\widehat{V}_{j-1} \frac{1}{1 + \alpha^2 \gamma^{-2} \lambda} d\widehat{V}_{j-1}^* \phi \\ &= \frac{\alpha^2}{1 + \alpha^2 \gamma^{-2} \lambda} \widehat{C}_{j-1} \phi = \frac{\alpha^2 \lambda}{1 + \alpha^2 \gamma^{-2} \lambda} \phi. \end{aligned}$$

Since the map $\lambda \mapsto \frac{\alpha^2 \lambda}{1 + \alpha^2 \gamma^{-2} \lambda}$ is monotonically increasing, we obtain the relation between the minimum eigenvalues

$$(3.20) \quad \lambda_{j-1}^{min} = \frac{\alpha^2 \widehat{\lambda}_{j-1}^{min}}{1 + \frac{\alpha^2}{\gamma^2} \widehat{\lambda}_{j-1}^{min}}.$$

Combining [\(3.19\)](#) and [\(3.20\)](#), we obtain the inequality for $\widehat{\lambda}_j^{min}$.

$$\widehat{\lambda}_j^{min} \geq \frac{e^{-ah} \alpha^2 \widehat{\lambda}_{j-1}^{min}}{1 + \frac{\alpha^2}{\gamma^2} \widehat{\lambda}_{j-1}^{min}}.$$

We now consider the following discrete dynamical system of the eigenvalue

$$\lambda_{n+1} = g(\lambda_n), \quad \lambda_0 > 0,$$

where $g(\lambda) = \frac{e^{-ah} \alpha^2 \lambda}{1 + \frac{\alpha^2}{\gamma^2} \lambda}$. Note that $\lambda_n > 0$ for $n \in \mathbb{N} \cup \{0\}$. Let $\lambda_\infty = \frac{\gamma^2}{\alpha^2} (e^{-ah} \alpha^2 - 1)$ be a fixed point of the dynamical system, i.e., $\lambda_\infty = g(\lambda_\infty)$. Then, if $e^{-ah} \alpha^2 > 1$, the ratio $\frac{g(\lambda)}{\lambda}$ satisfies $\frac{g(\lambda)}{\lambda} \geq 1$ (resp. < 1) for $\lambda \leq \lambda_\infty$ (resp. $\lambda > \lambda_\infty$). Hence, we have $\lim_{n \rightarrow \infty} \lambda_n = \lambda_\infty$. On the other hand, $\lim_{n \rightarrow \infty} \lambda_n = 0$ if $e^{-ah} \alpha^2 \leq 1$. Therefore, we obtain the lower bound

$$(3.21) \quad \widehat{\lambda}_j^{min} \geq \min \left\{ \widehat{\lambda}_0^{min}, \frac{\gamma^2}{\alpha^2} (e^{-ah} \alpha^2 - 1) \right\} = \min \left\{ e^{-ah} \lambda_0, \frac{\gamma^2}{\alpha^2} (e^{-ah} \alpha^2 - 1) \right\} = \lambda_* > 0$$

if and only if $e^{-ah} \alpha^2 > 1$.

Finally, we establish the recurrence inequality for $\mathbb{E}[|e_j|^2]$. From [Assumption 3](#), the equation (3.11) is reduced to

$$(I + \alpha^2 \gamma^{-2} \widehat{C}_j) \bar{v}_j = \bar{\tilde{v}}_j + \alpha^2 \gamma^{-2} \widehat{C}_j y_j.$$

As in the proof of [Theorem 2](#), the error is divided into the two terms $e_j = r_1 + r_2$, where

$$(3.22) \quad r_1 = (I_{\mathcal{U}} + \alpha^2 \gamma^{-2} \widehat{C}_j)^{-1} \widehat{e}_j,$$

$$(3.23) \quad r_2 = (I_{\mathcal{U}} + \alpha^2 \gamma^{-2} \widehat{C}_j)^{-1} \alpha^2 \gamma^{-2} \widehat{C}_j (y_j - u_j),$$

and $\widehat{e}_j = \bar{\tilde{v}}_j - u_j$. From [\(A.2\)](#), we have $|(I_{\mathcal{U}} + \alpha^2 \gamma^{-2} \widehat{C}_j)^{-1} \alpha^2 \gamma^{-2} \widehat{C}_j|_{\mathcal{L}} \leq 1$, and hence,

$$(3.24) \quad |r_2| \leq |y_j - u_j| = |\xi_j|.$$

From this lower bound of the minimum eigenvalue [\(3.21\)](#), we have $|(I_{\mathcal{U}} + \alpha^2 \gamma^{-2} \widehat{C}_j)^{-1}|_{\mathcal{L}} \leq (1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^{-1}$. Hence,

$$(3.25) \quad |r_1|^2 \leq \frac{1}{(1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^2} |\widehat{e}_j|^2 \leq \frac{e^{2(\beta+\epsilon)h}}{(1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^2} (|e_{j-1}|^2 + D) - \frac{D}{(1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^2},$$

where [Lemma 3.3](#) is used for any $\epsilon > 0$.

As in the proof of [Theorem 2](#), we treat r_1 and r_2 separately when computing the expectation. Therefore, we obtain

$$\begin{aligned} \mathbb{E}[|e_j|^2] &= \mathbb{E}[|r_1|^2] + \mathbb{E}[|r_2|^2] \leq \mathbb{E}[|r_1|^2] + \mathbb{E}[|\xi_j|^2] = \mathbb{E}[|r_1|^2] + m\gamma^2 \\ &\leq \frac{e^{2(\beta+\epsilon)h}}{(1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^2} (\mathbb{E}[|e_{j-1}|^2] + D) - \frac{D}{(1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^2} + m\gamma^2 \\ &\leq \theta (\mathbb{E}[|e_{j-1}|^2] + D) - \frac{D}{(1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^2} + m\gamma^2 = \theta (\mathbb{E}[|e_{j-1}|^2] + D) + \mathcal{E}, \end{aligned}$$

where $\mathcal{E} = m\gamma^2 - \Theta D$. Then, we have

$$\mathbb{E}[|e_j|^2] + D \leq \theta (\mathbb{E}[|e_{j-1}|^2] + D) + \mathcal{E} + D.$$

Applying this inequality repeatedly, we have

$$\begin{aligned} \mathbb{E}[|e_j|^2] &\leq \theta^j (\mathbb{E}[|e_0|^2] + D) + (\mathcal{E} + D) \frac{1 - \theta^j}{1 - \theta} - D \\ &= \theta^j (\mathbb{E}[|e_0|^2] + D) + (m\gamma^2 + (1 - \Theta)D) \frac{1 - \theta^j}{1 - \theta} - D \\ &= \theta^j (\mathbb{E}[|e_0|^2] + D) + m\gamma^2 \frac{1 - \theta^j}{1 - \theta} + \left(\frac{(1 - \theta^j)(1 - \Theta)}{1 - \theta} - 1 \right) D. \end{aligned}$$

Moreover, if $\theta < 1$, [\(3.16\)](#) holds in the limit of $j \rightarrow \infty$. ■

Remark 3.4. The explicit condition about α_0 so that $\theta = (1 + \frac{\alpha^2}{\gamma^2} \lambda_*)^{-2} e^{2(\beta+\epsilon)h} < 1$ is given by

$$(3.26) \quad \alpha_0 = \max\{\lambda_0^{-\frac{1}{2}} \gamma e^{ah} (e^{(\beta+\epsilon)h} - 1)^{\frac{1}{2}}, e^{\frac{1}{2}(a+\beta+\epsilon)h}\}.$$

Indeed, the condition $\theta < 1$ is simplified to

$$e^{(\beta+\epsilon)h} - 1 < \frac{\alpha^2}{\gamma^2} \lambda_* = \min\left\{\frac{\alpha^2}{\gamma^2} e^{-ah} \lambda_0, e^{-ah} \alpha^2 - 1\right\}.$$

Hence, (3.26) is obtained by solving the inequalities for α . Note that this lower bound is only a sufficient condition to obtain the error bound (3.16) and hence smaller α can provide an accurate state estimation numerically.

As noted in Section 1, Luo and Hoteit [27] derive both upper and lower bounds for the multiplicative inflation parameter α to ensure the residual error $e_n^r = y_n - H\bar{v}_n$ remains within a prescribed interval. The bound for α is adaptively computed using the prediction residual error $\hat{e}_n^r = y_n - H\hat{v}_n$ before the analysis step at each time. They do not impose any assumptions on the model dynamics. However, their theory only guarantees the bound for the residual error e_n^r , not the actual error e_n . In contrast, our theory ensures the bound for the actual error e_n , provided that the constants ρ and β for the model dynamics are estimated. This estimation is a separate task from obtaining the bound for α .

Compared to the additive inflation in [20], the multiplicative inflation in this paper inflates the spread of the ensemble. As a result, we cannot obtain the error bound for all members in the analysis ensemble V_j . Instead, we establish the error bound for the analysis mean $\bar{v}_j$ using Lemma 3.3, which introduces an additional constant D . The constant D gives rise to the second term in the error bound (3.16). However, this additional term becomes negligible in a certain limit as shown in Corollary 3.5.

Corollary 3.5. Under the same assumptions of Theorem 3, we consider the accurate observation limit (i.e., $\gamma \rightarrow 0$). Then, the filtering error is the order of the observation noise.

$$\limsup_{j \rightarrow \infty} \mathbb{E}[|e_j|^2] = O(\gamma^2).$$

Proof. It is clear that $\Theta = \left(\frac{\gamma^2}{\gamma^2 + \alpha^2 \lambda_*}\right)^2 = O(\gamma^4)$ and $\theta = O(\gamma^4)$. Then, we have

$$\begin{aligned} \frac{1 - \Theta}{1 - \theta} - 1 &= (1 - \Theta)(1 + \theta + O(\theta^2)) - 1 = 1 - \Theta + \theta + O(\theta^2) - 1 \\ &= \Theta(e^{2(\beta+\epsilon)h} - 1) + O(\theta^2) = O(\gamma^4). \end{aligned}$$

Therefore, it follows that

$$\limsup_{j \rightarrow \infty} \mathbb{E}[|e_j|^2] = \frac{m\gamma^2}{1 - \theta} + \left(\frac{1 - \Theta}{1 - \theta} - 1\right) D = m\gamma^2(1 + O(\gamma^4)) + O(\gamma^4) = O(\gamma^2). \quad \blacksquare$$

4. Numerical example. We show a numerical example applying the ETKF with multiplicative inflation to the Lorenz 96 (L96) equation, which is a chaotic phenomenological model in meteorology [25, 26]. For a given number of components $m \in \mathbb{N}$, the governing equation of a state vector $u = (u^1, \dots, u^m)^* \in \mathbb{R}^m$ is given by

$$\frac{du^i}{dt} = (u^{i+1} - u^{i-2})u^{i-1} - u^i + f, \quad i = 1, \dots, m,$$

where $u^{-1} = u^{m-1}$, $u^0 = u^m$, and $u^{m+1} = u^1$ and $f \in \mathbb{R}$ is external forcing. It is easy to check that the L96 satisfies [Assumption 1](#) as Proposition 3.2 in [23] and [Assumption 4](#) (locally Lipschitz) since the derivative of its vector field is bounded on any bounded domain.

For a numerical example, we set $m = 40$ and $f = 8$. For a time interval $h = 0.05$, we simulate the forward map Ψ_h of the L96 by using the fourth-order Runge-Kutta method with a time step size $dt = 0.01$. First, we generate the hidden true states $(u_j)_{j=1}^J$ for $J = 480$ as the solutions after computing the forward map 1440 times so that u_1 is in an absorbing ball associated with the L96. Second, we consider the full observation $H = I_m$ and $\Gamma = \gamma^2 I_m$ with $\gamma = \sqrt{0.1}$ so that [Assumption 3](#) holds. Then, we generate a time series of observations $(y_j)_{j=1}^J$ by adding Gaussian noises to the true states $y_j = u_j + \xi_j$, $\xi_j \sim N(0, \Gamma)$ for $j = 1, \dots, J$. For the ETKF, we set an ensemble size $N = m + 1$ and define an initial ensemble by $V_0 = [e_1, e_2, \dots, e_m, -\sum_{i=1}^m e_i] \in \mathbb{R}^{m \times N}$ for the standard basis e_i ($i = 1, \dots, m$) of $\mathbb{R}^m$ so that the assumption $\lambda_{\min}(C_0) > 0$ in [Theorem 3](#) holds. From V_0 and $(y_j)_{j=1}^J$, we compute a time series of the prediction $(\hat{V}_j)_{j=1}^J$ and the analysis $(V_j)_{j=1}^J$ ensembles using the ETKF with the multiplicative covariance inflation for each $\alpha = 1.0, 1.1, 5.0$.

In [Figure 1\(a\)](#), we show the expectation of the square error (SE) of the analysis mean $\mathbb{E}[|e_j|^2]$ for $\alpha = 1.0, 1.1, 5.0$. The expectation is approximated by averaging over 20 random seeds to generate observation noises. We also show the theoretical error bound (3.16) for the case when the error growth is sufficiently suppressed by the inflation, i.e., $\theta \ll 1$. In this case, we approximate the bound by $m\gamma^2$. Additionally, in [Figure 1\(b\)](#), we show the minimum eigenvalue of the prediction covariance $\hat{\lambda}_{\min} = \lambda_{\min}(\hat{C}_j)$ for $\alpha = 1.0, 1.1, 5.0$. Without the inflation ($\alpha = 1.0$), the analysis SE is larger than the theoretical bound ([Figure 1\(a\)](#)), and the $\hat{\lambda}_{\min} \sim 10^{-10}$ is negligibly small compared to the variance of observation noise $\gamma^2 = 0.1$ ([Figure 1\(b\)](#)). On the other hand, a large inflation parameter ($\alpha = 5.0$) leads to a lower bound of $\hat{\lambda}_{\min}$ with a value $\sim 10^{-2}$, which is about 8 orders of magnitude larger than that with $\alpha = 1.0$ ([Figure 1\(b\)](#)). Then, the analysis SE is smaller than or has the same order with the theoretical bound except at the initial time ([Figure 1\(a\)](#)). This result supports [Theorem 3](#), considering the approximation error of the theoretical bound and numerical errors. With a relatively small inflation parameter ($\alpha = 1.1$), the analysis SE is even smaller than when $\alpha = 5.0$ although $\hat{\lambda}_{\min}$ still takes a small value. This implies that the optimal α is not characterized by our estimate. Nevertheless, this numerical result does not contradict our theory since a large inflation parameter and a lower bound of the minimum eigenvalues are sufficient conditions to obtain the error bound (3.16). The codes are available in https://github.com/KotaTakeda/etkf_inflation.

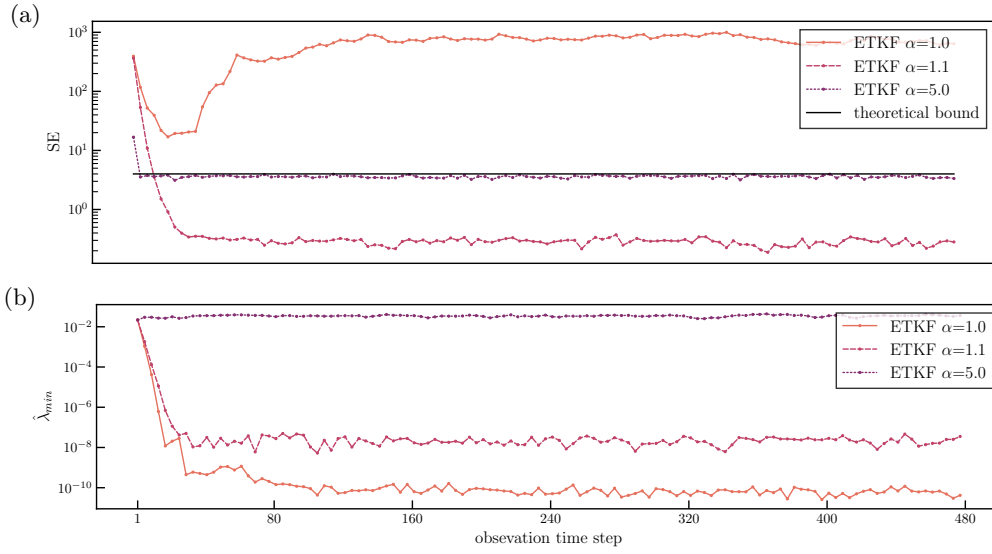


Figure 1. Plots of (a) the SE and (b) $\hat{\lambda}_{min}$ vs. observation time step j for the ETKF with the multiplicative inflation parameter $\alpha = 1.0, 1.1, 5.0$.

5. Summary and discussion. The theoretical aspects of the ETKF are investigated for chaotic dynamics on Hilbert spaces. We prove that the filtering error of the ETKF is bounded for any finite time even on an infinite-dimensional state space. In addition, an appropriate multiplicative covariance inflation to the ETKF ensures the uniform-in-time error bound on a finite-dimensional state space. Furthermore, we determine the minimum value of the inflation parameter α sufficient to obtain the uniform-in-time error bound. As a corollary, the leading order of the error bound is equal to the variance γ^2 of observation noises, indicating that the accuracy of the state estimation is effectively improved by the accurate observation. These results are relevant to those for the PO method with the additive inflation [20]. Besides, numerical results exemplify a scenario indicated by Theorem 3. The numerical results also imply that our bound does not provide the optimal α .

Let us discuss future directions. Since our result establishes only a sufficient condition of α to achieve the error bound, we need to investigate the optimal value of α . Additionally, our analysis is limited to linear and full observation as in Assumption 3 for a technical reason, although such situations are rare in many real-world applications. For a non-trivial observation operator H and Γ , the term $\hat{C}H^*\Gamma^{-1}H$ is not generally symmetric, which makes the error analysis complicated. In the case of nonlinear observation operators, the term $\hat{C}H^*\Gamma^{-1}H$ and the matrix transformation (2.14) even make no sense. Furthermore, our analysis is limited to when the ensemble size N is larger than or equal to the state space dimension m to prevent a degenerated ensemble covariance. Hence, we need to establish the error bound for the ETKF when N is smaller than $m \leq \infty$. To this end, we consider detailed properties of the model dynamics in the present analysis. For instance, it is useful to identify unstable directions of the filtering error where the error grows through the evolution of the model dynamics. It is known that many dissipative dynamical systems, including the two-dimensional Navier-Stokes equations, have a finite number of such unstable directions [8, 9, 33]. Hence, it is reasonable to

assume that the dynamics have a finite number of unstable directions, and perturbations to the trajectory in the other directions decay through its evolution. In addition, we assume the N is larger than or equal to the finite number of unstable directions. If the ETKF appropriately reduces the error of these directions under these assumptions, the dynamics reduces the error of the other directions, and we obtain a better error bound.

Acknowledgments. We would like to thank the reviewers for their valuable feedback and constructive comments. We are also grateful to Dr. Miyoshi for his insightful comments on the numerical results. We used an AI tool to edit or polish the authors' written text for spelling, grammar, or general style.

REFERENCES

- [1] O. AL-GHATTAS AND D. SANZ-ALONSO, *Non-asymptotic analysis of ensemble Kalman updates: Effective dimension and localization*, Inf. Inference, 13 (2024), p. iaad043, <https://doi.org/10.1093/imaiai/iaad043>.
- [2] J. ANDERSON, *Spatially and temporally varying adaptive covariance inflation for ensemble filters*, Tellus A, 61 (2009), pp. 72–83, <https://doi.org/10.1111/j.1600-0870.2007.00361.x>.
- [3] J. L. ANDERSON, *An Ensemble Adjustment Kalman Filter for Data Assimilation*, Mon. Weather Rev., 129 (2001), pp. 2884–2903, [https://doi.org/10.1175/1520-0493\(2001\)129<2884:AEAKFF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2).
- [4] J. L. ANDERSON, *An adaptive covariance inflation error correction algorithm for ensemble filters*, Tellus Dyn. Meteorol. Oceanogr., 59 (2007), pp. 210–224, <https://doi.org/10.1111/j.1600-0870.2006.00216.x>.
- [5] J. L. ANDERSON AND S. L. ANDERSON, *A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts*, Mon. Weather Rev., 127 (1999), pp. 2741–2758, [https://doi.org/10.1175/1520-0493\(1999\)127<2741:AMCIOT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2).
- [6] C. H. BISHOP, B. J. ETHERTON, AND S. J. MAJUMDAR, *Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects*, Mon. Weather Rev., 129 (2001), pp. 420–436, [https://doi.org/10.1175/1520-0493\(2001\)129<0420:ASWTET>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2).
- [7] G. BURGERS, P. J. VAN LEEUWEN, AND G. EVENSEN, *Analysis Scheme in the Ensemble Kalman Filter*, Mon. Weather Rev., 126 (1998), pp. 1719–1724, [https://doi.org/10.1175/1520-0493\(1998\)126<1719:ASITEK>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2).
- [8] P. CONSTANTIN AND C. FOIAS, *Global lyapunov exponents, kaplan-yorke formulas and the dimension of the attractors for 2D navier-stokes equations*, Commun. Pure Appl. Math., 38 (1985), pp. 1–27, <https://doi.org/10.1002/cpa.3160380102>.
- [9] P. CONSTANTIN AND C. FOIAS, *Navier-Stokes Equations*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, 1988.
- [10] J. B. CONWAY, *A Course in Functional Analysis*, vol. 96 of Graduate Texts in Mathematics, Springer New York, New York, 2007, <https://doi.org/10.1007/978-1-4757-4383-8>.
- [11] J. DE WILJES, S. REICH, AND W. STANNAT, *Long-time stability and accuracy of the ensemble kalman-bucy filter for fully observed processes and small measurement noise*, SIAM J. Appl. Dyn. Syst., 17 (2018), pp. 1152–1181, <https://doi.org/10.1137/17m1119056>.
- [12] G. EVENSEN, *Data Assimilation: The Ensemble Kalman Filter*, Springer, Berlin, Heidelberg, 2009, <https://doi.org/10.1007/978-3-642-03711-5>.
- [13] T. M. HAMILL, J. S. WHITAKER, AND C. SNYDER, *Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter*, Mon. Weather Rev., 129 (2001), pp. 2776–2790, [https://doi.org/10.1175/1520-0493\(2001\)129<2776:DDFOBE>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2).
- [14] M. HLADNIK AND MATJA. OMLADIC, *Spectrum of the Product of Operators*, Proc. Amer. Math. Soc., 102 (1988), pp. 300–302, <https://doi.org/10.2307/2045879>.
- [15] P. L. HOUTEKAMER AND F. ZHANG, *Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation*, Mon. Weather Rev., 144 (2016), pp. 4489–4532, <https://doi.org/10.1175/MWR-D-15-0440.1>.
- [16] B. R. HUNT, E. J. KOSTELICH, AND I. SZUNYOGH, *Efficient data assimilation for spatiotemporal chaos:*

- A local ensemble transform Kalman filter*, Physica D, 230 (2007), pp. 112–126, <https://doi.org/10.1016/j.physd.2006.11.008>.
- [17] T. JANJIĆ, L. NERGER, A. ALBERTELLA, J. SCHRÖTER, AND S. SKACHKO, *On Domain Localization in Ensemble-Based Kalman Filter Algorithms*, Mon. Weather Rev., 139 (2011), pp. 2046–2060, <https://doi.org/10.1175/2011MWR3552.1>.
 - [18] R. E. KALMAN, *A New Approach to Linear Filtering and Prediction Problems*, J. Basic Eng., 82 (1960), pp. 35–45, <https://doi.org/10.1115/1.3662552>.
 - [19] E. KALNAY, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, 2002.
 - [20] D. T. B. KELLY, K. J. H. LAW, AND A. M. STUART, *Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time*, Nonlinearity, 27 (2014), pp. 2579–2603, <https://doi.org/10.1088/0951-7715/27/10/2579>.
 - [21] E. KWIATKOWSKI AND J. MANDEL, *Convergence of the square root ensemble kalman filter in the large ensemble limit*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 1–17, <https://doi.org/10.1137/140965363>.
 - [22] T. LANGE AND W. STANNAT, *Mean field limit of Ensemble Square Root filters - discrete and continuous time*, Found. Data Sci., 3 (Thu Jan 21 19:00:00 EST 2021), pp. 563–588, <https://doi.org/10.3934/fods.2021003>.
 - [23] K. J. H. LAW, D. SANZ-ALONSO, A. SHUKLA, AND A. M. STUART, *Filter accuracy for the Lorenz 96 model: Fixed versus adaptive observation operators*, Physica D, 325 (2016), pp. 1–13, <https://doi.org/10.1016/j.physd.2015.12.008>.
 - [24] K. J. H. LAW, A. M. STUART, AND K. C. ZYGALAKIS, *Data Assimilation: A Mathematical Introduction*, Springer Cham, Cham, 2015.
 - [25] E. N. LORENZ, *Predictability: A problem partly solved*, Proc. Semin. Predict., 1 (1996), pp. 1–18.
 - [26] E. N. LORENZ AND K. A. EMANUEL, *Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model*, J. Atmos. Sci., 55 (1998), pp. 399–414, [https://doi.org/10.1175/1520-0469\(1998\)055<0399:OSFSWO>2.0.CO;2](https://doi.org/10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2).
 - [27] X. LUO AND I. HOTEIT, *Covariance Inflation in the Ensemble Kalman Filter: A Residual Nudging Perspective and Some Implications*, Mon. Weather Rev., 141 (2013), pp. 3360–3368, <https://doi.org/10.1175/MWR-D-13-00067.1>.
 - [28] A. MAJDA AND X. WANG, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*, Cambridge University Press, Cambridge, 2006, <https://doi.org/10.1017/CBO9780511616778>.
 - [29] A. J. MAJDA AND J. HARLIM, *Filtering Complex Turbulent Systems*, Cambridge University Press, Cambridge, 2012.
 - [30] J. MANDEL, L. COBB, AND J. D. BEEZLEY, *On the convergence of the ensemble Kalman filter*, Appl. Math., 56 (2011), pp. 533–541, <https://doi.org/10.1007/s10492-011-0031-2>.
 - [31] T. MIYOSHI, *The Gaussian Approach to Adaptive Covariance Inflation and Its Implementation with the Local Ensemble Transform Kalman Filter*, Mon. Weather Rev., 139 (2011), pp. 1519–1535, <https://doi.org/10.1175/2010MWR3570.1>.
 - [32] S. REICH AND C. COTTER, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, Cambridge, 2015, <https://doi.org/10.1017/CBO9781107706804>.
 - [33] R. TEMAM, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer New York, New York, 1997.
 - [34] X. T. TONG, A. J. MAJDA, AND D. KELLY, *Nonlinear stability and ergodicity of ensemble based Kalman filters*, Nonlinearity, 29 (2016), pp. 657–691, <https://doi.org/10.1088/0951-7715/29/2/657>.
 - [35] X. T. TONG, A. J. MAJDA, AND D. KELLY, *Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation*, Comm. Math. Sci., 14 (2016), pp. 1283–1313, <https://doi.org/10.4310/CMS.2016.v14.n5.a5>.
 - [36] J. S. WHITAKER AND T. M. HAMILL, *Ensemble Data Assimilation without Perturbed Observations*, Mon. Weather Rev., 130 (2002), pp. 1913–1924, [https://doi.org/10.1175/1520-0493\(2002\)130<1913:EDAWPO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2).

Appendix A. Linear operators on Hilbert space. Let $\mathcal{H}$ be a Hilbert space and $I = I_{\mathcal{H}}$.

Lemma A.1. *Let $A \in \mathcal{L}(\mathcal{H})$. If $I + A$ is invertible, then we have*

$$(A.1) \quad (I + A)^{-1} = I - (I + A)^{-1}A.$$

Especially, if $A \in \mathcal{L}_{sa}(\mathcal{H})$ and $A \succ 0$,

$$(A.2) \quad 0 \preceq A(A + I)^{-1} = (A + I)^{-1}A \preceq I, \quad 0 \preceq (A + I)^{-1} \preceq I.$$

Proof. (A.1) is easily confirmed by

$$LHS = (I + A)^{-1}(I + A - A) = I - (I + A)^{-1}A.$$

Next, for (A.2), (A.1) yields $A(A + I)^{-1} = (A + I)^{-1}A$. The inequalities hold from the spectral mapping theorem. ■

Lemma A.2. *Let $\Gamma : \mathcal{H} \rightarrow \mathcal{H}$ be invertible and $V \in \mathcal{H}^N$. Then, the operator $VV^* + \Gamma$ is invertible, and*

$$(A.3) \quad (I + V^*\Gamma^{-1}V)^{-1}V^*\Gamma^{-1} = V^*(VV^* + \Gamma)^{-1}.$$

Furthermore,

$$(A.4) \quad (I + V^*\Gamma^{-1}V)^{-1} = I - V^*(VV^* + \Gamma)^{-1}V.$$

Proof. $VV^* + \Gamma$ is invertible owing to $VV^* \succeq 0$ and $\Gamma \succ 0$. Then, we have

$$V^*\Gamma^{-1}(VV^* + \Gamma) = V^*\Gamma^{-1}VV^* + V^* = (I + V^*\Gamma^{-1}V)V^*.$$

This is equivalent to (A.3). For (A.4), the assertion (A.3) yields

$$(I + V^*\Gamma^{-1}V)^{-1}V^*\Gamma^{-1}V = V^*(VV^* + \Gamma)^{-1}V.$$

Hence, we have

$$I - V^*(VV^* + \Gamma)^{-1}V = I - (I + V^*\Gamma^{-1}V)^{-1}V^*\Gamma^{-1}V = (I + V^*\Gamma^{-1}V)^{-1},$$

where the last equality holds from (A.1). ■