Journal of Computational Physics 545 (2026) 114465

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Data assimilation models for computing probability distributions of complex multiscale systems

Di Qi [a,*], Jian-Guo Liu [b]

[a] Department of Mathematics, Purdue University, 47907, West Lafayette, IN, USA
[b] Department of Mathematics, Department of Physics, Duke University, 27708, Durham, NC, USA

ARTICLE INFO

Keywords:
Nonlinear data assimilation
Statistical modeling
Multiscale system
Ensemble method

ABSTRACT

We introduce a data assimilation strategy aimed at accurately capturing key non-Gaussian structures in probability distributions using a small ensemble size. A major challenge in statistical forecasting of nonlinearly coupled multiscale systems is mitigating the large errors that arise when computing high-order statistical moments. To address this issue, a high-order stochastic-statistical modeling framework is proposed that integrates statistical data assimilation into finite ensemble predictions. The method effectively reduces the approximation errors in finite ensemble estimates of non-Gaussian distributions by employing a filtering update step that incorporates observation data in leading moments to refine the high-order statistical feedback. Explicit filter operators are derived from intrinsic nonlinear coupling structures, allowing straightforward numerical implementations. Performance of the proposed method is first demonstrated through extensive numerical experiments on a prototype triad system, which offers an instructive and computationally manageable platform mimicking essential aspects of nonlinear turbulent dynamics. Further experiments on the Lorenz 96 system are conducted to assess potential generalization to high-dimensional systems. The numerical results show that the statistical data assimilation algorithm consistently captures the mean and covariance, as well as various non-Gaussian probability distributions exhibited in various statistical regimes. The modeling framework can serve as a useful tool for efficient sampling and reliable forecasting of complex probability distributions commonly encountered in a wide variety of applications involving multiscale coupling and nonlinear dynamics.

1. Introduction

Predicting the distinct statistical behaviors observed in nonlinear dynamical systems involving multiple spatial and temporal scales remains a fundamental challenge across various natural and engineering problems [1–4]. One primary difficulty arises from accurately quantifying the multiscale nonlinear interactions between the large-scale mean state and small-scale stochastic fluctuations amplified by inherent instability. Such interactions often lead to non-Gaussian probability distributions characterized by high-order statistics and intermittent extreme events, driven by the intricate multiscale coupling mechanism [5–7]. Developing efficient computational algorithms capable of capturing these critical non-Gaussian probabilistic features remains a central issue in practical applications [8–10]. Ensemble-based methods together with data assimilation strategies [8,11–13] have been successfully applied for recovering leading-order statistics in linear dynamical systems from noisy and partial observations. However, as nonlinear coupling effects become dominant, low-order approaches such as Kalman filters using only the leading two moments often suffer inherent difficulties

* Corresponding author.
E-mail addresses: qidi@purdue.edu (D. Qi), jian-guo.liu@duke.edu (J.-G. Liu).

https://doi.org/10.1016/j.jcp.2025.114465
Received 28 March 2025; Received in revised form 22 August 2025; Accepted 13 October 2025
Available online 26 October 2025

and fail to capture the essential higher-order moment statistics [14–16]. As a result, accurate and efficient methods for quantification and prediction of these high-order statistics and the associated non-Gaussian probability distributions are still needed for reliable forecasting of the complex phenomena.

We consider the nonlinear statistical forecast problem formulated as the following general stochastic dynamical equation (SDE) [14,17] describing the uncertainty evolution of the random state $u \in \mathbb{R}^d$ starting from $u(0;\omega) \sim \mu_0$ according to the initial distribution μ_0 and driven by external forcing and nonlinear interactions

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \Lambda u + B(u,u) + F(t) + \sigma(t)\dot{W}(t;\omega). \tag{1.1}$$

On the right hand side of the above Eq. (1.1), the first term, $\Lambda = L - D$, represents linear dispersion and dissipation effects, where $L^* = -L$ is an energy-conserving skew-symmetric operator; and $D < 0$ is a negative definite operator. Inhomogeneous forcing effects are introduced in a deterministic component, F, and a stochastic component represented by a Gaussian random process, $\sigma(t)\dot{W}(t;\omega)$. Most importantly, nonlinear coupling effect has a non-negligible contribution in the dynamical system introduced via a quadratic form, $B(u,u)$, which satisfies the energy conservation law by $u \cdot B(u,u) = 0$. The model structures in (1.1) are representative in a wide variety of multiscale systems found in many fields [3,18,19]. In computing key statistical predictions of the model state u, the low-order moments become intricately connected to the high-order statistical information due to the nonlinear coupling $B(u,u)$. In this case even with low dimensionality d, finite ensemble approximation frequently suffers from collapse of particles, with the group of particles concentrating in the center region of the PDF and failing to capture the outliers charactering the key non-Gaussian statistics and extreme events [6,16,20]. Thus effective algorithms require to capture the entire probability density functions (PDFs) including high-order information using a moderate ensemble size to maintain the affordable computational cost.

1.1. Related works in data assimilation

Sequential data assimilation strategies [11,21,22] have long been used for finding the optimal probability estimate of a stochastic state based on observation data. Among them, *ensemble Kalman filters* [23,24] based on Gaussian or near-Gaussian assumptions provide effective tools for state and parameter estimations in relatively high-dimensional settings. In accommodating nonlinear systems involving highly non-Gaussian statistics, *particle-type filters* [25–27] are proposed to approximate the probability distribution of model state through a set of weighted particles. A rich variety of techniques have been introduced in recent advances of nonlinear filtering methods for modeling nonlinearly coupled signals, such as feedback particle filters [28,29], variational mapping filters [30], and particle flow filters [31], just to name a few. A hybrid ensemble Kalman and particle filter is also proposed [32] aiming to combine the benefits of both filters. Furthermore, *learning filters* [33,34] have emerged recently exploiting score-based generative models [35,36] and variational Bayes techniques [37,38] to aid the model forecast and analysis steps with machine learning strategies to achieve better filter behaviors. Despite wide applications, major difficulties persist for accurate statistical forecast of stochastic states especially when non-Gaussian statistics are present combined with inherent model instability. Conventional ensemble-based approaches often suffer difficulties in accurately capturing the crucial higher-order moments information thus become insufficient to maintain stable and accurate prediction with a finite number of samples [39,40].

1.2. Contributions and paper outline

In this paper, we introduce a practical modeling and computational strategy designed to accurately capture the probability distributions and key statistical characteristics of the solution to (1.1). Based on the theoretical framework presented in [17] and the filtering approach using statistical observations, we develop a new data assimilation algorithm aimed at achieving accurate statistical prediction in finite ensemble approximation of potentially highly non-Gaussian probability distributions. To cope with the computational limitation in practical applications, the ensemble of simulated samples needs to be constrained in a small size. We propose to correct the large fluctuating errors that commonly appear using small ensemble size by exploiting partial observation data of the low-order statistical moments. An effective data assimilation algorithm is then formulated to improve the model accuracy by capturing the higher-order moment information and reduce the high computational cost at the same time. In particular, we conduct detailed numerical study on the proposed ensemble data assimilation scheme based on the systematic statistical modeling framework and applied on a representative triad system [41] with multiple distinctive statistical regimes.

The main ideas in constructing the data assimilation model using statistical observation data is illustrated in the flow chart in Fig. 1. We propose to compute the probability distribution ρ of the model state u in (1.1) using a set of more tractable stochastic-statistical Eqs. (2.2). The target probability distribution will be approximated by an empirical probability distribution ρ^N through an interacting particle simulation of the stochastic coefficients Z. However, in practice this ensemble-based approach will often become insufficient to accurately capture the essential PDF structures when only a small sample size N is available. To study the evolution of uncertainty from a finite sample size, we consider the continuous distribution $\rho(\cdot; y^N)$ of each sample Z^i as a $\mathcal{P}(\mathbb{R}^d)$-valued random field defined by (2.7). The randomness is introduced due to the finite sample estimation of the leading moments $y^N(\omega)$ in the statistical Eq. (2.9). This leads to a natural filtering problem to find the optimal state estimation of the random field ρ based on the observation data $\mathcal{G}_t$ generated by the low-order statistical moments observed up to the time t. The optimal filter solution $\hat{\rho}$ then can be found through the projection on the space of $\mathcal{G}_t$-measurable square-integrable random fields and is given by the Kalman-Bucy filter (2.11) as an infinite-dimensional functional equations. To propose an efficient computational strategy to solve $\hat{\rho}$, a new stochastic process $\tilde{Z} \sim \tilde{\rho}$ is designed so that they can provide consistent high-order statistics $\tilde{\mathbb{E}}H = \mathcal{H}\hat{\rho}$ according to the nonlinear observation function H and

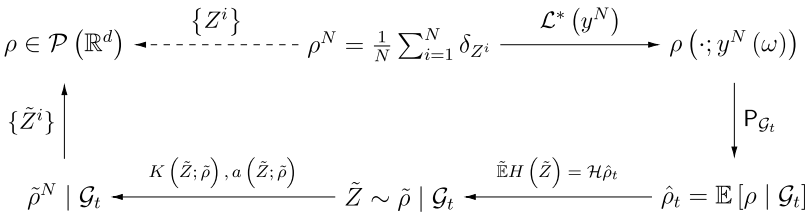


Fig. 1. Flow chart illustrating ideas in constructing the data assimilation model for statistical forecast.

Table 1
Key notations for the data assimilation model in this paper.

target probability density distribution of model state	ρ
empirical approximation of the target distribution	ρ^N
sample trajectories generated by the target distribution	Z_t^i
observations and filtration of observation data	$y_t^N, \mathcal{G}_t$
observation function and associated observation operator	$H, \mathcal{H}$
surrogate stochastic process and the filter distribution	$\tilde{Z} \sim \tilde{\rho} \mid \mathcal{G}_t$
Kalman gain and drift operator in the filtering process	K, a

the associated observation operator $\mathcal{H}$. In particular, the governing SDE for the new process $\tilde{Z}$ (2.12) is derived with explicit forms of the filtering coefficients a, K in (3.9). Finally, this new probability distribution $\tilde{\rho}$ can be computed efficiently by a finite sample approximation $\tilde{\rho}^N$ which provides accurate high-order statistical consistency and generates samples giving a better representation of the target distribution of the model. For clarity in presentation, we list the key notations used in this paper in Table 1.

In the structure of this paper, we first discuss the general multiscale modeling framework and develop the data assimilation model based on the coupling structure in the stochastic-statistical model in Section 2. Then, the detailed the ensemble data assimilation equations involving the explicit filtering operators and practical computational algorithms are constructed in Section 3. The performance of the new data assimilation model and its skill in recovering both leading-order mean and covariance and the crucial higher-order statistical feedbacks are extensively tested under a representative prototype triad system demonstrating different statistical regimes in Section 4, and are further examined on the higher-dimensional Lorenz 96 system in Section 5. A summarizing discussion and potential future research directions are given in Section 6. Additional proofs of the results presented in the main text are provided in Appendix A and detailed equations and properties of the triad system with direct link to realistic applications are listed in Appendix B.

2. An integrated multiscale modeling framework with data assimilation

We start with describing the multiscale modeling strategy for solving the statistical solution to the general system (1.1). In particular, we propose the coupled stochastic-statistical equations and the associated ensemble approximation that can be naturally combined with data assimilation for improved sampling of the target probability distributions.

2.1. The coupled stochastic-statistical formulation for multiscale dynamics

To characterize the uncertainty in the stochastic model state, the solution u is represented as a random field (denoted by ω) and decomposed into the multiscale composition of a statistical mean state $\bar{u} = \mathbb{E}(u)$ and stochastic fluctuations u' in a high-dimensional representation under a proper orthonormal basis $\{\hat{v}_k\}_{k=1}^d$, that is,

$$u(t;\omega) = \bar{u}(t) + u'(t;\omega) = \bar{u}(t) + \sum_{k=1}^{d} Z_k(t;\omega)\hat{v}_k. \tag{2.1}$$

Above, $\bar{u}$ represents the statistical mean field of the dominant largest scale structure; and $Z(t;\omega) = \{Z_k(t;\omega)\}_{k=1}^d$ are the stochastic processes characterizing the uncertainty in the fluctuation processes u' on each eigenmode $\hat{v}_k$. Such decomposition is commonly used, for example, in describing the zonal jets in geophysical turbulence and the coherent radial flow in fusion plasmas [42,43].

Under the decomposition (2.1), we can reformulate the full statistics in the original stochastic state u as the leading two statistical moments $\bar{u}$, R and a mean-zero stochastic process Z governed by coupled statistical and stochastic dynamical equations. In particular, the *statistical dynamical equations* describing the evolution of the mean $\bar{u}(t) \in \mathbb{R}^d$ and covariance $R(t) \in \mathbb{R}^{d\times d}$ can be found to satisfy the following equations

$$\begin{aligned} \frac{\mathrm{d}\bar{u}}{\mathrm{d}t} &= \Lambda\bar{u} + B(\bar{u},\bar{u}) + F + \sum_{k,l=1}^{d} B\big(\hat{v}_k,\hat{v}_l\big)\mathbb{E}\big(Z_k Z_l\big), \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= L(\bar{u})R + RL^T(\bar{u}) + Q_\sigma + Q_F(\mathbb{E}(Z\otimes Z\otimes Z)). \end{aligned} \tag{2.2a}$$

Accordingly, the stochastic process $Z(t;\omega) \in \mathbb{R}^d$ satisfies the following *stochastic differential equation* coupled with the leading-order statistical moments $(\bar{u}, R)$ solved from (2.2a)

$$\mathrm{d}Z = L(\bar{u})Z\mathrm{d}t + Q_v(Z \otimes Z - R)\mathrm{d}t + \sigma \mathrm{d}W. \tag{2.2b}$$

In the statistical Eq. (2.2a), we define the stochastic forcing $Q_\sigma = \sigma\sigma^T$ from the white noise process, the coupling coefficients $L(\bar{u})$ due to interactions between the mean and covariance, and Q_F due to the higher-order moments feedback from the triad modes $Z \otimes Z \otimes Z = \{Z_m Z_n Z_k\}$. In stochastic Eq. (2.2b), Q_v represents the stochastic quadratic coupling between modes from $Z \otimes Z = \{Z_m Z_n\}$. The explicit expressions for these coefficients can be found according to the nonlinear coupling function $B(u,u)$ defined by all the modes $k, l = 1, \cdots, d$ as

$$\begin{aligned} L_{kl}(\bar{u}) =& \hat{v}_k \cdot \left[\Lambda \hat{v}_l + B(\bar{u}, \hat{v}_l) + B(\hat{v}_l, \bar{u})\right], \\ Q_{F,kl} =& \sum_{m,n=1}^{d} \left[\gamma_{kmn}\mathbb{E}(Z_m Z_n Z_l) + \gamma_{lmn}\mathbb{E}(Z_m Z_n Z_k)\right], \\ Q_{v,k} =& \sum_{m,n=1}^{d} \gamma_{kmn}(Z_m Z_n - R_{mn}), \end{aligned} \tag{2.3}$$

with the coupling coefficients $\gamma_{kmn} = \hat{v}_k \cdot B(\hat{v}_m, \hat{v}_n)$ due to the quadratic nonlinear coupling.

The above *coupled stochastic-statistical equations* (2.2) provide a self-consistent closed formulation for recovering the statistical solution of u. The leading-order moments $\bar{u}$ and R are solved by the statistical Eq. (2.2a) involving the higher-order moments of the stochastic coefficients Z, and all high-order statistical information is recovered through the law ρ of the stochastic process Z from (2.2b) dependent on the solutions $\{\bar{u}, R\}$. More detailed discussions on the derivation and advantages of this new formulation can be found in [17]. It demonstrates that this new set of equations provides consistent statistical solutions with the original system (1.1), while enjoys additional advantages that are more adaptive to various model reduction and data assimilation strategies [14,41].

2.2. Predicting probability density functions using statistical observation data

A practical approach for numerically implementing the coupled stochastic-statistical Eqs. (2.2) is to adopt a particle approximation to the probability distribution of the stochastic process Z. Thus the expectations required in the statistical Eq. (2.2a) can be estimated through an empirical average of the samples $\mathbf{Z} = \{Z^i\}_{i=1}^N$, that is

$$\rho^N(z,t) = \frac{1}{N}\sum_{i=1}^{N} \delta(z - Z^i(t)), \quad \mathbb{E}^N(f(\mathbf{Z})) = \frac{1}{N}\sum_{i=1}^{N} f(Z^i). \tag{2.4}$$

Therefore, the statistical solution can be computed by solving the following equations as an interacting particle system by evolving the ensemble $\mathbf{Z}$ coupled with the moments $\bar{u}^N, R^N$

$$\begin{aligned} \frac{\mathrm{d}Z^i}{\mathrm{d}t} =& L(\bar{u}^N)Z^i + Q_v(Z^i \otimes Z^i - R^N) + \sigma\dot{W}^i, \quad i = 1, \cdots, N, \\ \frac{\mathrm{d}\bar{u}^N}{\mathrm{d}t} =& \Lambda\bar{u}^N + B(\bar{u}^N, \bar{u}^N) + \sum_{k,l} B(\hat{v}_k, \hat{v}_l)\mathbb{E}^N(\mathbf{Z}_k \otimes \mathbf{Z}_l) + F, \\ \frac{\mathrm{d}R^N}{\mathrm{d}t} =& L(\bar{u}^N)R^N + R^N L^T(\bar{u}^N) + Q_F(\mathbb{E}^N(\mathbf{Z} \otimes \mathbf{Z} \otimes \mathbf{Z})) + Q_\sigma \\ & + \epsilon^{-1}(\mathbb{E}^N[\mathbf{Z} \otimes \mathbf{Z}] - R_t^N). \end{aligned} \tag{2.5}$$

Several modifications are introduced in the numerical model (2.5) compared to the original Eqs. (2.2). Instead of computing the exact law $\rho(z,t)$ of the stochastic process Z by solving the following PDE (2.6), a finite particle approximation in the form of (2.4) is used to estimate the crucial higher moments feedback in the mean and covariance equations. The samples are generated by a McKean-Vlasov SDE implicitly dependent on all the sample trajectories through the statistical solutions $\bar{u}^N, R^N$. In addition, a relaxation term with an additional parameter $\epsilon > 0$ is added to the covariance equation for R^N to enforce consistency in the finite particle approximation of the covariance. It is found that this term is essential for maintaining stable numerical especially with strong mean-fluctuation coupling from the term $L(\bar{u})$ (see Fig. 4 in Section 4.1).

It can be shown [44] that the empirical measure ρ^N converges to the law of each Z^i, $\rho^N \to \rho$, as $N \to \infty$. The solution of the probability distribution ρ of Z^i is given by the corresponding Fokker-Planck equation

$$\frac{\partial\rho}{\partial t} = \mathcal{L}^*(\bar{u}, R)\rho := -\nabla \cdot \left[L(\bar{u})z\rho + Q_v(z \otimes z - R)\rho\right] + \frac{1}{2}\nabla \cdot \left[\nabla \cdot (Q_\sigma \rho)\right], \tag{2.6}$$

where $\mathcal{L}^*$ is the adjoint of the generator $\mathcal{L}$ that is dependent on the mean $\bar{u}$ and covariance R. Still, a major difficulty remains if only a very small number of samples N is affordable to estimate the empirical distribution ρ^N. Furthermore, the internal instability (that is, the positive eigenvalues in $L(\bar{u})$ due to the mean-fluctuation coupling $B(\bar{u}, \hat{v}_l) \cdot \hat{v}_k$) may lead to fast growth of the sample errors and quick divergence of the numerical solutions (see [41,45] and Fig. 3 in Section 4.1.2). This sets an inherent obstacle for efficient computation of the statistical solutions in the multiscale coupling system. To address this inherent difficulty, we assume that additional observation data $y = \{\bar{u}, R\}$ containing only the first two moments is available to improve the prediction of the probability

distribution ρ^N from the stochastic samples. Especially when the nonlinear coupling plays a dominant role in the dynamics, many non-Gaussian features will emerge in the probability distribution ρ. Our goal is then to filter the non-Gaussian PDF ρ of Z containing crucial higher-order statistics by taking into account observations from the leading two moments that are often easy to access.

In order to introduce the data assimilation strategy to improve the prediction of the computational model, we rewrite the coupled stochastic-statistical Eq. (2.5) as a conditional linear system about the probability distribution ρ of the samples Z^i coupled with the empirical moments $\bar{u}^N, R^N$ also as stochastic processes

$$\begin{aligned} \partial_t \rho &= \mathcal{L}^*\big(\bar{u}^N, R^N\big)\rho, \\ \mathrm{d}\bar{u}^N &= \big[\mathbb{E}H^m(Z) + h_m\big(\bar{u}^N\big)\big]\mathrm{d}t + \Gamma_m^N \mathrm{d}B_m, \\ \mathrm{d}R^N &= \big[\mathbb{E}H^v(Z) + h_v\big(\bar{u}^N, R^N\big)\big]\mathrm{d}t + \Gamma_v^N \mathrm{d}B_v. \end{aligned} \tag{2.7}$$

Above, $\mathcal{L}(\bar{u}, R)$ is given by the infinitesimal generator in (2.6) and ρ is the continuous density solution. In the equations for $\bar{u}^N, R^N$, we summarize all the deterministic terms (that is, all terms in (2.5) beside the ones with $\mathbb{E}^N$) in the functions h_m and h_v respectively. Higher-order moment feedbacks can be then written as expectations with respect to the continuous probability distribution ρ. We introduce the *observation functions* $H^m \in \mathbb{R}^d$ and $H^v \in \mathbb{R}^{d\times d}$ with explicit expressions found in (2.3) as quadratic and cubic functions about z

$$H_k^m(z) = \sum_{p,q=1}^{d} \gamma_{kpq} z_p z_q, \quad H_{kl}^v(z) = \sum_{p,q=1}^{d} \big(\gamma_{kpq} z_p z_q z_l + \gamma_{lpq} z_p z_q z_k\big). \tag{2.8}$$

Importantly, the additional noise terms with coefficients Γ_m^N and Γ_v^N are introduced in the observed states $\bar{u}^N, R^N$ to calibrate the errors from finite ensemble approximation. In fact, we can assume that the empirical average in the mean and covariance equations can be both decomposed into the expectation with ρ and the additional noise as an unbiased correction to the finite sample estimation

$$\mathbb{E}^N H(Z)\mathrm{d}t \cong \mathbb{E}H(Z)\mathrm{d}t + \Gamma^N \mathrm{d}B. \tag{2.9}$$

Thus, the additional noise term $\Gamma^N \mathrm{d}B$ is used to represent the fluctuating error from the N samples approximation to the true expectation $\mathbb{E}H(Z)$. It is confirmed from the numerical tests in Section 4.2.1 that (2.9) offers desirable characterization of the approximation errors in practical applications. In this way, the coupled system (2.7) sets up a standard linear filtering problem given by an infinite-dimensional signal process ρ as a $\mathcal{P}\big(\mathbb{R}^d\big)$-valued stochastic process and the observation process $\mathcal{G}_t = \sigma\{y(s), s \le t\}$ with $y(t) = \big\{\bar{u}^N(t), R^N(t)\big\}$ satisfying the linear equation with respect to the signal process ρ

$$\mathrm{d}y = [\mathcal{H}\rho + h(y)]\mathrm{d}t + \Gamma \mathrm{d}B, \quad \mathcal{H}\rho = \mathbb{E}H(Z) = \int H(z)\rho(z)\mathrm{d}z, \tag{2.10}$$

where $\mathcal{H}$ becomes a linear operator acting on the probability density ρ with $H = [H^m, H^v]$, $h = \big[h_m, h_v\big]$, and $\Gamma \mathrm{d}B = \big[\Gamma_m \mathrm{d}B_m, \Gamma_v \mathrm{d}B_v\big]$.

Applying the Kalman-Bucy filter in the infinite-dimensional Hilbert space [46,47] for the stochastic process ρ in (2.7) conditional on the observation processes in (2.10), we find the *optimal high-order filter solution* $\hat{\rho} = \mathbb{E}\big[\rho \mid \mathcal{G}_t\big]$ satisfying the following closed system of functional equations

$$\begin{aligned} \mathrm{d}\hat{\rho} &= \mathcal{L}^*\big(y\big)\hat{\rho}\mathrm{d}t + \hat{C}\mathcal{H}^*\Gamma^{-2}\big\{\mathrm{d}y_t - \big[\mathcal{H}\hat{\rho} + h\big(y\big)\big]\mathrm{d}t\big\}, \\ \mathrm{d}\hat{C} &= \big[\mathcal{L}^*\big(y\big)\hat{C} + \hat{C}\mathcal{L}(y)\big]\mathrm{d}t - \hat{C}\mathcal{H}^*\Gamma^{-2}\mathcal{H}\hat{C}\mathrm{d}t, \end{aligned} \tag{2.11}$$

where $\hat{C}(\omega) : L^2\big(\mathbb{R}^d\big) \to L^2\big(\mathbb{R}^d\big)$ is the self-adjoint covariance operator with $\hat{C}^* = \hat{C}$. The idea of filtering the probability distributions starts from the Fokker-Planck filter in [13], and a systematic filtering model is developed in [17] based on the specific nonlinear coupling structure in the stochastic-statistical model (2.2). Still, it remains intractable to directly solve the functional system (2.11). The final step is to construct effective ensemble solvers for the above optimal filter solution $\hat{\rho}$.

2.3. The approximate ensemble data assimilation with consistent high-order statistics

As a final step, we introduce a practical strategy to efficiently compute the optimal filtering solution $\hat{\rho}$ based on the observed statistics. The idea is to construct a surrogate process $\tilde{Z}$ so that its corresponding probability distribution of $\tilde{Z} \sim \tilde{\rho}$ can serve as an effective representation of the optimal filter solution $\hat{\rho}$. Then, efficient particle approaches can be adopted to capture the probability distribution of $\tilde{Z}$ instead of solving the infinite-dimensional equations (2.11).

Associated with the forecast Eq. (2.5) for the stochastic process Z, given N particles for the sampling solution of the stochastic state, $\tilde{\mathbf{Z}} = \big\{\tilde{Z}^i\big\}_{i=1}^N$, we can construct the following filtering updating equation with an additional update according to the observation in the second line

$$\begin{aligned} \mathrm{d}\tilde{Z}^i &= L\big(\bar{u}^N\big)\tilde{Z}^i\mathrm{d}t + Q_v\big(\tilde{Z}^i \otimes \tilde{Z}^i - R^N\big)\mathrm{d}t + \sigma \mathrm{d}\tilde{W}^i \\ &\quad + a^m\big(\tilde{Z}^i;\tilde{\rho}\big)\mathrm{d}t + K^m\big(\tilde{Z}^i;\tilde{\rho}\big)\mathrm{d}I^m + a^v\big(\tilde{Z}^i;\tilde{\rho}\big)\mathrm{d}t + K^v\big(\tilde{Z}^i;\tilde{\rho}\big)\mathrm{d}I^v, \end{aligned} \tag{2.12}$$

where the innovations I^m, I^v for the statistical observations are defined based on the observation data $\mathrm{d}y = (\mathrm{d}\bar{u}, \mathrm{d}R)$ and the ensemble statistical dynamics as

$$\begin{aligned} \mathrm{d}I^m(t) &= \mathrm{d}\bar{u} - \big[H^m\big(\tilde{Z}^i\big) + h_m\big(\bar{u}^N\big)\big]\mathrm{d}t, \\ \mathrm{d}I^v(t) &= \mathrm{d}R - \big[H^v\big(\tilde{Z}^i\big) + h_v\big(\bar{u}^N, R^N\big)\big]\mathrm{d}t. \end{aligned} \tag{2.13}$$

In the new filter Eq. (2.12) for the stochastic process $\tilde{Z}$, the first line follows the same dynamical equation (2.2b) as the forecast model, while the second line introduces additional control correction based on the observation data. In the filtering equation, two new functionals known as the *Kalman gain* K and the *drift* a are defined based on the probability distribution of $\tilde{Z}^i \sim \tilde{\rho}$, and needs to be solved by the Eq. (2.15) shown below. Notice that the filter Eq. (2.12) is also dependent on the first two moments $\left(\bar{u}^N, R^N\right)$ which can be solved by integrating the two statistical equations in (2.5). In addition, we need to introduce continuous observations to estimate $\mathrm{d}y_t$ in the filtering scheme. Assume that the observation data $y_n = y(t_n)$ comes at times $t_n = n\Delta t$ with a short observation interval Δt. We can approximate the increment at t_n in the observation data from the linear interpolation for $t \in \left(t_n, t_{n+1}\right]$

$$\mathrm{d}y(t) \cong \Delta y_n = y_{n+1} - y_n. \tag{2.14}$$

Furthermore, even though consecutive observations are required in a short interval for the estimate of Δy_n, we may not require to have continuous observation data at each time updating step. The updating step of data assimilation in the second line of the equation (2.12) can be applied only at the steps when observation data Δy_n is received.

Finally, the development of effective filtering scheme relies on the construction of the Kalman gain operators (K^m, K^v) and the drift terms (a^m, a^v) in the second line of the filter update in (2.12) based on the observation functions H^m, H^v. In general, these terms should be chosen so that the probability distribution $\tilde{\rho}$ of the constructed stochastic $\tilde{Z}^i$ can correctly reflect the optimal filter solution $\hat{\rho}$ in (2.11). From the standard procedure of developing the mean field evolution equations [27,29], they can be solved by the following equations according to the probability distribution $\tilde{\rho}$ of $\tilde{Z}$

$$-\nabla \cdot \left(K^T \tilde{\rho}\right) = \tilde{\rho}\Gamma^{-2}\left[H\left(\tilde{Z}\right) - \tilde{\mathbb{E}}(H)\right], \quad a = \nabla \cdot \left(K\Gamma^2 K^T\right) - K\Gamma^2 \nabla \cdot K^T. \tag{2.15}$$

Detailed analysis in [17] shows that the high-order filtering Eq. (2.12) generates consistent statistics with the optimal filter solution (2.11) in the analysis update if the above conditions (2.15) are satisfied. Importantly, high-order statistics according to the nonlinear observation operators H in (2.8), $\tilde{\mathbb{E}}\left[H\left(\tilde{Z}\right)\right] = \mathcal{H}\hat{\rho}$, are preserved in the filter updating equation. This indicates that the new filtering model for $\tilde{Z}$ is able to capture the crucial high-order statistics in the optimal filter solution $\hat{\rho}$ rather than only the first two moments in the linear ensemble Kalman filters. However, efficient ways to compute the Kalman gain and drift operator through (2.15) are still needed without losing the essential high-order moments information. In the next section, we will propose an easy-to-implement scheme to compute these key filter operators K and a without the need to solve the distribution function $\tilde{\rho}$.

3. Ensemble data assimilation schemes maintaining high-order statistics

In this section, we construct a practical numerical scheme for implementing the ensemble filtering Eq. (2.12). The goal is to generate a better empirical representation (2.4) of the probability distribution using only a small ensemble size. The accurate computation of the model statistics requires that the non-Gaussian statistics involved in the observation function H^m and H^v are properly represented through the filtering update. This leads to several key treatments in the construction of the filtering operators. In particular, we exploit the detailed structures of the observation functions to derive explicit expressions for the functions a^m, K^m and a^v, K^v according to the mean and covariance observations in the filter equation.

3.1. Construction of explicit filter operators with nonlinear observation functions

Assuming that the first s components of the mean and covariance are observed, we can derive the filter operators (a, K) by exploiting the specific quadratic and cubic structures of the observation functions $H^m \in \mathbb{R}^s$ and $H^v \in \mathbb{R}^{s\times s}$ from (2.8) for all the observed modes $1 \le k, l \le s$

$$\begin{aligned} H_k^m(z) &= \sum_{p,q} \gamma_{kpq} z_p z_q, \\ H_{kl}^v(z) &= \sum_{p,q} \gamma_{kpq} z_p z_q z_l + \gamma_{lpq} z_p z_q z_k, \end{aligned} \tag{3.1}$$

where we define the coefficient $\gamma_{kpq} = \hat{v}_k \cdot B\left(\hat{v}_p, \hat{v}_q\right)$. The following property of the observation functions H^m, H^v can be found from direct computation using the above quadratic and cubic structures.

Lemma 1. *The observation functions H^m and H^v defined in (3.1) satisfy the relation*

$$\begin{aligned} z \cdot \nabla H^m(z) &= \sum_j z_j \partial_j H^m = 2H^m(z), \\ z \cdot \nabla H^v(z) &= \sum_j z_j \partial_j H^v = 3H^v(z). \end{aligned} \tag{3.2}$$

With the above symmetry in the observation functions (3.2), we are able to find explicit expressions for the Kalman gain and drift operators that enable efficient computation of these terms. In the following, we summarize the useful results and put detailed proofs and derivations of the formulas in Appendix A.

3.1.1. The explicit forms of the Kalman gain and drift operators

With the above explicit relations for the observation functions, we can first find special solutions to the Eq. (2.15) to recover the Kalman gain operator $K(z;\tilde{\rho})$, that is,

$$-\nabla\cdot\left(K^T\tilde{\rho}\right)=\tilde{\rho}\Gamma^{-2}\left[H\left(\tilde{Z}\right)-\bar{H}\right],$$

where $H=H^m$ or H^v is the stretched vector and $\bar{H}=\tilde{\mathbb{E}}\left[H\left(\tilde{Z}\right)\right]$. Still, we would like to avoid directly solving the above equation since the probability distribution $\tilde{\rho}$ is usually intractable and can be only estimated from an ensemble approach. By multiplying H on both sides, the identity for K implies a necessary condition

$$\tilde{\mathbb{E}}\left[K^T\nabla H\right]=\Gamma^{-2}C^H, \tag{3.3}$$

where $C^H=\tilde{\mathbb{E}}\left[\left(H-\bar{H}\right)\left(H-\bar{H}\right)^T\right]$ is the second-order moment of H with respect to $\tilde{\rho}$. Then we may solve instead the above equation to find proper candidate for the gain function K_t. In this way, we can compute the detailed expressions for the Kalman gain operators based on the explicit forms and their particular symmetry in the observation functions (3.2).

Proposition 2. *The following expressions for Kalman gain operators $K^m\in\mathbb{R}^{d\times s}$ and $K^v\in\mathbb{R}^{d\times s^2}$ give solutions to the Eq. (3.3) corresponding to observation functions H^m and H^v*

$$\begin{aligned}K^m(Z)&=\frac{1}{2}Z\left[H^m(Z)-\bar{H}^m\right]^T\Gamma_m^{-2},\\ K^v(Z)&=\frac{1}{3}Z\left[H^v(Z)-\bar{H}^v\right]^T\Gamma_v^{-2},\end{aligned} \tag{3.4}$$

with the state vector $Z\in\mathbb{R}^d$, observation noises $\Gamma_m\in\mathbb{R}^{s\times s}_{\mathrm{sym}}$, $\Gamma_v\in\mathbb{R}^{s^2\times s^2}_{\mathrm{sym}}$, and $\bar{H}=\tilde{\mathbb{E}}[H(Z)]$.

Next, the function $a(z;\tilde{\rho})$ can be directly solved using the explicit expressions of K in (3.4) according to

$$a=\nabla\cdot\left(K\Gamma^2K^T\right)-K\Gamma^2\nabla\cdot\left(K^T\right). \tag{3.5}$$

We can also find the explicit forms of the drift terms through direct computation using (3.5) and again the specific structures in observation functions (3.2).

Proposition 3. *With the solutions of $K^m$and K^v in (3.4), the corresponding drift terms satisfying (3.5) can be found as*

$$\begin{aligned}a^m(Z)&=\frac{1}{4}Z\left[H^m(Z)-\bar{H}^m\right]^T\Gamma_m^{-2}\left[3H^m(Z)-\bar{H}^m\right],\\ a^v(Z)&=\frac{1}{9}Z\left[H^v(Z)-\bar{H}^v\right]^T\Gamma_v^{-2}\left[4H^v(Z)-\bar{H}^v\right].\end{aligned} \tag{3.6}$$

Notice that the solution to (3.3) only satisfies a necessary condition for the original equation for the Kalman gain. Still, it already accounts for the crucial high-order statistics with respect to $\tilde{\rho}$ involving in the nonlinear observation functions (3.1). Therefore, the achieved explicit forms of Kalman gain and drift terms (3.4) and (3.6) can serve as suitable candidate for the construction of high-order filtering schemes. It shows to be a better choice than that in the standard EnKF scheme (shown next in Section 3.2 and the numerical comparisons in Section 4) which only considers Gaussian projection of $\tilde{\rho}$ thus neglects the crucial high-order statistics information in H.

3.1.2. Numerical implementation of the filter operators

Based on the explicit expressions of the filtering operators in (3.4) and (3.6), we are able to construct direct algorithms for effective implementation of the filtering scheme. At each time updating step t_n, the mean and covariance can be computed by integrating the statistical Eq. (2.5)

$$\begin{aligned}\Delta\bar{u}_n^N&=\bar{u}_{n+1}^N-\bar{u}_n^N=\int_{t_n}^{t_{n+1}}\left[\mathbb{E}^N H^m\left(\tilde{\mathbf{Z}}(s)\right)+h_m\left(\bar{u}^N(s)\right)\right]\mathrm{d}s,\\ \Delta R_n^N&=R_{n+1}^N-R_n^N=\int_{t_n}^{t_{n+1}}\left[\mathbb{E}^N H^v\left(\tilde{\mathbf{Z}}(s)\right)+h_v\left(\bar{u}^N(s),R^N(s)\right)\right]\mathrm{d}s.\end{aligned} \tag{3.7}$$

Above, the empirical expectation $\mathbb{E}^N(\cdot)$ is computed as in (2.4) using the ensemble average of all the simulated samples from the filter Eq. (2.12). Then the filter updating step combines the observation data $(\Delta\bar{u},\Delta R)$ in (2.14) and the model forecast (3.7) to find an optimal estimate for the ensemble distribution for $\tilde{Z}^i$ in the following two-step updating procedure

$$\begin{aligned}\tilde{Z}_{n+1}^i&=\hat{Z}_{n+1}^i+(a^m\Delta t+K^m\Delta I^m)+(a^v\Delta t+K^v\Delta I^v),\\ \hat{Z}_{n+1}^i&=\tilde{Z}_n^i+L\left(\bar{u}_n^N\right)\tilde{Z}_n^i\Delta t+Q_v\left(\tilde{Z}_n^i\otimes\tilde{Z}_n^i-R^N\right)\Delta t+\sigma\Delta\tilde{W}_n^i.\end{aligned} \tag{3.8}$$

Above, we split the filtering procedure in the standard two-step process, where $\hat{Z}^i$ gets the forecast step update for the stochastic state then the prior forecast is corrected through the filtering operators when the observation data is available. The following proposition provides the explicit expressions for directly computing the filter updates using the samples and observation data.

Proposition 4. *Given the observations* $(\Delta\bar{u}, \Delta R)$ *and the model predicted increments* $\left(\Delta\bar{u}^N, R^N\right)$*, the filter update in the filtering Eq. (3.8) can be computed directly based on the samples* $\tilde{Z}^i$ *as*

$$\begin{aligned} a^m\Delta t + K^m\Delta I^m &= \frac{1}{2}\left[\tilde{Z}^i H_m'^T\left(\tilde{Z}^i\right)\Gamma_m^{-2}\right]\left(\Delta\bar{u} - \Delta\bar{u}^N\right) \\ &\quad + \frac{\Delta t}{2}\left[\tilde{Z}^i H_m'^T\left(\tilde{Z}^i\right)\Gamma_m^{-2}\right]\bar{H}_m + \frac{\Delta t}{4}\left[\tilde{Z}^i H_m'^T\left(\tilde{Z}^i\right)\Gamma_m^{-2}H_m'\left(\tilde{Z}^i\right)\right], \\ a^v\Delta t + K^v\Delta I^v &= \frac{1}{3}\left[\tilde{Z}^i H_v'^T\left(\tilde{Z}^i\right)\Gamma_v^{-2}\right]\left(\Delta R - \Delta R^N\right) \\ &\quad + \frac{\Delta t}{3}\left[\tilde{Z}^i H_v'^T\left(\tilde{Z}^i\right)\Gamma_v^{-2}\right]\bar{H}_v + \frac{\Delta t}{9}\left[\tilde{Z}^i H_v'^T\left(\tilde{Z}^i\right)\Gamma_v^{-2}H_v'\left(\tilde{Z}^i\right)\right]. \end{aligned} \tag{3.9}$$

with $\bar{H} = \mathbb{E}^N\left[H\left(\tilde{\mathbf{Z}}\right)\right]$ *and* $H' = H\left(\tilde{Z}\right) - \bar{H}$.

Using the explicit formulas in (3.8), we can directly update each filtering sample $\tilde{Z}^i$ during the time updating interval containing the higher-order moments information in the observation functions H^m and H^v. Notice that the most expensive part in computing the filter update is the observation functions $H^m \in \mathbb{R}^s$ and $H^v \in \mathbb{R}^{s\times s}$ in (3.1), where $s \le d$ is the size of the observed modes. Luckily, these two functions are already computed in solving the forecast Eq. (2.5). The additional computational cost for solving the Kalman gain and drift operators in (3.9) thus can be estimated in the order $O\left(N(s+s^2)\right)$. Therefore, no large computational demand will be required from the additional filtering update step and the algorithm scales well with dimension as long as the number of particles N is kept in moderate size. To summarize, the dynamical equations for the particles are coupled through the empirical average among all the samples according to Algorithm 1.

Algorithm 1 Ensemble probability filter with statistical observations.

Model Setup: Given the discrete time step Δt with $M\Delta t = T$, the sequence of statistical observations are generated by the increments of the mean and covariance $\Delta y_n = \left\{\Delta\bar{u}_n, \Delta R_n\right\}$ measured at time instants $t_n = n\Delta t$.

Initial condition: At initial time $t = 0$, draw an ensemble of samples $\left\{\tilde{Z}_0^i\right\}_{i=1}^N$ from the initial distribution $\tilde{\rho}_0$.

1: **for** $n = 0$ while $n < M$, during the time updating interval $t \in \left[t_n, t_{n+1}\right]$. **do**
2: Integrate the samples to the next time step $\left\{\hat{Z}_{n+1}^i\right\}$ using the forecast model given by the second equation in (3.8).
3: Integrate the statistical mean and covariance to the next time step $\left\{\bar{u}_{n+1}^N, R_{n+1}^N\right\}$ by (3.7) using the average of all samples.
4: Compute the filtering update terms using the explicit formulas in (3.9) and the observation data Δy_n.
5: Update the samples $\left\{\tilde{Z}_{n+1}^i\right\}$ from the prior states $\left\{\hat{Z}_{n+1}^i\right\}$ using the first equation in (3.8).
6: **end for**

Remark 1. 1. In practical implementations, it is observed that some outliers of the samples $\tilde{Z}^i$ may occasional introduce large errors by creating some extremely large values in the high-order terms in (3.9). To improve stability in the highly unstable regime, it is found useful to use the expectation values $\mathbb{E}^N\left[\tilde{Z}H'^T\left(\tilde{Z}\right)\Gamma^{-2}\right]$ and $\mathbb{E}^N\left[\tilde{Z}H'^T\left(\tilde{Z}\right)\Gamma^{-2}H'\left(\tilde{Z}\right)\right]$ instead of each ensemble evaluation to improve filter stability without sacrificing too much accuracy.

2. High computational cost may become a major issue in solving realistic problems involving a high-dimensional SDE (2.12). This difficulty could be mitigated by adopting reduced-order and data-driven algorithms such as the random batch methods [15,48]. We aim to combine the efficient forecast models with the data assimilation strategy in high-dimensional problems in the following-up research.

3.2. Comparison with ensemble Kalman filter schemes

For comparison, we also describe the strategy commonly used in ensemble Kalman filters [27]. Assuming that the Kalman gain K is a deterministic matrix with no randomness, this leads to the following choice of the deterministic Kalman gain matrix from the overall moment of $\tilde{Z}$, and a zero drift term due to the constant Kalman gain according to the Eq. (2.15)

$$K = \tilde{\mathbb{E}}\left[\tilde{Z}\left(H\left(\tilde{Z}\right) - \bar{H}\right)^T\right]\Gamma^{-2} = \tilde{C}^{ZH}\Gamma^{-2}, \quad \text{and} \quad a = 0. \tag{3.10}$$

where the covariance matrix $\tilde{C}^{ZH}$ is given by the cross-covariance between the process $\tilde{Z}$ and the observation function $H\left(\tilde{Z}\right)$. The ensemble Kalman filter scheme then yields the following filter equations

$$\begin{aligned} \tilde{Z}_{n+1}^i &= \tilde{Z}_n^i + L\left(\bar{u}_n^N\right)\tilde{Z}_n^i\Delta t + Q_v\left(\tilde{Z}_n^i \otimes \tilde{Z}_n^i - R^N\right)\Delta t + \sigma\Delta\tilde{W}_n^i \\ &\quad + \tilde{C}^{ZH^m}\Gamma_m^{-2}\left\{\Delta\bar{u} - \left[H^m\left(\tilde{Z}_n^i\right) + h_m\right]\Delta t\right\} + \tilde{C}^{ZH^v}\Gamma_v^{-2}\left\{\Delta R - \left[H^v\left(\tilde{Z}_n^i\right) + h_v\right]\Delta t\right\}, \end{aligned} \tag{3.11}$$

where the covariance $\hat{C}^{ZH} = \tilde{\mathbb{E}}^N\left[\hat{\mathbf{Z}}H'\left(\hat{\mathbf{Z}}\right)^T\right]$ is computed according to the empirical average among all the sample forecast $\tilde{\mathbf{Z}}_{n+1}$ from the second equation of (3.8). The above updating scheme (3.11) is usually referred to as the ensemble Kalman filter (EnKF). It has been shown that the EnKF approach can effectively drive the probability density functions to the equilibrium such as using the ensemble Fokker-Planck filter [13].

However, in our modeling framework consisting of the coupled stochastic-statistical equations, the high-order moments are playing a central role as the high-order feedbacks in the statistical Eq. (2.2a) for accurate statistical prediction. Notice that in the EnKF approach, constant Kalman gains K_m and K_v matrices are used independent of each sample realization. It adopts the Gaussian projection on the stochastic process Z thus only statistics up to the second-order moments are considered in the filter update. As a result, this approximation deliberately neglected the crucial high-order statistics contained in the observation functions H^m and H^v. Compared with the more precisely calibrated filter operators (3.9), directly applying the EnKF in the coupled stochastic-statistical model (2.5) may miss the crucial high-order statistical information in the sampled observation $H(\tilde{Z}^i)$ thus lead to larger errors and instability in the filter updates. The degeneracy of particles failing to capture the essential high-order information in the EnKF prediction is demonstrated in Figs. 10 and 11 from direct numerical tests.

3.3. Convergence of the statistical ensemble filter approximation

In this final part, we discuss the convergence of the discrete numerical scheme for the statistical estimates using finite ensemble approximation. Let $\tilde{\rho}^N(z,t) = \frac{1}{N}\sum_{i=1}^N \delta(z - \tilde{Z}^i(t))$ be the random field from the finite ensemble approximation of the N stochastic samples in (3.8), and $\tilde{\rho}(z,t)$ is the corresponding continuous distribution of the $\mathcal{P}(\mathbb{R}^d)$-valued random field from the law of each stochastic process $\tilde{Z}$ in (2.12) conditional on the observations $\mathcal{G}_t$.

First, assume that the initial samples $\{\tilde{Z}_0^i\}_{i=1}^N$ are drawn i.i.d. from the initial distribution $\tilde{\rho}_0$, and a unique solution exists for the McKean-Vlasov SDE (2.12) for each sample $\tilde{Z}^i \sim \tilde{\rho}$. From established conclusions from the limit of the N interacting particle system [44,49], the empirical probability distribution $\tilde{\rho}^N$ estimated with finite samples will converge to the continuous measure $\tilde{\rho}$ as the ensemble size $N \to \infty$. In addition, we will need the following assumptions on the structures of the coupled stochastic-statistical Eq. (2.2).

Assumption 5. The model structure functions $B : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and $L : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ in the mean and covariance equations (2.2a) are Lipschitz continuous, that is, there is a constant $\beta > 0$ so that

$$|B(u,u) - B(v,v)| \le \beta|u - v|, \quad \|L(u) - L(v)\| \le \beta|u - v|.$$

In addition, the nonlinear coupling coefficients $\gamma_{kmn} = \hat{v}_k \cdot B(\hat{v}_m, \hat{v}_n)$ in (2.2b) are uniformly bounded, that is, there exists a constant $C > 0$, so that for all k, m, n

$$|\gamma_{kmn}| \le C.$$

Given the observations $\mathcal{G}_t$, we have for any test function $\varphi \in C_b^2(\mathbb{R}^d)$ the empirical measure $\tilde{\rho}^N$ converges to the continuous distribution for each sample $\tilde{\rho}$ in the sense

$$\langle \tilde{\rho}^N, \varphi \rangle = \frac{1}{N}\sum_{i=1}^N \varphi(\tilde{Z}^i) \to \mathbb{E}[\varphi(\tilde{Z}) \mid \mathcal{G}_t] = \langle \tilde{\rho}, \varphi \rangle, \tag{3.12}$$

a.s. as $N \to \infty$. Furthermore, there is the error estimate for the empirical estimate $\langle \tilde{\rho}_t^N, \varphi \rangle := \mathbb{E}^N \varphi = \frac{1}{N}\sum \varphi(\tilde{Z}^i(t))$ and $\langle \tilde{\rho}_t, \varphi \rangle := \tilde{\mathbb{E}}[\varphi(\tilde{Z}(t))]$ for $T > 0$

$$\mathbb{E}\left[\sup_{0\le t\le T} \left|\langle \tilde{\rho}_t^N, \varphi \rangle - \langle \tilde{\rho}_t, \varphi \rangle\right|^2\right] \le \frac{C_T}{N}\|\varphi\|_\infty^2. \tag{3.13}$$

Proofs on (3.12) and (3.13) follow directly from the law of large numbers [see, for example in 50, Thm. 9.18].

Next, we consider the finite ensemble and discrete time estimation of the statistical mean and covariance states in the data assimilation model. The statistical equations for the continuous solutions $(\bar{u}, R)$ can be written based on their coupling dynamics in (2.2a) and the high-order terms according to the observation functions H^m, H^v in (2.8) with respect to the continuous probability $\tilde{Z} \sim \tilde{\rho}$

$$\begin{aligned} \frac{d\bar{u}}{dt} &= \Lambda\bar{u}(t) + B(\bar{u}(t),\bar{u}(t)) + F + \tilde{\mathbb{E}}H^m(\tilde{Z}(t)), \\ \frac{dR}{dt} &= L(\bar{u}(t))R(t) + R(t)L(\bar{u}(t))^T + Q_\sigma + \tilde{\mathbb{E}}H^v(\tilde{Z}(t)). \end{aligned} \tag{3.14}$$

On the other hand, the numerical mean and covariance estimates $(\bar{u}^{N,\delta}, R^{N,\delta})$ from the discrete time numerical updates and with the ensemble approximation $\{\tilde{Z}^i\}$ are computed from the equations (3.7) with respect to the discrete empirical distribution $\tilde{\rho}^N$

$$\begin{aligned} \frac{d\bar{u}^{N,\delta}}{dt} &= \Lambda\bar{u}^{N,\delta}(\tau(t)) + B(\bar{u}^{N,\delta}(\tau(t)),\bar{u}^{N,\delta}(\tau(t))) + F + \frac{1}{N}\sum_{i=1}^N H^m(\tilde{Z}^i(\tau(t))), \\ \frac{dR^{N,\delta}}{dt} &= L(\bar{u}^{N,\delta}(\tau(t)))R^{N,\delta}(\tau(t)) + R^{N,\delta}(\tau(t))L(\bar{u}^{N,\delta}(\tau(t)))^T + Q_\sigma + \frac{1}{N}\sum_{i=1}^N H^v(\tilde{Z}^i(\tau(t))), \end{aligned} \tag{3.15}$$

where the forward Euler scheme is adopted here with the discrete time update using a constant $\tau(t) = n\Delta t$ during the time interval $t \in [t_n, t_{n+1}]$. Above in (3.14) and (3.15), we neglect the last relaxation term with ϵ since it will automatically vanish with the resulting consistency. Notice that $\bar{u}^{N,\delta}, R^{N,\delta}$ and $\bar{u}, R$ are stochastic processes due to the random samples $\{\tilde{Z}^i\}$ and the conditional expectation dependent on the observations $\mathcal{G}_t$. We have the following result for the convergence of finite ensemble N and finite time step Δt approximation to the continuous model prediction.

Theorem 6. *If Assumption 5 is satisfied and under the same initial condition, the statistical solution* $\left(\bar{u}_n^{N,\delta}, R_n^{N,\delta}\right) = \left(\bar{u}^{N,\delta}(t_n), R^{N,\delta}(t_n)\right)$ *of the finite ensemble model (3.15) with discrete time step* Δt *converges to the true statistical solution* $(\bar{u}_n, R_n) = \left(\bar{u}(t_n), R(t_n)\right)$ *of the continuous model (3.14) with the error estimates*

$$\begin{aligned}
\mathbb{E}\left[\sup_{n\Delta t\le T}\left|\bar{u}_n^{N,\delta}-\bar{u}_n\right|^2\right] &\le \left(C_{1,T}\Delta t+\frac{C_{2,T}}{N}\right)\|H^m\|_\infty,\\
\mathbb{E}\left[\sup_{n\Delta t\le T}\left\|R_n^{N,\delta}-R_n\right\|^2\right] &\le \left(C'_{1,T}\Delta t+\frac{C'_{2,T}}{N}\right)\left(\|H^m\|_\infty+\|H^v\|_\infty\right),
\end{aligned} \tag{3.16}$$

where $C_{1,T}, C_{2,T}, C'_{1,T}, C'_{2,T}$ *are constants depending on the final time* T.

Proof. First, considering the mean equations in (3.14) and (3.15) from the same initial state, we have

$$\bar{u}^{N,\delta}(t)-\bar{u}(t)=\int_0^t\left[M\left(\bar{u}^{N,\delta}(\tau(s))\right)-M(\bar{u}(s))\right]\mathrm{d}s+\int_0^t\left[\left\langle\tilde{\rho}^N_{\tau(s)},H^m\right\rangle-\langle\tilde{\rho}_s,H^m\rangle\right]\mathrm{d}s,$$

where we define $M(u)=\Lambda u+B(u,u)+F$ and assume that the forcing F and Q_σ are constants for simplicity. Using the Lipschitz condition for M from Assumption 5 and applying Cauchy-Schwarz inequality, there is

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\le T}\left|\bar{u}^N(t)-\bar{u}(t)\right|^2\right] &\le 2T\beta^2\mathbb{E}\int_0^T\left|\bar{u}^{N,\delta}(\tau(s))-\bar{u}(s)\right|^2\mathrm{d}s+2T\mathbb{E}\int_0^T\left|\left\langle\tilde{\rho}^N_{\tau(s)},H^m\right\rangle-\langle\tilde{\rho}_s,H^m\rangle\right|^2\mathrm{d}s\\
&\le C_1T\int_0^T\mathbb{E}\left[\sup_{s'\le s}\left|\bar{u}^{N,\delta}(s')-\bar{u}(s')\right|^2\right]\mathrm{d}s+C_2T^2\Delta t\|H^m\|^2_\infty+\frac{C_3T^2}{N}\|H^m\|^2_\infty.
\end{aligned} \tag{3.17}$$

Above in the first term of the last inequality, we estimate the error by comparing the discretized time solution $\bar{u}^{N,\delta}(\tau(t))$ following (3.15) with the corresponding continuous time solution $\bar{u}^{N,\delta}(t)$

$$\left|\bar{u}^{N,\delta}(\tau(t))-\bar{u}^{N,\delta}(t)\right|^2\le|t-\tau(t)|^2\left[\left|M\left(\bar{u}^{N,\delta}(\tau(s))\right)\right|^2+\left|\left\langle\tilde{\rho}^N_{\tau(t)},H^m\right\rangle\right|^2\right]\le\Delta t^2\left(\|M\|^2_\infty+\|H^m\|^2_\infty\right).$$

Thus the error estimation follows

$$\begin{aligned}
\left|\bar{u}^{N,\delta}(\tau(s))-\bar{u}(s)\right|^2 &\le 2\left|\bar{u}^{N,\delta}(\tau(s))-\bar{u}^{N,\delta}(s)\right|^2+2\left|\bar{u}^{N,\delta}(s)-\bar{u}(s)\right|^2\\
&\le C\Delta t^2\|H^m\|^2_\infty+2\sup_{s'\le s}\left|\bar{u}^{N,\delta}(s')-\bar{u}(s')\right|^2.
\end{aligned}$$

And for the second term involving expectation of H^m, the convergence of the empirical measure (3.13) gives

$$\begin{aligned}
\mathbb{E}\left[\left|\left\langle\tilde{\rho}^N_{\tau(t)},H^m\right\rangle-\langle\tilde{\rho}_t,H^m\rangle\right|^2\right] &\le 2\mathbb{E}\left[\left|\left\langle\tilde{\rho}^N_{\tau(t)},H^m\right\rangle-\left\langle\tilde{\rho}^N_t,H^m\right\rangle\right|^2\right]+2\mathbb{E}\left[\left|\left\langle\tilde{\rho}^N_t,H^m\right\rangle-\langle\tilde{\rho}_t,H^m\rangle\right|^2\right]\\
&\le\left(C\Delta t+\frac{C_T}{N}\right)\|H^m\|^2_\infty.
\end{aligned}$$

Finally applying Grönwall's inequality to (3.17), we get the mean state estimate in (3.16).

Next, under a similar fashion, we have for the covariance equation

$$\begin{aligned}
R^{N,\delta}(t)-R(t)=&\int_0^t L(\bar{u}(s))\left[R^{N,\delta}(\tau(s))-R(s)\right]\mathrm{d}s+\int_0^t\left[L\left(\bar{u}^{N,\delta}(\tau(s))\right)-L(\bar{u}(s))\right]R(s)\mathrm{d}s\\
&+\int_0^t\left[L\left(\bar{u}^{N,\delta}(\tau(s))\right)-L(\bar{u}(s))\right]\left[R^{N,\delta}(\tau(s))-R(s)\right]\mathrm{d}s+c.c.\\
&+\int_0^t\left[\left\langle\tilde{\rho}^N_{\tau(s)},H^v\right\rangle-\langle\tilde{\rho}_s,H^v\rangle\right]\mathrm{d}s.
\end{aligned}$$

Above, $c.c.$ represents the symmetric terms from the transposes $RL(\bar{u})^T$. Again, using the Lipschitz condition of L in Assumption 5, $\|L(u)\|\le\beta|u|+\beta_1$, we can compute errors from the covariance equation

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\le T}\left\|R^{N,\delta}(t)-R(t)\right\|^2\right] \le{}& C_1T\beta^2\mathbb{E}\left[\sup_{t\le T}|\bar{u}|^2\int_0^T\left\|R^{N,\delta}(\tau(s))-R(s)\right\|^2\mathrm{d}s\right]\\
&+C_2T\beta^2\mathbb{E}\left[\sup_{t\le T}\left|\bar{u}^{N,\delta}(t)-\bar{u}(t)\right|^2\sup_{t\le T}\|R(t)\|^2\right]\\
&+C_3\mathbb{E}\int_0^T\left|\left\langle\tilde{\rho}^N_{\tau(s)},H^v\right\rangle-\langle\tilde{\rho}_s,H^v\rangle\right|^2\mathrm{d}s+C_T\Delta t^2.
\end{aligned}$$

Using the uniform boundedness of $\bar{u}, R$ and (3.13) for H^v together with the previous error estimate of the mean state for $\mathbb{E}\left[\sup_{t\le T}\left|\bar{u}^{N,\delta}(t)-\bar{u}(t)\right|^2\right]$, we reach the final covariance error estimate in (3.16).

Theorem 6 guarantees that the discrete numerical scheme of the approximating ensemble data assimilation model can recover the leading-order statistics in mean and covariance. It implies that the performance of the ensemble filter estimation relies on the

accurate approximations of the expectation of the observation functions H^m, H^v. Usually, the higher-order moments in H^m and H^v in (3.1) become extremely difficult to capture with a small sample size. This leads to the rapidly growing model errors (as in Fig. 3 shown in the numerical tests). On the other hand, the design of the new high-order filtering scheme guarantees consistent statistics $\tilde{\mathbb{E}}\left[H\left(\tilde{Z}_t\right)\right]$ in the observation functions with the optimal filter solution $\hat{\rho}$ (see Theorem 7 in [17]). This leads to the much improved performance of the statistical forecasts using the new filter model.

4. Numerical performance on the prototype triad system

Using the explicit filtering Eq. (3.9), we first test the performance of the proposed data assimilation algorithm on a prototype triad system with instructive implications to many practical applications. The triad system is given by a three-dimensional ODE system [14] for the state $\mathbf{u} = \left(u_1, u_2, u_3\right)^T$ with both linear and nonlinear coupling combined with stochastic forcing

$$\begin{aligned}
\frac{\mathrm{d}u_1}{\mathrm{d}t} &= \lambda_2 u_3 - \lambda_3 u_2 - d_1 u_1 + B_1 u_2 u_3 + \sigma_1 \dot{W}_1, \\
\frac{\mathrm{d}u_2}{\mathrm{d}t} &= \lambda_3 u_1 - \lambda_1 u_3 - d_2 u_2 + B_2 u_3 u_1 + \sigma_2 \dot{W}_2, \\
\frac{\mathrm{d}u_3}{\mathrm{d}t} &= \lambda_1 u_2 - \lambda_2 u_1 - d_3 u_3 + B_3 u_1 u_2 + \sigma_3 \dot{W}_3.
\end{aligned} \tag{4.1}$$

It can be seen that the above triad system (4.1) fits into our general formulation (1.1), where the model coefficients defined by

$$\Lambda = \begin{bmatrix} -d_1 & -\lambda_3 & \lambda_2 \\ \lambda_3 & -d_2 & -\lambda_1 \\ -\lambda_2 & \lambda_1 & -d_3 \end{bmatrix}, \quad B(\mathbf{u}, \mathbf{u}) = \begin{bmatrix} B_1 u_2 u_3 \\ B_2 u_3 u_1 \\ B_3 u_1 u_2 \end{bmatrix},$$

contain the linear skew-symmetric (off-diagonal) and dissipation (diagonal) operator Λ, together with the nonlinear quadratic coupling $B(\mathbf{u}, \mathbf{u})$ satisfying energy conservation with $B_1 + B_2 + B_3 = 0$. The triad system can serve as an elementary building block of many more general turbulent systems emphasizing the key energy conserving nonlinear interactions. Though low-dimensional, this system can demonstrate a wide variety of different statistical regimes (as shown next in Fig. 2), making it a nice first test model for a thorough study of the prediction skill of the proposal ensemble data assimilation strategy in dealing with different statistical features.

4.1. Typical statistical regimes in the triad system

One attractive feature of the triad system (4.1) as a prototype test model is that it is able to generate a wide variety of dynamical regimes demonstrating distinctive statistical features ranging from near-Gaussian to highly non-Gaussian probability distributions. This sets up a desirable testbed for examining the skill of different statistical prediction methods in dealing with vastly different statistical dynamics.

4.1.1. Statistical regimes with distinctive statistical features

The triad systems (4.1) constitutes the generic linear and nonlinear coupling mechanism between any three modes in larger systems with quadratic nonlinearity. A direct three-dimensional Galerkin truncation of many complex turbulent dynamics possesses the energy-conserving nonlinearity as in the general formulation (1.1). For example, a direct link can be built to interpret the triad system as a prototype three-mode interaction with forward and backward energy cascades in geophysical turbulence (see Appendix B.1). The random forcing together with the damping term simulates the inhomogeneous effects of the interaction with other modes that are not resolved in the projected three dimensional subspace. Thus, the stochastic triad system can serve as a qualitative model for a wide variety of turbulent phenomena regarding energy exchange and cascades and supply important intuition for many phenomena [1,51,52]. They also provide elementary test models with subtle features for prediction and uncertainty quantification. Additional dynamical and statistical properties of the triad system are summarized in Appendix B.2 showing an explicit equilibrium invariant measure and detailed nonlinear energy exchanging mechanism.

In our testing cases, we consider the following three typical dynamical regimes of the triad system (4.1) containing representative statistical structures. Model parameters used for the three test regimes are listed in Table 2.

- *Regime I: Near-Gaussian regime with equipartition of energy.* This regime considers the convergence to a Gaussian equilibrium distribution with the competition of linear and nonlinear effects. The equipartition of energy, that is, $\frac{\sigma_1^2}{2d_1} = \frac{\sigma_2^2}{2d_2} = \frac{\sigma_3^2}{2d_3} = \sigma_{\mathrm{eq}}^2$, is designed so that a Gaussian distribution, $p_{\mathrm{eq}} \sim \exp\left(-\frac{1}{2}\sigma_{\mathrm{eq}}^{-2}|\mathbf{u}|^2\right)$, will be reached at the final equilibrium state. The linear and nonlinear parameters are chosen to have comparable values in this case to induce strong interactions during the transient state;
- *Regime II: Nonlinear regime with forward energy cascade.* This regime focuses on strong quadratic coupling with weak linear damping and forcing effects. Skew-symmetric linear terms are set to be zero and only small damping and noise effects are added. The first mode u_1 is set to have large initial mean and covariance while the other two modes u_2, u_3 only have small initial values. This induces strong energy cascades from u_1 to the other two less energetic modes u_2, u_3 driven by the dominant nonlinear coupling;
- *Regime III: Unstable regime with dual energy cascades.* This regime is used to simulate the inherent internal instability observed in turbulent systems. The instability is introduced by a negative damping $d_1 = -0.4$ in the first mode u_1, while the other two modes u_2, u_3 are stable with positive damping. On the other hand, the first mode is weakly forced by stochastic forcing while the other

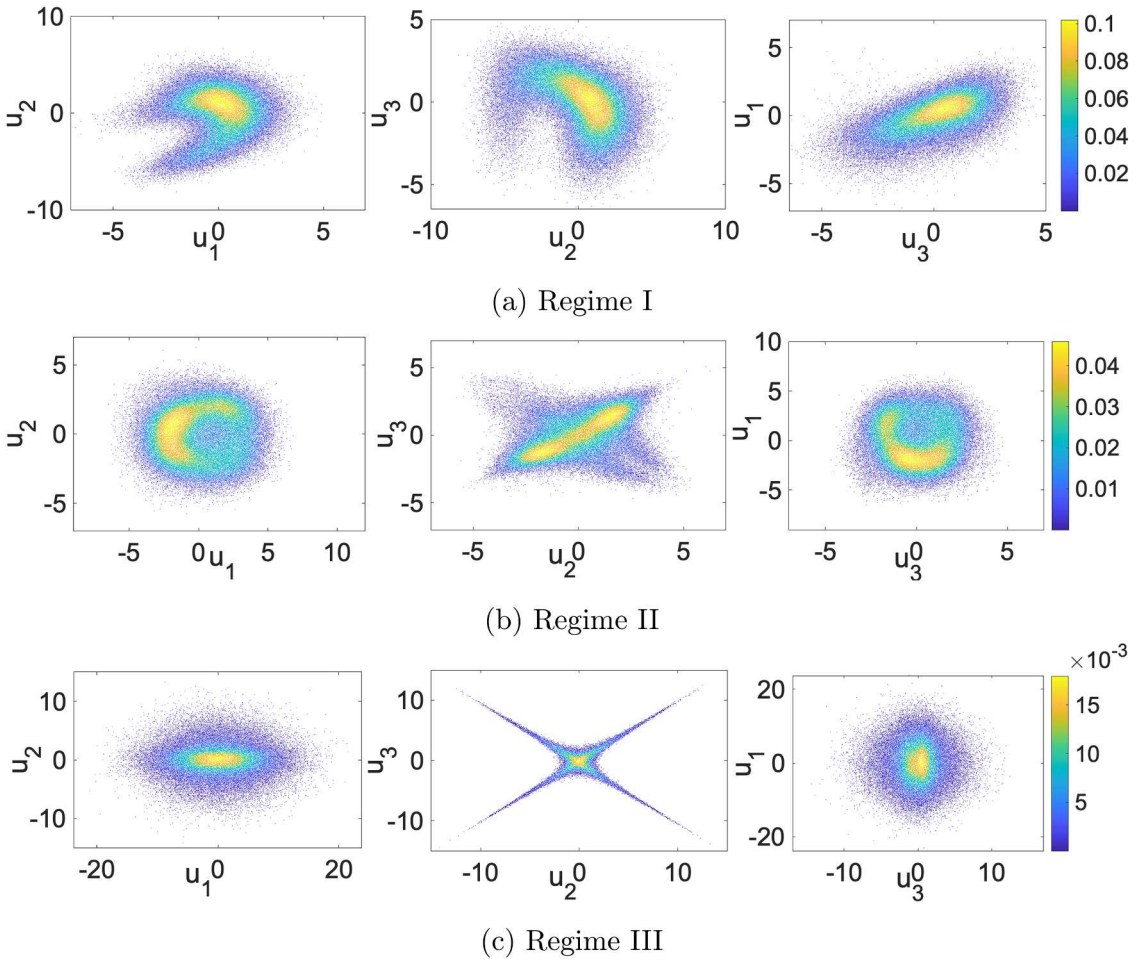


Fig. 2. Joint PDFs of the triad modes u_1, u_2, u_3 at $t = 5$ in the three test regimes shown in scatter plots from a direct MC simulation using MC $= 1 \times 10^5$ samples. The density of particles is represented by colors in the scatter plots.

Table 2
Parameters for the triad system (4.1) in the three test regimes.

	(B_1, B_2, B_3)	$(\lambda_1, \lambda_2, \lambda_3)$	(d_1, d_2, d_3)	$(\sigma_1, \sigma_2, \sigma_3)$	$\bar{\mathbf{u}}_0$	$\mathbf{r}_0$
regime I	$(1, -0.6, -0.4)$	$(3, -2, -1)$	$(0.2, 0.1, 0.1)$	$(1.58, 1.12, 1.12)$	$(2, 1.6, -2)$	$(0.5, 0.5, 1)$
regime II	$(1, -0.6, -0.4)$	$(0, 0, 0)$	$(0.02, 0.01, 0.01)$	$(0.5, 0.35, 0.35)$	$(3, -0.1, 0.1)$	$(0.5, 0.01, 0.01)$
regime III	$(2, -1, -1)$	$(0.09, 0.06, -0.03)$	$(-0.4, 2, 2)$	$(0.1, 0.32, 0.32)$	$(2, 1, 1.5)$	$(0.5, 5, 10)$

two are strongly excited by random noises. The nonlinear coupling first makes that energy cascades forwardly from mode u_1 to the other less energetic modes u_2, u_3 then backwardly from the excited modes u_2, u_3 back to u_1.

The initial state $\mathbf{u}_0 \sim \mathcal{N}(\bar{\mathbf{u}}_0, \mathbf{r}_0)$ is set to satisfy an independent Gaussian distribution with mean $\bar{\mathbf{u}}_0$ and variance $\mathbf{r}_0$. The true statistical solutions of the triad system (4.1) in the above dynamical regimes are solved through direct Monte-Carlo simulations. We run an ensemble of MC $= 1 \times 10^5$ particles, which shall be enough for capturing the statistics in a three-dimensional phase space. A fourth-order Runge-Kutta scheme with time step $\Delta t = 1 \times 10^{-3}$ is used to integrate the system in time (in practice, other numerical integrators could be also adopted with no major difference as long as the discrete time step offers stable numerical update). The stochastic forcing is simulated through the standard Euler-Maruyama scheme. The initial ensemble is chosen from a standard Gaussian random sampling with the mean $\bar{\mathbf{u}}_0$ and variance $\mathbf{r}_0$ listed in Table 2. In particular, we choose $B_1 > 0$ and $B_2, B_3 < 0$ to induce nonlinear instability (see the stability analysis in (B.6)). The model is run up to a final time $T = 10$ where near equilibrium state is reached.

The projected probability distributions of the triad state $p(\mathbf{u}, t)$ captured by direct Monte-Carlo simulations are demonstrated in Fig. 2. Representative non-Gaussian probability distributions are observed among all test regimes with distinctive statistical structures. The first test regime is the simplest but nevertheless representative showing the route of transient convergence to equipartition of energy. Still, as we will show in the following numerical tests in Section 4.2, higher-order moments are playing a pivoting role

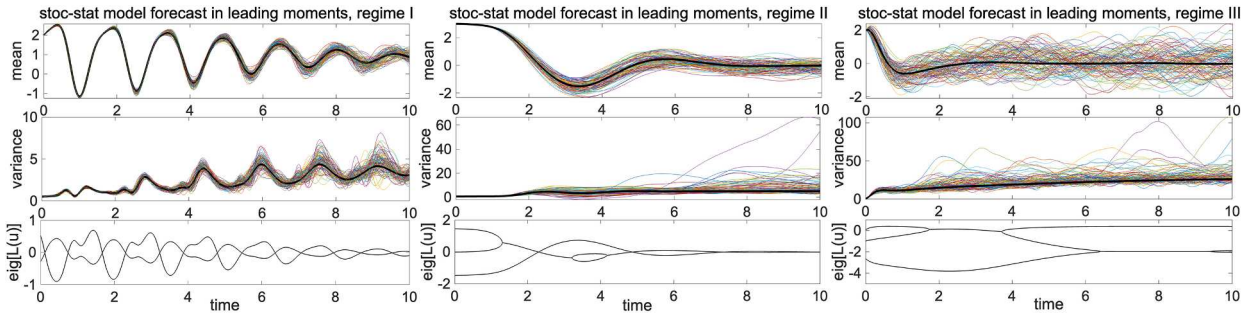


Fig. 3. Statistical forecasts using the stochastic-statistical model with $N = 100$ samples. First two rows: different realizations of the predicted mean $\bar{u}_1$ and variance r_1 in comparison with the truth in black lines; Third row: the Lyapunov exponent of the system indicating the unstable growth rate.

in this case and cannot be simply ignored in determining the correct final near-Gaussian equilibrium distribution. The second test regime emphasizes the nonlinear quadratic coupling between the three modes, leading to more complicated non-Gaussian probability distributions. In particular, we observe the wider spread of the samples in the scatter plots representing extreme events that are crucial but difficult to capture with a small ensemble. The third test regime introduces stronger interactions and dual energy cascades between the interacting modes. This leads to a strange attractor with star-shaped joint-distribution, showing a strongly nonlinear regime dominated by highly non-Gaussian statistics. Especially in this regime, the negative damping $d_1 = -0.4$ in the first mode introduces persistent internal instability into the system. In addition, the skew-symmetric linear interaction terms add extra emphasis on the cross-covariances. This regime becomes especially interesting and challenging because of the strong and persistent instability.

4.1.2. Small ensemble prediction with the coupled stochastic-statistical model

We start with testing direct forecast of the coupled stochastic-statistical model (2.5) to capture key model statistics. Using this coupled modeling framework, the statistical equations will be used to compute the leading moments $\bar{u}^N$ and R^N, combined with an ensemble simulation for the stochastic coefficients $\{Z^i\}_{i=1}^N$ aiming to capture the high-order moments feedback. To cope with the realistic scenario where only a small number of samples are affordable, we check the model forecast skill using a moderate ensemble size $N = 100$, in contrast to the truth in the previous section generated by MC $= 1 \times 10^5$ samples. Due to the dominant nonlinear coupling terms, the higher-order moments are involved in the statistical Eq. (2.2a), requiring accurately capturing the non-Gaussian statistics from the limited samples even only to predict the leading-order mean and covariance. This sets an especially challenging task demanding good characterization of the non-Gaussian distributions (including the extreme outliers observed in the PDFs in Fig. 2) using only the small number of samples.

First, the direct numerical predictions by running the coupled stochastic-statistical model (see the detailed equations for the triad model in Appendix B.3) in the three test regimes are shown in Fig. 3. To demonstrate the unavoidable large amount uncertainty induced through the small ensemble forecast, we plot multiple realizations of the mean and variance trajectories $(\bar{u}^N, R^N)$ from different randomly sampled initial stochastic states $\{Z^i\}$ using the small sample size $N = 100$. Model errors in the mean and variance forecasts are shown to rapidly grow in time among all three test regimes starting from accurate initial states. To further illustrate the development of such errors, we also plot the Lyapunov exponent, that is, the real parts of eigenvalues of the linearized matrix $L(\bar{u})$ in (2.2a), characterizing the inherent instability due to the interaction with the mean state. Positive eigenvalues indicate the unstable growth rate that amplifies small uncertainties in the variance. It can be observed clearly that persistent instability maintains in time amplifying the spread of different realizations of the solutions in all three test regimes, especially in regimes II and III which are experiencing stronger unstable growth during longer time periods. In the exact Eq. (2.2a), such unstable growth rate will be marginally balanced by the higher moments terms through the nonlinear coupling between the states. However, due to the insufficient representation of the highly non-Gaussian structures (as illustrated later in Fig. 10) with the limited number of samples, larger errors are introduced into the system. As a result, the trajectories of the mean and variance fail to track the truth and quickly diverge from the initial state. The accuracy of the predicted mean and variance will improve if we increase the number of samples N as indicated in Theorem 6. However, this will usually require an enormous sample size (due to the T dependence in the coefficients in (3.16)) even in this low-dimensional example, making any direct numerical approach impractical. These simple examples offer a typical illustration of the inherent difficulty in accurate prediction of model statistics when only a small sample size is affordable.

In addition, to further enforce the convergence of the numerical scheme, we show that the additional relaxation term added in the covariance equation of the numerical model (2.5) is essential especially in the regimes with stronger instability. In Fig. 4, we plot the model prediction of the variance in the most unstable mode u_1 in regime III by directly applying the forecast model. It shows that without the relaxation term $\epsilon^{-1} = 0$ enforcing the consistency between the sample approximation $\mathbb{E}^N\left[ZZ^T\right]$ and the covariance R_t, large numerical errors will start to develop in time even using an extremely large sample size. This is due to the persistent model instability among the states (see also the last row of Fig. 3 for the internal growth rate). On the other hand, it shows that the numerical errors can be effectively corrected by just introducing a small relaxation term with a small parameter $\epsilon^{-1} = 0.1$, thus accurate statistical convergence is guaranteed for the long-term time integration. In all the following numerical tests, we adopt this small relaxation parameter $\epsilon^{-1} = 0.1$.

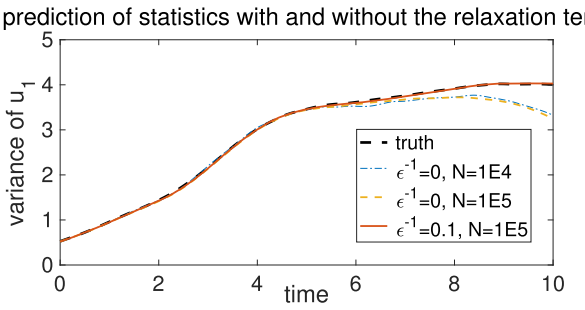


Fig. 4. Model prediction of the variance in the most unstable mode u_1 with and without the additional relaxation term in the numerical model (2.5).

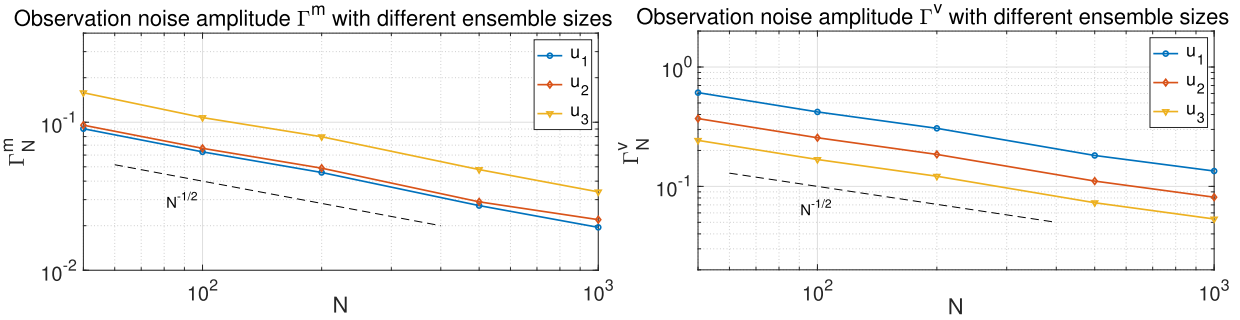


Fig. 5. Estimate of the observation noise with different sample sizes N. The noise parameters for the mean Γ^m and covariance Γ^v are computed according to the approximation errors in the three modes of the triad system.

4.2. Numerical performance using the data assimilation model

Next, we demonstrate that the proposed data assimilation model can effectively improve both stability and accuracy in the prediction of the key statistics in the triad system. Furthermore, it shows that maintaining the high-order correction terms constructed in the new filtering Eq. (3.9) is essential to achieve stable statistical prediction compared with the ensemble Kalman filters (3.11) where only the low-order moments information is used.

4.2.1. Calibration of observation noises

In setting up the filtering equations, we need to first estimate the observation noises Γ_m^N and Γ_v^N in (2.7) based on the finite ensemble size N. From the direct model simulations for $\bar{u}^N$ and R^N in Fig. 3, it shows that it is reasonable to treat $y^N = (\bar{u}^N, R^N)$ as a stochastic process and the randomness generated from the errors in the finite ensemble approximation in (2.9), such that the empirical estimate $\mathbb{E}^N H\mathrm{d}t = \mathbb{E}H\mathrm{d}t + \Gamma^N \mathrm{d}B$. In general, we can only expect upper bounds for the errors in the empirical averages as in (3.16) regarding to the sample size N and observation function H. Still in practical implementations, it is sufficient to get an estimate of the noise levels of Γ_m^N and Γ_v^N. In particular, we propose the following equations for the observation states according to (2.10) where we assume that error from the finite sample is dominant in the observation equation

$$\begin{aligned} \mathrm{d}y &= \mathbb{E}H\mathrm{d}t, \\ \mathrm{d}y^N &= \mathbb{E}^N H\mathrm{d}t = \mathbb{E}H\mathrm{d}t + \Gamma^N \mathrm{d}B. \end{aligned}$$

Above, y is the true deterministic observed states $\bar{u}$, R, and y^N is the stochastic observation process modeled with an additional noise term accounting for the randomness with finite ensemble estimation. Therefore, we find the following way to estimate the observation errors by assuming that the noise amplitude remains a constant in time for simplicity

$$\mathbb{E}\left(\left\|y^N - y\right\|^2\right) \approx \int_0^t \left(\Gamma^N\right)^2 \mathrm{d}s \approx t\left(\Gamma^N\right)^2. \tag{4.2}$$

In Fig. 5, we plot the estimated noise amplitudes in the observed states of mean and variance using (4.2) using different sample sizes N. The numerical results confirm the $N^{-\frac{1}{2}}$ convergence rate (3.16) found in Theorem 6 depending on the sample size N. It also provides a systematic way to estimate the observation noise level in different filtering model simulations according to the scaling law without repeating the different ensemble simulations many times.

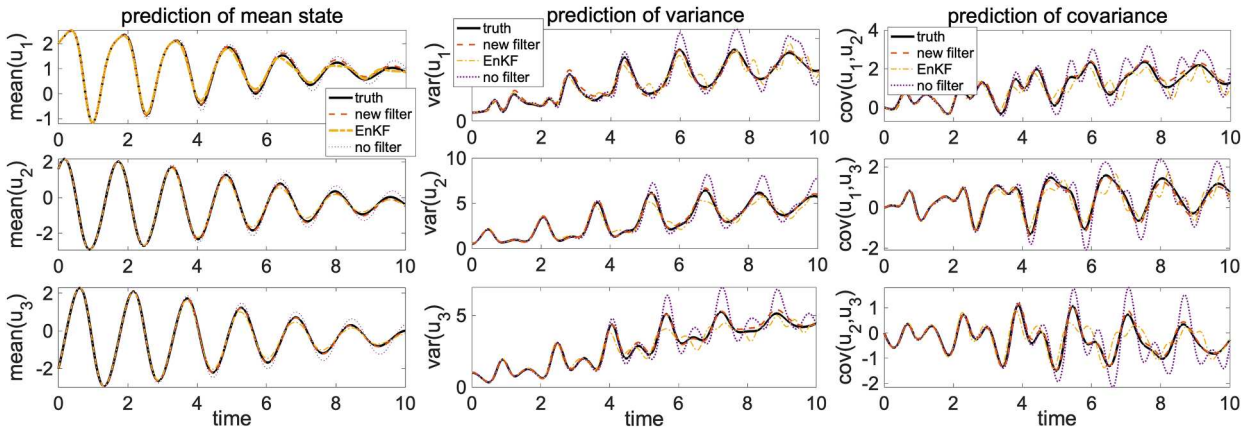


Fig. 6. Statistical prediction of the mean, variance, and covariance in regime I of the triad system. Results from the high-order data assimilation model are compared with the direct prediction without filter (2.5) and the EnKF (3.11) with $N = 100$ samples. The truth is generated with a direct MC approach with MC $= 1 \times 10^5$ samples.

4.2.2. *Prediction of the statistical mean and covariance*

Then, we compare the performance of the high-order data assimilation model in the triad system. To be adaptive to the general high-dimensional systems, we focus on testing the forecast skill of the models using a small sample size. As illustrated in Fig. 3, this small sample size has already become insufficient to capture the key statistical features in the simple triad system, and leads to large fluctuating errors in the prediction of mean and covariance by directly running the forecast model without using filtering. The additional observation data $y_n = \{\bar{u}_n, R_n\}$ is then introduced aiming to correct the errors in the finite ensemble forecast of the stochastic coefficients $\tilde{\mathbf{Z}}_n = \{\tilde{Z}_n^i\}$ in (2.12). The goal is to generate representative samples $\tilde{\mathbf{Z}}_n$ that can accurately characterize the high-order moments and PDF structure of model states. Notice that in the filter updating scheme only the information of derivatives $\mathrm{d}\bar{u}, \mathrm{d}R$ of leading-order moments are used from the observation data for the updates of stochastic samples. The model forecasts of the mean and covariance $\bar{u}^N$ and R^N are still directly updated through the statistical Eq. (2.5), thus the observed mean and covariance are not involved in updating the forecast mean and covariance. Therefore, their accuracy closely replies on the finite sample estimates of the higher-order moments due to their nonlinear dynamics. In the following, we check the prediction of mean and covariance using different models as an indicator for the model skill to capture key high-order statistics in $\tilde{\mathbf{Z}}_n$.

In Figs. 6–8, we plot the model predictions of the mean, variance, and cross-covariance between the three modes u_1, u_2, u_3 in the three typical test regimes respectively. The true statistics are compared with the forecast model without filter (2.5) and two data assimilation models. The first model is the EnKF (3.11) using only the low-order information and a constant Kalman gain in the filter update, while higher-order moments are considered according to the nonlinear observation operators in the new high-order filter model (3.9). The truth is captured by running the original triad system (4.1) using a very large ensemble size MC $= 1 \times 10^5$. Only a small ensemble size $N = 100$ is used in the model forecasts for all the tests. Frequent observation data is generated with the time integration step $\Delta t = 1 \times 10^{-3}$. First in regime I, the model state will converge to the final near-Gaussian equilibrium probability distribution. However, this regime demonstrates strong interactions between the linear operator L and the quadratic nonlinear operator B. This can be illustrated by the persistent positive growth rate in Fig. 3. The competing effects lead to strong oscillatory motions between the three modes indicating frequent exchange of energy between the scales. Non-Gaussian distributions will also be generated during the transient evolution of the states. As a result, even starting from accurate initial value large errors will gradually develop in the direct forecast model without filter in both the mean and covariance. The low-order EnKF model can correct the errors a bit from the forecast but still largely deviates from the true statistical values. In contrast, the high-order filter maintains the high accuracy in the predictions during the entire evolution time. In regime II, we focus on the nonlinear effect in the model driving strong cascade from mode u_1 to u_2, u_3. In this case, the system is dominated by the nonlinear coupling, and an accurate characterization of the high-order feedbacks in the statistical equations will play a central role in achieving good prediction result. As illustrated in Fig. 7, the predictions from the direct forecast model and low-order EnKF quickly diverge from the truth due to their insufficient sampling of the target probability distributions. This indicates that the stochastic samples in these models failed to correctly recover the higher-order moments in the nonlinear feedback terms in the statistical equations. Again, the high-order data assimilation scheme keeps stable and accurate predictions in both the mean and covariance up to the final prediction time. Finally, regime III sets a most challenging test case containing inherent internal instability from the linear operator. The nonlinear term then is needed to introduce the stabilizing effect that needs to be accurately quantified. Similarly to the other two test cases, we observe that the direct forecast model and EnKF fail to track the target trajectories of the mean and covariance with quick divergence to the truth due to the strong instability quickly amplifying the errors, while the high-order data assimilation model maintains its high skill against the persistently instability using only a very small sample size.

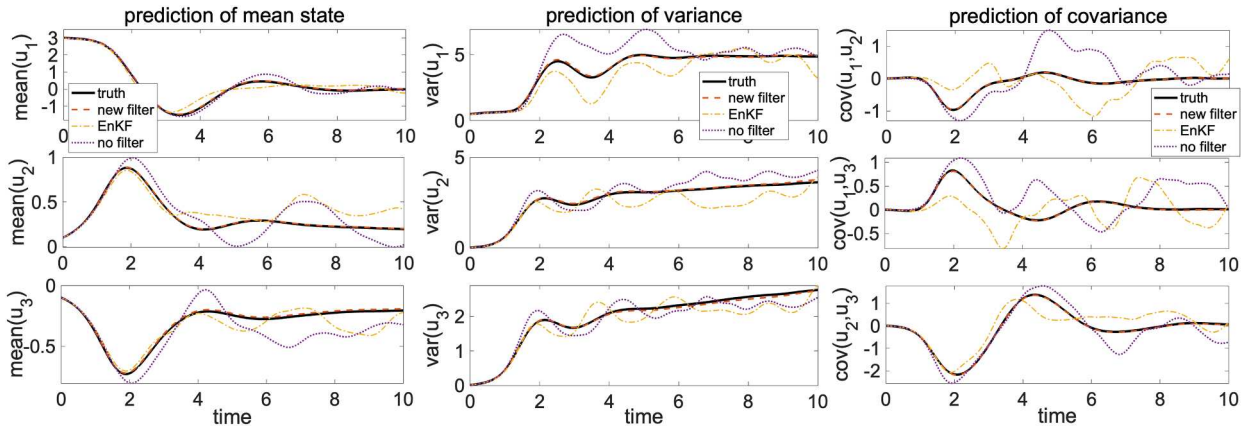


Fig. 7. Statistical prediction of the mean, variance, and covariance in regime II of the triad system, with the same setup as in Fig. 7.

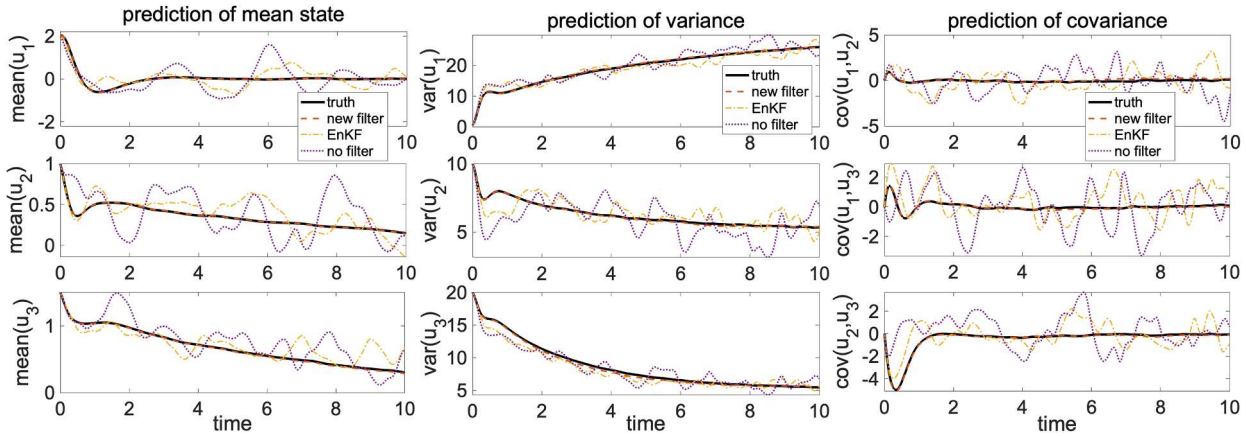
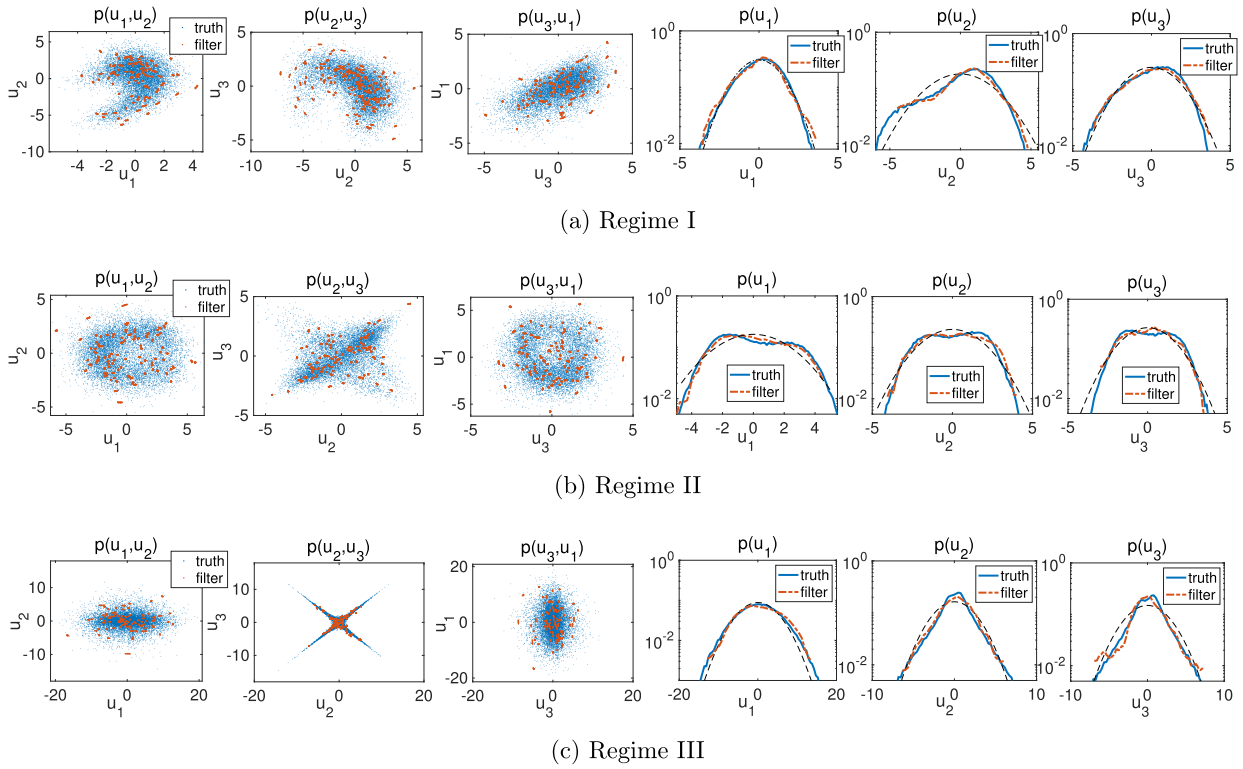


Fig. 8. Statistical prediction of the mean, variance, and covariance in regime III of the triad system, with the same setup as in Fig. 7.

4.3. Prediction of probability distributions and non-Gaussian features

The successful prediction of the mean and covariance confirms that the high-order data assimilation model is able to generate accurate samples covering the entire spread of the probability distribution containing essential non-Gaussian statistical features. As a more detailed illustration of the model forecast skill, we show the sampled probability distributions in Fig. 9. The projected distributions of the joint states are plotted in scatter plots together with the marginal PDFs of the three states u_1, u_2, u_3. First, it can be observed that the typical non-Gaussian features are generated in all the three test regimes demonstrating highly skewed or fat-tailed PDFs. These features make important contribution in the high-order feedback terms in the statistical equations, thus failing to accurately characterizing their effects in the finite ensemble approximation will lead to quick divergence that is shown in the statistical prediction. This explains the large errors and unstable performance observed in the direct model forecast and EnKF shown in Figs. 6–8 due to the insufficient sampling of these key probability distribution structures. On the other hand, it shows that the high-order data assimilation model drives enough samples to the suitable extreme locations so that the entire non-Gaussian PDFs are will represented among all the test regimes. This guarantees the high skill of the data assimilation model to recover the key model statistics including high-order moment information without requiring a large ensemble size. The uniformly high accuracy and stability of the new high-order filtering model among all the three test regimes with distinctive statistical features demonstrate the universal skill and robustness of the proposed filtering model.

To demonstrate more clearly how the direct forecast model and EnKF approach fail to reach the accurate time-series predictions in Figs. 6–8, we show one snapshot of the finite ensemble estimate of the probability distributions. In particular, Fig. 10 gives the scatter plots of samples representing the joint PDF of u_2 and u_3 in the most non-Gaussian regime III. It is shown that the extended four branches of the PDF tail structures are largely missed in the two models. In the direct forecast model, the small number of samples cannot sufficiently cover the regions containing extreme events, and only a few samples can reach the extended wings of the distribution. The corrections from the EnKF however draw the samples even closer to a Gaussian distribution rather than reaching the non-Gaussian features. In contrast, as shown in Fig. 9, the new high-order data assimilation model achieves a much better characterization of the key



(a) Regime I

(b) Regime II

(c) Regime III

Fig. 9. Probability distributions of the model states at time $t = 5$ from the high-order data assimilation model using $N = 100$ samples. The 2D scatter plots of the truth (blue) are compared with the ensemble filter forecast (red) as well as the 1D marginal distributions. Gaussian density functions with the same mean and variance are shown in dashed black lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

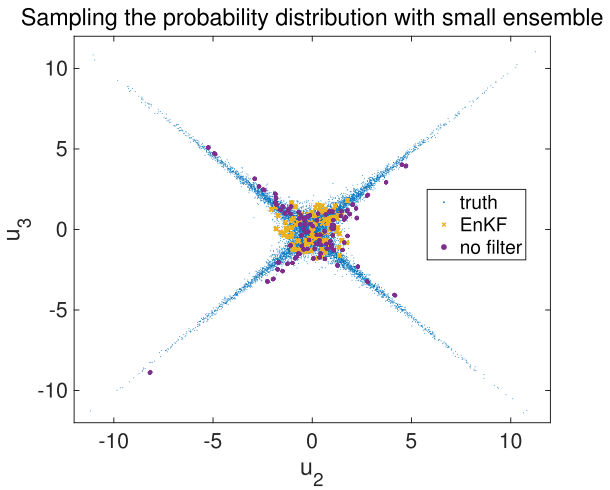


Fig. 10. Sampling of the target model distribution from a small ensemble forecast using the direct forecast model without filter (2.5) and the EnKF (3.11).

structures in the probability distribution, thus guarantees accurate prediction of the statistics. This example demonstrates the crucial role of accurately sampling the non-Gaussian PDFs in achieving accurate statistical prediction involving the nonlinear dynamics.

As a further illustration of the model prediction of higher-order moments, Fig. 11 plots the ensemble recovery of the third-order moments $M_3 = \mathbb{E}(u_1' u_2' u_3')$ in the three test regimes of the triad model. M_3 appears in the dynamical equations for the variances and plays a central role of balancing the instability from the linear couple terms (see the explicit statistical equations in (B.10)). First notice that non-zeros values in M_3 emerge in all three regimes, showing the non-negligible role of this high-order feedback term.

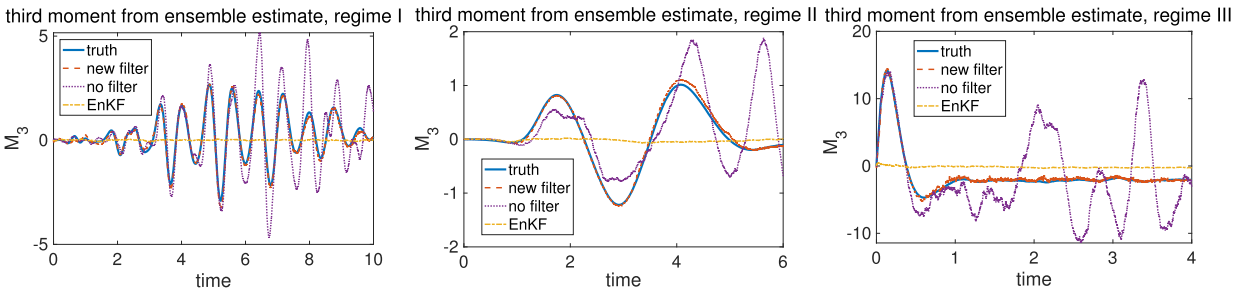


Fig. 11. Prediction of the third moment $M_3 = \mathbb{E}(u_1' u_2' u_3')$ through the ensemble approximation using the different models in the three test regimes.

Table 3
Prediction errors with different observation times Δt_{obs} compared with the direct forecast with no filter in three test regimes.

	regime I				regime II				regime III			
Δt_{obs}	0.001	0.01	0.05	no filter	0.001	0.01	0.05	no filter	0.001	0.01	0.05	no filter
RMSE in mean	0.034	0.058	0.084	0.309	0.022	0.130	0.142	0.480	0.021	0.142	0.240	0.632
RMSE in variance	0.263	0.437	0.572	1.129	0.234	0.631	0.768	1.099	0.560	1.897	2.911	3.970

Table 4
Prediction errors with different ensemble sizes N using the data assimilation model in three test regimes.

	regime I				regime II				regime III			
N	50	100	200	500	50	100	200	500	50	100	200	500
RMSE in $\bar{u}$	0.095	0.034	0.015	0.013	0.068	0.022	0.021	0.020	0.037	0.021	0.005	0.002
RMSE in R	0.701	0.263	0.172	0.126	0.373	0.234	0.105	0.077	1.347	0.560	0.441	0.238

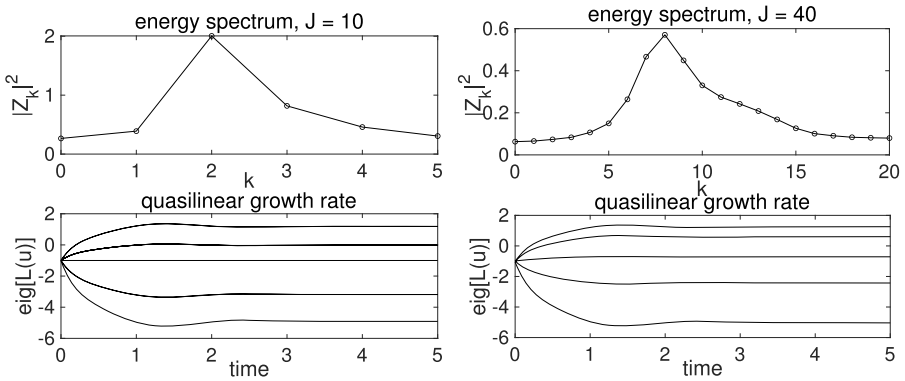


Fig. 12. Statistical energy spectra and quasilinear growth rate of the L-96 system with $J = 10$ (left) and $J = 40$ (right).

However, in the direct forecast model without filter, it can be observed that the sample estimates of M_3 are largely missed especially in the bursts of extreme values. This leads to the final large errors in the statistical predictions in the mean and covariance shown in Figs. 6–8 as well as confirming the biased estimate of the PDF in Fig. 10. From the EnKF prediction, on the other hand, near zero values are assigned to M_3 from the samples due to the Gaussian nature of this filter. This is also consistent with the PDF shown in Fig. 10 and explains the lack of skill in EnKF prediction of the key statistics. In contrast, the new high-order filter accurately tracks the true values of M_3 in the time-series, thus guarantees the successful predictions of the key statistics.

Finally, we also check the filter performance using different observation time frequencies Δt_{obs} and scaling with different ensemble sizes N. The total root mean square errors (RMSE) for the predicted mean and variance are listed in Tables 3 and 4. As expected, using shorter observation time Δt_{obs} and a larger number of sample N will increase the prediction accuracy, while the good performance is maintained even with less frequent observations and an even smaller ensemble size. In terms of computation time, the full MC simulation will run between 20–30 mins for the triad system with a large ensemble, while the data assimilation model simulations will all finish within in 1 minute on a laptop computer. This further confirms the robust performance of the high-order data assimilation model to successfully recover the leading statistics and generate samples that better represent the key non-Gaussian features in the probability distributions.

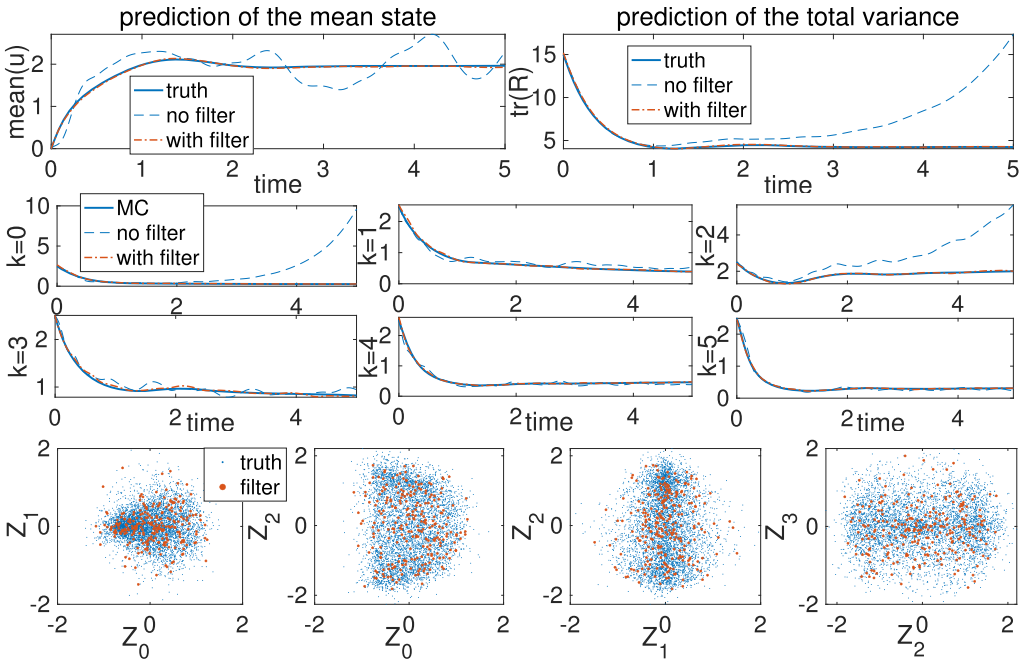


Fig. 13. Statistical prediction of the L-96 system with dimension $J = 10$. Statistical mean and variances from the high-order data assimilation model are compared with direct prediction using $N = 100$ samples. Joint probability distributions of the stochastic coefficients are compared in scatter plots. The truth is generated with a direct MC simulation using MC $= 5 \times 10^4$ samples.

5. Numerical performance on Lorenz 96 system

To further evaluate the performance on higher-dimensional systems, we apply the proposed data assimilation algorithm on the Lorenz 96 (L-96) system [53] that is used as a prototype model to examine data assimilation schemes. The L-96 system can be expressed as a J-dimensional ODE system as

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} = -u_j + F + \big(u_{j+1} - u_{j-2}\big)u_{j-1},\ j = 1, \cdots, J, \tag{5.1}$$

with periodic boundary condition $u_{J+1} = u_1$ and constant uniform forcing and damping effects. The model state u_j is defined to mimic geophysical waves at J equally distributed locations along a constant mid-latitude circle. Various representative statistical features can be found in the L-96 solution by varying the constant forcing F and state dimension J [41]. Notice that by taking the dimension of the system $J = 3$, the L-96 Eq. (5.1) shares similar dynamical structures as the triad system (4.1) with homogeneous linear terms and energy-conserving quadratic nonlinear coupling. In the numerical tests here, we adopt the constant forcing $F = 6$ that demonstrates strong non-Gaussian statistics in the state solution u_j. Two cases with moderate ($J = 10$) and high ($J = 40$) dimension are considered to test the model performance with increasingly high dimension and non-Gaussian PDFs.

Following the general stochastic-statistical formulation in (2.1), we introduce the mean-fluctuation decomposition for the L-96 state as

$$u_j = \bar{u}_j + \frac{1}{J}\sum_{|k|\le J/2} Z_k(t) e^{i2\pi k\frac{j}{J}}. \tag{5.2}$$

Above, Fourier basis $\hat{v}_k = \{\frac{1}{J}e^{i2\pi k\frac{j}{J}}\}_{j=1}^{J}$ is taken as a natural choice for the periodic boundary condition. To sufficiently resolve the true statistical solution, we run the above Eq. (5.1) using a large ensemble size MC $= 5 \times 10^4$. To show the adaptiveness to different numerical integrators, we use the first-order forward Euler scheme for the time integration with time step $\Delta t = 1 \times 10^{-3}$. The model state starts with an initial distribution with independent Gaussian distribution in each mode Z_k with small variances, while the internal instability will rapidly amplify the uncertainty among the modes. The model is simulated up to the final time $T = 5$ where the final equilibrium state has been reached. True statistical solutions with dimensions $J = 10$ and $J = 40$ are shown in Fig. 12. The L-96 system maintains a wide spectrum of energetic modes indicating nonlinear interactions between the stochastic spectral modes Z_k. Similar to the triad system case, we also plot the Lyapunov exponent as the eigenvalues of the linearized operator $L(\bar{u})$ in (2.3). It can be observed that strong internal instability automatically arises in the leading modes of both test cases with multiple positive growth rates that increase the uncertainty among the modes, leading to the wide-spread spectral structure. This indicates a challenging test case for stable statistical prediction that requires an accurate characterization of the third-order coupling terms in (2.3) that play the crucial role in balancing the unstable positive growth and driving the system to the final equilibrium state.

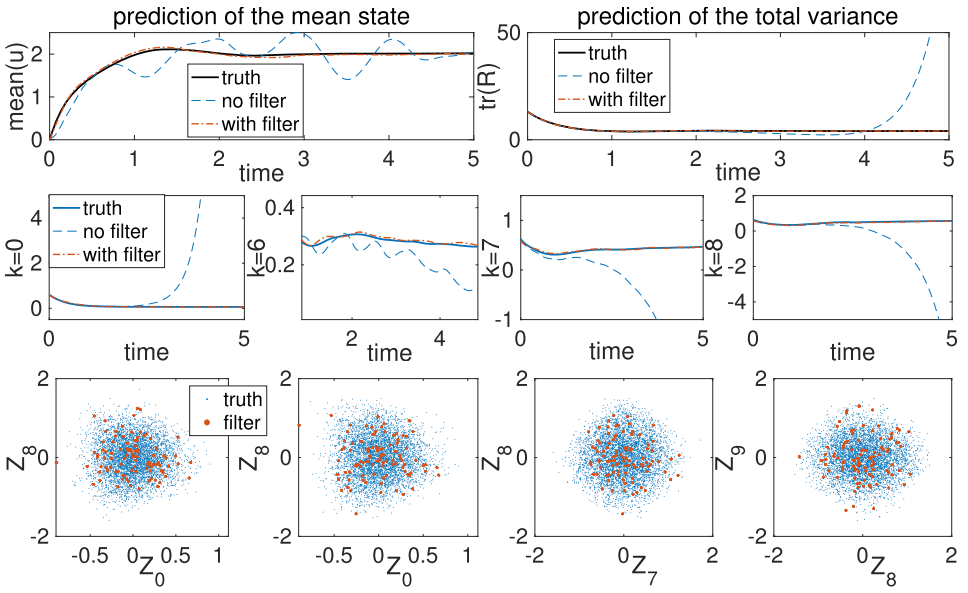


Fig. 14. Statistical prediction of the L-96 system with dimension $J = 40$. Statistical mean and variances from the high-order data assimilation model are compared with direct prediction using $N = 100$ samples. Joint probability distributions of the stochastic coefficients are compared in scatter plots. The truth is generated with a direct MC simulation using MC $= 5 \times 10^4$ samples.

Now, we apply the data assimilation model on the L-96 system containing a large number of internal unstable modes. Still, we aim to capture the key model statistics using a small number of samples $N = 100$. First, the prediction results for the mean and variances as well as the joint PDFs capured by the particles in the moderate dimension $J = 10$ case are plotted in Fig. 13. Similar to the triad model results, the direct numerical prediction of the stochastic-statistical model (2.5) using a small ensemble size fails to capture the statistical solution. Large numerical fluctuations are observed in the mean state and variance modes, and the predictions of variances in unstable modes (that is, $k = 0$ and $k = 2$) quickly diverge due to the insufficient characterization of the stabilizing third moments using small samples. Then, using the data assimilation model by incorporating additional statistical observations of low-order moments, the predictions of mean and variances are tracked accurately during the entire evolution time and the inherent instability is effectively balanced. The joint PDFs captured by the ensemble approximation are also shown with the truth from large ensemble simulation and filter prediction in small samples. Typical non-Gaussian distributions are observed in the scatter plots, confirming the crucial role the high-order statistics to guarantee accurate and stable prediction. Using only $N = 100$ samples, the data assimilation model successfully captures the key non-Gaussian structure in the probability distributions, thus achieve accurate statistical prediction. In the final test, Fig. 14 shows the prediction results in the high-dimensional case with $J = 40$. As illustrated in Fig. 12, this higher-dimensional case contains a wider spectrum of large unstable modes. This leads to the more rapid divergence in the direct approach. Again, the data assimilation model maintains the high skill in accurately capturing both the mean and variances with long time stability as well as the scattered probability distributions. In terms of computational time, the direct MC simulations of the L-96 system will take more than 1 hr to finish, while the data assimilation model runs around 5 mins for $J = 40$ case and below 1 min for $J = 10$ case on a laptop. The robust and efficient performance of the data assimilation algorithm implies the potential application of the method to more general systems with high dimensionality and strong non-Gaussian statistics.

6. Summary

In this paper, we developed an explicit high-order data assimilation framework for effective ensemble prediction of probability distributions exhibiting highly non-Gaussian statistics. By leveraging observation data from lower-order statistical moments, the stability and accuracy of statistical predictions are significantly enhanced using a computational affordable finite ensemble approach. Specifically, detailed filtering operators are derived utilizing the explicit quadratic and cubic structures of the nonlinear coupling terms, resulting in a straightforward numerical implementation without high computational cost. We performed comprehensive numerical experiments using an illustrative triad system and the more general L-96 system with different dimensions, which generates representative turbulent phenomena across different statistical regimes, to systematically evaluate the skill of the numerical scheme. Inherent computational barriers for accurate statistical prediction with the finite ensemble approaches are demonstrated under this simple test model. Direct numerical comparisons confirm that accurately capturing non-Gaussian distributions is essential for precise statistical prediction in highly nonlinear dynamics under restricted sampling constraint. The filtering updates within the proposed data assimilation model consistently show robust performance in capturing the various types of non-Gaussian features across multiple tested statistical regimes requiring only on a small sample size and observation data of leading-moments. In contrast, traditional ensemble Kalman filter approaches with near-Gaussian assumptions typically fail to capture such crucial high-order statistical information.

Limitations of the data assimilation model and future research directions

Still, many interesting problems remain open in both theoretical analysis and practical strategies of the approximate data assimilation model. In this paper, we aim to provide a thorough computational investigation of the new data assimilation strategy on typical non-Gaussian and nonlinear structures by experimenting on simple and highly tractable prototype models in the test examples. In practice, the systems may be subject to further constraints. For example, the Lipschitz conditions for the model coefficients in Assumption 5 may not be valid. This requires further exploration exploiting specific structures such as the detailed energy conservation laws [17] in the nonlinear terms to equip the computational strategy with a better understanding of its approximation skill and scope of application on different realistic scenarios. In addition, sensitivity to errors in the imperfect prediction model and the observation data is another issue that requires further investigation. From the numerical tests, we observe robust model performance against the large errors from the stochastic-statistical model prediction and noisy observation data. A detailed study is needed based on the errors from the finite ensemble approximation and a systematic strategy is required to calibrate the observation noises to achieve optimal prediction performance.

In the immediate application of this proposed computational scheme, additional model reduction strategies will be needed to further enhance the new data assimilation strategy for really high-dimensional realistic turbulent systems. In dealing with the potentially high computational cost of solving a really high-dimensional system, we plan to combine our approach with high-order moment closure methods incorporating the random batch approximations [15,48]. Another approach is to combine the stochastic-statistical modeling framework with the rapidly developing data-driven approaches [54,55] to automatically learn the unresolved high-order terms from data. The additional model reduction strategies will create practical and computational efficient algorithms suitable for high-dimensional problems. Immediate applications of these developments include statistical forecasting in geophysical turbulence [56] and modeling viscoelastic fluids [57]. Further research may also explore the extension and validation of the modeling and computational framework combined with the recent optimal transport and particle flow filtering strategies [58,59], potentially enabling more accurate predictions of complex turbulent phenomena in broader classes of multiscale dynamical systems.

CRediT authorship contribution statement

Di Qi: Writing – original draft, Validation, Methodology, Investigation, Conceptualization; **Jian-Guo Liu:** Writing – original draft, Validation, Methodology, Investigation, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research of J.-G. L. is partially supported by the NSF Grant DMS-2106988. D. Q. is partially supported by ONR Grant N00014-24-1-2192, and NSF Grant DMS-2407361.

Appendix A. Proofs of theorems

Proof of Lemma 1By taking partial derivatives using the explicit expressions in (3.1), we get

$$\begin{aligned}\partial_j H^m_k &= \gamma_{kjq} z_q + \gamma_{kpj} z_p,\\ \partial_j H^v_{kl} &= \big(\gamma_{kjq} z_q z_l + \gamma_{kpj} z_p z_l + \gamma_{kpq} z_p z_q \delta_{lj}\big)\\ &\quad + \big(\gamma_{ljq} z_q z_k + \gamma_{lpj} z_p z_k + \gamma_{lpq} z_p z_q \delta_{kj}\big).\end{aligned}$$

Above, for convenience double appearance of the subindex implies the summation about the index. Next, multiplying z_j and taking the summation about j yield

$$\begin{aligned}z^T \nabla H^m_k &= z_j \partial_j H^m_k = \gamma_{kjq} z_j z_q + \gamma_{kpj} z_p z_j = 2H^m_k,\\ z^T \nabla H^v_{kl} &= z_j \partial_j H^v_{kl} = \big(\gamma_{kjq} z_l + \gamma_{ljq} z_k\big) z_j z_q + \big(\gamma_{kpj} z_l + \gamma_{lpj} z_k\big) z_p z_j\\ &\quad + \big(\gamma_{kpq} z_l + \gamma_{lpq} z_k\big) z_p z_q = 3H^v_{kl}.\end{aligned}$$

Proof of Proposition 2 We can check the solution (3.4) by directly putting the expressions back into the Eq. (3.3). Therefore, for the mean observation we have

$$\tilde{\mathbb{E}}\Big[(K^m)^T \nabla H^m\Big] = \frac{1}{2}\Gamma_m^{-2}\tilde{\mathbb{E}}\big[\big(H^m - \bar{H}^m\big)\big(Z^T \nabla H^m\big)\big]$$

$$= \Gamma_m^{-2}\tilde{\mathbb{E}}\Big[\big(H^m - \bar{H}^m\big)\big(H^m\big)^T\Big]$$
$$= \Gamma_m^{-2}\tilde{\mathbb{E}}\Big[\big(H^m - \bar{H}^m\big)\big(H^m - \bar{H}^m\big)^T\Big] = \Gamma_m^{-2}C^{H^m}.$$

Above, the second equality uses the first identity in (3.2) and the third uses $\tilde{\mathbb{E}}\big[H^m - \bar{H}^m\big] = 0$. In the same way, we can check the case for variance observation only with a difference in the coefficient

$$\tilde{\mathbb{E}}\Big[\big(K^v\big)^T\nabla H^v\Big] = \frac{1}{3}\Gamma_v^{-2}\tilde{\mathbb{E}}\big[\big(H^v - \bar{H}^v\big)\big(Z^T\nabla H^v\big)\big]$$
$$= \Gamma_v^{-2}\tilde{\mathbb{E}}\Big[\big(H^v - \bar{H}^v\big)\big(H^v\big)^T\Big]$$
$$= \Gamma_v^{-2}\tilde{\mathbb{E}}\Big[\big(H^v - \bar{H}^v\big)\big(H^v - \bar{H}^v\big)^T\Big] = \Gamma_v^{-2}C^{H^v}.$$

Proof of Proposition 3 Using the explicit formula in (3.4), we can compute for the mean observation case with $\big(\Gamma_m^{-2}\big)_{pq} = \gamma_{pq}^{-2}$

$$\big(K^m\Gamma_m^2K^{mT}\big)_{ij} = \frac{1}{4}Z_iZ_j\sum_{p,q}\gamma_{pq}^{-2}H_p^{m\prime}(Z)H_q^{m\prime}(Z),$$

where we denote $H'(Z) = H(Z) - \bar{H}$. By taking the divergence on the above identity, we can compute

$$\nabla\cdot\big(K^m\Gamma_m^2K^{mT}\big)_i = \sum_j\partial_j\big(K^m\Gamma_m^2K^{mT}\big)_{ij} = \frac{1}{4}\sum_{j,p,q}\gamma_{pq}^{-2}\partial_j\Big[Z_iZ_jH_p^{m\prime}(Z)H_q^{m\prime}(Z)\Big]$$
$$= \frac{1}{4}\sum_{j,p,q}\gamma_{pq}^{-2}\Big[\delta_{ij}Z_jH_p^{m\prime}(Z)H_q^{m\prime}(Z) + Z_iH_p^{m\prime}(Z)H_q^{m\prime}(Z)$$
$$Z_iH_q^{m\prime}(Z)Z_j\partial_jH_p^m(Z) + Z_iH_p^{m\prime}(Z)Z_j\partial_jH_q^m(Z)\Big]$$
$$= \frac{1}{4}\sum_{p,q}\gamma_{pq}^{-2}Z_i\Big[H_p^{m\prime}(Z)H_q^{m\prime}(Z) + dH_p^{m\prime}(Z)H_q^{m\prime}(Z)\Big]$$
$$+ \frac{1}{4}\sum_{p,q}\gamma_{pq}^{-2}Z_i\Big[2H_p^m(Z)H_q^{m\prime}(Z) + 2H_q^m(Z)H_p^{m\prime}(Z)\Big]$$
$$= \frac{5+d}{4}Z_i\sum_{p,q}\gamma_{pq}^{-2}H_p^{m\prime}(Z)H_q^{m\prime}(Z) + \frac{1}{2}Z_i\sum_{p,q}\gamma_{pq}^{-2}\Big[\bar{H}_p^mH_q^{m\prime}(Z) + \bar{H}_q^mH_p^{m\prime}(Z)\Big]. \tag{A.1}$$

The second from last equality above again uses the identity (3.2), $\sum_j Z_j\partial_jH^m = 2H^m$. Notice that additional term in the last line above due to the mean term $\bar{H}^m$. In the same way, we can use (3.2) again and find

$$\Big(\nabla\cdot\big(K^m\big)^T\Big)_q = \sum_j\partial_j\Big(\frac{1}{2}ZH^{m\prime}(Z)^T\Gamma_m^{-2}\Big)_{jq} = \frac{1}{2}\sum_{j,p}\gamma_{pq}^{-2}\partial_j\Big[Z_jH_p^{m\prime}(Z)\Big]$$
$$= \frac{1}{2}\sum_{j,p}\gamma_{pq}^{-2}\Big[H_p^{m\prime}(Z) + Z_j\partial_jH_p^m(Z)\Big]$$
$$= \frac{1}{2}\sum_p\gamma_{pq}^{-2}\Big[dH_p^{m\prime}(Z) + 2H_p^m(Z)\Big]$$
$$= \frac{d+2}{2}\sum_p\gamma_{pq}^{-2}H_p^{m\prime}(Z) + \sum_p\gamma_{pq}^{-2}\bar{H}_p^m.$$

Thus the second term in (3.5) gives with the above identity

$$\Big(K^m\Gamma_m^2\nabla\cdot\big(K^m\big)^T\Big)_i = \frac{1}{2}Z_i\big(H^{m\prime}\big)^T\nabla\cdot\big(K^m\big)^T$$
$$= \frac{d+2}{4}Z_i\sum_{p,q}\gamma_{pq}^{-2}H_p^{m\prime}H_q^{m\prime} + \frac{1}{2}Z_i\sum_{p,q}\gamma_{pq}^{-2}\bar{H}_p^mH_q^{m\prime}. \tag{A.2}$$

Combining the final results in (A.1) and (A.2), we find

$$a_i^m = \nabla\cdot\big(K^m\Gamma_m^2K^{mT}\big)_i - K_m\Gamma_m^2\nabla\cdot\big(K^{mT}\big)_i$$
$$= \frac{3}{4}Z_i\sum_p\gamma_{pq}^{-2}H_p^{m\prime}H_q^{m\prime} + \frac{1}{2}Z_i\sum_{p,q}\gamma_{pq}^{-2}\bar{H}_q^mH_p^{m\prime}(Z)$$
$$= \frac{1}{4}Z_i\sum_p\gamma_{pq}^{-2}H_p^{m\prime}\Big(3H_q^{m\prime} + 2\bar{H}_q^m\Big)$$
$$= \frac{1}{4}Z_i\sum_p\gamma_{pq}^{-2}\Big(H_p^m - \bar{H}_p^m\Big)\Big(3H_q^m - \bar{H}_q^m\Big).$$

This gives the expression for the drift term of the mean in (3.6). Repeating the same procedure for the variance, we can arrive at the expression for a^v in a similar fashion.

Proof of Proposition 4 Using the explicit expressions derived in (3.4) and (3.6) as well as the discrete integration of (3.7), we can compute the filtering update terms from the observation of the mean as

$$\begin{aligned}
a^m\Delta t + K^m\Delta I^m =& \frac{1}{4}Z\left[H^m(Z)-\bar{H}^m\right]^T\Gamma_m^{-2}\left[3H^m(Z)-\bar{H}^m\right]\Delta t\\
&+\frac{1}{2}Z\left[H^m(Z)-\bar{H}^m\right]^T\Gamma_m^{-2}\left[\Delta\bar{u}-\left[H^m(Z)+h_m(\bar{u}^N)\right]\Delta t\right]\\
=&\frac{3}{4}ZH^{m\prime}(Z)^T\Gamma_m^{-2}H^{m\prime}(Z)\Delta t+\frac{1}{2}ZH^{m\prime}(Z)^T\Gamma_m^{-2}\bar{H}^m\Delta t\\
&+\frac{1}{2}ZH^{m\prime}(Z)^T\Gamma_m^{-2}\left[\Delta\bar{u}-\Delta\bar{u}^N-H^{m\prime}(Z)\Delta t\right]\\
=&\frac{1}{4}ZH^{m\prime}(Z)^T\Gamma_m^{-2}H^{m\prime}(Z)\Delta t+\frac{1}{2}ZH^{m\prime}(Z)^T\Gamma_m^{-2}\bar{H}^m\Delta t+\frac{1}{2}ZH^{m\prime}(Z)^T\Gamma_m^{-2}\left(\Delta\bar{u}-\Delta\bar{u}^N\right).
\end{aligned}$$

Above, we define the mean and fluctuation terms, $\bar{H}^m=\mathbb{E}^N\left[H^m(\tilde{\mathbf{Z}})\right]$ and $H^{m\prime}=H^m(\tilde{Z})-\bar{H}^m$, w.r.t. the empirical ensemble averages. Similarly, we can compute

$$\begin{aligned}
a^v\Delta t + K^v\Delta I^v =& \frac{1}{9}Z\left[H^v(Z)-\bar{H}^v\right]^T\Gamma_v^{-2}\left[4H^v(Z)-\bar{H}^v\right]\\
&+\frac{1}{3}Z\left[H^v(Z)-\bar{H}^v\right]^T\Gamma_v^{-2}\left[\Delta R-\left[H^v(Z)+h_v(\bar{u}^N,R^N)\right]\Delta t\right]\\
=&\frac{4}{9}ZH^{v\prime}(Z)^T\Gamma_v^{-2}H^{v\prime}(Z)\Delta t+\frac{1}{3}ZH^{v\prime}(Z)^T\Gamma_v^{-2}\bar{H}^v\Delta t\\
&+\frac{1}{3}ZH^{v\prime}(Z)^T\Gamma_v^{-2}\left[\Delta R-\Delta R^N-H^{v\prime}(Z)\Delta t\right]\\
=&\frac{1}{9}ZH^{v\prime}(Z)^T\Gamma_v^{-2}H^{v\prime}(Z)\Delta t+\frac{1}{3}ZH^{v\prime}(Z)^T\Gamma_v^{-2}\bar{H}^v\Delta t+\frac{1}{3}ZH^{v\prime}(Z)^T\Gamma_v^{-2}\left(\Delta R-\Delta R^N\right).
\end{aligned}$$

Appendix B. Details on the triad system

Here, we provide more details on the dynamical and statistical properties on the triad model (4.1).

B.1. A direct link to geophysical turbulent fluid

Consider the quasi-geostrophic (QG) potential vorticity equation with forcing and dissipation defined on a two-dimensional periodic domain $\mathbf{x}\in[-\pi,\pi]\times[-\pi,\pi]$

$$\frac{\partial q}{\partial t}+\nabla^\perp\psi\cdot\nabla q=\nu\Delta q,\quad \Delta\psi=q, \tag{B.1}$$

where $\nabla^\perp=\left(-\partial_y,\partial_x\right)$. Under projection to the Fourier spectral modes $\mathbf{k}=\left(k_x,k_y\right)$ inside a set of finite wavenumber truncation $\mathcal{K}$, the flow streamfunction ψ and potential vorticity q can be expressed as

$$\psi=\sum_{\mathbf{k}\in\mathcal{K}}\hat{\psi}_{\mathbf{k}}e^{i\mathbf{k}\cdot\mathbf{x}},\quad q=\sum_{\mathbf{k}\in\mathcal{K}}(-|\mathbf{k}|)\hat{\psi}_{\mathbf{k}}e^{i\mathbf{k}\cdot\mathbf{x}}.$$

The QG system (B.1) then can be expressed for each spectral mode $\hat{\psi}_{\mathbf{k}}$ under the above decomposition as

$$\frac{\mathrm{d}\hat{\psi}_{\mathbf{k}}}{\mathrm{d}t}+\sum_{\mathbf{k}=-\mathbf{m}-\mathbf{n}}\frac{|\mathbf{n}|^2}{|\mathbf{k}|^2}\mathbf{m}^\perp\cdot\mathbf{n}\hat{\psi}_{\mathbf{m}}\hat{\psi}_{\mathbf{n}}=-\nu|\mathbf{k}|^2\hat{\psi}_{\mathbf{k}}.$$

Therefore, we have the barotropic triads of three wavenumber components, $\hat{\psi}_{\mathbf{k}},\hat{\psi}_{\mathbf{m}},\hat{\psi}_{\mathbf{n}}$, obeying the selecting rule $\mathbf{k}+\mathbf{m}+\mathbf{n}=\mathbf{0}$. Consider an initial condition in which only these three components of a particular triad are excited, then these three modes will only interact with each other while no other modes will get excited due to the particular triad relations as the system evolves in time. By projecting the above equation to the active triad modes, we get the dynamical equations for the selected modes neglecting the forcing and dissipation terms on the right hand side

$$\frac{\mathrm{d}\hat{\psi}_{\mathbf{k}}}{\mathrm{d}t}+A_{\mathbf{kmn}}\hat{\psi}_{\mathbf{m}}\hat{\psi}_{\mathbf{n}}=0,\quad \mathbf{k}+\mathbf{m}+\mathbf{n}=\mathbf{0}, \tag{B.2}$$

where $A_{\mathbf{kmn}}=\frac{|\mathbf{n}|^2}{|\mathbf{k}|^2}\mathbf{m}^\perp\cdot\mathbf{n}$ is the triad interaction coefficient with the detailed symmetry $A_{kmn}+A_{mnk}+A_{nkm}=0$, showing the conservation of kinetic energy,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(|\mathbf{k}|^2|\hat{\psi}_{\mathbf{k}}|^2+|\mathbf{m}|^2|\hat{\psi}_{\mathbf{m}}|^2+|\mathbf{n}|^2|\hat{\psi}_{\mathbf{n}}|^2\right)=0.$$

The typical forward and backward cascades of energy and enstrophy in turbulent flow are characterized by the triad interactions between the three models. Hence from the above discussion, in the two-dimensional QG turbulence, the nonlinear energy transfer is exactly governed by the barotropic triads the same as (4.1) in the nonlinear interaction part. More detailed characterization of these coupling effects with link to geophysical turbulence can be found in [14].

B.2. Statistical and dynamical properties of the triad system

The triad system (4.1) is subject to stochasticity from the initial state and external forcing. The probability density function $p(\mathbf{u},t)$ associated with the triad equations satisfies the following Fokker-Planck equation

$$\partial_t p = -(B(\mathbf{u},\mathbf{u})+\Lambda\mathbf{u})\cdot\nabla_{\mathbf{u}}p + \sum_{i=1}^{3}\left(d_i p + \frac{1}{2}\sigma_i^2\partial_{u_i}^2 p\right), \tag{B.3}$$

with initial state $p(\mathbf{u},t)\,|_{t=0}= p_0(\mathbf{u})$. While the original triad system (4.1) is nonlinear, the statistical dynamics (B.3) becomes linear equation for the smooth PDF p. However, in general the explicit solution of the Fokker-Planck equation is still difficult to achieve due to the nonlinear interaction terms in the triad system.

B.2.1. Equilibrium invariant measure with equipartition of energy

Under a special arrangement about the damping and noise coefficients, one special solution of a Gaussian invariant measure, p_{eq}, can be reached at the equilibrium. Assume that the damping operator d_i and random noise forcing σ_i satisfy the following relation in each component

$$\sigma_{\mathrm{eq}}^2 = \frac{\sigma_1^2}{2d_1} = \frac{\sigma_2^2}{2d_2} = \frac{\sigma_3^2}{2d_3}. \tag{B.4}$$

Therefore, a Gaussian invariant measure can be found with equipartition of energy in each component, that is,

$$p_{\mathrm{eq}}(\mathbf{u}) = C_{\mathrm{eq}}^{-1}\exp\left(-\frac{1}{2}\sigma_{\mathrm{eq}}^{-2}|\mathbf{u}|^2\right). \tag{B.5}$$

Above σ_{eq}^2 is the equilibrium variance in the Gaussian invariant distribution p_{eq} that controls the variability in each mode. To see this, we can substitute the invariant measure (B.5) back into the Fokker-Planck Eq. (B.3). It is a special case from Theorem 3.1 in [60]. In the general case with additional external forcing and inhomogeneous structure, energy is injected into the modes and transferred to each other due to the nonlinear quadratic interaction through more complicated mechanism, thus strong nonlinear non-Gaussian statistics with energy cascade and internal instabilities can be generated. Detailed energy mechanism and stability for the triad system can be also found in [14,41].

B.2.2. Typical dynamical regimes in the triad system

Though simple in appearance, the triad system (4.1) has representative statistical features including energy cascade between modes and internal instabilities that can be created in this simple set-up. A fundamental factor in the triad system is the internal instabilities that make the mean unstable over various directions in phase space as is typical for anisotropic fully turbulent systems. Elementary intuition about energy transfer in such models can be gained by looking at the special situation with only the nonlinear interactions in (4.1). We examine the linear stability of the fixed point, $\bar{\mathbf{u}} = (\bar{u}_1,0,0)^T$. Elementary calculations show that the perturbation δu_1 satisfies $\frac{d\delta u_1}{dt}=0$ while the perturbations $\delta u_2,\delta u_3$ satisfy the second-order equations

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\delta u_2) = (B_2B_3\bar{u}_1^2)\delta u_2, \quad \frac{\mathrm{d}^2}{\mathrm{d}t^2}(\delta u_3) = (B_2B_3\bar{u}_1^2)\delta u_3,$$

so that we find that there is instability in the states u_2,u_3 from a non-zero $\bar{u}_1$ if $B_2B_3>0$. Combined with the energy conservation principle $B_1+B_2+B_3=0$, we find that from the initial state $(\bar{u}_1,0,0)$

$$\begin{aligned}&\text{the energy in } \delta u_2,\delta u_3 \text{ grows provided that}\\ &B_1 \text{ has the opposite sign with } B_2 \text{ and } B_3.\end{aligned} \tag{B.6}$$

The elementary analysis in (B.6) suggests that we can expect a flow or cascade of energy from u_1 to u_2 and u_3 where it is dissipated provided the interaction coefficient B_1 has the opposite sign from B_2 and B_3. Then energy cascades can be induced from the strongly forced unstable energetic mode to the stable less energetic modes with stronger damping effects.

B.3. Moment equations for the triad system

Here, we provide the detailed moment equations for the mean mean and covariances of the triad state $\mathbf{u}$. First, the mean state $\bar{\mathbf{u}} = (u_1,u_2,u_3)^T$ of the triad model can be written as

$$\begin{aligned}\mathrm{d}\bar{u}_1 &= \left[\left(-d_1\bar{u}_1-\lambda_3\bar{u}_2+\lambda_2\bar{u}_3\right)+B_1\bar{u}_2\bar{u}_3+B_1\left\langle u_2'u_3'\right\rangle\right]\mathrm{d}t,\\ \mathrm{d}\bar{u}_2 &= \left[\left(\lambda_3\bar{u}_1-d_2\bar{u}_2-\lambda_1\bar{u}_3\right)+B_2\bar{u}_1\bar{u}_3+B_2\left\langle u_1'u_3'\right\rangle\right]\mathrm{d}t,\\ \mathrm{d}\bar{u}_3 &= \left[\left(-\lambda_2\bar{u}_1+\lambda_1\bar{u}_2-d_3\bar{u}_3\right)+B_3\bar{u}_1\bar{u}_2+B_3\left\langle u_1'u_2'\right\rangle\right]\mathrm{d}t,\end{aligned} \tag{B.7}$$

where we use $\langle\cdot\rangle$ to represent the expectation. Correspondingly, the stochastic fluctuation $\mathbf{u}'=(u_1',u_2',u_3')^T$ of the triad model satisfies the following set of SDEs

$$\begin{aligned}\mathrm{d}u_1' &= \left[\left(-d_1u_1'-\lambda_3u_2'+\lambda_2u_3'\right)+B_1\left(\bar{u}_2u_3'+\bar{u}_3u_2'\right)+B_1\left(u_2'u_3'-c_1\right)\right]\mathrm{d}t+\sigma_1\mathrm{d}W_1,\\ \mathrm{d}u_2' &= \left[\left(\lambda_3u_1'-d_2u_2'-\lambda_1u_3'\right)+B_2\left(\bar{u}_1u_3'+\bar{u}_3u_1'\right)+B_2\left(u_1'u_3'-c_2\right)\right]\mathrm{d}t+\sigma_2\mathrm{d}W_2,\\ \mathrm{d}u_3' &= \left[\left(-\lambda_2u_1'+\lambda_1u_2'-d_3u_3'\right)+B_3\left(\bar{u}_1u_2'+\bar{u}_2u_1'\right)+B_3\left(u_1'u_2'-c_3\right)\right]\mathrm{d}t+\sigma_3\mathrm{d}W_3.\end{aligned} \tag{B.8}$$

Above, the stochastic equations are coupled with the cross-covariances $\mathbf{c} = (c_1, c_2, c_3)^T$ that satisfy the following statistical equations

$$\begin{aligned}
\mathrm{d}c_1 =&\big[-(d_2+d_3)c_1 + \lambda_3 c_2 - \lambda_2 c_3 + \lambda_1(r_2 - r_3) \\
&+(B_2\bar{u}_1 r_3 + B_3\bar{u}_1 r_2) + (B_2\bar{u}_3 c_2 + B_3\bar{u}_2 c_3) + (B_2\langle u_1' u_3'^2\rangle + B_3\langle u_1' u_2'^2\rangle)\big]\mathrm{d}t, \\
\mathrm{d}c_2 =&\big[-\lambda_3 c_1 - (d_1+d_3)c_2 + \lambda_1 c_3 + \lambda_2(r_3 - r_1) \\
&+(B_1\bar{u}_2 r_3 + B_3\bar{u}_2 r_1) + (B_1\bar{u}_3 c_1 + B_3\bar{u}_1 c_3) + (B_1\langle u_2' u_3'^2\rangle + B_3\langle u_1'^2 u_2'\rangle)\big]\mathrm{d}t, \\
\mathrm{d}c_3 =&\big[\lambda_2 c_1 - \lambda_1 c_2 - (d_1+d_2)c_3 + \lambda_3(r_1 - r_2) \\
&(B_1\bar{u}_3 r_2 + B_2\bar{u}_3 r_1) + (B_1\bar{u}_2 c_1 + B_2\bar{u}_1 c_2) + (B_1\langle u_2'^2 u_3'\rangle + B_2\langle u_1'^2 u_3'\rangle)\big]\mathrm{d}t.
\end{aligned} \tag{B.9}$$

And similarly, the statistical equations for the variances $\mathbf{r} = (r_1, r_2, r_3)^T$ satisfy the following equations

$$\begin{aligned}
\mathrm{d}r_1 =&2\big[(-d_1 r_1 + \lambda_2 c_2 - \lambda_3 c_3) + B_1(\bar{u}_2 c_2 + \bar{u}_3 c_3) + B_1\langle u_1' u_2' u_3'\rangle + \sigma_1^2\big]\mathrm{d}t, \\
\mathrm{d}r_2 =&2\big[(-\lambda_1 c_1 - d_2 r_2 + \lambda_3 c_3) + B_2(\bar{u}_1 c_1 + \bar{u}_3 c_3) + B_2\langle u_1' u_2' u_3'\rangle + \sigma_2^2\big]\mathrm{d}t, \\
\mathrm{d}r_3 =&2\big[(\lambda_1 c_1 - \lambda_2 c_2 - d_3 r_3) + B_3(\bar{u}_1 c_1 + \bar{u}_2 c_2) + B_3\langle u_1' u_2' u_3'\rangle + \sigma_3^2\big]\mathrm{d}t.
\end{aligned} \tag{B.10}$$

References

[1] U. Frisch, Turbulence: The Legacy of AN Kolmogorov, Cambridge university press, 1995.
[2] A. Majda, X. Wang, Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows, Cambridge University Press, 2006.
[3] J. Pedlosky, Geophysical Fluid Dynamics, Springer Science & Business Media, 2013.
[4] T.N. Palmer, Stochastic weather and climate models, Nat. Rev. Phys. 1 (7) (2019) 463–471.
[5] W. Cousins, T.P. Sapsis, Quantification and prediction of extreme events in a one-dimensional nonlinear dispersive wave model, Phys. D 280 (2014) 48–58.
[6] S. Tong, E. Vanden-Eijnden, G. Stadler, Extreme event probability estimation using PDE-constrained optimization and large deviation theory, with application to tsunamis, Commun. Appl. Math. Comput. Sci. 16 (2) (2021) 181–225.
[7] D. Qi, E. Vanden-Eijnden, Anomalous statistics and large deviations of turbulent water waves past a step, AIP Adv. 12 (2) (2022) 025016.
[8] M. Leutbecher, T.N. Palmer, Ensemble forecasting, J. Comput. Phys. 227 (7) (2008) 3515–3539.
[9] S.C. Surace, A. Kutschireiter, J.P. Pfister, How to avoid the curse of dimensionality: scalability of particle filters with and without importance weights, SIAM Rev. 61 (1) (2019) 79–91.
[10] Y. Gao, T. Li, X. Li, J.G. Liu, Transition path theory for Langevin dynamics on manifolds: optimal control and data-driven solver, Multiscale Model. Simul. 21 (1) (2023) 1–33.
[11] K. Law, A. Stuart, K. Zygalakis, Data Assimilation, Cham, Switzerland: Springer 214 (2015) 52.
[12] E. Cleary, A. Garbuno-Inigo, S. Lan, T. Schneider, A.M. Stuart, Calibrate, emulate, sample, J. Comput. Phys. 424 (2021) 109716.
[13] E. Bach, T. Colonius, I. Scherl, A. Stuart, Filtering dynamical systems using observations of statistics, Chaos 34 (3) (2024).
[14] A.J. Majda, D. Qi, Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems, SIAM Rev. 60 (3) (2018) 491–549.
[15] D. Qi, J.-G. Liu, High-order moment closure models with random batch method for efficient computation of multiscale turbulent systems, Chaos 33 (10) (2023) 103133.
[16] D. Qi, J. Harlim, A data-driven statistical-stochastic surrogate modeling strategy for complex nonlinear non-stationary dynamics, J. Comput. Phys. 485 (2023) 112085.
[17] D. Qi, J.G. Liu, Coupled stochastic-statistical equations for filtering multiscale turbulent systems, arXiv preprint arXiv:2407.04881 (2024).
[18] M. Lesieur, Turbulence in Fluids: Stochastic and Numerical Modelling, 488, Nijhoff Boston, MA, 1987.
[19] E. Kalnay, Atmospheric Modeling, Data Assimilation and Predictability, Cambridge university press, 2003.
[20] M.A. Mohamad, T.P. Sapsis, Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems, Proc. Natl. Acad. Sci. 115 (44) (2018) 11138–11143.
[21] J. Xiong, An Introduction to Stochastic Filtering Theory, 18, OUP Oxford, 2008.
[22] S. Reich, C. Cotter, Probabilistic Forecasting and Bayesian Data Assimilation, Cambridge University Press, 2015.
[23] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. 99 (C5) (1994) 10143–10162.
[24] P.L. Houtekamer, H.L. Mitchell, Data assimilation using an ensemble Kalman filter technique, Mon. Weather Rev. 126 (3) (1998) 796–811.
[25] P. Del Moral, Nonlinear filtering: interacting particle resolution, Comptes Rendus de l'Académie des Sciences-Series I-Mathematics 325 (6) (1997) 653–658.
[26] A. Doucet, N. De Freitas, N.J. Gordon, et al., Sequential Monte Carlo Methods in Practice, 1, Springer, 2001.
[27] E. Calvello, S. Reich, A.M. Stuart, Ensemble Kalman methods: a mean field perspective, Acta Numer. 34 (2025) 123–291.
[28] T. Yang, H.A.P. Blom, P.G. Mehta, The continuous-discrete time feedback particle filter, in: 2014 American Control Conference, IEEE, 2014, pp. 648–653.
[29] S. Pathiraja, S. Reich, W. Stannat, McKean–vlasov SDEs in nonlinear filtering, SIAM J. Control Optim. 59 (6) (2021) 4188–4215.
[30] M. Pulido, P.J. van Leeuwen, Sequential Monte Carlo with kernel embedded mappings: the mapping particle filter, J. Comput. Phys. 396 (2019) 400–415.
[31] C.C. Hu, P.J. van Leeuwen, A particle flow filter for high-dimensional system applications, Q. J. R. Meteorolog. Soc. 147 (737) (2021) 2352–2374.
[32] I. Grooms, G. Robinson, A hybrid particle-ensemble Kalman filter for problems with medium nonlinearity, PLoS ONE 16 (3) (2021) e0248266.
[33] J. Bröcker, D. Engster, U. Parlitz, Probabilistic evaluation of time series models: a comparison of several approaches, Chaos 19 (4) (2009).
[34] G.A. Gottwald, S. Reich, Supervised learning from noisy observations: combining machine-learning techniques with data assimilation, Phys. D 423 (2021) 132911.
[35] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, arXiv preprint arXiv:2011.13456 (2020).
[36] J.L. Wu, K. Kashinath, A. Albert, D. Chirila, H. Xiao, et al., Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems, J. Comput. Phys. 406 (2020) 109209.
[37] A. Campbell, Y. Shi, T. Rainforth, A. Doucet, Online variational filtering and parameter learning, Adv. Neural Inf. Process. Syst. 34 (2021) 18633–18645.
[38] J. Lu, Y. Wang, Guidance for twisted particle filter: a continuous-time perspective, arXiv preprint arXiv:2409.02399 (2024).
[39] C. Snyder, T. Bengtsson, P. Bickel, J. Anderson, Obstacles to high-dimensional particle filtering, Mon. Weather Rev. 136 (12) (2008) 4629–4640.
[40] G.A. Gottwald, A.J. Majda, A mechanism for catastrophic filter divergence in data assimilation for sparse observation networks, Nonlinear Process. Geophys. 20 (5) (2013) 705–712.
[41] A.J. Majda, D. Qi, Linear and nonlinear statistical response theories with prototype applications to sensitivity analysis and statistical control of complex turbulent dynamical systems, Chaos 29 (10) (2019).
[42] R. Salmon, Lectures on Geophysical Fluid Dynamics, Oxford University Press, USA, 1998.
[43] D.R. Nicholson, D.R. Nicholson, Introduction to Plasma Theory, 1, Wiley New York, 1983.

[44] C. Graham, T.G. Kurtz, S. Méléard, P.E. Protter, M. Pulvirenti, D. Talay, S. Méléard, Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and boltzmann models, Probabilistic Models for Nonlinear Partial Differential Equations: Lectures given at the 1st Session of the Centro Internazionale Matematico Estivo (CIME) held in Montecatini Terme, Italy, May 22–30, 1995 (1996) 42–95.
[45] D. Qi, Unambiguous models and machine learning strategies for anomalous extreme events in turbulent dynamical system, Entropy 26 (6) (2024) 522.
[46] R.S. Liptser, A.N. Shiryaev, Statistics of Random Processes II: Applications, 6, Springer Science & Business Media, 2013.
[47] R.F. Curtain, Infinite-dimensional filtering, SIAM J. Control 13 (1) (1975) 89–104.
[48] D. Qi, J.-G. Liu, A random batch method for efficient ensemble forecasts of multiscale turbulent systems, Chaos 33 (2) (2023) 023113.
[49] K. Oelschlager, A martingale approach to the law of large numbers for weakly interacting stochastic processes, Ann. Probab. 12 (2) (1984) 458–479.
[50] A. Bain, D. Crisan, Fundamentals of Stochastic Filtering, 3, Springer, 2009.
[51] A. Gluhovsky, E. Agee, An interpretation of atmospheric low-order models, J. Atmos. Sci. 54 (6) (1997) 768–773.
[52] A. Majda, I. Timofeyev, E. Vanden-Eijnden, A priori tests of a stochastic mode reduction strategy, Phys. D 170 (3–4) (2002) 206–252.
[53] E.N. Lorenz, Predictability: a problem partly solved, in: Proc. Seminar on Predictability, 1, Reading, 1996, pp. 1–18.
[54] D. Qi, J. Harlim, Machine learning-based statistical closure models for turbulent dynamical systems, Philos. Trans. Royal Soc. A 380 (2229) (2022) 20210205.
[55] Z. Wang, N. Chen, D. Qi, A closed-form nonlinear data assimilation algorithm for multi-layer flow fields, arXiv preprint arXiv:2412.11042 (2024).
[56] D. Qi, A.J. Majda, Rigorous statistical bounds in uncertainty quantification for one-layer turbulent geophysical flows, J. Nonlinear Sci. 28 (2018) 1709–1761.
[57] X. Bao, C. Liu, Y. Wang, A deterministic–particle–based scheme for micro-macro viscoelastic flows, J. Comput. Phys. 522 (2025) 113589.
[58] M. Pulido, P.J. van Leeuwen, D.J. Posselt, Kernel embedded nonlinear observational mappings in the variational mapping particle filter, in: International Conference on Computational Science, Springer, 2019, pp. 141–155.
[59] C.C. Hu, P.J. van Leeuwen, J.L. Anderson, An implementation of the particle flow filter in an atmospheric model, Mon. Weather Rev. 152 (10) (2024) 2247–2264.
[60] A.J. Majda, Introduction to Turbulent Dynamical Systems in Complex Systems, Springer, 2016.