Chaos
An Interdisciplinary Journal of Nonlinear Science

RESEARCH ARTICLE | OCTOBER 23 2023

High-order moment closure models with random batch method for efficient computation of multiscale turbulent systems FREE

Special Collection: Nonlinear Model Reduction From Equations and Data

Di Qi ; Jian-Guo Liu

Check for updates

Chaos 33, 103133 (2023)
https://doi.org/10.1063/5.0160057

View Online
Export Citation
CrossMark

29 October 2023 08:59:52

AIP Publishing

High-order moment closure models with random batch method for efficient computation of multiscale turbulent systems

Cite as: Chaos 33, 103133 (2023); doi: 10.1063/5.0160057
Submitted: 30 May 2023 · Accepted: 2 October 2023 ·
Published Online: 23 October 2023

View Online Export Citation CrossMark

Di Qi[1,a] and Jian-Guo Liu[2,b]

AFFILIATIONS

[1]Department of Mathematics, Purdue University, 150 North University Street, West Lafayette, Indiana 47907, USA
[2]Department of Mathematics and Department of Physics, Duke University, Durham, North Carolina 27708, USA

Note: This paper is part of the Focus Issue on Nonlinear Model Reduction From Equations and Data.
[a]**Author to whom correspondence should be addressed:** qidi@purdue.edu
[b]**Electronic mail:** jliu@phy.duke.edu

ABSTRACT

We propose a high-order stochastic–statistical moment closure model for efficient ensemble prediction of leading-order statistical moments and probability density functions in multiscale complex turbulent systems. The statistical moment equations are closed by a precise calibration of the high-order feedbacks using ensemble solutions of the consistent stochastic equations, suitable for modeling complex phenomena including non-Gaussian statistics and extreme events. To address challenges associated with closely coupled spatiotemporal scales in turbulent states and expensive large ensemble simulation for high-dimensional systems, we introduce efficient computational strategies using the random batch method (RBM). This approach significantly reduces the required ensemble size while accurately capturing essential high-order structures. Only a small batch of small-scale fluctuation modes is used for each time update of the samples, and exact convergence to the full model statistics is ensured through frequent resampling of the batches during time evolution. Furthermore, we develop a reduced-order model to handle systems with really high dimensions by linking the large number of small-scale fluctuation modes to ensemble samples of dominant leading modes. The effectiveness of the proposed models is validated by numerical experiments on the one-layer and two-layer Lorenz '96 systems, which exhibit representative chaotic features and various statistical regimes. The full and reduced-order RBM models demonstrate uniformly high skill in capturing the time evolution of crucial leading-order statistics, non-Gaussian probability distributions, while achieving significantly lower computational cost compared to direct Monte-Carlo approaches. The models provide effective tools for a wide range of real-world applications in prediction, uncertainty quantification, and data assimilation.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0160057

29 October 2023 08:59:52

Understanding and predicting key physical processes in complex turbulent systems remains a grand challenge in natural and engineering systems. A novel high-order moment closure strategy is developed addressing the nonlinear multiscale coupling mechanism as a characteristic feature in high-dimensional turbulent systems. Efficient computational algorithms are developed that significantly reduce the high computational cost in achieving accurate ensemble solutions. The central statistical quantities involving multiscale interactions and highly non-Gaussian features are fully recovered with accuracy through a random batch approximation of the wide spectrum of fluctuation states. The proposed strategy shows potential to facilitate the development of new effective methods in uncertainty quantification and data assimilation across a wide variety of complex turbulent systems.

I. INTRODUCTION

Turbulent dynamical systems encountered in science and engineering[1–4] exhibit distinguished characteristics including a high-dimensional state space with multiple spatiotemporal scales and strong internal instabilities.[5,6] Forecasting the intricate behaviors of such complex systems poses a significant challenge in uncertainty quantification and data assimilation problems.[7–10] One key aspect involves accurately quantifying the multiscale interaction between

Published under an exclusive license by AIP Publishing

the large-scale mean state and the many interacting small-scale fluctuations induced by internal instability. The interplay between multiscale coupling and instability gives rise to a diverse array of complex phenomena, such as bursting extreme structures and skewed non-Gaussian distributions.[11–14] Small randomness in initial conditions and external stochastic effects is also rapidly amplified and redistributed along the spectrum as the model evolves in time. Capturing these unique features with efficient algorithms remains a central issue in many practical problems.[15–18] For example, in data assimilation, accurate prediction of the mean and covariance statistics is essential, while the evolution of these low-order moments is closely linked to higher-order feedback due to the nonlinear coupling. This requires the accurate and efficient quantification for the extreme outliers and the related non-Gaussian statistics.

In developing efficient algorithms for accurate statistical prediction of complex turbulent systems featuring multiscale interactions and strong nonlinearity, a probabilistic approach is usually needed to quantify the uncertainty utilizing a probability density function (PDF) of the model states. Ensemble forecasting, through a Monte Carlo (MC) type method, is commonly used to estimate the PDF evolution by independently sampling an ensemble of trajectories from an initial distribution.[19] Ensemble-based approaches have been extensively applied to address uncertainties arising from various sources in real-world problems such as weather and climate forecast.[4,20,21] However, achieving accurate numerical prediction is hampered by the prohibitively high computational cost attributed to the strongly coupled nonlinear interactions among different scales in a high-dimensional space, known as "curse of dimensionality."[22,23] With insufficient samples, ensemble approximation often suffers from the collapse of samples, leading to an inadequate coverage of the entire probability measure in the high-dimensional phase space. Various strategies have been devised to increase the effective ensemble size and sufficiently characterize the probability distribution.[24,25] Parameterization is often used for efficient approximation of the effects from unresolved small-scale processes using information from the resolved large-scale states.[8,26–29] In the case of high-dimensional turbulent systems, model errors are significantly amplified by internal instability and careful calibrations of the feedbacks from the large number of unresolved processes are often required. Consequently, this often leads to an inherent difficulty for generalization to high-dimensional and strongly turbulent systems in realistic applications.

In this paper, we aim to develop a systematic modeling and computational strategy for statistical prediction and data assimilation of turbulent systems under a unified framework. The proposed formulation (1) is applicable to complex spatially extended nonlinear dynamical systems widely studied in many fields.[4,6,30–32] In particular, we construct a coupled stochastic–statistical model (7) that is suitable for both efficient prediction of leading-order statistics (in the statistical equations) and explicit quantification of higher-order non-Gaussian features (using the stochastic equations) at the same time. The stochastic and statistical equations are seamlessly coupled through the nonlinear interaction terms and using an essential relaxation term for consistency. The statistical equations admit a hierarchical structure requiring a closure form for the higher-order moment feedback induced by the quadratic nonlinear term. Instead of the parameterization methods that usually require the exhausting procedure of model calibration, we close the moment equations by explicitly modeling the higher-order feedback through the ensemble solution of the stochastic equations. This high-order moment closure through the stochastic–statistical model provides a statistically consistent formulation that is able to correctly represent the crucial features involving non-Gaussian and nonlinear phenomena.

The computational demand for solving the fully coupled system (7) remains substantial when the state variable resides in a high-dimensional space. For example, for a system with K-dimensional state space, one time step update of the covariance dynamics requires the computational cost of $O\left(K^4\right)$ by evolving $K \times K$ matrix-valued differential equations with K^2 quadratic coupling terms in each entry. The ensemble simulation for the stochastic dynamics further requires the computational cost of $O\left(MK^3\right)$ by updating the K stochastic coefficients using M samples (usually with exponential dependent on K). To overcome this issue, we use the effective random batch method (RBM)[33,34] for efficient computation of a large number of high-order coupling terms reaching a much reduced computational cost. The proposed method generalizes the idea developed for simple mean-fluctuation systems in Ref. 35 and provides an efficient computational strategy for the more universal formulation (7) valid for statistical prediction in a large group of problems. Using the RBM approximation, the high-dimensional modes are divided into small batches randomly drawn in each time updating interval. Then, the nonlinear interaction is solely computed among a small subset of modes within one batch so that the computational cost is maintained at a low level. During the time evolution, the batches are frequently resampled at the start of each time updating step. The accuracy of the RBM approximation is preserved by the ergodicity of the fast mixing fluctuation modes, which implies that the accumulated effect of the average among fractional fluctuation modes inside a small random batch at each time step has an equivalent overall statistical contribution as the original high-order moment feedback in the full model. Thus, accurate statistical feedback is guaranteed in the sense of consistent time-averaged statistics. The resulting RBM model (8) offers significant computational reduction with a cost of $O\left(p^2K^2\right)$ for the covariance equations and $O\left(M_1p^2K\right)$ for the ensemble forecast, where a considerable smaller ensemble size $M_1 \ll M$ is sufficient only requiring sampling the batch subspace consisting of $p = O(1)$ modes in one batch. The convergence of the mean and variance using RBM approximation is proved through detailed error estimates dependent on the discrete time step size. Furthermore, for really high-dimensional systems with an extended spectrum, a reduced-order model (10) is proposed to further reduce the computational cost to $O\left(M_1\left(K_1^3 + p^3\right)\right)$, making it independent of the full dimension K by focusing on the first K_1 leading modes and using RBM approximation for the rest of $K - K_1$ small-scale fluctuation modes.

The performance of the stochastic–statistical model with RBM approximations is examined under the one-layer and two-layer Lorenz '96 (L-96) systems.[36,37] The Lorenz '96 systems have been widely used as prototype models for the atmosphere involving rich chaotic phenomena with direct links to realistic systems.[15,25,38,39] In particular, the L-96 systems maintain a slow decay variance spectrum involving a large number of unstable modes (as illustrated in Figs. 1 and 7), setting up a very challenging testing case for effective ensemble prediction. The direct MC approach requires a very large

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

ensemble size of order $O(10^5)$ to resolve the highly non-Gaussian statistics. The strong internal instability leads to additional problems for easy divergence away from the final equilibrium state (shown in Fig. 2). Using the efficient RBM model, it is found that a small sample size of $O(100)$ is sufficient to fully recover the evolutions of statistical mean and variance as well as the non-Gaussian PDFs in the 40-dimensional one-layer L-96 system (see Figs. 4 and 5). In the genuinely high-dimensional two-layer L-96 system with full dimension $K = 264$. The RBM model becomes especially effective to capture the highly non-Gaussian statistics using at most $M_1 = 500$ samples (see Figs. 10 and 11). The computational cost is further reduced in the reduced-order model focusing on the leading $K_1 = 8$ large-scale modes enabling an even smaller ensemble for all the small-scale modes.

In the remainder part of this paper, we introduce the general formulation of the stochastic–statistical model for multiscale turbulent systems using high-order moment closure in Sec. II. The efficient algorithms for solving the coupled high-dimensional equations using the RBM approximation for ensemble prediction are developed in Sec. III together with the theoretical convergence analysis of the scheme. The performance of the methods is evaluated on the concrete examples of the Lorenz '96 systems in Sec. IV. The paper is closed with a summary and discussions on future directions in Sec. V.

II. A STATISTICALLY CONSISTENT MODELING FRAMEWORK FOR MULTISCALE TURBULENT SYSTEMS

Turbulent dynamical systems are characterized by multiscale nonlinear interactions, which redistribute energy across a broad spectrum of stable and unstable modes, ultimately leading to a complicated statistical equilibrium. The abstract mathematical formulation of complex turbulent systems projected from the continuous PDE systems can be introduced in the following canonical equations about the state variable $\mathbf{u} \in \mathbb{R}^d$ in a high-dimensional phase space (with $d \gg 1$)

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \Lambda\mathbf{u} + B(\mathbf{u},\mathbf{u}) + \mathbf{F}(t) + \boldsymbol{\sigma}(t)\dot{\mathbf{W}}(t). \tag{1}$$

The model state starts from an initial distribution $\mathbf{u}(0) \sim \mu_0(\mathbf{u})$ representing initial uncertainty. On the right-hand side of Eq. (1), the first component, $\Lambda = L - D$, represents linear dispersion and dissipation effects, where the dispersion $L^* = -L$ is an energy-conserving skew-symmetric operator; and the dissipation $D < 0$ is a negative-definite operator. The model (1) emphasizes the important role of nonlinear interactions in a bilinear quadratic form, $B(\mathbf{u},\mathbf{u})$. This typical structure of nonlinear interactions is inherited from a discretization of the continuous full system (for example, a spectral projection of the nonlinear advection in fluid models). The nonlinear interaction ensures the energy conservation invariance such that $\mathbf{u} \cdot B(\mathbf{u},\mathbf{u}) \equiv 0$ with the inner product defined according to the conserved quantity. In addition, the system is subject to external forcing effects that are decomposed into a deterministic component, $\mathbf{F}(t)$, and a stochastic component represented by a Gaussian random process, $\boldsymbol{\sigma}(t)\dot{\mathbf{W}}(t)$, used to model the unresolved processes.

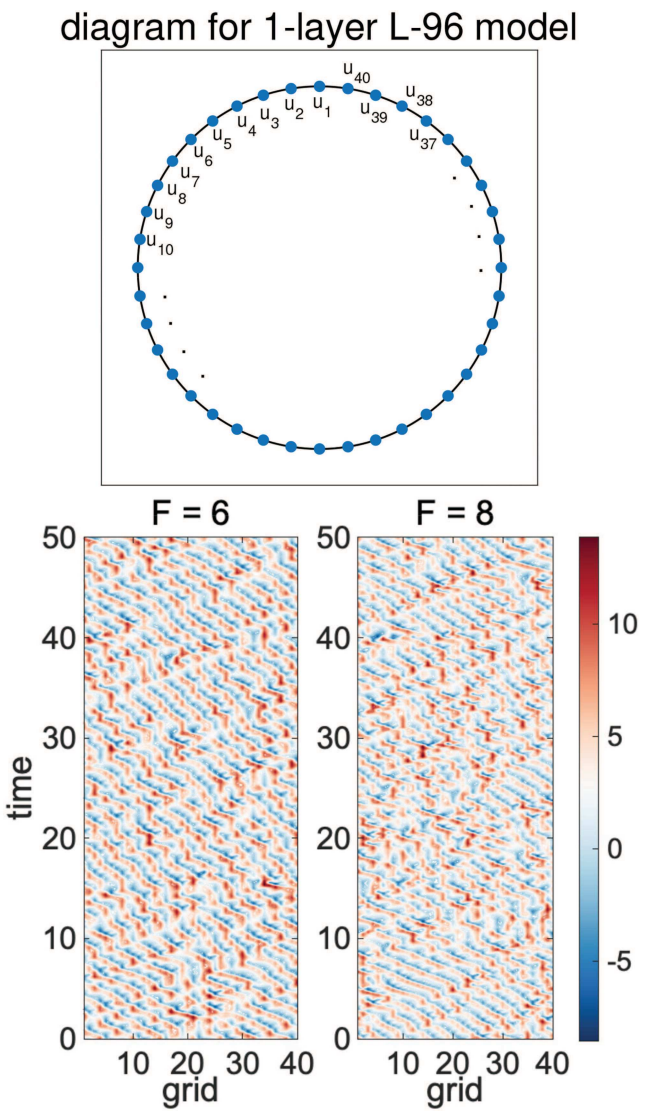

FIG. 1. Illustration of the model structure and typical solutions of the one-layer L-96 system (18).

The evolution of the model state $\mathbf{u}$ depends on sensitivity to the randomness in initial conditions and stochastic forcing effects. These uncertainties will be amplified in time by the inherent internal instability due to the nonlinear coupling term in (1). The associated Fokker–Planck equation (FPE),[40]

$$\frac{\partial p_t}{\partial t} = \mathcal{L}_{\mathrm{FP}} p_t := -\mathrm{div}_{\mathbf{u}}\left[\Lambda\mathbf{u} + B(\mathbf{u},\mathbf{u}) + \mathbf{F}\right] p_t + \frac{1}{2}\mathrm{div}_{\mathbf{u}}\nabla\left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T p_t\right), \tag{2}$$

describes the time evolution of the probability density function (PDF) $p_t(\mathbf{u}) = e^{t\mathcal{L}_{\mathrm{FP}}(\mathbf{u})}\mu_0$ starting from an initial distribution

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

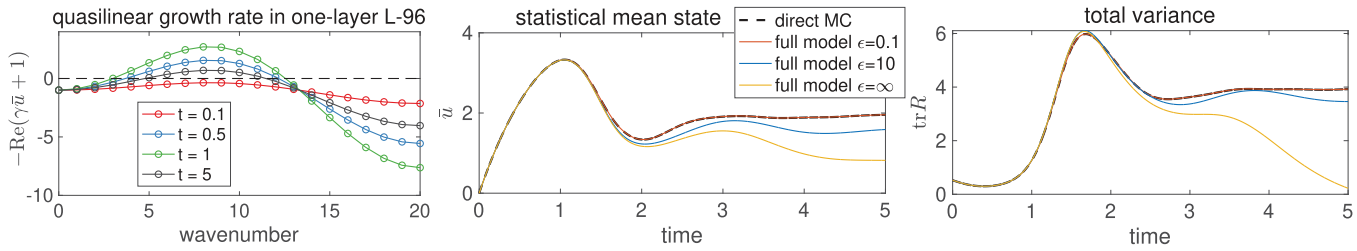

FIG. 2. Effects of instability in the one-layer L-96 system with $F = 6$. Left: quasilinear growth rate in each wavenumber $-\mathrm{Re}\left(\gamma_k^*\bar{u}+1\right)$ at several time instants; right: direct solutions from the full model with different relaxation strengths ϵ.

$p_{t=0}(\mathbf{u}) = \mu_0(\mathbf{u})$. However, it remains a challenging task for directly solving the FPE (2) as a high-dimensional PDE system. As an alternative approach, ensemble forecast by tracking the Monte-Carlo (MC) solutions estimates the model statistics through empirical averages among a group of independent samples drawn from the initial distribution $\mathbf{u}^{(i)}(0) \sim \mu_0(\mathbf{u}), i = 1, \ldots, M$ at the starting time $t = 0$. The PDF solution $p_t(\mathbf{u})$ and the associated statistical expectation of any function $\varphi(\mathbf{u})$ at each time instant $t > 0$ are then approximated by the empirical ensemble representation of the M samples, that is,

$$p_t(\mathbf{u}) \simeq p_t^{\mathrm{MC}}(\mathbf{u}) := \frac{1}{M}\sum_{i=1}^{M}\delta\left(\mathbf{u}-\mathbf{u}^{(i)}(t)\right),$$
$$\langle\varphi(\mathbf{u})\rangle_{p_t} \simeq \langle\varphi\rangle_{p_t^{\mathrm{MC}}} = \frac{1}{M}\sum_{i=1}^{M}\varphi\left(\mathbf{u}^{(i)}(t)\right), \tag{3}$$

where δ is the Dirac delta function and $\langle\cdot\rangle_p$ is the expectation about the probability measure p. Still, the *curse of dimensionality*[23,41] arises in systems of even moderate dimension d since the model errors grow significantly as the system dimension increases, while only a small ensemble size M is allowed in practical numerical methods due to the limited computational resources. Clearly, efficient strategies and algorithms are still needed to effectively reduce the computational cost and maintain high accuracy in sampling the high-dimensional systems using a small number of samples.

Remark. Many complex turbulent systems from nature and engineering can be categorized into the abstract mathematical framework in (1). The high-dimensional state $\mathbf{u}$ can be viewed as a finite-dimensional truncation of the corresponding continuous field with sufficient numerical resolution. One major group of examples comes from the fluid flows including the Navier–Stokes equation and turbulence at high Reynolds number[5,30] and applications to the geophysical models in coupled atmosphere and ocean systems involving rotation, stratification, and topography[1,2,4] and controlled fusion in magnetically confined plasma systems.[31,42] In particular, we will consider the prototype Lorenz '96 systems in (18) and (21)[36,37,43] that admit all representative dynamical structures in (1) as the main test model in this paper.

A. Statistical and stochastic formulations for multiscale systems

One major difficulty in complex turbulent systems is the fully coupled nonlinear interactions across scales. The multiscale interactions involve a large-scale mean state, which can destabilize the smaller scales, while the excited fluctuation energy contained in numerous small-scale modes can inversely impact the development of the coherent structure at the largest scale. Thus, disregarding contributions from small-scale modes through a simple high wavenumber truncation is not a viable approach. To address this central issue of coupled interactions with mixed scales, we start with a mean-fluctuation decomposition for the model state $\mathbf{u}$ so that interactions between different scales can be identified in detail. To achieve this, we view $\mathbf{u}$ as a random field (denoted by ω due to randomness in initial state and stochastic forcing) and separate it into the composition of a statistical mean state $\bar{\mathbf{u}}$ and a wide spectrum of stochastic fluctuations $\mathbf{u}'$ in a finite K-dimensional projected representation under a fixed-in-time, orthonormal basis $\{\mathbf{v}_k\}_{k=1}^K$ (usually with $K = d$ for the full model and $K < d$ for the reduced-order model)

$$\mathbf{u}(t;\omega) = \bar{\mathbf{u}}(t) + \mathbf{u}'(t;\omega) = \bar{\mathbf{u}}(t) + \sum_{k=1}^{K} Z_k(t;\omega)\,\mathbf{v}_k. \tag{4}$$

Here, $\bar{\mathbf{u}} = \langle\mathbf{u}\rangle_{p_t}$ represents the *statistical mean field* usually capturing the dominant large-scale structure; and $Z_k(t;\omega)$ is the *stochastic coefficient* measuring the uncertainty in multiscale fluctuation processes $\mathbf{u}'$ projected on the eigenmode $\mathbf{v}_k$. The state decomposition (4) provides a convenient way to identify different components in the multiscale interactions, and thus can be used to derive new effective multiscale models.

One way to avoid the high computational cost in directly solving the FPE (2) as well as the large ensemble MC simulation of the full SDE (1) for the probability distribution $p_t(\mathbf{u})$ is to seek a hierarchical statistical description of its moments $\langle\varphi(\mathbf{u})\rangle_{p_t}$ as the expectation with respect to the time-dependent probability measure p_t. In most situations, the primary interest lies in tracking the time evolution of the leading moments quantifying the most essential statistical characteristics. We can first derive the dynamics for the mean state $\bar{\mathbf{u}} = \langle\mathbf{u}\rangle_{p_t}$ and the covariance among fluctuation modes $R_{kl} = \langle Z_k Z_l^*\rangle_{p_t}$ governed by the following set of *deterministic*

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

statistical equations:

$$\frac{\mathrm{d}\bar{\mathbf{u}}}{\mathrm{d}t} = \Lambda\bar{\mathbf{u}} + B(\bar{\mathbf{u}},\bar{\mathbf{u}}) + \sum_{k,l=1}^{K} R_{kl} B(\mathbf{v}_k,\mathbf{v}_l) + \mathbf{F}, \tag{5a}$$

$$\begin{aligned}\frac{\mathrm{d}R_{kl}}{\mathrm{d}t} &= \sum_{m=1}^{K} L_{v,km}(\bar{\mathbf{u}}) R_{ml} + R_{km} L^{*}_{v,kl}(\bar{\mathbf{u}}) + Q_{\sigma,kl} \\ &\quad + \sum_{m,n=1}^{K} \langle Z_m Z_n Z_l\rangle_{p_t} B(\mathbf{v}_m,\mathbf{v}_n)\cdot\mathbf{v}_k \\ &\quad + \langle Z_m Z_n Z_k\rangle_{p_t} B(\mathbf{v}_m,\mathbf{v}_n)\cdot\mathbf{v}_l,\end{aligned} \tag{5b}$$

for all the wavenumbers $1 \le k,l \le K$. In (5b), the operator $L_{v,kl} = [\Lambda\mathbf{v}_l + B(\bar{\mathbf{u}},\mathbf{v}_l) + B(\mathbf{v}_l,\bar{\mathbf{u}})]\cdot\mathbf{v}_k$ characterizes quasilinear coupling between the statistical mean and stochastic modes; the positive-definite operator $Q_{\sigma,kl} = \sum_m (\mathbf{v}_k\cdot\boldsymbol{\sigma}_m)(\boldsymbol{\sigma}_m\cdot\mathbf{v}_l)$ expresses energy injection from the stochastic forcing. Notably, the nonlinear flux term involving all the third-order moments $\langle Z_m Z_n Z_k\rangle_{p_t}$ enters the equation for the second-order covariance R_{kl} describing nonlinear energy exchanges among fluctuation modes, ending up with a still unclosed set of equations.

Accordingly, the stochastic coefficients $\{Z_k\}_{k=1}^{K}$ in the decomposition (4) satisfy the associated *stochastic fluctuation equations*

$$\begin{aligned}\frac{\mathrm{d}Z_k}{\mathrm{d}t} &= \sum_{m=1}^{K} L_{v,km}(\bar{\mathbf{u}}) Z_m + \sigma(t)\dot{\mathbf{W}}(t)\cdot\mathbf{v}_k \\ &\quad + \sum_{m,n=1}^{K} (Z_m Z_n - R_{mn}) B(\mathbf{v}_m,\mathbf{v}_n)\cdot\mathbf{v}_k.\end{aligned} \tag{6}$$

The above equation is achieved by simply projecting the original equation (1) on each fluctuation mode $\mathbf{v}_k$ and removing the mean dynamics (5a). The second-order moment equation (5b) is then derived by applying Itô's formula to $\varphi(Z_k) = |Z_k|^2$ using the stochastic equation (6). Therefore, the above statistical and stochastic formulations are consistent for the evolution of uncertainty in multiscale fluctuations. The detailed derivation of Eqs. (5) and (6) from first principle can be found in Refs. 6 and 8.

Both the dynamical moment representation (5) and its stochastic counterpart (6) have their respective advantages, and they also suffer several difficulties when applied to resolve the key statistical quantities. The statistical moment equations (5) are easier to compute with its deterministic dynamics, while such hierarchical equations lead to a non-closed system of infinite-dimensional ODEs as each lower-order moment equation is coupled to the next higher-order moment. On the other hand, the stochastic equations (6) provide a closed formulation to include all the higher-order information. However, direct simulation of the SDE requires an MC approach of a large sample size exponentially dependent on the state dimension. In addition, the computational cost of both statistical and stochastic models remains prohibitive for the coupled high-dimensional systems characterized by an extended wide spectrum of fluctuation modes $K \gg 1$. The situation becomes especially challenging when an ensemble approach is required for accurate state estimation and data assimilation of extreme events represented by the extended PDF tails.

B. A stochastic–statistical closure model with explicit higher-order feedbacks

We introduce a seamless high-order closure model that integrates the statistical equations (5) with the stochastic counterpart (6) to effectively close the original non-closed equations. The resulting *coupled stochastic–statistical equations for the multiscale interacting model* become

$$\frac{\mathrm{d}\bar{\mathbf{u}}}{\mathrm{d}t} = \Lambda\bar{\mathbf{u}} + B(\bar{\mathbf{u}},\bar{\mathbf{u}}) + \sum_{k,l=1}^{K} R_{kl} B(\mathbf{v}_k,\mathbf{v}_l) + \mathbf{F}, \tag{7a}$$

$$\frac{\mathrm{d}Z_k}{\mathrm{d}t} = \sum_{m=1}^{K} L_{v,km}(\bar{\mathbf{u}}) Z_m + \sum_{m,n=1}^{K} \gamma_{mnk}(Z_m Z_n - R_{mn}) + \sigma_k \dot{W}_k, \tag{7b}$$

$$\begin{aligned}\frac{\mathrm{d}R_{kl}}{\mathrm{d}t} &= \sum_{m=1}^{K} L_{v,km}(\bar{\mathbf{u}}) R_{ml} + R_{km} L^{*}_{v,ml}(\bar{\mathbf{u}}) + Q_{\sigma,kl} \\ &\quad + \sum_{m,n=1}^{K} \left[\gamma_{mnk}\langle Z_m Z_n Z_l\rangle_{p_t} + \gamma_{mnl}\langle Z_m Z_n Z_k\rangle_{p_t}\right] \\ &\quad + \epsilon^{-1}\left(\langle Z_k Z_l\rangle_{p_t} - R_{kl}\right),\end{aligned} \tag{7c}$$

with the coupling coefficients $L_{v,km} = [\Lambda\mathbf{v}_m + B(\bar{\mathbf{u}},\mathbf{v}_m) + B(\mathbf{v}_m,\bar{\mathbf{u}})]\cdot\mathbf{v}_k$ and $\gamma_{mnk} = B(\mathbf{v}_m,\mathbf{v}_n)\cdot\mathbf{v}_k$ derived from the original equations. Here, the mean equation (7a) for the leading-order statistics $\bar{\mathbf{u}}$ is kept the same aiming to capture the dominant large-scale mean structures. The covariance equation (7c) for R is closed by computing the expectations of the cubic terms under the probability measure p_t discovered by the stochastic solution $\mathbf{Z}$ from (7b) so that the higher-order moment feedbacks are explicitly modeled. In addition, a relaxation term is added with a small control parameter ϵ to guarantee statistical consistency between the statistical and stochastic equations. This term will become crucial if the system contains strong internal instability (for example, see Fig. 2). Importantly, both the statistical and stochastic equations become indispensable for the modeling of the fully coupled multiscale system: (i) the stochastic equations for the fluctuation modes Z_k are introduced to provide exact closure for the covariance R and (ii) the covariance dynamics for R serves as an auxiliary equation to facilitate the explicit interactions between the mean $\bar{\mathbf{u}}$ and stochastic modes Z_k and can deal with the inherent instability in the turbulent systems.

In developing an effective strategy to compute the high-order expectation in the covariance equation (7c) according to the PDF p_t of the stochastic solutions in (7b), the stochastic equations for the random fluctuation modes Z_k are solved through an ensemble approach using $\mathbf{Z}^{(i)} = \left\{Z_k^{(i)}\right\}$ with sample index $i = 1,\ldots,M$. The higher-order feedbacks are then approximated through the empirical average of the ensemble as in (3), $\langle\varphi(\mathbf{Z})\rangle_{p_t} \approx \frac{1}{M}\sum_{i=1}^{M}\varphi\left(\mathbf{Z}^{(i)}\right)$. Compared with the direct MC approach of the original system (1),

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

the new coupled model (7) adopting the explicitly coupled stochastic–statistical dynamics presenting an attractive equivalent formulation that enjoys the flexibility of developing efficient reduced-order models.

The new formulation provides a desirable framework that is suitable for the development of efficient computational methods and reduced-order models as described in Sec. III B of this paper. It can deal with the inherent difficulties raised in the original formulation with irreducible equations. Instead of adding *ad hoc* approximations for the unresolved higher moments (such as the data-driven model in Ref. 29), the crucial third moments are captured explicitly through the ensemble estimation of the stochastic modes. In addition, we aim to control the computational cost by only running a very small ensemble for limited samples $M_1 \ll M$ without sacrificing accuracy (through the efficient random batch method introduced in Sec. III) compared with the original direct MC approach which demands a large sample size M. This combined framework enables flexible modeling of both key leading-order moments through the statistical equation and achieves accurate non-Gaussian higher-order statistics and fat-tailed PDFs in extreme events. It is also more advantageous than the mean-fluctuation model using only (7a) and (7b) (as proposed in Ref. 35), which often leads to large numerical errors and unstable dynamics due to the inherent instability in the turbulent systems (see examples in Fig. 2 of Sec. IV A).

III. METHODS FOR EFFICIENT ENSEMBLE FORECAST USING RANDOM BATCH APPROXIMATION

We propose new efficient computational methods to address the inherent difficulties in complex turbulent systems involving multiscale interaction terms by solving the coupled stochastic–statistical equations (7). Here, we describe the general strategy in the new approach for efficient statistical prediction by running a very small ensemble simulation of the stochastic fluctuation coefficients enabled by the random batch method (RBM) approximation. The ideas will be further illustrated using concrete examples from the L-96 systems in Sec. IV.

A. Random batch method for coupled turbulent systems with a wide spectrum

It is realized that the most computational demanding part in solving the coupled equations comes from getting accurate quantification for the combined nonlinear feedbacks in the stochastic coefficients and statistical covariance equations (7b) and (7c), which involves high-order coupling terms involving a wide spectrum of stochastic modes $\mathbf{Z} = \{Z_k\}_{k=1}^K, K \gg 1$ for each single trajectory. Furthermore, in order to resolve the necessary high-order statistics with desirable accuracy, an extremely large ensemble $\left\{\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}\right\} \in \mathbb{R}^{K\times M}$ (with the sample size M) is required for the full high-dimensional modes. Thus, using the direct MC method for the stochastic equations ends up with a computational cost of $O\left(MK^3\right)$ for the M samples and K modes and cost $O\left(K^4\right)$ for the covariance equation. The K^3 cost is due to the quadratic interactions (of total number K^2 for each mode) and then for all K modes. In addition, the required ensemble size M to maintain sufficient accuracy in the empirical statistical estimation (3) grows with an exponential rate dependent on the dimension K. This is known as the curse of dimensionality[22,41] and sets an inherent obstacle for effective ensemble prediction of high-dimensional systems, especially when non-Gaussian statistics amplifies the demand for an even larger ensemble to capture PDF tails.

ALGORITHM 1. Full RBM approximation for the coupled multiscale stochastic–statistical model.

Initial condition: At initial time $t = 0$, assign the initial mean and covariance for $\{\bar{\mathbf{u}}_0, R_0\}$ and draw samples $\mathbf{Z}_0$ from the initial distribution μ_0.
1: **for** $s = 1$ while $s \le \lceil T/\Delta t \rceil$, at the start of the time interval $t \in (t_s, t_{s+1}]$ with time step $\Delta t = t_{s+1} - t_s$ **do**
2: Partition the K modes into S batches (with $pS = K$) randomly as $\mathcal{Z}_q^s = \left\{Z_k(t_s), k \in \mathcal{I}_q^s\right\}$ with $\cup_{q=1}^S \mathcal{Z}_q = \mathbf{Z}$.
3: Update $\{Z_k(t_{s+1}), R_{kl}(t_{s+1})\}$ for modes in batch $k, l \in \mathcal{I}_q^s$ independently according to (8a)) and (8b).
4: Update $\bar{\mathbf{u}}(t_{s+1})$ according to the original mean equation (7a) using all the batch outputs.
5: **end for**

Here, we propose to design a computational efficient model using the random batch method (RBM) developed in Refs. 33 and 34. The idea is proposed for the special mean-fluctuation systems in Ref. 35 concerning only the coupling between the mean and fluctuation modes. Now, we aim to develop an effective practical strategy for the fully coupled multiscale states using the general stochastic–statistical formulation (7). The crucial issue in constructing the RBM approximation lies in devising an efficient estimation of the nonlinear cross-interaction terms $Z_m Z_n$ in the stochastic and covariance equations (7b) and (7c) without running a very large ensemble. This becomes especially important in modeling high-dimensional turbulent systems where a key feature is the nonlocal coupling of multiscale states involving a large number of fluctuation modes $\mathbf{Z}$.

In the main idea of the new RBM approach, we no longer compute the expensive nonlinear interactions among the entire stochastic coefficients $1 \le k \le K$ in the stochastic and covariance equations. At the start of each time updating step $t = t_s$, a partition $\left\{\mathcal{I}_q^s\right\}$ of the mode index is introduced with $\cup_q \mathcal{I}_q^s = \{k : 1 \le k \le K\}$, where each batch only contains a small portion of $\left|\mathcal{I}_q^s\right| = p$ elements randomly drawn from the total K indices (ending up with $\left\lceil \frac{K}{p} \right\rceil$ batches). Accordingly, the full spectrum of modes is also randomly divided into small batches $\mathcal{Z}_q^s = \left\{Z_k(t), k \in \mathcal{I}_q^s\right\}$ with $\cup_q \mathcal{Z}_q^s = \left\{Z_k(t) : 1 \le k \le K\right\}$ in the time interval $t \in (t_s, t_{s+1}]$. Then, instead of taking summation over all the wavenumbers m, n in the summation terms of (7b) and (7c), only a small portion of the modes $Z_k \in \mathcal{Z}_q^s$ with indices in the batch $k \in \mathcal{I}_q^s$ are used to update the mode Z_k during the time updating interval. The exhausting procedure to resolve all high-order feedback is effectively avoided through this simple RBM decomposition. The entire high-dimensional system is

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

then decomposed into smaller subsystems for modes $\{Z_k, R_{kl}\}$ constrained inside each small batch $\{k,l\} \in \mathcal{I}_q^s$ rather than the entire spectral space. The resulting *RBM model for the stochastic and statistical equations* during the time interval $t \in (t_s, t_{s+1}]$ becomes

$$\frac{dZ_k^{(i)}}{dt} = \sum_{m,n\in\mathcal{I}_q^s} \tilde{L}_{v,km}(\bar{\mathbf{u}}) Z_m^{(i)} + \tilde{\gamma}_{mnk}\left(Z_m^{(i)} Z_n^{(i)} - R_{mn}\right) + \sigma_k \dot{W}_k^{(i)}, \quad (8a)$$

$$\begin{aligned}\frac{dR_{kl}}{dt} &= \sum_{m,n\in\mathcal{I}_q^s} \tilde{L}_{v,km}(\bar{\mathbf{u}}) R_{ml} + R_{km}\tilde{L}_{v,ml}^*(\bar{\mathbf{u}}) + Q_{\sigma,kl} \\ &+ \sum_{m,n\in\mathcal{I}_q^s} \tilde{\gamma}_{mnk} \frac{1}{M_1}\sum_{i=1}^{M_1} Z_m^{(i)} Z_n^{(i)} Z_l^{(i)} \\ &+ \epsilon^{-1}\left(\frac{1}{M_1}\sum_{i=1}^{M_1} Z_k^{(i)} Z_l^{(i)} - R_{kl}\right). \end{aligned} \quad (8b)$$

Here, all the wavenumbers $k, l, m, n \in \mathcal{I}_q^s$ belong to the same self-consistent random batch $\mathcal{Z}_q^s$. Importantly, new coupling coefficients, $\tilde{L}_{v,km} = c_p^1 L_{v,km}, \tilde{\gamma}_{mnk} = c_p^2 \gamma_{mnk}$ [with the original coefficients $L_{v,km}$ and γ_{mnk} defined in (7)], are introduced in the new equations to guarantee consistent statistics with the scaling factors $c_p^1 = \frac{K}{p}$ and $c_p^2 = \frac{K(K-1)}{p(p-1)}$. It is based on the idea that the random batch approximation should yield an equivalent effect in the summation of nonlinear coupling terms under probability expectation on the batch partition $\mathcal{I}^s$. Detailed explanation for the rescaling coefficients $c_p^{1,2}$ can be found in the proof of Theorem 1 and Ref. 35. A much smaller ensemble size is used for the stochastic modes $\mathbf{Z}^{(i)}, i = 1, \ldots, M_1$ in (8a) to give empirical estimation of the higher-order moments in (8b). Notice that in (8a) within the time interval, each sample is updated independently. After this time updating interval $(t_s, t_{s+1}]$, the batches are resampled at the start of the new time step $t = t_{s+1}$ to repeat the same procedure, so the modes from different batches get mixed.

The statistical consistency of the RBM approximation (8) is guaranteed by the ergodicity and fast mixing of the high wavenumber modes. First, considering the typical property of the turbulent modes, the energy inside the single small-scale mode $\langle |Z_k|^2 \rangle$ decays fast as k grows large and de-correlates rapidly in time. Second, ergodicity of the stochastic fluctuation modes $\mathbf{Z}$ implies that updating the key statistics using fractional fluctuation modes at each time step with consistent time-averaged feedback can provide an equivalent total contribution. Therefore, the total K spectral modes are divided into smaller batches to be updated individually during each time interval. As a result, rather than running a large ensemble of high-dimensional solutions of the full fluctuation modes $\mathbf{Z}^{(i)}$ as in the direct MC approach of (7), only a small number of stochastic trajectories are needed as long as it is sufficient to sample the p-dimensional modes inside each batch rather than the full K-dimensional space. We show the rigorous convergence in leading-order statistics of this RBM approximation in Sec. III C.

As the first major reduction of the above RBM model, instead of using the entire spectrum of modes $1 \le k \le K$ to update each wavenumber mode, we only consider the nonlinear interactions between modes in a very small subset $\mathcal{I}_q^s$ of size p. As shown in the numerical examples in Sec. IV A, the batch size p can be picked as a very small number $p = O(1)$. This leads to the effective computational reduction from $O(K^3)$ in direct MC to $O(p^2K)$ for time evolution of each single stochastic trajectory and from $O(K^4)$ to $O(p^2K^2)$ for the covariance equation. In addition, through the RBM, only a very small number of modes are contained in each batch for the estimation of the nonlinear coupling. Consequently, we do not need to compute all the cubic terms $Z_mZ_nZ_l$ but only the modes inside one batch of size $p \ll K$ independent of the full dimension. This leads to a further significant reduction in the required sample size from $O(MK^3)$ to $O(M_1p^2K)$ in the ensemble prediction. Specifically, the ensemble size M_1 only needs to sample the p-dimensional batch subspace in contrast to M in the full MC approach to sample the K-dimensional full space. Thus, the required sample size will not grow as the dimension of the full system increases, which shows the potential to overcome the curse of dimensionality. The algorithm using random batch method for ensemble simulation of the high-dimensional turbulent system under the coupled stochastic–statistical closure model is summarized in Algorithm 1.

Remark. In the RBM model (8), we choose the random batches $\mathcal{I}_q^s$ by treating all the fluctuation modes equally. As a more detailed calibration of the energy cascade in turbulent flows with localized interactions, the spectral modes can be first grouped into clustered samples of neighboring modes and random batches can be drawn among each localized group with the strongest interactions. The size of the random batches can also change according to the energy in each mode.

B. Reduced-order model for efficient ensemble simulations

In the above model with the RBM approximation, we still need to run an ensemble simulation for the entire spectral modes $\{Z_k^{(i)}\}_{k=1}^K$ even though with a much small number of samples $i = 1, \ldots, M_1$, which reaches the final computational cost dependent on the full dimension K. On the other hand, in practical situations, we are mostly interested in the statistics in a much smaller number of leading modes $1 \le k \le K_1 \ll K$ (such as the largest scales or the most energetic modes) rather than the entire spectrum. This leads to the second major approximation to introduce the effective reduced-order modeling strategy focusing on the leading dominant modes while still taking into account the contributions from the large number of small-scale fluctuation modes through the effective RBM approximation allowing an even smaller sample size.

To achieve the model reduction, we utilize the idea introduced in Ref. 35 and extend it to the general coupled stochastic–statistical model (7). The key idea is still to notice that a large number of fast mixing small-scale modes make an equivalent contribution through time average; thus, we can decompose them into batches for updating different ensemble samples of the central large-scale modes. We further introduce an explicit large–small scale decomposition for the original fluctuation modes $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = \{X_k\}$ represents the small number of leading modes (such as $\{Z_k, k \le K_1\}$) while $\mathbf{Y} = \{Y_l\}$ are all the rest large number of smaller scale fluctuation modes (such as all the rest modes $\{Z_l, K_1 < l \le K\}$). We aim to use

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

an ensemble empirical approximation for the marginal distribution of $\mathbf{X}$, that is,

$$p_t^{\mathrm{RBM}}(\mathbf{X}) := \frac{1}{M_1}\sum_{i=1}^{M_1}\prod_{k=1}^{K_1}\delta\left(X_k - X_k^{(i)}(t)\right). \tag{9}$$

By focusing on the leading modes in a much lower dimension $K_1 \ll K$, the required ensemble size M_1 is further reduced. Still, each large-scale mode $X_k^{(i)}$ is coupled to the unresolved small-scale modes Y_l through the linear and nonlinear coupling terms. Using the model reduction strategy to be combined with the RBM approximation, we no longer run ensemble simulation for the large number of small-scale fluctuation modes $\mathbf{Y}$. Instead, only one (or at most a small number) of stochastic trajectory $\mathbf{Y}(t)$ is solved in time. The total $K - K_1$ spectral modes in $\mathbf{Y}$ are then divided into small batches to update different ensemble members of $\mathbf{X}^{(i)}, i = 1, \ldots, M_1$. This leads to M_1 batches from the small-scale modes with the relation $pM_1 = K - K_1$. The idea here is to use a large number of small-scale modes to update different large-scale ensemble samples at each time step.

In this way, we decompose the original stochastic equation of independent samples (8a) into a coupled system with interacting samples in large-scale states $\mathbf{X}^{(i)} = \left\{Z_k^{(i)}\right\}_{k=1}^{K_1}, i = 1, \ldots, M_1$ and one single trajectory of the small-scale state $\mathbf{Y} = \{Z_l\}_{l=K_1}^{K}$ (with $K_1 \ll K$). At the start of time step $t = t_s$, each large-scale sample $\mathbf{X}^{(i)}$ is grouped with one batch of the small-scale modes $\mathcal{Y}_i^s = \{Y_l(t_s)\}_{l\in\mathcal{I}_i^s}$, where $\mathcal{I}_i^s$ is the RBM partition of the small-scale modes with $\cup_i \mathcal{I}_i^s = \left\{l : K_1 \le l \le K\right\}$. Therefore, we get the *coupled reduced-order RBM equations* for the ith large-scale ensemble sample grouped with a batch of the small-scale modes $\left\{X_k^{(i)}, Y_l\right\}_{k\le K_1, l\in\mathcal{I}_i^s}$ during the time updating interval $t \in (t_s, t_{s+1}]$,

$$\begin{aligned}
\frac{\mathrm{d}X_k^{(i)}}{\mathrm{d}t} &= F_k\left(\mathbf{X}^{(i)}, \mathbf{Y}\right) + \sum_{m,n\le K_1}\tilde{L}_{v,km}(\bar{\mathbf{u}})\, X_m^{(i)} + \tilde{\gamma}_{mnk}\left(X_m^{(i)}X_n^{(i)} - R_{mn}\right) \\
&\quad + \sigma_k \dot{W}_k^{(i)}, \quad 1 \le k \le K_1, \\
\frac{\mathrm{d}Y_l}{\mathrm{d}t} &= G_l\left(\mathbf{X}^{(i)}, \mathbf{Y}\right) + \sum_{m,n\in\mathcal{I}_i^s}\tilde{L}_{v,lm}(\bar{\mathbf{u}})\, Y_m \\
&\quad + \tilde{\gamma}_{mnl}\left(Y_m Y_n - R_{mn}\right) + \sigma_l \dot{W}_l, \quad l \in \mathcal{I}_i^s.
\end{aligned} \tag{10}$$

Here, F_k and G_l contain the residual terms that represent the cross coupling between the large- and small-scale states. Usually, the state decomposition and the coupling dynamics will become straightforward according to the specific dynamical structure of the multiscale system. The batches are then resampled each time at the start of the new time updating cycle $t = t_{s+1}$ same as the previous RBM strategy. It needs to be noticed that through the above model approximation (10), different ensemble samples $\mathbf{X}^{(i)}$ are no longer independent since they are linked by the shared small-scale trajectory of $\mathbf{Y}$. This is a reasonable assumption from the common observation in turbulent systems that a large number of small-scale fluctuation modes can be viewed as almost independent random processes with a rapidly decaying energy spectrum. In Sec. IV B, we will show the construction of the reduced-order model through one explicit example from the two-layer L-96 system.

ALGORITHM 2. Reduced-order RBM approximation for the coupled multiscale stochastic–statistical model.

Model Setup: Decompose the stochastic modes $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ into large scales $\mathbf{X} \in \mathbb{R}^{K_1}$ and small scales $\mathbf{Y} \in \mathbb{R}^{K-K_1}$ with $K_1 \ll K$.
Initial condition: At initial time $t = 0$, assign the initial mean and covariance for $\{\bar{\mathbf{u}}_0, R_0\}$.
Draw samples for the large-scale modes $\mathbf{X}_0^{(i)}, i = 1, \ldots, M_1$ and small-scale modes $\mathbf{Y}_0$ from the initial distribution.
1: **for** $s = 1$ while $s \le \lceil T/\Delta t \rceil$, at the start of the time interval $t \in (t_s, t_{s+1}]$ with time step $\Delta t = t_{s+1} - t_s$ **do**
2: Partition the $K - K_1$ small-scale modes into M_1 batches randomly. The ith large-scale sample is grouped with the small-scale modes in one batch as $\mathcal{Z}_i^s = \left\{\mathbf{X}^{(i)}, Y_l\right\}_{l\in\mathcal{I}_i^s}$.
3: Update $\left\{\mathbf{X}^{(i)}(t_{s+1}), Y_l(t_{s+1})\right\}$ for each batch $l \in \mathcal{I}_i^s$ according to (10).
4: Update the statistical mean and covariance to $\bar{\mathbf{u}}(t_{s+1}), R(t_{s+1})$ accordingly based on (7a) and (7b) using all batch outputs.
5: **end for**

Using the above decomposition, we can effectively reduce the computational cost by only sampling a much smaller space of dimension K_1, thus enabling an even smaller sample size. The final computational cost is then reduced to $O\left(M_1 K_1^3 + (K - K_1)p^2\right) = O\left(M_1\left(K_1^3 + p^3\right)\right)$ (using the above partition relation for small-scale modes, $pM_1 = K - K_1$) rather than the cost $O\left(M_1 K p^2\right)$ in the full RBM model involving a very large number of K. The ensemble size M_1 is then only sampling the very small large-scale modes of dimension $K_1 = O(1)$ rather than the full state dimension K, together with the rest dimension $K - K_1$ only related to the small batch size $p = O(1)$. In this way, the curse of dimensionality is effectively avoided. Similarly, we summarize the reduced-order RBM strategy in Algorithm 2.

C. Error analysis for the RBM model approximation

Here, we provide convergence analysis on the RBM model Algorithm 1 in (8) compared with the solution obtained from the full stochastic–statistical equations (7). Through this analysis, we rigorously demonstrate the effectiveness of the proposed RBM as a precise approximation to the direct MC solutions. Besides, it also provides error estimates and guidelines for selecting appropriate model parameters in implementing the computational schemes.

First, we consider the convergence of the RBM approximation $\tilde{Z}_k$ of the stochastic equation (8a) to the full stochastic equation (7b) of the coefficients Z_k

$$\frac{\mathrm{d}\tilde{Z}_k}{\mathrm{d}t} = \sum_{m,n\in\mathcal{I}_k^s}\tilde{L}_{v,km}(\bar{u})\,\tilde{Z}_m + \tilde{\gamma}_{mnk}\left(\tilde{Z}_m\tilde{Z}_n - R_{mn}\right) + \sigma_k\dot{\tilde{W}}_k, \tag{11a}$$

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

$$\frac{\mathrm{d}Z_k}{\mathrm{d}t} = \sum_{m,n\le K} L_{v,km}\left(\bar{u}\right) Z_m + \gamma_{mnk}\left(Z_m Z_n - R_{mn}\right) + \sigma_k \dot{W}_k, \quad (11b)$$

with the RBM parameters $\tilde{L}_{v,km} = \frac{K}{p} L_{v,km}$ and $\tilde{\gamma}_{mnk} = \frac{K(K-1)}{p(p-1)}\gamma_{mnk}$ and p being the batch size. We need the following structure assumption for the model parameters:

Assumption. Suppose that the quasilinear coupling coefficients in the stochastic equations are uniformly bounded

$$\max_{k,m} L_{km}\left(\bar{u}\right) \le C, \quad \max_{k,m,n} \gamma_{mnk} \le C. \quad (12)$$

The quadratic coupling term satisfies the symmetry

$$B\left(\mathbf{u}_k, \mathbf{u}_k\right) = 0, \quad \left[B\left(\mathbf{u}_k, \mathbf{u}_l\right) + B\left(\mathbf{u}_l, \mathbf{u}_k\right)\right] \cdot \mathbf{u}_k = 0, \ \text{for any } k, l \le K. \quad (13)$$

In the above equations (11a) and (11b), we assume that the full and RBM model have a consistent mean and covariance $\bar{u}, R$ from the statistical equations. In this way, we are able to focus on the approximation due to the random batches $\mathcal{I}_k^s$. Conditional on the accurate statistical mean state and covariance $\bar{u}$ and R, the RBM approximation adds additional randomness through the indices of the modes. The model dynamics fit into the SDE systems discussed in Refs. 33 and 34, where the interacting particles are replaced by the coupled spectral modes in the above turbulent multiscale model. We can find the statistical convergence in the stochastic coefficients following the similar argument as in Ref. 35.

Theorem 1. *Under the assumptions (12), the statistical estimation of the RBM model (11a) with time step Δt converges to the statistics of the full model (11b) up to the final time T as*

$$\sup_{s\Delta t\le T}\left|\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\varphi\left(\tilde{Z}_k^s\right) - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\varphi\left(Z_k^s\right)\right| \le C_\varphi\left(T\right)\Delta t, \quad (14)$$

with the test function $\varphi \in C_b^2$, and $\tilde{Z}_k^s = \tilde{Z}_k\left(t_s\right), Z_k^s = Z_k\left(t_s\right)$ the solutions at $t_s = s\Delta t$. C_φ is the coefficient independent of the model dimension K.

The proof of Theorem 1 compares the backward equations for the two equations (11a) and (11b) and uses the fact that the expectation on the random batch samples gives back the original equation (which leads to the precise forms of the rescaling factors $\tilde{L}_v, \tilde{\gamma}$). The statistical consistency in the stochastic and statistical equations is guaranteed by the additional relaxation term. We give the proof of the theorem in Appendix A. Specifically, using the above conclusion and choosing $\varphi\left(Z\right) = Z^2$, we find the convergence of the total variance in the ensemble estimate of the stochastic equations.

Corollary 2. *The total variance estimation in the RBM model (8) converges to the true statistics as*

$$\frac{1}{N}\sum_{s=1}^{N}\left|\mathrm{tr}\tilde{R}\left(t_s\right) - \mathrm{tr}R\left(t_s\right)\right| \le \sup_{s\Delta t\le T}\left|\mathrm{tr}\tilde{R}_s - \mathrm{tr}R_s\right| \le C_v\left(T\right)K\Delta t, \quad (15)$$

where $\mathrm{tr}\tilde{R}\left(t_s\right) = \sum_k \mathbb{E}\left|\tilde{Z}_k^s\right|$ and $\mathrm{tr}R\left(t_s\right) = \sum_k \mathbb{E}\left|Z_k^s\right|$ are the traces of the variance from the second moments of the stochastic coefficients.

Second, we consider the error in the statistical mean state using the simplified version of the scalar mean equation for RBM model and the full model

$$\frac{\mathrm{d}\tilde{\bar{u}}}{\mathrm{d}t} = -\lambda\tilde{\bar{u}} + B\left(\tilde{\bar{u}}, \tilde{\bar{u}}\right) + \sum_{k,l\le K}\tilde{R}_{kl}B\left(\mathbf{v}_k, \mathbf{v}_l\right) + F, \quad (16a)$$

$$\frac{\mathrm{d}\bar{u}}{\mathrm{d}t} = -\lambda\bar{u} + B\left(\bar{u}, \bar{u}\right) + \sum_{k,l\le K}R_{kl}B\left(\mathbf{v}_k, \mathbf{v}_l\right) + F. \quad (16b)$$

In the above statistical mean equations, we assume only linear damping in the linear term and consider the error from the RBM approximation of the second moments. R and $\tilde{R}$ are the covariance matrices estimated in (14). The idea is to use the total statistical energy equation,[8] which provides a balanced estimation for the statistical in both mean and total variance, $E = \bar{u}^2 + \frac{1}{K}\mathrm{tr}R$. The assumption (13) provides a very simple equation describing the evolution of the total statistical energy E. Using the estimate for the total variance, we find the error in the RBM prediction in the mean state.

Theorem 3. *Under the assumptions (13) and the estimate for the total variance (15), the statistical mean estimation for the RBM model (16a) converges to the full model solution (16b) with time step Δt and final time T as*

$$\left\|\tilde{\bar{u}} - \bar{u}\right\|_{[0,T]} \le \sup_{s\Delta t\le T}\left|\tilde{\bar{u}}\left(t_s\right) - \bar{u}\left(t_s\right)\right| \le C_m\left(T\right)\Delta t. \quad (17)$$

Here, the error in the mean state is taking over the time average $\|f\|_{[0,T]}^2 = \frac{1}{N}\sum_{s=1}^{N} f^2\left(t_s\right)$ among solutions at each time evaluation step $t_s = s\Delta t$ up to the final time $T = N\Delta t$.

Again, we put the proof of the above theorem in Appendix A. The estimates in (15) and (17) quantify the approximation errors dependent on the time step size Δt. In practice, it is always easier to apply a smaller time step to effectively increase the prediction accuracy. The results in the mean and covariance convergence still rely on the consistency in the stochastic and statistical equations (7b) and (7c) in the coupled model. This is implicitly guaranteed by the essential relaxation term as $\epsilon \to 0$. We leave the complete analysis of the coupled stochastic–statistical equations (7) as a mean field system[10] in future research.

IV. NUMERICAL PERFORMANCE ON PROTOTYPE MODELS: THE LORENZ '96 SYSTEMS

In evaluating the performance of the proposed methods, we start with a simple prototype model, which is still able to capture key properties of the general turbulent systems. The Lorenz '96 (L-96) systems provide a desirable testbed, exhibiting a range of representative statistical features such as the interaction of variables from different scales and non-Gaussian statistics. The model was originally introduced to study the mid-latitude weather in a simpler one-layer form[36] and is later generalized to a two-layer form to include strong multiscale spatiotemporal coupling.[37] In this section, we illustrate the application of the full and reduced-order RBM models for predicting leading statistics and non-Gaussian PDFs using the one-layer and two-layer L-96 systems.

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

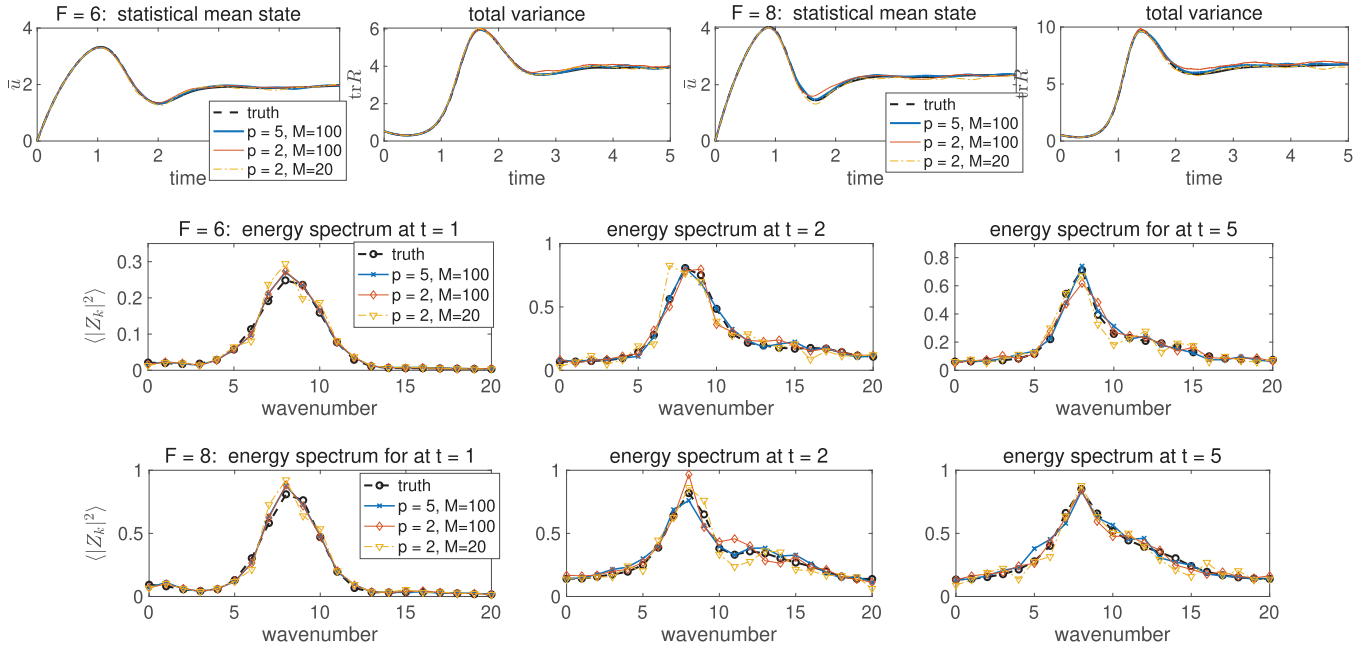

FIG. 3. Time evolutions of the mean and total variance as well as the variance spectra from the RBM prediction with batch size $p = 2, 5$ and ensemble size $M_1 = 100, 20$. Results from the two test regimes $F = 6$ and $F = 8$ are shown and the truth is recovered from a direct MC simulation with a large sample size $M = 1 \times 10^5$.

A. The one-layer L-96 system

First, we consider the one-layer L-96 system[36] that can be expressed as a 40-dimensional ODE system with homogeneous damping and forcing

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} = \left(u_{j+1} - u_{j-2}\right) u_{j-1} - u_j + F,\ j = 1, \ldots, J = 40, \tag{18}$$

with $u_{J+1} = u_1$ and constant uniform damping and forcing. The model state is defined with periodic boundary condition $u_{j+J} = u_j$ mimicking geophysical waves at J equally distributed locations along a constant mid-latitude circle (see the diagram in Fig. 1). The ODE system has a moderate dimension $J = 40$. Various representative statistical features comparable with the real data from observations can be generated in the L-96 solutions (18) by simply changing the constant forcing F. A smaller forcing F is related to a weaker mean state and stronger non-Gaussian statistics, while a large forcing value F leads to stronger mixing and near-Gaussian statistics.[44] As illustrated in the typical solution trajectories plotted in Fig. 1, the one-layer L-96 solutions display the distinctive dynamical transition from more regular wave patterns (with $F = 6$) to more chaotic flow features (with $F = 8$). The model also maintains a spectrum of a large number of unstable and energetic stochastic modes (see Figs. 2 and 3). Accordingly, the PDFs of spectral modes demonstrate the statistical transition from non-Gaussian distributions to near-Gaussian ones between the two regimes (see Figs. 4 and 5 for the marginal and joint PDFs). Thus, it established a challenging and important test case for accurate prediction of model statistics and PDFs.

1. Explicit formulation and RBM approximation for the one-layer L-96 system

Here, we derive the explicit stochastic–statistical formulation for the one-layer L-96 system as an example to illustrate the main ideas and performance in the proposed strategy. In the L-96 system (18), randomness comes only from the initial condition and is amplified due to the internal instability. We seek the probability solution of the model state $p_t\left(\{u_j\}\right)$ starting from an initial distribution $p_0\left(\{u_j\}\right)$. Clearly, the L-96 system fits into the general framework (1). The periodic boundary condition implies the mean-fluctuation decomposition as in (4) by a Galerkin projection on the Fourier basis

$$u_j = \bar{u}_j + \frac{1}{J} \sum_{|k| \le J/2} Z_k\,(t)\ e^{i2\pi k \frac{j}{J}}, \tag{19}$$

where $\bar{u}_j = \langle u_j \rangle_{p_t}$ is the statistical mean and $\{Z_k\}$ are stochastic coefficients with $\langle Z_k \rangle_{p_t} = 0$. Due to the constant forcing and damping terms, the resulting statistics in each moment of the solution state is translation invariant.[8,45] It implies that the mean state, $\langle u_j \rangle_{p_t} = \bar{u}\,(t)$, is uniform; and the off-diagonal covariance entries are all vanishing, $R\,(t) = \langle \mathbf{Z}\mathbf{Z}^* \rangle_{p_t} = \mathrm{diag}\,\{r_k\,(t)\}$. Therefore, we can focus on the dynamical equations for the scalar mean $\bar{u}$ and variance r_k of each Fourier mode together with the stochastic coefficients Z_k to recover the entire statistics in this system.

The explicit full stochastic–statistical formulation (7) and the corresponding RBM model (8) for the L-96 system are listed in Appendix B 1. Due to the still relatively low dimension $J = 40$ of

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

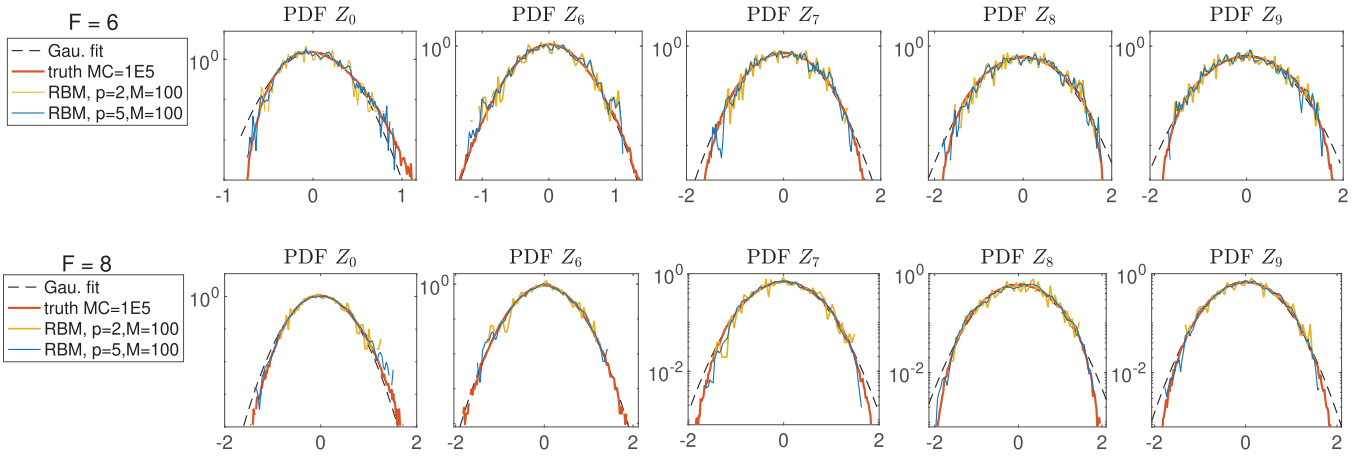

FIG. 4. 1D marginal PDFs of the stochastic coefficients Z_k in the two test regimes $F = 6, 8$ of the one-layer L-96 system. The RBM predictions with sample size $M_1 = 100$ are compared with the truth from direct MC simulation with $M = 1 \times 10^5$. The Gaussian fits of the PDFs with the same mean and variance are also plotted in dashed lines.

the one-layer model, we apply the full RBM model in Algorithm 1 and test the random batch approximation for the full spectrum prediction involving a large number of highly unstable and energetic modes. We summarize several main features observed from using the RBM model on the one-layer L-96 system before showing the detailed discussions on the numerical results in Sec. IV A 2:

- The computational cost for the full model (B1) is $O\left(J+MJ^2+MJ^2\right) = O\left(J\left(1+2MJ\right)\right)$, where J is the full dimension of the system and M is the ensemble size to sufficiently sample the J-dimensional space. In the RBM approximation (B2), the computational cost is effectively reduced to $O\left(J\left(1+2M_1p\right)\right)$ with p being the batch size and $M_1 \ll M$ is the sample size only required to sample the much smaller p-dimensional subspace in each batch. Specifically, p does not increase as the dimensional J increases.
- The coupled formulation combining the variance equations for r_k and stochastic equations for Z_k with the relaxation factor ϵ is essential to reach the correct final equilibrium state. Using purely the stochastic equations for Z_k is insufficient to recover the correct statistics when the sample size becomes small due to the strong model instability.
- The RBM approximation relies on the ergodicity and fast mixing of the small-scale modes. The L-96 system possesses a wide spectrum with a large number of fluctuation modes showing distinctive time scales, making it a difficult test case for the RBM model. Still, accurate statistical prediction is achieved even with an extremely small batch size $p = 2$. Prediction results can be further improved by reducing the time step size Δt.

2. Numerical tests on the one-layer L-96 model

In the numerical tests, we take two typical regimes of the L-96 system with $F = 6$ (showing non-Gaussian statistics) and

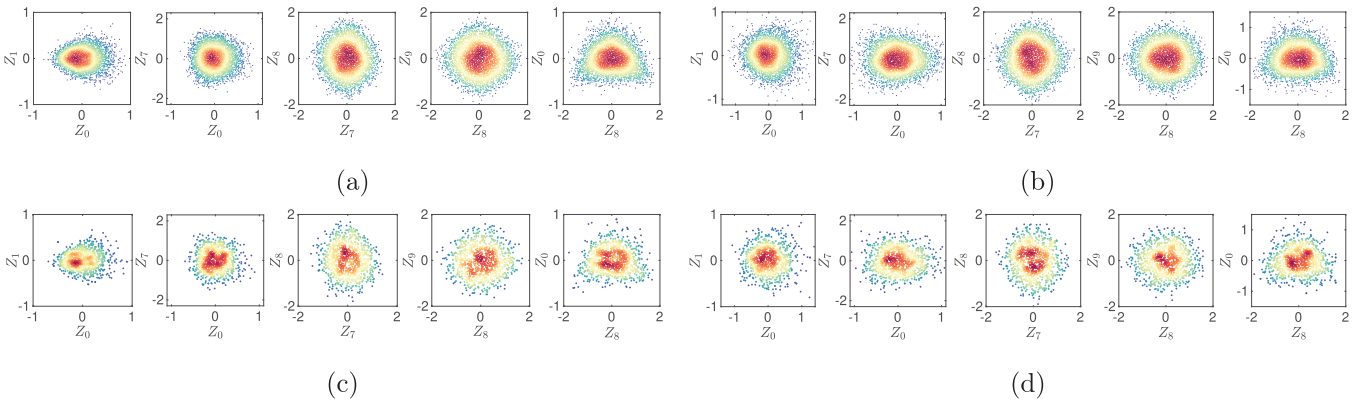

FIG. 5. 2D joint PDFs of the leading modes in the two test regimes $F = 6, 8$ of the one-layer L-96 system. The RBM predictions with sample size $M_1 = 100$ and batch size $p = 5$ are compared with the truth from direct MC simulation with $M = 1 \times 10^5$.

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

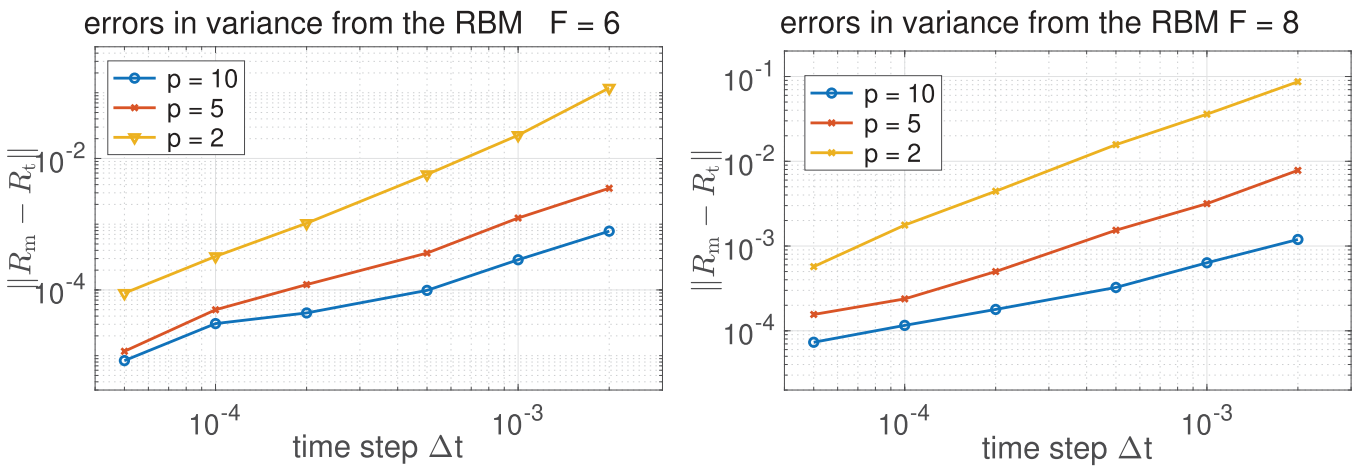

FIG. 6. Errors in the total variance from the RBM prediction with different batch sizes $p = 2, 5, 10$ in the two test regimes $F = 6, 8$ of the one-layer L-96 system.

$F = 8$ (showing near-Gaussian statistics). The standard fourth-order Runge–Kutta scheme is adopted for the time integration with time step size $\Delta t = 1 \times 10^{-4}$. A large ensemble size $M = 1 \times 10^5$ is needed to capture the true model statistics accurately from the direct MC simulation. In the RBM prediction, different batch sizes $p = 2, 5, 10$ are used to compute the high-order interactions compared with the full model using the entire $J = 40$ modes. Therefore, only a very small sample size $M_1 = 100$ becomes sufficient, allowing for efficient computation.

First, we illustrate the inherent difficulty in running ensemble prediction for turbulent systems with instability. In the full model (B1), the variance equations for r_k (as well as the stochastic equations for Z_k) are subject to internal instability due to the real part of the coupling coefficient $-\left(\gamma_k^* \bar{u} + 1\right)$ representing interaction with the mean state $\bar{u}$. As illustrated in the left panel of Fig. 2, positive growth rates are induced in a large number of modes as the system evolves in time. This indicates the crucial role of the combined third-order moments $\langle Z_m Z_n^* Z_k^* \rangle$ acting as a balancing factor for these internally unstable modes. On the other hand, with insufficient sample size, large errors can be introduced to the empirical estimation of the higher-order feedback term. The inherent instability in the turbulent model formulation can also be seen in the zeroth mode equation (20). It shows that the equation is only marginally stable, thus small errors in the ensemble estimation will lead to large errors.

In the practical simulation of the ensemble scheme, the internal instabilities will amplify the small errors and lead to disastrous results. The situation will become increasingly serious if we only want to use a small ensemble size. As a typical example shown on the right panel of Fig. 2, we get the truth from the direct MC simulation of (18) with an extremely large ensemble size $M = 1 \times 10^5$. In comparison, we run the full model (B1) with several different values ϵ in the additional relaxation term $\epsilon^{-1}\left(\langle |Z_k|^2 \rangle - r_k\right)$. It shows that even with this moderate dimension $J = 40$ and very large ensemble size, the solution for the statistical mean and variance will diverge without the relaxation term $\epsilon = \infty$, while equilibrium consistency is guaranteed when a smaller value of ϵ is added. It demonstrates the crucial role of the additional relaxation term to guarantee equilibrium converge. It also shows that the model is robust with consistent final statistics for a wide range of values of ϵ as long as it is not too large.

Remark. As a simple example to illustrate the internal instability, the explicit equation for the zeroth mode Z_0 in the full model (B1) gives

$$\frac{\mathrm{d}Z_0}{\mathrm{d}t} = \frac{1}{J}\sum_k \gamma_k \left(|Z_k|^2 - r_k\right) - Z_0, \tag{20}$$

with $\gamma_k = \cos\frac{4\pi k}{J} - \cos\frac{2\pi k}{J}$. The consistent final equilibrium requires $\langle Z_0 \rangle = \frac{1}{J}\sum_k \gamma_k \left(\langle |Z_k|^2 \rangle - r_k\right) = 0$, while instability will be introduced through errors due to the empirical average $\langle |Z_k|^2 \rangle \sim r_k$. This will lead to inherent difficulty even with a very large sample size as shown in Fig. 2.

Next, we check the performance of the RBM prediction with different batch sizes. In the one-layer L96 model with $J = 40$, we will focus on the fully resolved RBM model (B2) described in Algorithm 1. In Fig. 3, the time series of the mean $\bar{u}$ and total variance $\mathrm{tr}R = \sum r_k$ as well as the prediction for the detailed variance spectrum in each mode during the time evolution are plotted. Two batch sizes $p = 5, 2$ with $M_1 = 100$ samples from the RBM model prediction are compared with the truth from the direct MC simulation using $M = 1 \times 10^5$ samples for the full dimension $J = 40$. Furthermore, we also test an extreme case with only $p = 2$ modes (that is, using only one term in the high-order feedback for the strongly unstable dynamics) in each batch and a very small sample size $M_1 = 20$ in computing the statistics. It shows that all the RBM results accurately track the truth statistics with lines overlapping on each other while saving much computational costs using the extremely small ensemble. In particular, the L-96 system sets a challenging test model for the RBM model since it contains a wider spectrum of energetic modes containing large degrees of instabilities. Still, as shown in the energy spectra from the starting transient stage to the final

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

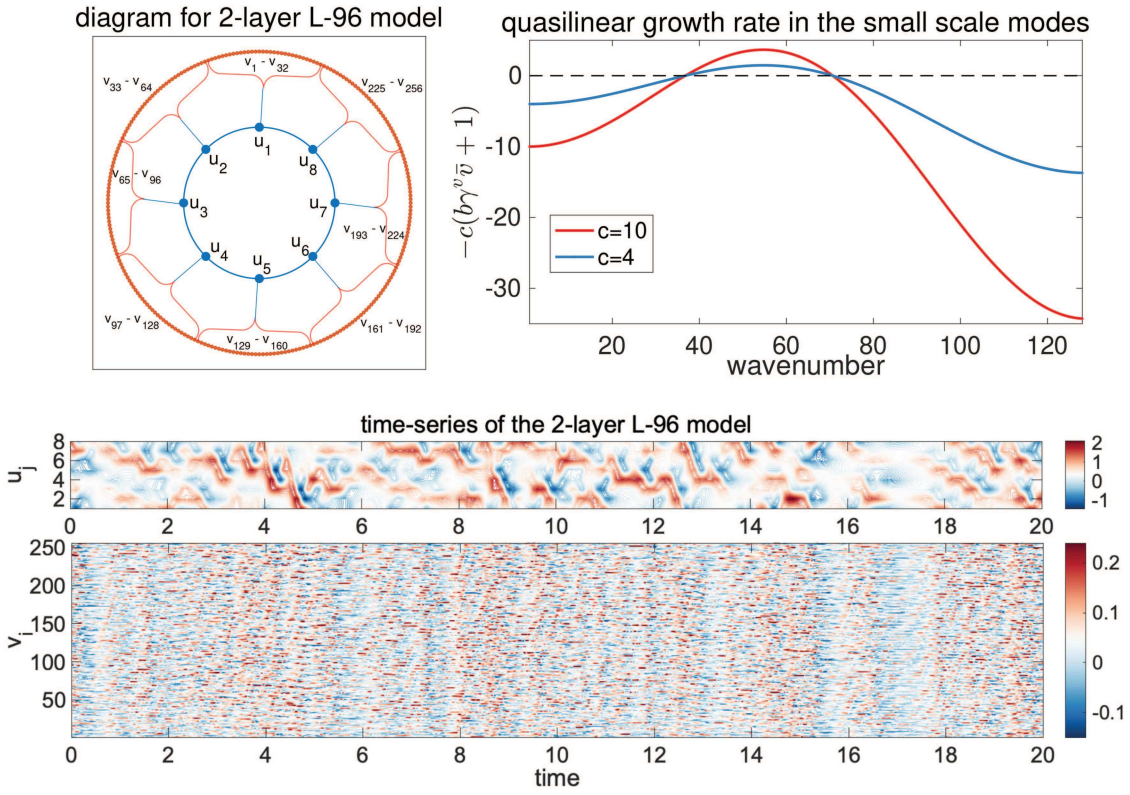

FIG. 7. Illustration of the two-layer L-96 system with large-scale size $J = 8$ and small-scale size $JL = 256$. Upper: diagram for the model structure and growth rates in the small-scale modes; Lower: a typical time-series solution for the large- and small-scale states with $c = 4, b = 10, h = 1$.

equilibrium, the variances in all the modes are captured with high precision.

Then, another important goal of the RBM model is to accurately characterize the PDFs of the model states. This is especially important in application to uncertainty quantification and data assimilation, where an accurate estimate of PDFs using a limited number of samples is crucial. In the RBM approach, the PDFs are captured by the empirical distribution (3) of the stochastic coefficients Z_k. Figure 4 first plots the 1D marginally PDFs in the leading Fourier modes Z_k of the one-layer L-96 system in the two test regimes. It shows that the $F = 6$ regime generates stronger non-Gaussian PDFs while the $F = 8$ regime is closer to Gaussian but still contains non-negligible non-Gaussian features. Usually, a very large ensemble is essential to capture such non-Gaussian statistics in the direct MC approach. Using the efficient RBM approximation, the shapes of PDFs including the skewed and sub-Gaussian structures are accurately characterized using a very small ensemble size $M_1 = 100$. As a more precise calibration of the prediction of PDFs, we plot in Fig. 5 the 2D joint PDFs between the most important modes containing the largest variances. The non-Gaussian structures can be observed more clearly in the joint distributions. The outliers in the sampled PDFs play an important role in characterizing the occurrence of extreme events and have a crucial impact in many practical applications with limited samples. The RBM model successfully captures the joint PDFs especially recovers the representative non-Gaussian shapes in the outlier regions by using only a very small number of samples.

Finally, we provide a quantitative quantification of the model errors of the RBM prediction with different batch sizes p. In particular, one of the central quantities to estimate in the RBM is the second moments; thus, we consider the time-averaged errors in the variance prediction $\|R_m - R_t\| = \frac{1}{N}\sum_{s=1}^{N}\sum_k \left|r_{m,k}(t_s) - r_{t,k}(t_s)\right|$, where R_m is the RBM prediction and R_t is the truth. Figure 6 plots the errors with the batch sizes $p = 2, 5, 10$ and sample size $M_1 = 500$. We use a relatively larger sample size to reduce the fluctuation errors from the small ensemble. The convergence of RBM estimates depends on the average of the random batches (thus equivalently on time average of the fast mixing modes); therefore, the errors grow as the time step size Δt increases. A smaller batch size will lead to larger errors due to the fewer high-order terms explicitly modeled at each time updating step. When the time step decreases to smaller values, the

TABLE III. Model parameters used for the two-layer L-96 model (21).

J	L	F	c	b	h	Δt	M	M_1
8	32	20	4, 10	10	1	1×10^{-4}	5×10^5	500, 100

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

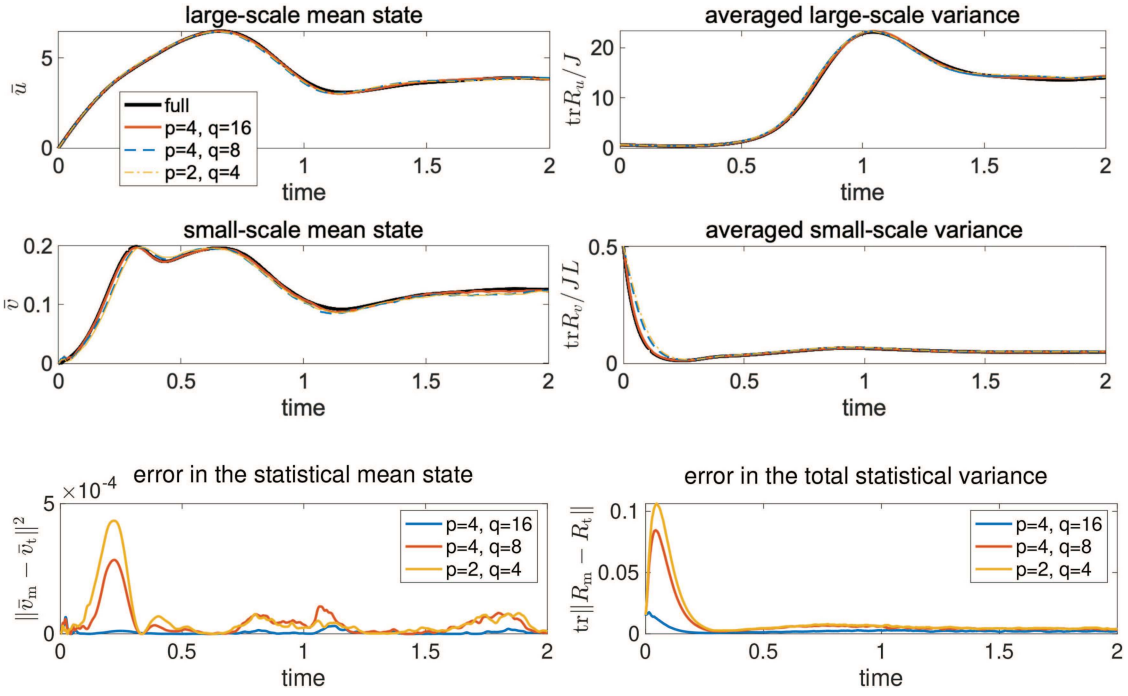

FIG. 8. Prediction of the mean and averaged variance in both large- and small-scale modes from full RBM model with different batch sizes (p, q) compared with the full MC model with $JL = 256$ modes in small scales and sample size $M = 5 \times 10^5$. $M_1 = 500$ samples are used for all the RBM models. The trajectory errors in the mean and variance are also compared.

major source of errors is taken over by the fluctuation errors from the small ensemble size, so the decay of the errors starts to slow down. Also, the larger batch size $p = 10$ shows a slower decay rate since it needs to sample a relatively larger dimensional batch subspace in computing the nonlinear coupling terms. Overall, the error plot shows that the RBM model has a uniformly high prediction skill in accurately recovering the key statistics with a much lower affordable computational cost. It also implies that it is a good strategy to improve the prediction accuracy by taking a smaller time step Δt without increasing much of the computational cost.

B. The two-layer L-96 system

In the second test model, we examine the model effectiveness in handling truly high-dimensional systems with multiscale structures by considering the more complicated two-layer L-96 system. As a further generalization of the original one-layer system, the two-layer L-96 system[37,43] introduces an additional second layer state v_i, $i = 1, \ldots, JL$ to the first layer state $u_j, j = 1, \ldots, J$ such that

$$\begin{aligned} \frac{du_j}{dt} &= \left(u_{j+1} - u_{j-2}\right) u_{j-1} - u_j + F - \frac{hc}{b} \sum_{i=L(j-1)+1}^{jL} v_i, \\ \frac{\mathrm{d}v_i}{\mathrm{d}t} &= -cb\left(v_{i+2} - v_{i-1}\right) v_{i+1} - cv_i + \frac{hc}{b} u_{\left[\frac{i-1}{L}\right]+1}, \end{aligned} \tag{21}$$

both with periodic boundary conditions, $u_{j+J} = u_j$ and $v_{i+JL} = v_i$. Here, u is usually referred to as the large-scale slow variables and v as the small-scale fast variables. On the right-hand sides of (21), the double layer states follow the same energy-conserving nonlinear self-coupling structure as well as uniform linear damping effect as in the one-layer case. The two states of different scales are then coupled through three additional model parameters, c, b, h: c signifies the timescale separation; b controls the ratio between the amplitudes of two layer states u_j and v_i; and h characterizes the coupling strength between the two states. We illustrate the coupling structure of the two-layer system in the upper diagram of Fig. 7. In particular, the second layer states v_i for $i = 1, \ldots, JL$ are locally coupled with one corresponding first layer state u_j with the index $j = \left[\frac{i-1}{L}\right] + 1$ (where $[a]$ takes the integer part of a), and are globally linked with each other by the nonlinear self-interactions. Inversely, each first layer state u_j receives the combined feedback from a sequence of L second layer states v_i, $i = L\left(j-1\right) + 1, \ldots, jL$. This leads to a fully coupled high-dimensional system including the two-level states $\left\{u_j, v_i\right\}$ with a total of $J\left(L+1\right)$ state variables. The multiscale structure of the dynamical solution is demonstrated by a typical time series of the large- and small-scale processes in the lower panel of Fig. 7. It is clear to observe the distinctive time and spatial scales and close correlation between the two scales. The small scales v_i possess a much faster time scale and more rapidly changing spatial structures. On the other hand, the slow dynamics in the large-scale states u_j are closely correlated to a large number of fast- and

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

small-scale processes, leading to much more complicated turbulent structures.

The model parameters used in the numerical tests are listed in Table I taken the standard model setup as in Refs. 25, 37, and 43. Specifically, we consider two typical parameter regimes with a mild timescale separation $c = 4$ and a strong timescale separation $c = 10$. Again, the two-layer L-96 system displays strong internal instability containing a large number of unstable modes with positive growth rates as shown in Fig. 7. To recover the true model statistics with sufficient accuracy, we need to take a small time step $\Delta t = 1 \times 10^{-4}$ in order to completely resolve the extremely fast time scales in the second layer variables v_i. A fourth-order Runge–Kutta scheme is adopted for the time integration in both the direct MC simulation and the RBM approaches. A very large ensemble size $M = 5 \times 10^5$ is required considering the very high-dimension $d = J(L+1) = 264$ of the system. In contrast, it is sufficient to only use $M_1 = O(100)$ samples in both the full and reduced-order RBM approaches.

1. Stochastic–statistical formulation for the two-layer L-96 system

Similar to the one-layer L-96 system, we can project the large- and small-scale states to the Fourier modes due to the periodic boundary condition for both u_j and v_i as

$$u_j = \bar{u} + \frac{1}{J}\sum_{|k|\le J/2} Z_k(t)\, e^{i2\pi k\frac{j}{J}}, \quad v_i = \bar{v} + \frac{1}{JL}\sum_{|l|\le JL/2} Y_l(t)\, e^{i2\pi l\frac{i}{JL}}, \tag{22}$$

where Z_k and Y_l are the stochastic coefficients for the large- and small-scale states correspondingly. The two-layer L-96 system (21) also accepts the general system structure (1). Using the translation invariance in both large and small scales, the statistical prediction aims to recover the homogeneous mean states $\bar{u} \equiv \langle u_j\rangle_{p_t}$ and $\bar{v} \equiv \langle v_i\rangle_{p_t}$ and the variances in large- and small-scale modes $r_k^u = \langle |Z_k|^2\rangle_{p_t}$, $r_l^v = \langle |Y_l|^2\rangle_{p_t}$ according to the model probability measure p_t. The detailed equations involving various linear and nonlinear coupling effects between different scales become very complicated. We list the explicit mean and covariance equations involving complicated higher-order coupling terms across the scales in Appendix B 2.

The most important ideas concerning the RBM approximation of the two-layer L96 system (21) can be illustrated in the stochastic equations for the large- and small-scale fluctuation modes

$$\begin{aligned}\frac{\mathrm{d}Z_k}{\mathrm{d}t} &= \frac{1}{J}\sum_{m-n=k}\left(Z_m Z_n^* - r_m^u\delta_{mn}\right)\gamma_{mn}^u - d_u Z_k - \frac{1}{L}\sum_{|s|\le L/2}\lambda_{k+sJ}^* Y_{k+sJ},\\ \frac{\mathrm{d}Y_l}{\mathrm{d}t} &= \frac{1}{JL}\sum_{p-q=l}\left(Y_p Y_q^* - r_p^v\delta_{pq}\right)\gamma_{pq}^v - d_v Y_l + \lambda_l Z_{\mathrm{mod}(l,J)}.\end{aligned} \tag{23}$$

Here, we have the coupling coefficients $\gamma_{mn}^u = e^{2\pi i\frac{m+n}{J}} - e^{-2\pi i\frac{2m-n}{J}}$, $\gamma_{pq}^v = cb\left(e^{-2\pi i\frac{p+q}{JL}} - e^{2\pi i\frac{2p-q}{JL}}\right)$, $\lambda_l = \frac{hc}{b}\frac{1-e^{-2\pi i\frac{l}{J}}}{1-e^{-2\pi i\frac{l}{JL}}}$, and the quasilinear operator for interaction with the mean states, $d_u = 1 + \gamma_k^{u*}\bar{u}$, $d_v = c\left(1 + b\gamma_l^{v*}\bar{v}\right)$ [see the detailed model parameters in (B4)]. Notice that in the large-scale modes, the indices in the summation for nonlinear coupling go through all the wavenumbers, $|m| \le J/2$, while in the small-scale modes the indices for nonlinear coupling include all the small-scale wavenumbers, $|p| \le JL/2$. This leads to an extremely high computational overload considering the high dimension of the small-scale modes. Finally, the large- and small-scale modes are coupled through the last linear terms. Specifically, the small-scale modes $Y_{k+sJ}, |s| \le L/2$ give a combined feedback to the large-scale mode Z_k, while each Z_k is acting on a sequence of small-scale modes Y_{k+sJ}. It is realized that in (23) the most computational demanding part comes from the summation terms going through all the wavenumbers representing both linear and nonlinear coupling between scales.

Next, we present the performance of the RBM models on the two-layer L-96 with genuinely high-dimensional and multiscale processes. In particular, we first consider the full RBM model in Algorithm 1, then further reduce the computational cost by applying the reduced-order RBM model in Algorithm 2 exploiting the large number of fast-mixing small-scale modes.

2. Numerical results for the full RBM model

First, we consider the full RBM approximation, that is, to introduce random batch decomposition in both the large- and small-scale stochastic modes $\left\{Z_k^{(i)}, Y_l^{(i)}\right\}$, while still run independent samples for ensemble simulation of the entire stochastic equations during each time updating interval. This leads to the *full RBM model* for the stochastic coefficients of the two-layer L-96 system following the general formulation (8)

$$\begin{aligned}\frac{\mathrm{d}Z_k^{(i)}}{\mathrm{d}t} &= c_p\sum_{m\in\mathcal{I}_k^s}\left(Z_m^{(i)}Z_{m-k}^{(i)*} - r_m^u\delta_{m,m-k}\right)\gamma_{m,m-k}^u - d_u Z_k^{(i)}\\ &\quad - c_L\sum_{s:k+sJ\in\mathcal{J}_l^s}\lambda_{k+sJ}^* Y_{k+sJ}^{(i)},\\ \frac{\mathrm{d}Y_l^{(i)}}{\mathrm{d}t} &= c_q\sum_{p\in\mathcal{J}_l^s}\left(Y_p^{(i)}Y_{p-l}^{(i)*} - r_p^v\delta_{p,p-l}\right)\gamma_{p,p-l}^v - d_v Y_l^{(i)} + \lambda_l Z_{\mathrm{mod}(l,J)}^{(i)},\end{aligned} \tag{24}$$

with independent samples $i = 1,\ldots,M_1$. In the RBM approximation at each time updating step $t = t_s$, we choose the batches $\cup_k\mathcal{I}_k^s = \left\{k : |k| \le \frac{J}{2}\right\}$ with batch size $p = \left|\mathcal{I}_k^s\right|$ for the large-scale modes Z_k and batches $\cup_l\mathcal{J}_l^s = \left\{l : |l| \le \frac{JL}{2}\right\}$ with $q = \left|\mathcal{J}_l^s\right|$ for the small-scale modes Y_l. Then, the original summation terms taking over the entire spectrum space are reduced to the summation for modes restricted inside one batch of very small size (p, q) (with the pair elements corresponding to large- and small-scale batch sizes). The new normalization factors can be found as $c_p = \frac{1}{J}\frac{J-1}{p-1}$, $c_q = \frac{1}{JL}\frac{JL-1}{q-1}$, and $c_L = \frac{1}{L}\frac{L-1}{q-1}$ following the same principle as the one-layer case. The corresponding RBM equations for the associated covariance equations can be derived accordingly. We list the explicit equations in Appendix B 2. In this way, the computational cost for the nonlinear coupling terms, particularly the small scales with a large number of modes, is greatly saved by constraining the high-order interactions only inside the very small batch. This leads to the effective

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

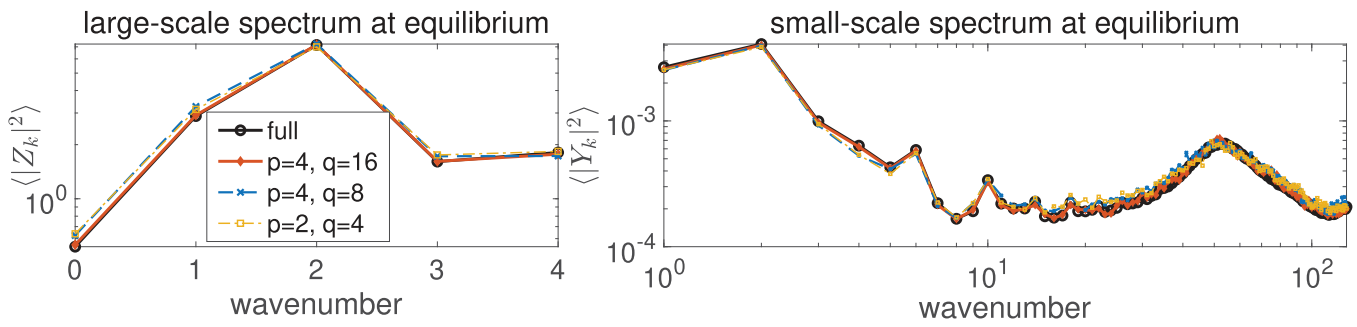

FIG. 9. Prediction of the variance spectra for large- and small-scale modes in full RBM model with different batch sizes (p, q) compared with the full MC model with $JL = 256$ modes in small scales and sample size $M = 5 \times 10^5$.

reduction in computational cost to $O\left(M_1 J\left(p + Lq\right)\right)$ in solving the stochastic equations compared to the original cost $O\left(MJ^2\left(1 + L^2\right)\right)$ in the direct MC approach in computing the full ensemble of size $M \gg M_1$.

In the numerical test of the full RBM model, we take the model parameters for the standard regime $c = 10, b = 10, h = 1, F = 20$, and $J = 8, L = 32$. The resulting coupled large and small-scale processes form a very high total dimension $d = J\left(L + 1\right) = 264$. In particular, the small-scale variable v_i has a much faster time scale and the large-scale state u_j. This leads to a more challenging multiscale problem since we have to resolve a large number of small-scale modes even we are only interested in the large-scale state. In order to get accurate statistical prediction resolving all the multiscale features, we use a very large ensemble size $M = 5 \times 10^5$ for the direct MC simulation to accurately recover the true reference solution. In the RBM model, three different batch pairs $\left(p, q\right) = \left(4, 16\right), \left(4, 8\right), \left(2, 4\right)$ are tested in contrast to the total number of modes $\left(J, JL\right) = \left(8, 256\right)$. A much smaller ensemble size $M_1 = 500$ is used and this is made possible by sampling only the q-dimensional subspace instead of the JL-dimensional full space of the small scales.

In Fig. 8, we first show the RBM model prediction for the mean $\bar{u}, \bar{v}$ and average total variance $\sum r_k^u/J, \sum r_l^v/JL$ with three different batch sizes, as well as the pointwise errors in mean and total variance at each time step. It shows accurate recovery of the key statistics in both the mean and variance, and in both the starting transient stage and final equilibrium. The high accuracy is

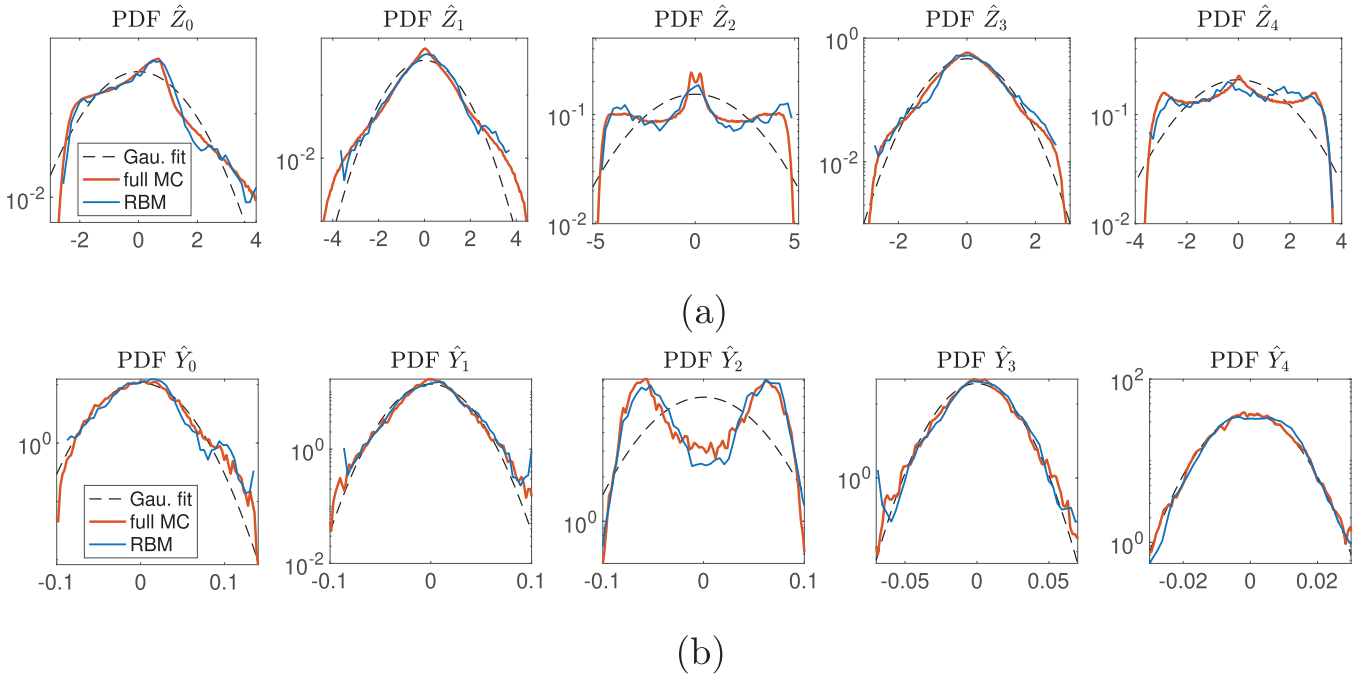

FIG. 10. 1D marginal PDFs of the leading large- and small-scale stochastic modes Z_k, Y_l of the two-layer L-96 system. The RBM predictions are compared with the truth from direct MC simulation. The Gaussian fits of the PDFs with the same mean and variance are also plotted in dashed lines.

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

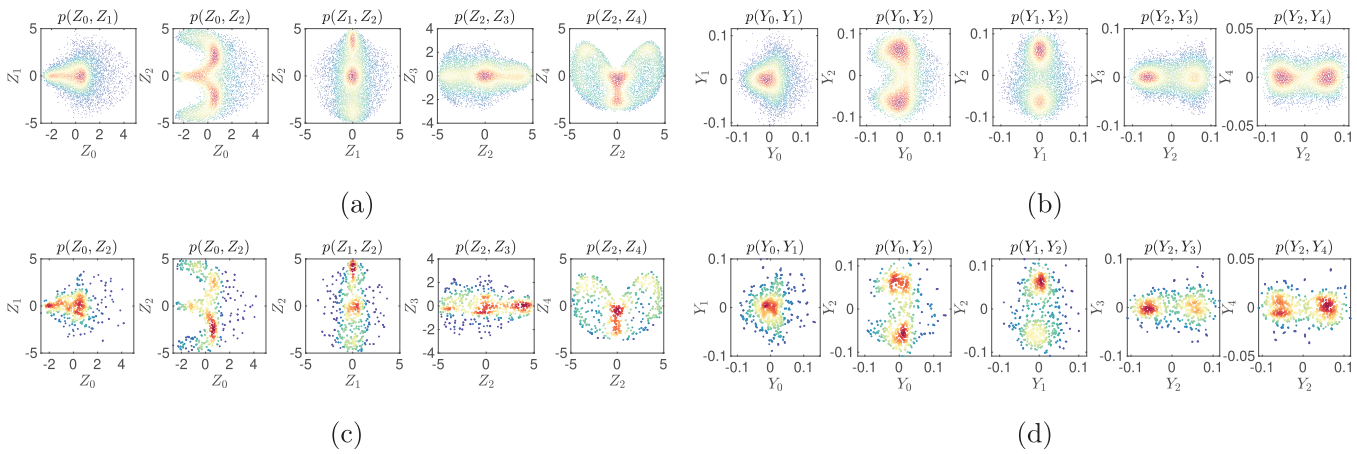

FIG. 11. 2D joint PDFs of the leading large- and small-scale stochastic modes Z_k, Y_l of the two-layer L-96 system. The RBM predictions are compared with the truth from direct MC simulation.

maintained with almost indistinguishable results as we reduce the batch sizes (p, q). Relatively larger errors are observed in the starting transient stage when a small batch size is applied. Notice that significant computational reduction is achieved where the smallest batch size refers using only $q = 4$ nonlinear coupling terms for each Y_l out of the total $JL = 256$ terms. In addition, the detailed model prediction for the variance spectra in each large- and small-scale mode is shown in Fig. 9. A wide range of the small-scale modes with high wavenumbers are excited due to the instability in small scales, as illustrated in Fig. 7. This presents a highly challenging scenario as we only resolve a very small batch of modes for the nonlinear coupling term to stabilize a large number of unstable modes. Again, the energy spectrum is recovered exactly with uniformly high skill using the three extremely small batch sizes. This highlights the robust skill of the RBM model to achieve computational efficiency and maintain high prediction accuracy at the same time.

Next, we proceed to check the ability of the RBM model to capture the PDFs of the dominant modes. Figure 10 shows the marginal PDFs of the leading stochastic modes in both large and small scales. In this case of the two-layer L-96 system, even stronger non-Gaussian statistics are displayed with highly skewed and fat-tailed PDFs. Accurate characterization of such non-Gaussian PDFs becomes an even more critical issue in statistical prediction of such high-dimensional systems. Usually, the small number of samples tend to concentrate near the central part of the PDF and miss the crucial extreme events featuring the edge regions of the large phase space. Using the RBM model with a very small ensemble size, the strongly non-Gaussian PDFs are successfully captured confirming the very high skill of the RBM approach in recovering the true model statistics not only in the leading moments but also in the more challenging higher-order statistics. In addition, we also compare the joint PDFs in the predicted leading modes in Fig. 11. The non-Gaussian structures become more pronounced in the 2D distributions, revealing bimodal PDFs as already indicated in the marginal PDFs. The RBM predictions maintain high accuracy in recovering the highly non-Gaussian statistics despite using a very small ensemble size.

Finally, we provide a quantitative comparison of the RBM model prediction errors as the key model parameters change. In Fig. 12, we show the averaged errors in the total variance $\|R_{\mathrm{m}} - R_{\mathrm{t}}\| = \frac{1}{N}\sum_{s=1}^{N}\sum_{k}\left|r_{\mathrm{m},k}(t_s) - r_{\mathrm{t},k}(t_s)\right|$ and the mean $\|\bar{v}_{\mathrm{m}} - \bar{v}_{\mathrm{t}}\|^2 = \frac{1}{N}\sum_{s=1}^{N}|\bar{v}_{\mathrm{m}}(t_s) - \bar{v}_{\mathrm{t}}(t_s)|^2$ according to the time integration step $\Delta t = t_{s+1} - t_s$ with two batch sizes $(p, q) = (4, 16)$, $(4, 8)$ and two ensemble sizes $M_1 = 500, 100$. The errors grow with larger time step size and with smaller ensemble and batch sizes agreeing with the intuition. Among the range of large to moderate time step sizes, reducing the time step can effectively improve the prediction accuracy since it is corresponding to more frequent resampling (thus more equivalent samples in the time average) in the RBM approximation. With very small time step size though, the error will saturate because the small ensemble size will account for the major error in the statistical estimates. It also implies that it might be useful to use a smaller batch size q with a relatively larger ensemble to achieve accurate prediction with minimum computational cost.

3. Numerical results for the reduced-order RBM model

Next, we consider a further reduction of computation cost by applying the reduced-order RBM model as described in Algorithm 2. In the reduced RBM model, we still keep the statistical mean and covariance equations (B3) and (B5) the same as the full RBM model case. In the two-layer L-96 system, most of the computational cost comes from the ensemble simulation for the stochastic coefficients in (23) for the wide range of small-scale modes Y_l. In most situations, we are mostly interested in the statistics in the main large-scale modes Z_k while the small-scale modes give crucial combined feedback to affect the large-scale state, thus their contributions cannot be simply ignored. The unstable growth (shown in Fig. 7) and excited small scales in equilibrium energy spectrum (in Fig. 9) emphasize

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

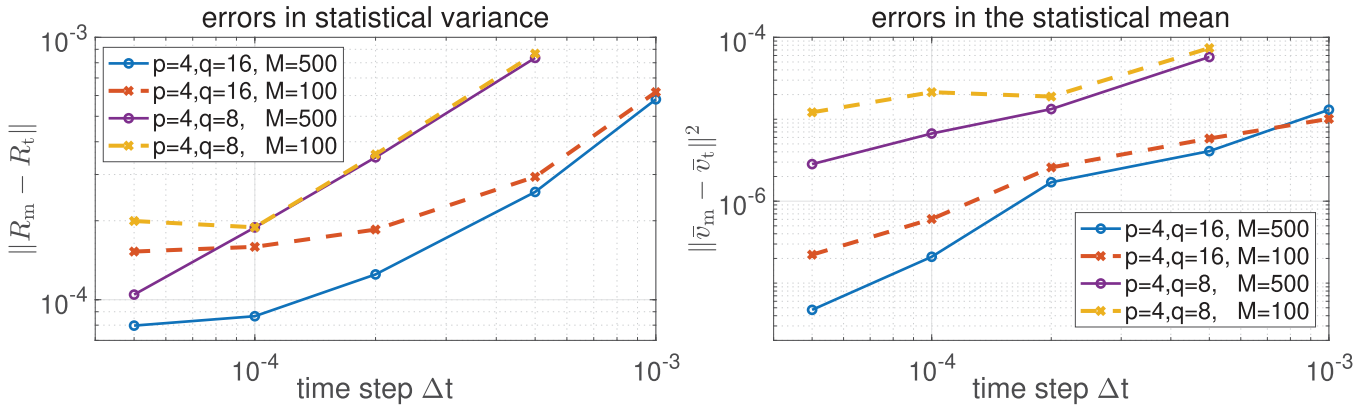

FIG. 12. Errors in the statistical mean and total variance from the RBM prediction with different batch and ensemble sizes for the two-layer L-96 system.

the non-negligible role of these small-scale modes. This sets the inherent obstacle in developing effective reduced-order models for multiscale systems.

To enable further computational reduction for the small-scale modes, following the idea in (10) we group the ensemble members in the large-scale modes $\mathbf{Z}^{(i)}, i = 1, \ldots, M_1$ with one single small-scale state $\mathbf{Y}$. In applying the RBM approximation, the same partition $\mathcal{I}_k^s$ is applied to the small number of large-scale modes. The model reduction strategy is applied to the expensive small-scale equations. The large number of small-scale modes is then divided into the batches $\mathcal{J}_i^s$, satisfying $\cup_{i=1}^{M_1} \mathcal{J}_i^s = \{l : |l| \le JL/2\}$. According to the modes in the batches, the full system is decomposed into much smaller subsystems with one large-scale sample $\mathbf{Z}^{(i)}$ together with a portion of the small-scale modes $Y_l, l \in \mathcal{J}_i^s$. The model reduction is made possible by the very large number of fast-mixing small-scale modes. This leads to the coupled equations for the stochastic coefficients in subsets $\{\mathbf{Z}^{(i)}, Y_l\}_{l\in\mathcal{J}_i^s}$ during the time interval $t \in (t_s, t_{s+1}]$,

$$\frac{\mathrm{d}Z_k^{(i)}}{\mathrm{d}t} = c_p \sum_{m\in\mathcal{I}_k^s} \left(Z_m^{(i)} Z_{m-k}^{(i)*} - r_m^u \delta_{m,m-k}\right) \gamma_{m,m-k}^u - d_u Z_k^{(i)} - c_L \sum_{s:k+sJ\in\mathcal{J}_i^s} \tilde{\lambda}_{k+sJ}^* Y_{k+sJ}, \tag{25}$$

$$\frac{\mathrm{d}Y_l}{\mathrm{d}t} = c_q \sum_{p\in\mathcal{J}_i^n} \left(Y_p Y_{p-l}^* - r_p^v \delta_{p,p-l}\right) \gamma_{p,p-l}^v - d_v Y_l + \tilde{\lambda}_l Z_{\mathrm{mod}(l,J)}^{(i)}, \quad l \in \mathcal{J}_i^n.$$

Here, the ensemble is used only to sample the low-dimensional large-scale state $\mathbf{Z}^{(i)} \in \mathbb{R}^J$, while only a small portion of small-scale modes $\{Y_l\}_{l\in\mathcal{I}_i}$ are grouped with the ith sample $\mathbf{Z}^{(i)}$ together for the time update at $t = t_s$. The union of all the randomly sampled groups forms the entire spectrum of small-scale modes. In this way, we no longer need to run a very large ensemble for the small scales Y_l by exploiting its wide spectrum of modes acting at different large-scale samples. Notice that the samples $\mathbf{Z}^{(i)}$ will no longer stay independent and will be linked by the small-scale modes through the random batches. Still, the important correlations between the small- and large-scale modes are maintained through this splitting of small-scale modes.

Remark. In practice, we may still want to run a small number of small-scale modes $\mathbf{Y}^{(j)}, j = 1, \ldots, M_2$. This is equivalent to introduce an ensemble of size M_2 to the above block model (25) so that the model parameters have the relation $\kappa = \frac{M_1}{M_2} = \frac{JL}{2q}$. For consistent statistics, the large- and small-scale

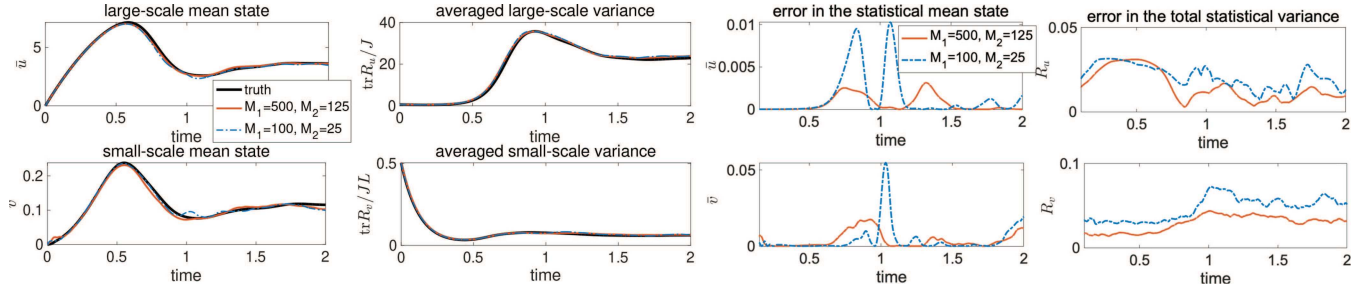

FIG. 13. Prediction of the mean and variances in the reduced-order RBM model with batch size $(p, q) = (4, 32)$ and sample size $(M_1, M_2) = (500, 125), (100, 25)$ for M_1 samples of large-scale state and M_2 of small-scale state. The time evolutions of the errors in mean and variance are also compared.

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

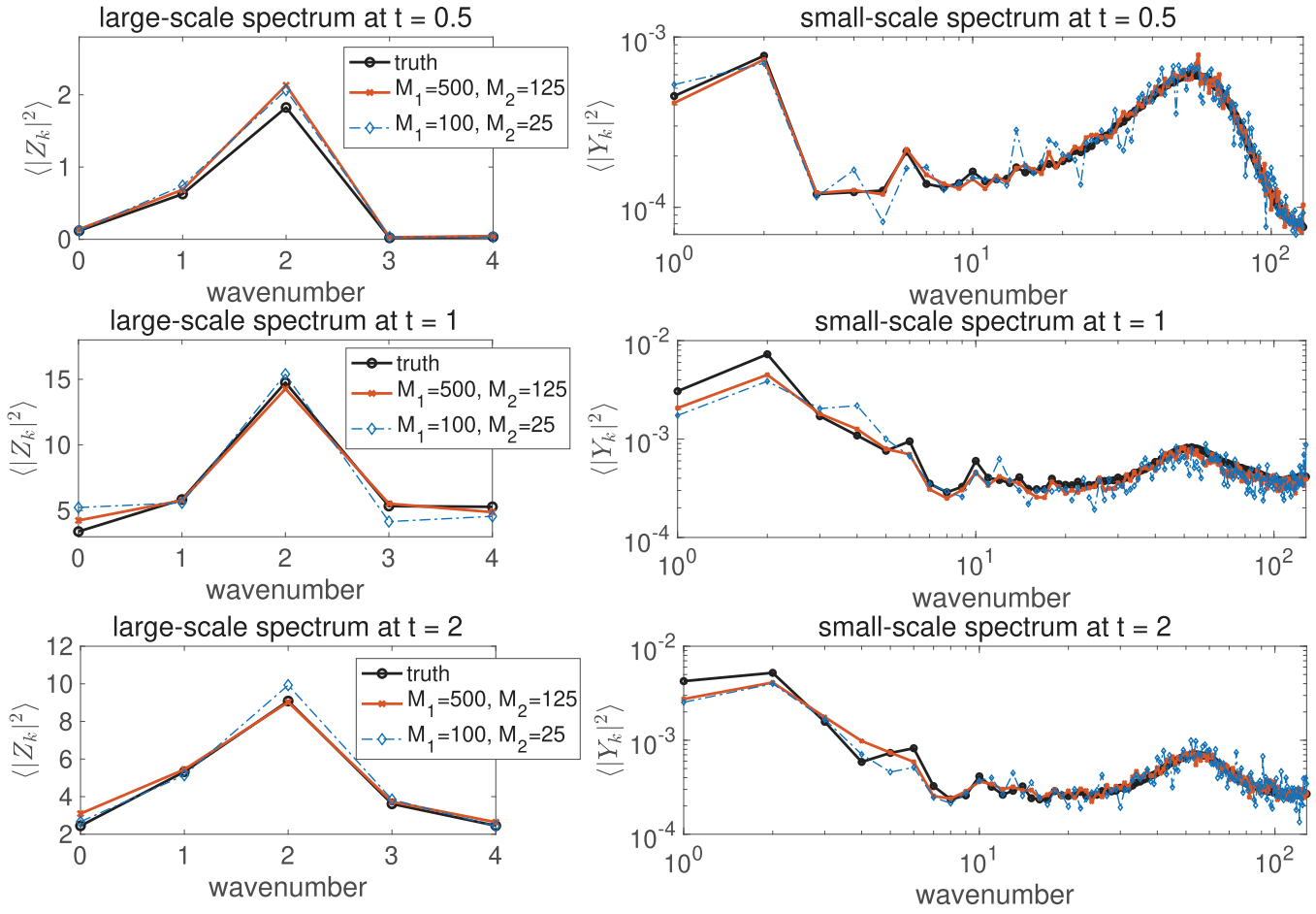

FIG. 14. Reduced-order RBM model prediction of the variance spectra in both large- and small-scale modes at several time instants $t = 0.5, 1, 2$.

coupling coefficient needs to be updated as $\tilde{\lambda}_l = \sqrt{\kappa}\lambda_l$. This leads to the computational reduction from $O\left(M_1 J\left(p + Lq\right)\right)$ in the full RBM model to $O\left(M_1 Jp + M_2 JLq\right) = O\left(M_1\left(q^2 + pJ\right)\right)$, which is only dependent on the dimension J of the large-scale state and independent of the small-scale dimension JL.

In the numerical test of the reduced-order model, we consider the parameter regime of the two-layer L-96 system with $c = 4, b = 10, h = 1, F = 20$. This regime gives a weaker scale separation between the large and small scales, thus setting a more important role for the small-scale processes with a non-negligible contribution to the large-scale modes. We fix the batch size as $(p, q) = (4, 32)$. The relatively large batch size $q = 32$ (compared with the full dimension $JL = 256$) is used to ensure a larger allowed sample size $M_2 = \frac{2qM_1}{JL}$ for the small-scale modes. We pick a moderate sample size $M_1 = 500$ to sample the large-scale mode of dimension J and the large number of small-scale modes only requires a small sample size $M_2 = 125$. Another extreme sample size $M_1 = 100, M_2 = 25$ is also tested. The reduced-order model enables even more efficient computation compared with the very high dimension of the full system $d = J\left(L + 1\right) = 264$.

We show the prediction results in the reduced-order RBM model in Fig. 13. It can be seen that the reduced model maintains the high skill to capture the leading statistics in both the large- and small-scale states while saving additional computational costs by avoiding running the large ensemble simulation. The pointwise errors in mean and variance further confirm the accurate statistical prediction even with a small and reduced sample size. Notice that there are bursts of relatively larger errors in the mean state prediction during the transition regime. This is due to the insufficient instant sampling of the full statistics with a very small ensemble in the RBM approximation. Accordingly, we show the detailed prediction of the energy spectra at several time instants in both large- and small-scale modes in Fig. 14. In the case of reduced-order model due to the very small ensemble size in small scales, it is expected that the small-scale modes will become very noisy and have larger fluctuation errors, especially for the large wavenumber modes. Still, the reduced-order RBM model successfully captures the structure of the

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

entire spectrum in both large and small scales. These encouraging results suggest further applications of the RBM models to more practical and realistic systems exhibiting stronger multiscale coupling and a very wide spectrum covering a large number of scales.

V. SUMMARY

We present a systematic closure modeling framework that enables efficient ensemble prediction of leading-order statistics and non-Gaussian PDFs in complex turbulent systems, which are characterized by strong internal instability and interactions of coupled spatiotemporal scales. A general stochastic–statistical formulation is established derived from a generic multiscale nonlinear system, capable of modeling a wide range of complex phenomena observed in natural and engineering systems. A mean and fluctuation decomposition of the original model state is introduced to deal with the irreducible dynamics in the high-dimensional equations. The fluctuation modes, which capture large uncertainty in multiscale modes, are modeled using stochastic equations that depend on the mean state and covariance of fluctuation modes. The leading-order statistical equations for the mean and covariance incorporate higher-order moment feedback from different scales. A precise high-order closure is introduced using an empirical average of the ensemble prediction of the stochastic equation solution, which explicitly captures detailed multiscale interactions. The coupling between stochastic and statistical equations is further reinforced by a relaxation term for statistical consistency. This approach effectively balances the strong internal instability often encountered in turbulent systems, mitigating the inherent problem of numerical divergence in the statistical solution. The model achieves high skill in capturing non-Gaussian statistics and extreme outliers by explicitly resolving crucial higher-order moments instead of relying on insufficient low-order parameterizations. The stochastic–statistical formulation offers a flexible approach for recovering essential model statistics, making it applicable to a wide range of problems in uncertainty quantification and data assimilation.[7,10,46]

To enable efficient computation using a small ensemble size for the stochastic equation and limit the number of high-order nonlinear coupling terms during each time update, we generalize the idea in RBM approximation for mean-fluctuation equations.[35] The approach decomposes the wide spectrum involving a large number of multiscale modes into small random batches. A much smaller ensemble size is made possible by just sampling the nonlinear coupling terms involving modes in a low-dimensional subspace inside one batch. Simultaneously, the contribution from all the other modes is fully modeled by resampling the batches at each time step, exploiting the ergodicity of the stochastic modes. Consequently, high prediction accuracy concerning all high-order feedback is achieved. For systems with exceptionally high dimensionality, a model reduction strategy is proposed to further reduce the computational cost by linking the large number of small-scale fluctuation modes to the ensemble samples of large-scale state. The resulting algorithms are straightforward to implement and are well-suited for a wide range of multiscale turbulent systems. We evaluate the performance of the proposed RBM models using the representative one-layer and two-layer L-96 systems, which exhibit strong multiscale coupling and a wide spectrum of energetic unstable modes. The models demonstrate uniform high prediction skills for the leading order mean and variance, and more notably, capture the highly non-Gaussian PDFs and extreme events using a very small ensemble. As a result, the computational cost is reduced to an affordable level for genuinely high-dimensional systems. In future research directions, we aim to develop a complete theory to analyze the general stochastic–statistical modeling framework building upon the preliminary estimates presented in this paper. This can provide precise guidelines for the choice of model parameters such as the random batch size and the relaxation parameter to achieve optimal model performance. The promising results suggest potential applications to more realistic high-dimensional systems such as the quasi-geostrophic and Boussinesq equations where the required sample size in the prediction methods will not increase with the dimension of the system. The reduced-order RBM model shows potential in overcoming the curse of dimensionality and providing an effective tool for a wide range of practical problems related to prediction and data assimilation.

ACKNOWLEDGMENTS

The research of J.-G.L. is partially supported by the National Science Foundation (NSF) under Grant No. DMS-2106988. The authors would like to thank the reviewers for the very helpful suggestions that improve the quality of this paper.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

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

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

APPENDIX A: PROOF OF THEOREMS IN SEC. III C

Proof. Define the following functions according to solutions to (11a) and (11b)

$$\tilde{w}(\mathbf{x},t)=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{x}}\varphi\left(\tilde{Z}_k(t)\right),$$
$$w(\mathbf{x},t)=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{x}}\varphi\left(Z_k(t)\right), \tag{A1}$$

with the test function $\varphi\in C_b^2$ and initial state $\mathbf{Z}(0)=\tilde{\mathbf{Z}}(0)=\mathbf{x}\in\mathbb{R}^K$. The functions (A1) defined in the time interval $t\in(t_s,t_{s+1}]$ are governed by the backward Kolmogorov equation[40] as

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

$$\partial_t \tilde{w} = \tilde{\mathcal{L}}_{\mathcal{I}^s} \tilde{w} = \sum_{k=1}^{K} \left[\sum_{m,n} I_k^s(m) \tilde{L}_{v,km}(\bar{u}) x_m + I_k^s(m) I_k^s(n) \tilde{\gamma}_{mnk} (x_m x_n - R_{mn}) \right] \cdot \partial_{x_k} \tilde{w} + \frac{1}{2} \sum_{k=1}^{K} \sigma_k^2 \partial_{x_k}^2 \tilde{w},$$
$$\partial_t w = \mathcal{L} w = \sum_{k=1}^{K} \left[\sum_{m,n} L_{v,km}(\bar{u}) x_m + \gamma_{mnk} (x_m x_n - R_{mn}) \right] \cdot \partial_{x_k} w + \frac{1}{2} \sum_{k=1}^{K} \sigma_k^2 \partial_{x_k}^2 w. \tag{A2}$$

In (A2), the backward equation for the RBM model $\tilde{w}$ is subject to the additional randomness due to the partition $\mathcal{I}^s = \{\mathcal{I}_k^s\}$ during the time updating interval. We introduce the index function defined in time $t \in (t_s, t_{s+1}]$

$$I_k^s(m) = \begin{cases} 1, & \text{if } m \in \mathcal{I}_k^s, \\ 0, & \text{otherwise.} \end{cases}$$

Notice that the function I_k^s is kept constant during each time interval $(t_s, t_{s+1}]$ and will change values subject to the resampling of the random batches. According to the conclusion in Ref. 35, we have the expectation of the partition functions by counting the ordered combinations of the random batches

$$\mathbb{E} I_k^s(m) = \frac{p}{K}, \quad \mathbb{E} I_k^s(m) I_k^s(n) = \frac{p}{K} \frac{p-1}{K-1}.$$

This leads to the consistent condition on the expectation with the partition using the proper choice of the coupling coefficients

$$\mathbb{E}^{\mathcal{I}^s} \tilde{\mathcal{L}}_{\mathcal{I}^s} = \mathcal{L}.$$

Using the semigroup operator $\tilde{\mathcal{S}}$ acting on the function $w(\mathbf{x}, t_s)$ and the above identity, we have

$$\begin{aligned} &\tilde{\mathcal{S}} w(\mathbf{x}, t_s) - w(\mathbf{x}, t_{s+1}) \\ &= \mathbb{E}^{\mathcal{I}^s} e^{\Delta t \tilde{\mathcal{L}}_{\mathcal{I}^s}} w(\mathbf{x}, t_s) - w(\mathbf{x}, t_{s+1}) \\ &= \int_0^{\Delta t} (\Delta t - \tau) \left[\mathbb{E}^{\mathcal{I}^s} \left(\tilde{\mathcal{L}}_{\mathcal{I}^s} \right)^2 e^{\tau \tilde{\mathcal{L}}_{\mathcal{I}^s}} - \mathcal{L}^2 e^{\tau \mathcal{L}} \right] w(\mathbf{x}, t_s) \, \mathrm{d}\tau. \end{aligned}$$

By the assumptions (12), the residual terms on the last equality are uniformly bounded

$$\left\| \left(\tilde{\mathcal{L}}_{\mathcal{I}^s} \right)^2 e^{\tau \tilde{\mathcal{L}}_{\mathcal{I}^s}} w(\cdot, t) \right\|_\infty < C, \quad \left\| \mathcal{L}^2 e^{\tau \mathcal{L}} w(\cdot, t) \right\|_\infty < C.$$

Therefore, the one-step error for between the RBM solution $\tilde{\mathcal{S}} w(\mathbf{x}, t_s)$ and the full model $w(\mathbf{x}, t_{s+1})$ can be estimated as

$$\left\| \tilde{\mathcal{S}} w(\cdot, t_s) - w(\cdot, t_{s+1}) \right\|_\infty \le C \Delta t^2.$$

Finally, by applying $\tilde{\mathcal{S}}$ on the initial function $w(\mathbf{x}, 0) = \varphi(\mathbf{x})$ s times and recurrently using the above contraction property, we compute the total error at $t = t_s$ as

$$\begin{aligned} \left\| \tilde{\mathcal{S}}^{(s)} \varphi(\cdot) - w(\cdot, t_s) \right\|_\infty &\le \left\| \tilde{\mathcal{S}} \left[\tilde{\mathcal{S}}^{(s-1)} \varphi(\cdot) - w(\cdot, t_{s-1}) \right] \right\|_\infty + \left\| \tilde{\mathcal{S}} w(\cdot, t_{s-1}) - w(\cdot, t_s) \right\|_\infty \\ &\le \left\| \tilde{\mathcal{S}}^{(s-1)} \varphi(\cdot) - w(\cdot, t_{s-1}) \right\|_\infty + \left\| \tilde{\mathcal{S}} w(\cdot, t_{s-1}) - w(\cdot, t_s) \right\|_\infty \\ &\le \sum_{i=1}^{s} \left\| \tilde{\mathcal{S}} w(\cdot, t_{i-1}) - w(\cdot, t_i) \right\|_\infty \le C(t_s) \Delta t. \end{aligned}$$

This completes the proof of the theorem. □

Proof of Theorem 3. Under the structure assumptions of the bilinear term (13), the total statistical energy $E = \bar{u}^2 + \frac{1}{K} \mathrm{tr} R$ from the solutions of the mean and covariance equations in (7) satisfies

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -2 dE + \bar{u} \cdot F.$$

Similarly, the RBM model (8) also satisfies the corresponding energy equation with $\tilde{E} = \tilde{\bar{u}}^2 + \frac{1}{K} \mathrm{tr} \tilde{R}$ with the consistent structure symmetry

$$\frac{\mathrm{d}\tilde{E}}{\mathrm{d}t} = -2 d\tilde{E} + \tilde{\bar{u}} \cdot F.$$

By taking the difference between the above two equations, we can write the solution for the difference between the statistical energy in truth and RBM approximation formally as

$$\delta E(t) = \delta \bar{u}^2 + \frac{1}{K} \delta \mathrm{tr} R = \int_0^t e^{-2d(t-s)} \delta \bar{u}(s) \cdot F(s) \, \mathrm{d}s,$$

where we assume the initial states are the same, $E(0) = \tilde{E}(0)$, and $\delta E = E - \tilde{E}$, $\delta \bar{u} = \bar{u} - \tilde{\bar{u}}$. Then, using the statistical estimation for the total variance derived in (14), we find

$$\begin{aligned} M |\delta \bar{u}| \le \left| \bar{u}^2 - \tilde{\bar{u}}^2 \right| &\le \frac{1}{K} \left| \mathrm{tr} R - \mathrm{tr} \tilde{R} \right| + \int_0^t e^{-2d(t-s)} |\delta \bar{u}| \cdot |F| \, \mathrm{d}s \\ &\le C(T) \Delta t + \int_0^t |\delta \bar{u}| \cdot |F| \, e^{-2d(t-s)} \, \mathrm{d}s. \end{aligned} \tag{A3}$$

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

On the left-hand side of the above equation, we use the assumption that the mean states $\left|\bar{u}+\tilde{\bar{u}}\right| > M$ have a uniformly common bound from below. This is observed in the numerical simulations. Finally, applying Grönwall's inequality to (A3), we have the uniformly bound for the convergence of the statistical mean state

$$|\delta\bar{u}|(t) \le C(T)\,\Delta t\, e^{\int_0^t |F| e^{-2d(t-s)}\,\mathrm{d}s} \le C_1(T)\,\Delta t.$$

The final result (17) is achieved by taking the supremum for $t \in [0, T]$. □

APPENDIX B: DETAILED RBM EQUATIONS FOR THE L-96 SYSTEMS

1. Explicit equations for the one-layer L-96 system

According to the general equations (7), the *full stochastic–statistical formulation* for the L-96 system (18) can be derived as

$$\begin{aligned}
\frac{\mathrm{d}\bar{u}}{\mathrm{d}t} &= \frac{1}{J^2}\sum_k \gamma_k r_k - \bar{u} + F,\\
\frac{\mathrm{d}r_k}{\mathrm{d}t} &= \frac{1}{J}\sum_{m-n=k}\left[\left\langle Z_m Z_n^* Z_k^*\right\rangle_p \gamma_{mn} + \left\langle Z_m^* Z_n Z_k\right\rangle_p \gamma_{mn}^*\right]\\
&\quad - 2\left(\mathrm{Re}\gamma_k\bar{u}+1\right) r_k + \epsilon^{-1}\left(\left\langle |Z_k|^2\right\rangle_p - r_k\right), \qquad \text{(B1)}\\
\frac{\mathrm{d}Z_k}{\mathrm{d}t} &= \frac{1}{J}\sum_{m-n=k}\left(Z_m Z_n^* - r_m\delta_{mn}\right)\gamma_{mn} - \left(\gamma_k^*\bar{u}+1\right) Z_k.
\end{aligned}$$

Here, we have the coupling coefficients $\gamma_{mn} = \exp\left(i2\pi\frac{m+n}{J}\right) - \exp\left(-i2\pi\frac{n-2m}{J}\right)$ and $\gamma_k := \gamma_{kk}$. The first two equations are deterministic providing the statistical mean and variance dynamics. The third equation characterizes the stochastic evolution of the coefficients Z_k (with randomness from the initial ensemble). The higher-order moments in the variance equations are recovered by the stochastic equation containing non-Gaussian information, $\langle f\rangle_p = \int f\,\mathrm{d}p$. Therefore, the above equations (B1) provide a closed system by coupling the statistical and stochastic equations including all the higher-order feedback. Specifically, an additional relaxation term is added to the variance equation for r_k to guarantee the consistent statistical mean and variance in the model prediction. In one time update of the above full equations, the mean equation requires J operations for the summation of all the variance. Each component of the J variances requires J operations in the summation together with M operations needed in computing the empirical average of the third moments. Finally, the ensemble simulation of the stochastic coefficients requires J^2 operations for each of the M samples.

Correspondingly, we can derive the *RBM model* for the one-layer L-96 system (B1) in the time updating interval $t \in (t_s, t_{s+1}]$ as

$$\begin{aligned}
\frac{\mathrm{d}\bar{u}}{\mathrm{d}t} &= \frac{1}{J^2}\sum_k \gamma_k r_k - \bar{u} + F,\\
\frac{\mathrm{d}r_k}{\mathrm{d}t} &= c_p\sum_{m\in\mathcal{I}_k^s}\left\langle Z_m Z_{m-k}^* Z_k^*\right\rangle_{pM}\gamma_{m,m-k} + \left\langle Z_m^* Z_{m-k} Z_k\right\rangle_{pM}\gamma_{m,m-k}^*\\
&\quad - 2\mathrm{Re}\gamma_k\bar{u}r_k - 2r_k + \epsilon^{-1}\left(\left\langle |Z_k|^2\right\rangle_{pM} - r_k\right), \qquad \text{(B2)}\\
\frac{\mathrm{d}Z_k^{(i)}}{\mathrm{d}t} &= c_p\sum_{m\in\mathcal{I}_k^s}\left(Z_m^{(i)} Z_{m-k}^{(i)*} - r_m\delta_{mn}\right)\gamma_{m,m-k} - \left(\gamma_k^*\bar{u}+1\right) Z_k^{(i)}.
\end{aligned}$$

Here, the deterministic solutions of the statistical mean $\bar{u}$ and variances r_k are closed by the stochastic equation for the coefficients Z_k, which are sampled by an ensemble simulation using a small sample size of trajectories $i = 1, \ldots, M_1$. The empirical statistics in the dynamics are computed from the ensemble average of the sample realizations at each time updating step

$$\langle f\rangle_{pM} = \frac{1}{M_1}\sum_{i=1}^{M_1} f\left(Z^{(i)}\right).$$

Importantly, it is crucial to use the consistent scaling factor $c_p = \frac{1}{J}\frac{J-1}{p-1}$ according to the batch size p in the summation for higher-order feedback. Through the RBM approach, the set of total spectral modes $\{Z_k\}_{|k|\le J/2}$ are randomly divided into small batches $\mathcal{I}_k^s$ with size p at each time step $t = t_s$. This enables the large computational reduction from $O(J^2)$ to $O(Jp)$ for each single sample trajectory and reduces the sample size from a very large M to $M_1 \ll M$ sampling only the p-dimensional batch subspace.

2. Explicit equations for the two-layer L-96 system

Following the same procedure using the general formulation (7), we first derive the *mean equations for the large- and small-scale states* of the two-layer L-96 system (21)

$$\begin{aligned}
\frac{\mathrm{d}\bar{u}}{\mathrm{d}t} &= \frac{1}{J^2}\sum_{|k|\le J/2}\gamma_k^u r_k^u - \bar{u} + F - \frac{hc}{b}L\bar{v},\\
\frac{\mathrm{d}\bar{v}}{\mathrm{d}t} &= \frac{cb}{J^2L^2}\sum_{|l|\le JL/2}\gamma_l^v r_l^v - c\bar{v} + \frac{hc}{b}\bar{u}.
\end{aligned} \qquad \text{(B3)}$$

Next, we have the set of *covariance equations* associated with the large- and small-scale stochastic coefficients from the Fourier decomposition in (22)

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

$$\begin{aligned}
\frac{\mathrm{d}r_k^u}{\mathrm{d}t} &= \frac{1}{J}\sum_{m-n=k}\left[\left\langle Z_m Z_n^* Z_k^*\right\rangle \gamma_{mn}^u + \left\langle Z_m^* Z_n Z_k\right\rangle \gamma_{mn}^{u*}\right] - 2\left(1+\mathrm{Re}\gamma_k^u \bar{u}\right) r_k^u - \frac{hc}{b}\frac{1}{L}\sum_{s=-L/2+1}^{L/2}\lambda_{k+sJ}^* r_{k+sJ}^x + \epsilon^{-1}\left(\left\langle |Z_k|^2\right\rangle - r_k^u\right), \quad |k| \le J/2,\\
\frac{\mathrm{d}r_l^v}{\mathrm{d}t} &= \frac{cb}{JL}\sum_{p-q=l}\left[\left\langle Y_p Y_q^* Y_l^*\right\rangle \gamma_{pq}^v + \left\langle Y_p^* Y_q Y_l\right\rangle \gamma_{pq}^{v*}\right] - 2c\left(1+b\mathrm{Re}\gamma_l^v \bar{v}\right) r_l^v + \frac{hc}{b}\lambda_l r_l^x + \epsilon^{-1}\left(\left\langle |Y_l|^2\right\rangle - r_l^v\right),\\
\frac{\mathrm{d}r_l^x}{\mathrm{d}t} &= \frac{1}{J}\sum_{m-n=k}\gamma_{mn}^u\left\langle Z_m Z_n^* Y_l^*\right\rangle + \frac{cb}{JL}\sum_{p-q=l}\gamma_{pq}^{v*}\left\langle Y_p^* Y_q Z_{\mathrm{mod}(l,J)}\right\rangle - \left(\gamma_{\mathrm{mod}(l,J)}^{u*}\bar{u} + cb\gamma_l^v\bar{v} + 1 + c\right) r_l^x\\
&\quad + \frac{hc}{b}\lambda_l^*\left(r_{\mathrm{mod}(l,J)}^u - \frac{1}{L}r_l^v\right) + \epsilon^{-1}\left(\left\langle Z_{\mathrm{mod}(l,J)} Y_l^*\right\rangle - r_l^x\right), \quad |l| \le JL/2,
\end{aligned} \tag{B4}$$

where $r_k^u = \langle |Z_k|^2\rangle$, $r_l^v = \langle |Y_l|^2\rangle$ are the variances for the large and small scales, and $r_l^x = \left\langle Z_{\mathrm{mod}(l,J)} Y_l^*\right\rangle$ gives the cross-covariance. The coupling coefficients are $\gamma_{mn}^u = e^{2\pi i\frac{m+n}{J}} - e^{-2\pi i\frac{2m-n}{J}}$, $\gamma_{pq}^v = e^{-2\pi i\frac{p+q}{JL}} - e^{2\pi i\frac{2p-q}{JL}}$, $\lambda_l = \frac{1-e^{-2\pi i\frac{l}{J}}}{1-e^{-2\pi i\frac{l}{JL}}}$, and $\gamma_k^u = e^{4\pi i\frac{k}{J}} - e^{-2\pi i\frac{k}{J}}$, $\gamma_l^v = e^{-4\pi i\frac{l}{JL}} - e^{2\pi i\frac{l}{JL}}$.

Correspondingly, the RBM approximation for the covariance equations is derived according to the full and reduced-order stochastic equations proposed in (24) and (25). The explicit *RBM model equations* during the time interval $t \in (t_s, t_{s+1}]$ can be found as

$$\begin{aligned}
\frac{\mathrm{d}r_k^u}{\mathrm{d}t} &= c_p\sum_{m\in\mathcal{I}_k^s}\left[\left\langle Z_m Z_{m-k}^* Z_k^*\right\rangle \gamma_{m,m-k}^u + \left\langle Z_m^* Z_{m-k} Z_k\right\rangle \gamma_{m,m-k}^{u*}\right] - 2\left(1+\mathrm{Re}\gamma_k^u\bar{u}\right) r_k^u - c_L\frac{hc}{b}\sum_{k+sJ\in\mathcal{I}_k^s}\lambda_{k+sJ}^* r_{k+sJ}^x + \epsilon^{-1}\left(\left\langle |Z_k|^2\right\rangle - r_k^u\right),\\
\frac{\mathrm{d}r_l^v}{\mathrm{d}t} &= c_q cb\sum_{p\in\mathcal{J}_l^s}\left[\left\langle Y_p Y_{p-l}^* Y_l^*\right\rangle \gamma_{p,p-l}^v + \left\langle Y_p^* Y_{p-l} Y_l\right\rangle \gamma_{p,p-l}^{v*}\right] - 2c\left(1+b\mathrm{Re}\gamma_l^v\bar{v}\right) r_l^v + \frac{hc}{b}\lambda_l r_l^x + \epsilon^{-1}\left(\left\langle |Y_l|^2\right\rangle - r_l^v\right),\\
\frac{\mathrm{d}r_l^x}{\mathrm{d}t} &= c_p\sum_{m\in\mathcal{I}_k^s}\gamma_{m,m-k}^u\left\langle Z_m Z_{m-k}^* Y_l^*\right\rangle + c_q cb\sum_{p\in\mathcal{J}_l^s}\gamma_{p,p-l}^{v*}\left\langle Y_p^* Y_{p-l} Z_{\mathrm{mod}(l,J)}\right\rangle - \left(\gamma_{\mathrm{mod}(l,J)}^{u*}\bar{u} + cb\gamma_l^v\bar{v} + 1 + c\right) r_l^x\\
&\quad + \frac{hc}{b}\lambda_l^*\left(r_{\mathrm{mod}(l,J)}^u - \frac{1}{L}r_l^v\right) + \epsilon^{-1}\left(\left\langle Z_{\mathrm{mod}(l,J)} Y_l^*\right\rangle - r_l^x\right),
\end{aligned} \tag{B5}$$

with the important rescaling parameters $c_p = \frac{1}{J}\frac{J-1}{p-1}$, $c_q = \frac{1}{JL}\frac{JL-1}{q-1}$, and $c_L = \frac{1}{L}\frac{L-1}{q-1}$. In the same way as the one-layer case, the higher-order moments are computed through the empirical average of the computed samples.

REFERENCES

[1] A. Majda and X. Wang, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows* (Cambridge University Press, 2006).
[2] J. Pedlosky, *Geophysical Fluid Dynamics* (Springer Science & Business Media, 2013).
[3] D. R. Nicholson and D. R. Nicholson, *Introduction to Plasma Theory* (Wiley, New York, 1983), Vol. 1.
[4] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge University Press, 2003).
[5] U. Frisch, *Turbulence* (Cambridge University Press, 1995).
[6] A. J. Majda, *Introduction to Turbulent Dynamical Systems in Complex Systems* (Springer, 2016).
[7] S. Reich and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation* (Cambridge University Press, 2015).
[8] A. J. Majda and D. Qi, SIAM Rev. **60**, 491 (2018).
[9] A. J. Majda and D. Qi, Chaos **29**, 103131 (2019).
[10] E. Calvello, S. Reich, and A. M. Stuart, arXiv:2209.11371 (2022).
[11] E. Vanden Eijnden, Commun. Pure Appl. Math. **54**, 1146 (2001).
[12] W. Cousins and T. P. Sapsis, Phys. Rev. E **91**, 063204 (2015).
[13] S. Tong, E. Vanden-Eijnden, and G. Stadler, Commun. Appl. Math. Comput. Sci. **16**, 181 (2021).
[14] D. Qi and E. Vanden-Eijnden, AIP Adv. **12**, 025016 (2022).
[15] N. Chen and A. J. Majda, Chaos **30**, 033101 (2020).
[16] Y. Gao, W. Li, and J.-G. Liu, arXiv:2212.05675 (2022).
[17] Y. Gao, J.-G. Liu, and N. Wu, Appl. Comput. Harmon. Anal. **62**, 261 (2023).
[18] Y. Gao, T. Li, X. Li, and J.-G. Liu, Multiscale Model. Simul. **21**, 1 (2023).
[19] M. Leutbecher and T. N. Palmer, J. Comput. Phys. **227**, 3515 (2008).
[20] P. L. Houtekamer, L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, Mon. Weather Rev. **124**, 1225 (1996).
[21] T. Palmer, Q. J. R. Meteorol. Soc. **145**, 12 (2019).
[22] J. H. Friedman, Data Min. Knowl. Discovery **1**, 55 (1997).
[23] D. L. Donoho *et al.*, AMS Math. Challenges Lect. **1**, 32 (2000).
[24] C. H. Bishop, B. J. Etherton, and S. J. Majumdar, Mon. Weather Rev. **129**, 420 (2001).
[25] B. Giggins and G. A. Gottwald, Q. J. R. Meteorol. Soc. **145**, 642 (2019).
[26] T. P. Sapsis and A. J. Majda, Proc. Natl. Acad. Sci. **110**, 13705 (2013).
[27] R. Maulik, O. San, A. Rasheed, and P. Vedula, J. Fluid Mech. **858**, 122 (2019).
[28] Y. Luo, X. Li, and W. Hao, Chaos **32**, 081103 (2022).
[29] D. Qi and J. Harlim, J. Comput. Phys. **485**, 112085 (2023).
[30] M. Lesieur, *Turbulence in Fluids: Stochastic and Numerical Modelling* (Nijhoff Boston, MA, 1987), Vol. 488.
[31] P. H. Diamond, S.-I. Itoh, and K. Itoh, *Modern Plasma Physics* (Cambridge University Press, Cambridge, 2010).

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing

[32]A. J. Majda, M. Moore, and D. Qi, Proc. Natl. Acad. Sci. **116**, 3982 (2019).
[33]S. Jin, L. Li, and J.-G. Liu, J. Comput. Phys. **400**, 108877 (2020).
[34]S. Jin, L. Li, and J.-G. Liu, SIAM J. Numer. Anal. **59**, 746 (2021).
[35]D. Qi and J.-G. Liu, Chaos **33**, 023113 (2023).
[36]E. N. Lorenz, in *Proceedings of Seminar on Predictability* (Reading, 1996), Vol. 1.
[37]D. S. Wilks, Q. J. R. Meteorol. Soc. **131**, 389 (2005).
[38]L. A. Smith, "Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems," in *Nonlinear Dynamics and Statistics* (Springer, 2001), pp. 31–64.
[39]D. Orrell, J. Atmos. Sci. **60**, 2219 (2003).
[40]S. S. Varadhan, *Stochastic Processes* (American Mathematical Society, 2007), Vol. 16.
[41]F. Daum and J. Huang, in *2003 IEEE Aerospace Conference Proceedings (Cat. No. 03TH8652)* (IEEE, 2003), Vol. 4, pp. 4_1979–4_1993.
[42]R. L. Dewar and R. F. Abdullatif, in *Frontiers in Turbulence and Coherent Structures* (World Scientific, 2007), pp. 415–430.
[43]H. Arnold, I. Moroz, and T. Palmer, Philos. Trans. R. Soc. **371**, 20110479 (2013).
[44]A. J. Majda and D. Qi, J. Nonlinear Sci. **26**, 233 (2016).
[45]D. Qi and J. Harlim, Philos. Trans. R. Soc. A **380**, 20210205 (2022).
[46]A. M. Stuart, Acta Numer. **19**, 451 (2010).

29 October 2023 08:59:52

Published under an exclusive license by AIP Publishing