Chaos
An Interdisciplinary Journal of Nonlinear Science

A random batch method for efficient ensemble forecasts of multiscale turbulent systems

Cite as: Chaos **33**, 023113 (2023); https://doi.org/10.1063/5.0129127
Submitted: 03 October 2022 • Accepted: 16 January 2023 • Published Online: 13 February 2023

Di Qi and Jian-Guo Liu

A random batch method for efficient ensemble forecasts of multiscale turbulent systems

Cite as: Chaos **33**, 023113 (2023); doi: 10.1063/5.0129127
Submitted: 3 October 2022 · Accepted: 16 January 2023 ·
Published Online: 13 February 2023

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

[a)]**Author to whom correspondence should be addressed:** qidi@purdue.edu
[b)]jliu@phy.duke.edu

ABSTRACT

A new efficient ensemble prediction strategy is developed for a multiscale turbulent model framework with emphasis on the nonlinear interactions between large and small-scale variables. The high computational cost in running large ensemble simulations of high-dimensional equations is effectively avoided by adopting a random batch decomposition of the wide spectrum of the fluctuation states, which is a characteristic feature of the multiscale turbulent systems. The time update of each ensemble sample is then only subject to a small portion of the small-scale fluctuation modes in one batch, while the true model dynamics with multiscale coupling is respected by frequent random resampling of the batches at each time updating step. We investigate both theoretical and numerical properties of the proposed method. First, the convergence of statistical errors in the random batch model approximation is shown rigorously independent of the sample size and full dimension of the system. Next, the forecast skill of the computational algorithm is tested on two representative models of turbulent flows exhibiting many key statistical phenomena with a direct link to realistic turbulent systems. The random batch method displays robust performance in capturing a series of crucial statistical features with general interests, including highly non-Gaussian fat-tailed probability distributions and intermittent bursts of instability, while requires a much lower computational cost than the direct ensemble approach. The efficient random batch method also facilitates the development of new strategies in uncertainty quantification and data assimilation for a wide variety of general complex turbulent systems in science and engineering.

It remains a grand challenge to obtain accurate statistical strategies for the understanding and forecast of key physical processes in many natural and engineering systems concerning complex turbulent dynamics, such as climate forecast. A new efficient ensemble forecast algorithm is developed dealing with the nonlinear multiscale coupling mechanism as a characteristic feature in high-dimensional turbulent systems. The high computational cost in large ensemble simulations is greatly reduced in the efficient algorithm, while the central statistical quantities involving multiscale interactions are fully recovered with accuracy exploiting a random batch decomposition of the wide spectrum of fluctuation states. The proposed strategy also shows potential to facilitate the development of new effective methods in uncertainty quantification and data assimilation in a wide variety of complex turbulent systems.

I. INTRODUCTION AND BACKGROUND

Turbulent dynamical systems appearing in many natural and engineering fields[1–5] are characterized by a wide range of spatiotemporal scales in a high-dimensional phase space. Small uncertainties in the multiscale high-dimensional states can be rapidly amplified through the nonlinearly coupled dynamics and inherent instability possessed by the turbulent flow. These distinctive features give rise to a wide variety of complex phenomena, such as intermittent bursts of extreme flow structures and strongly non-Gaussian probability density functions (PDFs) in the key state variables.[6–9] A probability representation for the evolution of the major flow states is, thus, essential to accurately quantify the uncertainty in the practical prediction of such turbulent systems. The ensemble forecast through a Monte Carlo (MC) type approach estimates the evolution of the PDFs by tracking an ensemble of trajectories solved

independently from an initial distribution.[10–12] The empirical statistics of the ensemble solutions are used to approximate the model uncertainty due to randomness from various internal and external sources. A particular issue with large societal impact is to accurately capture the non-Gaussian PDFs related to the extreme event outliers[13–15] using the finite size ensemble. However, the "curse-of-dimensionality" forbids direct MC simulations of such high-dimensional systems especially in cases, including strongly coupled multiscale nonlinear interactions.[4,11] A very large ensemble size is usually needed to sufficiently sample the entire coupled fluctuation modes in a wide energy spectrum, while only a small number of solutions are affordable in many practical situations, such as climate forecast.[3,16] New efficient methods are developed for improving ensemble forecasts in multiscale systems without a clear scale separation[17,18] that offer a promising approach to generate diverse samples. Still, it remains a grand challenge to obtain accurate statistical estimates for the key physical quantities from the multiscale interaction between the large-scale mean flow and the interacting high-dimensional small-scale fluctuations.

In this paper, we propose an efficient method for the understanding and ensemble forecast of a general group of complex turbulent systems accepting coupled multiscale dynamics.[19,20] We use the ideas in the random batch method (RBM) originally developed for interacting particle systems[21–23] and apply it to the very different problem of multiscale turbulent systems. Inspired by the stochastic gradient descent[24–26] in machine learning, the RBM randomly divides and constrains the large number of interacting particles into small batches in each time interval. The RBM can greatly reduce the computational cost in large particle systems and has many successful applications, such as on manifold learning[14,23] and quantum simulations.[27] In the ensemble prediction of turbulent systems, we focus on the dominant flow structure in the largest scale; thus, the required ensemble size to sample a low-dimensional subspace can be controlled. This reduced modeling strategy usually suffers difficulties in practice since the large scale is closely coupled with all the unresolved small-scale fluctuations; thus, it is impossible to only perform ensemble simulation inside the large-scale subspace. Using random batches, this difficulty is effectively overcome by regrouping the large number of small-scale fluctuating modes into small batches each containing only a few modes. Then, the batches from a single simulation of the fluctuation modes are used for updating different ensemble members of the large-scale state separately during a short time step update. This approximation is based on the important observation that the small-scale fluctuations often decorrelate much faster in time and contain less energy than the mean-flow state on the largest scale. The batches of different modes are randomly resampled before each time updating step; thus, contributions from all scales are well represented during the evolution in time. In this way, we achieve an efficient algorithm that gains high skill in capturing the fully non-Gaussian statistical feature in the most important large-scale mean-flow state, while greatly reduce the high computational cost independent of the dimensionality of the full system.

In order to achieve a detailed analysis of the method for the complex turbulent system, which usually combines various complex effects from different sources, we develop the new RBM algorithm on a class of simplified turbulent systems with emphasis on the explicit coupling between large and small scales, while the unresolved nonlinear coupling among small scales is approximated by linear damping and white noise forcing. The simplified formulation reveals the most important key physical processes on the nonlinear interaction between the large-scale mean-flow component and the smaller scale fluctuation components. On the other hand, the extra complexity due to the nonlinear self-interactions among the less important small scales is avoided to provide a cleaner model setup. The model framework is shown to have many representative applications in physical and engineering problems.[1,28–31] Precise error estimations are derived for the RBM approach based on this general framework using a finite number of samples. The convergence of the statistical quantities is proved using the semigroups generated by the backward Kolmogorov equations[32] of the RBM model. The RBM performance is then verified numerically on two representative prototype turbulent models. Different non-Gaussian statistics are observed in the two models inferring strong intermittency and extreme events induced from distinct physical mechanisms. The numerical tests show accurate prediction of both transient and equilibrium PDFs recovering various non-Gaussian features under a much lower computational cost requiring a much smaller ensemble size.

First in the following, we start with a general formulation, including dynamical structures that are representative in the turbulent systems. Then, a simplified multiscale model is derived from the general framework as the central model for the study in this paper to enable an accurate ensemble prediction strategy with precise analysis of the probability solutions.

A. General formulation for turbulent systems and challenges in efficient statistical forecast

The complex turbulent systems discussed above can be written as the following canonical equation about the state variable $\mathbf{u}$ in a high-dimensional phase space:

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

On the right hand side of Eq. (1), the first component, $\Lambda = -D + L$, represents linear dissipation and dispersion effects (with a negative-definite dissipation operator $-D < 0$ and a skew-symmetric dispersion operator $L^T = -L$ as in Ref. 19). One representative feature of such complex systems is the nonlinear energy conserving interaction that transports energy across scales. The nonlinear effect is introduced through a bilinear quadratic form, $B(\mathbf{u},\mathbf{u})$, that satisfies the conservation law $\mathbf{u}\cdot B(\mathbf{u},\mathbf{u}) \equiv 0$. External forcing effects are decomposed into a deterministic component, $\mathbf{F}(t)$, and a stochastic component represented by a Gaussian random process, $\boldsymbol{\sigma}\dot{\mathbf{W}}$.

The evolution of the model state $\mathbf{u}$ depends on the sensitivity to the randomness in initial conditions and external stochastic effects. Combined with the inherent internal instability due to the nonlinear coupling term in (1), small perturbations are amplified in time, thus requiring a probabilistic description to completely characterize the development of uncertainty in the model state $\mathbf{u}$. The time evolution of the PDF $p(\mathbf{u},t)$ can be found directly from the solution of the

associated Fokker–Planck equation (FPE),[32]

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

with an initial condition $p\,|_{t=0} = \mu_0$. However, it is still a challenging task to directly solve the FPE (2) as a high-dimensional PDE system. As an alternative approach, ensemble forecast by tracking the Monte Carlo solutions[11] estimates the essential statistics through empirical averages among a group of independently sampled trajectories of (1). In particular, the ensemble members $\{\mathbf{u}^{(i)}\}$ are sampled at the initial time $t = 0$ according to the initial distribution μ_0. The PDF solution $p\left(\mathbf{u}, t\right)$ at each time instant $t > 0$ is then approximated by evolving each sample independently in time according to the same dynamical equation.

Though simple to implement, a direct ensemble forecast running the original model (1) suffers several difficulties in accurately recovering the key model statistics and PDFs in a high-dimensional space. First, the ensemble size required to achieve desirable accuracy will grow exponentially in direct ensemble simulation of the full model as the dimension of the system increases. Second, the turbulent systems often contain strong internal instability and multiscale coupling along the entire spectrum. Thus, reduced models by simply truncating the stabilizing small-scale modes[19] are not feasible to correctly represent the true model dynamics. Besides, different orders of statistical characteristics are fully coupled in the general formulation (1); therefore, it is difficult to identify the contributions of different scales especially when highly non-Gaussian statistics appear. These are the central difficulties we will address in this paper using ideas in the RBM approximation.

In the following part of the paper, we first propose a more tractable model framework with emphasis on the explicit coupling between large and small scales in Sec. II. Based on this model framework, the RBM algorithm for ensemble prediction is developed in Sec. III. Detailed error estimation and convergence analysis of the new RBM method are obtained in Sec. IV. The performance of the method is verified on two prototype turbulence models with practical importance in Sec. V. A summary of this paper is given in Sec. VI.

II. A TURBULENT MODEL FRAMEWORK WITH AN EXPLICIT MULTISCALE COUPLING MECHANISM

One major difficulty in complex turbulent systems is the strong nonlinear coupling across scales where the large-scale state can destabilize the smaller scales with a small variance, while the increased fluctuation energy contained in a large number of small-scale states can inversely impact the development of the coherent largest-scale structure. To address this central issue of coupling with mixed scales in modeling turbulent systems, we introduce the *simplified multiscale model with explicit large–small scale interactions*,

$$\begin{aligned} \frac{\mathrm{d}\bar{\mathbf{u}}}{\mathrm{d}t} &= V\left(\bar{\mathbf{u}}\right) + \frac{1}{K} \sum_{k,l=1}^{K} Z_k Z_l \bar{B}\left(\mathbf{e}_k, \mathbf{e}_l\right) + \mathbf{F}, \\ \frac{\mathrm{d}Z_k}{\mathrm{d}t} &= \frac{1}{K} \sum_{l=1}^{K} \gamma_{kl}\left(\bar{\mathbf{u}}\right) Z_l \; - d_k Z_k + \sigma_k \dot{W}_k, \qquad k = 1, \ldots, K, \end{aligned} \tag{3}$$

focusing on the multiscale interaction between the large-scale state $\bar{\mathbf{u}}$ and the fluctuation coefficients Z_k defined next by the proper decomposition in (4). The simplified model structure enables us to focus on the key nonlinear large- and small-scale coupling mechanism in the general system (1) and will serve as the main model for the construction of efficient ensemble forecast methods.

Here, we illustrate the derivation of the central governing model (3) according to a mean-fluctuation decomposition of the original model state $\mathbf{u}$ in (1) so that the multiscale interactions can be identified in a natural way. To achieve this, we view $\mathbf{u}$ as a random field and separate it into the composition of *a large-scale random mean state* and *stochastic fluctuations* in a finite-dimensional representation under a basis $\{\mathbf{e}_k\}_{k=1}^k$ (which can be pre-determined based on the specific problem, for example, the Fourier basis as a natural choice for periodic boundary conditions),

$$\mathbf{u}\left(t;\omega\right) = \bar{\mathbf{u}} + \mathbf{u}' := \bar{\mathbf{u}}\left(t;\omega\right) + \frac{1}{\sqrt{K}} \sum_{k=1}^{K} Z_k\left(t;\omega\right) \mathbf{e}_k, \tag{4}$$

where the overbar "$\bar{\bullet}$" denotes coarse-graining from a spatial average operator. In this way, the mean state $\bar{\mathbf{u}}\left(t;\omega\right)$ represents the dominant spatial structure with randomness consisting of a collection of large-scale modes (for example, the zonal jets in geophysical turbulence[28] or the coherent radial flow in fusion plasmas[33]); and $\mathbf{Z}\left(t;\omega\right) = \{Z_k\left(t;\omega\right)\}_{k=1}^K$ are stochastic coefficients measuring the uncertainty in multiscale fluctuation processes $\mathbf{u}'$ on the basis $\mathbf{e}_k$ (with a zero averaged mean $\overline{\mathbf{u}'} = 0$). Usually, $\bar{\mathbf{u}}$ can be represented in a much lower dimension d than the dimension K of full stochastic modes $\mathbf{Z}$ representing fluctuations along a wide spectrum of scales. The state decomposition (4) enables us to analyze the individual contributions from different scale modes to the large- and small-scale dynamics. Similar ideas for proper state decomposition have been widely used for modeling turbulent processes, such as the closed stochastic models[34] and the Fourier decimated system.[35]

First, by averaging over the original equation (1) and applying the mean-fluctuation decomposition (4), the evolution equation of the large-scale mean state$\bar{\mathbf{u}}$ is given by the following dynamics:

$$\frac{\mathrm{d}\bar{\mathbf{u}}}{\mathrm{d}t} = \Lambda \bar{\mathbf{u}} + B\left(\bar{\mathbf{u}}, \bar{\mathbf{u}}\right) + \frac{1}{K} \sum_{k,l=1}^{K} Z_k Z_l \bar{B}\left(\mathbf{e}_k, \mathbf{e}_l\right) + \mathbf{F}. \tag{5a}$$

Above, the small-scale nonlinear fluctuating feedback to the large-scale mean dynamics is represented by the quadratic coupling $Z_k Z_l$ with the coupling coefficients $\bar{B}\left(\mathbf{e}_k, \mathbf{e}_l\right)$. The term $B\left(\bar{\mathbf{u}}, \bar{\mathbf{u}}\right)$ represents the self-interaction within the mean state. Next, by projecting the fluctuation equation to each orthogonal basis element $\mathbf{e}_i$, we obtain the evolution equation for the stochastic fluctuation coefficients

$$\begin{aligned} \frac{\mathrm{d}Z_k}{\mathrm{d}t} = \frac{1}{K} \sum_{l=1}^{K} \gamma_{kl}\left(\bar{\mathbf{u}}\right) Z_l + \frac{1}{K^{3/2}} \sum_{m,n=1}^{K} Z_m Z_n B'\left(\mathbf{e}_m, \mathbf{e}_n\right) \cdot \mathbf{e}_k \\ + \frac{\sigma\left(t\right)}{K^{1/2}} \dot{\mathbf{W}}\left(t;\omega\right) \cdot \mathbf{e}_k, \end{aligned} \tag{5b}$$

where $\gamma_{kl}\left(\bar{\mathbf{u}}\right) = \left[\Lambda \mathbf{e}_l + B'\left(\bar{\mathbf{u}}, \mathbf{e}_l\right) + B'\left(\mathbf{e}_l, \bar{\mathbf{u}}\right)\right] \cdot \mathbf{e}_k$ characterizes the quasilinear coupling from the mean state $\bar{\mathbf{u}}$ in the fluctuation modes $\mathbf{Z}$. The interactions between the fluctuation modes in different scales are summarized in the second term on the right hand side of (5b)

with the fluctuation coefficients $B' = B - \bar{B}$. In addition, without loss of generality, we assume that the deterministic forcing $\mathbf{F}$ exerts on the large-scale mean state, while the fluctuation modes are subject to the coupled white noise forcing.

Still, the fully coupled mean-fluctuation model (5) contains multiple linear and nonlinear interaction components involving both large-scale mean $\bar{\mathbf{u}}$ and small-scale fluctuations $\mathbf{u}'$; thus, it may not be a desirable starting model for identifying the central dynamics in multiscale interactions. Rather, we would like to propose a further simplified model based on a mean-fluctuation interaction mechanism, which only maintains the key large and small-scale interaction explicitly while eliminates the large number of small-scale self-interaction terms. Considering this, we assume that the combined nonlinear feedback among different small-scale modes (5b) can be parameterized by independent damping and stochastic forcing,

$$\frac{\mathrm{d}Z_k}{\mathrm{d}t} = \frac{1}{K}\sum_{l=1}^{K}\gamma_{kl}\left(\bar{\mathbf{u}}\right)Z_l \; - d_k Z_k + \sigma_k \dot{W}_k. \tag{6}$$

Above, the linear and nonlinear effects within the mean state, $\Lambda\bar{\mathbf{u}}$ and $B\left(\bar{\mathbf{u}},\bar{\mathbf{u}}\right)$, are summarized in a single term $V\left(\bar{\mathbf{u}}\right)$. The model simplification occurs in the small-scale dynamics (5b) for Z_k which replaces the combined nonlinear small-scale coupling $Z_m Z_n B'\left(\mathbf{e}_m,\mathbf{e}_n\right)\cdot\mathbf{e}_k$ by statistically equivalent damping and noise with parameters d_k, σ_k. In fact, this replaced term represents the higher-order moment feedback to the covariance dynamics from a detailed statistical analysis of the moment equations and can be generalized to include inhomogeneous effects as d_{kl}, σ_{kl}.[19] The simplification is derived from the important observation that these small-scale modes are fast mixing (thus decorrelate fast in time); therefore, the average of the large number of modes plays an equivalent role as the linear damping and white noise as in the homogenization theory. The introduced model damping parameter d_k is usually picked according to the decorrelation time of the mode Z_k, and the noise parameter σ_k is determined by the equilibrium statistical spectrum $E_k = \frac{\sigma_k^2}{2d_k} = E_0\left|k\right|^{-s}$[28] with s determining the energy decaying rate in the fluctuation modes.

The resulting simplified model (3) recovers the most essential coupling mechanism between the large-scale mean state $\bar{\mathbf{u}}$ and the small-scale fluctuation modes Z_k, which is explicitly modeled through the quasi-linear operator $\gamma_{kl}\left(\bar{\mathbf{u}}\right)$ in the fluctuation equations and globally linked to the mean equation through the fully coupled quadratic feedback term. Notice that the strong internal instability, which is the key feature of turbulent systems, is maintained in both the mean and fluctuation modes by coupling terms V and γ_{kl} in (5a) and (5b). The only model approximation comes from parameterization of the complicated self-interactions of small-scale fluctuations. Using this simplified model avoids the various sources of uncertainties from the fluctuation scales so that the dominant large–small scale interaction is identified. On the other hand, the simplified model lacks the skill to capture the detailed interactions among the small scales; thus, it is not sufficient to model the locally coupled states, such as what occurs in the multiscale Lorenz model.[17] The thorough analysis of the fully coupled nonlinear fluctuation model (5b) will be left for the future study.

As a final comment, the multiscale model formulation (3) enjoys the advantage of more flexibility to run ensemble simulations for statistical forecast, uncertainty quantification, and data assimilation in practical applications. A wide variety of turbulent systems (as well as a number of approximation reduced-order models)[30,36] can be categorized into this framework so that it has wide validity in developing the efficient ensemble forecast methods. Two typical prototype models accepting the dynamical structure (3) with a wide multiscale spectrum will be discussed in Sec. V displaying a wide variety of different turbulent features. In Secs. III–IV, we will develop efficient ensemble forecast strategies based on this representative turbulent model formulation.

III. RANDOM BATCH METHOD FOR ENSEMBLE FORECAST OF TURBULENT MODELS

Next, we propose an efficient ensemble forecast method to describe the time evolution of the probability distribution of the state $\mathbf{u}$. The direct Monte Carlo approach runs an ensemble simulation using N independent samples $\mathbf{u}^{(i)} = \left\{\bar{\mathbf{u}}^{(i)}, \mathbf{Z}^{(i)}\right\}, i = 1,\ldots,N$, with $\bar{\mathbf{u}} \in \mathbb{R}^d$ being the large-scale mean state and $\mathbf{Z} = \{Z_k\}_{k=1}^K \in \mathbb{R}^K$ the entire small-scale fluctuation modes in a high dimensional space $K \gg d$. The samples are drawn from the initial distribution $\mathbf{u}^{(i)}(0) \sim \mu_0(\mathbf{u})$ at the starting time $t = 0$, and the time-dependent solution of each sample $\mathbf{u}^{(i)}(t)$ is achieved by solving Eq. (3) independently in time. The resulting PDF at each time instant t is approximated by the empirical ensemble representation,

$$p\left(\mathbf{u},t\right) \simeq p^{\mathrm{MC}}\left(\mathbf{u},t\right) := \frac{1}{N}\sum_{i=1}^{N}\delta\left(\mathbf{u}-\mathbf{u}^{(i)}\left(t\right)\right), \quad \mathbf{u}\in\mathbb{R}^{d+K}, \tag{7}$$

with N being the total number of samples for the entire $(d+K)$-dimensional space. Associated with the PDF, the statistical expectation of any function $\varphi(\mathbf{u})$ can be estimated by the empirical average of the samples according to (7),

$$\mathbb{E}^p\varphi_t\left(\mathbf{u}\right) \simeq \mathbb{E}^{\mathrm{MC}}\varphi_t\left(\mathbf{u}\right) = \frac{1}{N}\sum_{i=1}^{N}\varphi\left(\mathbf{u}^{(i)}\left(t\right)\right).$$

In particular, for the model (3), all the small-scale modes contribute to the mean state equation as a combined feedback. As a result, even though we are mostly interested in the statistics from the mean state samples $\bar{\mathbf{u}}^{(i)}$ in the relatively low-dimensional subspace, the solution of entire K small-scale modes $\mathbf{Z}^{(i)}$ must be computed. The K-dimensional equations for fluctuation modes also need to be solved repeatedly N times for all the samples $i = 1,\ldots,N$. Thus, the direct ensemble method reaches a high computational cost of $O\left(NK^2\left(d+1\right)\right)$ for one time step update. Furthermore, the required number of samples N to maintain accuracy in the empirical PDF (7) is also dependent on the system dimension $(d+K)$ and will grow exponentially as K increases (known as the curse-of-dimensionality[37,38]). Therefore, this direct MC approach will quickly become computational unaffordable as a larger N is needed to resolve all the detailed small-scale fluctuations.

Here, we propose to reduce the computational cost in ensemble simulations of the turbulent models using the idea in the effective random batch method.[21,22] We focus on the ensemble sampling of

the most dominant mean state statistics $\bar{\mathbf{u}} \in \mathbb{R}^d$ in a much lower dimensional subspace $d \ll K$. Thus, accurate empirical estimation of the marginal probability distribution of $\bar{\mathbf{u}}$ in (3) can be reached using a much smaller ensemble size $N_1 \ll N$ sampling only the dominant large-scale state,

$$p^{\mathrm{RBM}}(\bar{\mathbf{u}}) = \frac{1}{N_1}\sum_{i=1}^{N_1}\delta\left(\bar{\mathbf{u}} - \bar{\mathbf{u}}^{(i)}\right), \quad \bar{\mathbf{u}} \in \mathbb{R}^d. \tag{8}$$

Accordingly, the expectation in the resolved mean state is computed by the empirical estimation, $\mathbb{E}^{\mathrm{RBM}}\varphi_t(\bar{\mathbf{u}}) = \frac{1}{N_1}\sum_i \varphi\left(\bar{\mathbf{u}}^{(i)}(t)\right)$. In the main idea of the RBM model, we no longer run the large ensemble simulation of the full fluctuation modes $\{\mathbf{Z}^{(i)}\} \in \mathbb{R}^{K\times N}$ associated with each mean state sample $\bar{\mathbf{u}}^{(i)}$. Instead, only one stochastic trajectory of $\mathbf{Z}(t)$ is solved in time, while the K spectral modes are divided into smaller subgroups (batches) for updating different ensemble members $\bar{\mathbf{u}}^{(i)}$. The RBM approximation is suitable considering the typical property of the turbulent system where the energy inside the single small-scale mode $\mathbb{E}|Z_k|^2, k \gg 1$ decays fast and decorrelates rapidly in time (see examples in Figs. 1 and 4 of Sec. V). On the other hand, ergodicity in the stochastic fluctuation modes[39,40] implies that updating the mean state $\bar{\mathbf{u}}$ using fractional fluctuation modes at each time step with consistent time-averaged feedback can provide an equivalent total contribution without altering the original statistical equations.

Next, we describe the detailed RBM approach for modeling turbulent systems. To accurately quantify the small-scale feedback in the mean state dynamics, we introduce a partition $\mathcal{I}^n = \{\mathcal{I}_i^n\}$ of the mode index $k = 1, \ldots, K$ at the start of each time updating step $t = t_{n-1}$. Thus, the full spectrum of modes is divided into N_1 small batches of size $p = |\mathcal{I}_i^n|$ according to the total number of samples $i = 1, \ldots, N_1$ of $\bar{\mathbf{u}}^{(i)}$,

$$\cup_{i=1}^{N_1}\mathcal{I}_i^n = \{k : 1 \le k \le K\}. \tag{9}$$

Then, the ith sample of the mean state $\bar{\mathbf{u}}^{(i)}$ is updated only using the fluctuation modes $\{Z_k\}$ whose indices belong to the batch $k \in \mathcal{I}_i^n$ during the time interval $t \in (t_{n-1}, t_n]$ with time integration step $\tau = t_n - t_{n-1}$,

$$\frac{\mathrm{d}\bar{\mathbf{u}}^{(i)}}{\mathrm{d}t} = V\left(\bar{\mathbf{u}}^{(i)}\right) + \sum_{k,l\in\mathcal{I}_i^n} c_{kl}Z_k Z_l \bar{B}(\mathbf{e}_k, \mathbf{e}_l) + \mathbf{F}. \tag{10a}$$

Correspondingly, only the modes $\{Z_k\}$ whose indices k are in the batch $\mathcal{I}_i^n$ are updated using the ith mean state solution,

$$\frac{\mathrm{d}Z_k}{\mathrm{d}t} = \frac{1}{p}\sum_{l\in\mathcal{I}_i^n}\gamma_{kl}\left(\bar{\mathbf{u}}^{(i)}\right)Z_l - d_k Z_k + \sigma_k \dot{W}_k, \quad k \in \mathcal{I}_i^n. \tag{10b}$$

Above, the coupled feedback of small-scale modes in the batch is rescaled by the new combining coefficients c_{kl}

$$c_{kl} = \begin{cases} \frac{1}{p}, & k = l, \\ \frac{1}{p}\frac{K-1}{p-1}, & k \ne l, \end{cases} \tag{11}$$

which will be explained in Sec. IV. Equations (10a) and (10b) combining all the batches $i = 1, \ldots, N_1$ give the complete formulation of the *random batch model for statistical ensemble forecast of turbulent systems* during the time interval $(t_{n-1}, t_n]$. Then, the batches are regrouped again at the start of the next time step $t = t_n$ to repeat this procedure. Through the RBM reduction, only a very small number of modes p is needed in each batch i,

$$p = \frac{K}{N_1} \ge 2.$$

Notice that the batch size p should be at least 2 to include one quadratic interaction of modes in the mean equation. In practice, it is shown that the batch size p can be sufficiently small [for example, we take $p = 5$ or even $p = 2$ compared with $K = O(100)$ for the numerical tests in Sec. V]. Through the shared modes among batches, the computational cost of the RBM model (10) is effectively reduced to $O\left(N_1 dp^2\right) = O\left(dpK\right)$. Particularly, the cost will not have the exponential growth depending on the full dimension $d + K$ of the system since the samples of size N_1 are only used to capture statistics in the low-dimensional state $\bar{\mathbf{u}} \in \mathbb{R}^d$.

We summarize the random batch method for ensemble simulation of high-dimensional turbulent system in the following algorithm:

Remark:

1. The total number of batches N_1 is associated with the number of small-scale fluctuation modes K. In the RBM model (10), we assume that the number of fluctuation modes is large enough $K \gg 1$ so that it can offer sufficient samples $N_1 = K/p$ to approximate the PDF of $\bar{\mathbf{u}}$. Sometimes, it is useful to expand the sample size to $N_1 = n_2 K/p$ by running a small number n_2

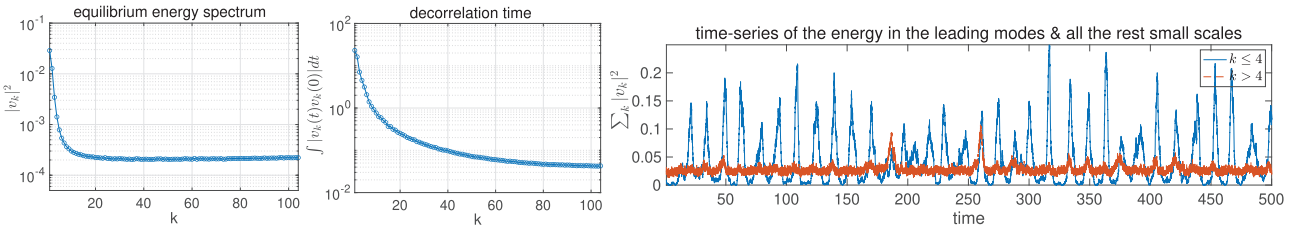


FIG. 1. Statistics of the conceptual turbulence model (21). Left: equilibrium energy spectrum and decorrelation time in the fluctuation modes. Right: time series of the energy in the first four leading modes and energy in all the rest fluctuation modes.

ALGORITHM 1. RBM model for ensemble forecast of turbulent systems.

Initial condition: At initial time $t = t_0$, draw random samples from the initial distribution $\bar{\mathbf{u}}^{(i)}(t_0) \sim \mu_0(\bar{\mathbf{u}})$ with $i = 1, \ldots, N_1$, and one associated sample for $\{Z_k(t_0)\}_{k=1}^K$.

1: **for** $n = 1$ while $n < T/\tau$, at the start of the time interval $t \in (t_{n-1}, t_n]$ with time step $\tau = t_n - t_{n-1}$ **do**
2: Collect samples $\bar{\mathbf{u}}^{(i)}(t_{n-1})$ and fluctuation modes $\{Z_k(t_{n-1})\}_{k=1}^K$ from all output batches of previous time step;
3: Partition the modes into N_1 batches $\mathcal{Z}_i = \{Z_k\}_{k\in\mathcal{I}_i^n}$ with $\cup_{i=1}^{N_1}\mathcal{Z}_i = \{Z_k\}_{k=1}^K$ according to (9);
4: Update the samples $\bar{\mathbf{u}}^{(i)}(t_n)$ and the spectral modes $\{Z_k(t_n)\}_{k\in\mathcal{I}_i^m}$ in each batch i to the next time step by solving Eq. (10).
5: **end for**

of samples for the fluctuation modes $Z_k^{(j)}$ with $j = 1, \ldots, n_2$. The sampled solutions are divided into small batches of size p according to the index j in the same way as before. The batches can also change size $p_i = |\mathcal{I}_i^n|$ according to the energy in each mode.

2. The RBM algorithm can model the internal instability in the leading fluctuation modes as a crucial feature in the turbulent dynamics. In practical applications to estimate the marginal distribution (8), the number of the resolved modes for the ensemble simulation can be extended to also include the leading fluctuation modes $Z_k, k \le K_1$ displaying unstable dynamics $\gamma_{kl} > 0$. The numerical scheme can be easily generalized to this case following exactly the same steps of Algorithm 1. Examples are shown in the unstable test models in Sec. V and explicit formulations in (B1) and (B2) in Appendix B.

IV. ERROR ESTIMATE OF THE RANDOM BATCH ALGORITHM FOR ENSEMBLE PREDICTION

In this section, we analyze the approximation error in the RBM model in Algorithm 1 compared with the direct Monte Carlo approach for the turbulent model. Ensemble simulation of the coupled large–small scale system (3) is considered for probabilistic prediction of the large-scale mean state $u \in \mathbb{R}^d$ (we modified the previous notation for the mean state $\bar{\mathbf{u}}$ in this section for a cleaner representation in the proofs). For the direct MC model, the governing equations for each ensemble member of the full states $\{u^{(i)}, Z_k^{(i)}\}, i = 1, \ldots, N$ can be expressed as

$$\begin{aligned} \frac{\mathrm{d}u^{(i)}}{\mathrm{d}t} &= V\left(u^{(i)}\right) + \frac{1}{K}\sum_{k,l} Z_k^{(i)} Z_l^{(i)} B_{kl} + F, \\ \frac{\mathrm{d}Z_k^{(i)}}{\mathrm{d}t} &= \gamma_k\left(u^{(i)}\right) Z_k^{(i)} - d_k Z_k^{(i)} + \sigma_k \dot{W}_k^{(i)}, \quad 1 \le k \le K, \end{aligned} \tag{12}$$

with the multiscale coupling coefficient $B_{kl} = B(\mathbf{e}_k, \mathbf{e}_l)$ and self-interaction inside the mean state summarized in $V(u)$. For simplicity, we adopt the diagonal coupling coefficients γ_k in the fluctuation equation for Z_k, which can be achieved by a proper change of basis from (3). An ensemble of samples $i = 1, \ldots, N$ is drawn from the initial distribution in the direct MC simulation. For each sample i, the mean state $u^{(i)}$ is coupled with all the small-scale fluctuation modes $\{Z_k^{(i)}\} \in \mathbb{R}^K$ of the entire spectrum $k \le K$ (with $K \gg 1$ in a high-dimensional phase space). The samples are updated independently for each i during the time evolution of solutions. The randomness in the model (12) comes from the small-scale white noise and the uncertainty from the initial samples $u^{(i)}(0) \sim \mu_0$.

In contrast, the RBM model (10) updates each mean state sample $\tilde{u}^{(i)}$ using only a portion of small-scale modes inside the corresponding batch $\mathcal{I}_i^m \subseteq \{k : 1 \le k \le K\}$ at the time updating step t_m. For distinction in notations, we use states $\{\tilde{u}, \tilde{Z}_k\}$ with "tildes" to represent the RBM solutions compared with the full MC states $\{u, Z_k\}$. The dynamics of the ith batch $\{\tilde{u}^{(i)}, \tilde{Z}_k\}_{k\in\mathcal{I}_i^m}$ during the time interval $t \in (t_{m-1}, t_m]$ can be formulated as

$$\begin{aligned} \frac{\mathrm{d}\tilde{u}^{(i)}}{\mathrm{d}t} &= V\left(\tilde{u}^{(i)}\right) + \sum_{k,l\in\mathcal{I}_i^m} c_{kl}\tilde{Z}_k\tilde{Z}_l B_{kl} + F, \\ \frac{\mathrm{d}\tilde{Z}_k}{\mathrm{d}t} &= \gamma_k\left(\tilde{u}^{(i)}\right)\tilde{Z}_k - d_k\tilde{Z}_k + \sigma_k\dot{\tilde{W}}_k, \qquad k \in \mathcal{I}_i^m, \end{aligned} \tag{13}$$

with samples only drawn for the large-scale mean state $\tilde{u}^{(i)}$, $i = 1, \ldots, N$ (for simplicity, we take $N_1 = N$ in Algorithm 1). Only one realization of the fluctuation modes $\tilde{Z}_k, k \le K$ is needed, and the K modes are shared among the N mean state samples $\tilde{u}^{(i)}$. At the start of each updating time $t = t_{m-1}$, the modes $\left\{\tilde{Z}_k\right\}$ are randomly regrouped into new batches $\cup_i \mathcal{I}_i^m = \left\{k : k \le K\right\}$ of size $p\left(= \frac{K}{N} = O(1)\right)$. The random batches are chosen independent of the noises in the stochastic modes. Only the modes belonging to the ith batch, $\tilde{Z}_k, k \in \mathcal{I}_i^m$, are updated using the ith mean state $\tilde{u}^{(i)}$, and only these modes in batch i are used to update the mean state sample $\tilde{u}^{(i)}$. Due to the batch approximation, new coupling coefficients c_{kl} in (11) are introduced. Notice that in this RBM model (13), different samples $\tilde{u}^{(i)}$ are no longer independent with each other during the entire time evolution since they are linked by the batch resampling at the start of each time update. Therefore, the RBM model has the additional randomness from the random partition of modes $\{\mathcal{I}^m\}_{m=1}^{[T/\tau]}$ with $\tau = t_m - t_{m-1}$ the discrete time step size.

A. Main theorem for the approximation error in the RBM ensemble model

In the ensemble forecast, we are interested in recovering the statistics through the empirical PDF (7) rather than each individual trajectory solution. One effective way to calibrate the statistical error is to compare the difference between the ensemble averaged $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\varphi\left(u^{(i)}(t_n)\right)$ and $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\varphi\left(\tilde{u}^{(i)}(t_n)\right)$ from model (12) and (13) for any test function $\varphi \in C_b^2$. We propose the following

structural assumptions for the coupling parameters of the models (12) and (13):

Assumption 1: *In the mean state equation for u, suppose that the bilinear coupling coefficients B_{kl} in the mean state dynamics are uniformly bounded,*

$$\max_{1\le k,l\le K} B_{kl} \le C. \tag{14a}$$

The self-coupling term $V(u)$ and its derivatives up to the second order are uniformly bounded,

$$\|V\|_{C^2} = \sum_{|\alpha|\le 2} \|\nabla^\alpha V\|_\infty \le C. \tag{14b}$$

In the fluctuation equations for $\{Z_k\}$, we assume that there is no internal instability induced by u in all small-scale modes and the total noise amplitude is bounded. That is, there is a positive constant r independent of u so that

$$\min_{1\le k\le K}\{d_k - \gamma_k(u)\} \ge r > 0 \quad \text{and} \quad \sum_{k=1}^{K} \sigma_k^2 \le C. \tag{14c}$$

Assumption (14a) is natural from the definition of the quadratic bilinear form $B_{kl} = B(\mathbf{e}_k, \mathbf{e}_l)$, and assumption (14b) for the self-coupling term $V(u)$ makes sure that this term does not have a rapid growth (this can be guaranteed by imposing constraints on the maximum value of u). Assumption (14c) implies that the mean state induces no internal instability to the fluctuation equations through the coupling with the small-scale modes. This requires that we only partition the stable small-scale modes into random batches for the time updating in Algorithm 1. On the other hand, it needs to be emphasized that the important unstable dynamics in the leading fluctuation modes can still be modeled exactly in Algorithm 1 by adding them to the resolved mean state as illustrated later in the numerical examples.

Under these assumptions, we have the following main theorem characterizing the statistical error in the RBM ensemble prediction:

Theorem 2: *With Assumption 1 satisfied, the empirical statistical estimation of the random batch model (8) with discrete time step τ converges to that of the full model solution (7) up to final time $t = T$ as*

$$\sup_{n\tau\le T}\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\varphi\left(\tilde{u}_n^{(i)}\right) - \mathbb{E}\varphi(u_n)\right| \le C_\varphi(T)\,\tau, \tag{15}$$

with any $\varphi \in C_b^2$, $u_n = u(n\tau)$ by solving (12), and $\tilde{u}_n = \tilde{u}(n\tau)$ by solving (13). C_φ is independent of the sample size N and the fluctuation modes dimension K.

Above, (15) gives the error estimation of the averaged RBM prediction of the statistics $\mathbb{E}\varphi\left(\tilde{u}_n^{(i)}\right)$ compared with the full ensemble method $\mathbb{E}\varphi\left(u_n^{(i)}\right)$. Since the samples in the full ensemble model are independent and identical under expectation, we have $\mathbb{E}\varphi(u_n) = \mathbb{E}\varphi\left(u_n^{(i)}\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\varphi\left(u_n^{(i)}\right)$. Notice that this formula does not quantify the sufficient size of the ensemble N to reach accurate statistical prediction of $\mathbb{E}\varphi$. In particular, the full ensemble model (12) requires exhausting sampling of the full phase space of high dimension $d + K \gg 1$. It ends up with an exponential growth in the sample size N as the dimension K increases, thus suffers the curse-of-dimensionality. In contrast, the RBM model (13) only needs to sample the low-dimensional resolved subspace $d \ll d + K$. Thus, the required sample size N for $\tilde{u}^{(i)}$ can be controlled regardless of the high dimension K of the full fluctuation state.

B. Proof of the main theorem

In order to estimate the statistical error under the test function φ, we first introduce the function w_Z based on the solution of the full model (12) given the realization of the small-scale stochastic process $Z(\cdot)$,

$$w_Z(x,t) := w(x,t \mid Z) = \frac{1}{N}\sum_{i=1}^{N}\varphi^x\left(u^{(i)}(t) \mid Z^{(i)}(s), s < t\right). \tag{16}$$

Above, we use $x = \{x_i\}_{i=1}^N \in \mathbb{R}^{d\times N}$ to denote the initial values of all the samples [with $u^{(i)}(0) = x_i$ for each initial state $x_i = \left(x_i^1, \ldots, x_i^d\right) \in \mathbb{R}^d$]. The function $\varphi^x \in C_b^2\left(\mathbb{R}^d\right)$ evaluates the sample solution $u^{(i)}(t)$ starting at initial state $u^{(i)}(0) = x_i$. The function w_Z is defined according to the solutions of the first equation of (12) depending on the entire time sequence of the stochastic process $Z(s), s < t$, with $Z = \{Z_k\}_{k=1}^K \in \mathbb{R}^K$. Next, we define the deterministic function $w(x,t)$ after taking expectation $\mathbb{E}^Z$ [we use superscript for expectation on the stochastic process $Z(\cdot)$]

$$w(x,t) := \mathbb{E}^Z w_Z(x,t) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_x\varphi\left(u^{(i)}(t)\right), \tag{17}$$

with $\mathbb{E}_x = \mathbb{E}\left(\cdot \mid u^{(i)}(0) = x_i\right)$ being the expectation given sample initial values x. By definition, we have the initial condition $w_z(x,0) = \frac{1}{N}\sum_i \varphi(x_i) = w(x,0)$. The functions w and w_z satisfy the basic properties from the semigroup.[41,42] That is, let $\mathcal{L}_z$ be the generator for Eq. (12), $\partial_t w_z(x,t) = \mathcal{L}_z w_z(x,t)$ as in (19a). We have from the L^∞ contraction of the semigroup, $\left\|e^{t\mathcal{L}_z}\varphi\right\|_\infty \le \|\varphi\|_\infty$, so that

$$\begin{aligned}\|w(\cdot,t)\|_\infty = \left\|\mathbb{E}^Z e^{t\mathcal{L}_Z} w(\cdot,0)\right\|_\infty &\le \mathbb{E}^Z\left\|e^{t\mathcal{L}_Z} w(\cdot,0)\right\|_\infty \\ &\le \mathbb{E}^Z\|w(\cdot,0)\|_\infty = \|w(\cdot,0)\|_\infty \\ &\le C.\end{aligned}$$

Accordingly, we can define the associated function $\tilde{w}$ using the solution $\tilde{u}^{(i)}(t)$ of the RBM model (13) starting from $\tilde{u}_0^{(i)} = x_i$ as

$$\tilde{w}(x,t) := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_x\left[\varphi\left(\tilde{u}^{(i)}(t)\right)\right].$$

Here, the expectation $\mathbb{E}_x$ applies on the additional randomness subject to random partition $\mathcal{I}^m$ of the batches in $\{Z_k\}$ at each time update as well as the stochastic white noise. In order to characterize the statistical evolution of the samples, we construct the discrete semigroup $\tilde{\mathcal{S}}$ associated with the RBM model in three steps according to Algorithm 1 in the time interval $(t_{m-1}, t_m]$. First, starting with the initial function $\tilde{w}(x, t_{m-1})$, we fix the solution of the fluctuation modes during the next updating interval, $Z(s), t_{m-1} < s \le t_m$. Next, before the time update at $t = t_{m-1}$, we partition the modes $Z = \{Z_k\}$ into small batches $\mathcal{I}^m = \{\mathcal{I}_i^m\}$. Finally, we update the solution to the next time step $t = t_m$ by integrating the backward equation (19b) of

the mean state with time step $\tau = t_m - t_{m-1}$. The one-step integration of the RBM generator $\tilde{\mathcal{L}}_z^{\mathcal{I}m}$ conditional on the partition $\mathcal{I}^m$ and the fluctuation solution Z during the mth update time interval gives the conditional updating operator

$$\tilde{\mathcal{S}}_Z^{\mathcal{I}m}\tilde{w}(x,t_{m-1}) := e^{\tau\tilde{\mathcal{L}}_Z^{\mathcal{I}m}}\tilde{w}(x,t_{m-1}).$$

The one-step semigroup operator follows by taking expectation first on the random partition $\mathbb{E}^{\mathcal{I}^m}$ and then on the stochastic process $\mathbb{E}^Z$ in the updating interval $Z(s), t_{m-1} < s \le t_m$,

$$\tilde{\mathcal{S}}\tilde{w}(x,t_{m-1}) := \mathbb{E}^Z\mathbb{E}^{\mathcal{I}^m}\tilde{\mathcal{S}}_Z^{\mathcal{I}m}\tilde{w}(x,t_{m-1}) = \mathbb{E}^Z\tilde{w}_Z(x,t_m) = \tilde{w}(x,t_m), \tag{18}$$

where we define $\tilde{w}_Z = \tilde{\mathcal{S}}_Z\tilde{w} = \mathbb{E}^{\mathcal{I}^m}\tilde{\mathcal{S}}_Z^{\mathcal{I}m}\tilde{w}$, and $\tilde{\mathcal{S}} = \mathbb{E}^Z\tilde{\mathcal{S}}_Z$. Therefore, it can be shown that the semigroup is formed by

$$\begin{aligned}\tilde{\mathcal{S}}^{(m)}\varphi(x) &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_x\left[\varphi\left(\tilde{u}_m^{(i)}\right)\right]\\ &= \tilde{w}(x,t_m) = \tilde{\mathcal{S}}\tilde{w}(x,t_{m-1}) = \tilde{\mathcal{S}}\circ\tilde{\mathcal{S}}^{(m-1)}\varphi,\end{aligned}$$

with $\tilde{u}_m = \tilde{u}(t_m)$.

Applying the backward Kolmogorov equation to the full MC model and the reduced RBM model, we find the governing equations depending on the realization $z(\cdot)$,

$$\partial_t w_z = \mathcal{L}_z w_z = \sum_{i=1}^N\left[V(x_i) + \frac{1}{K}\sum_{k,l} z_k z_l B_{kl} + F\right]\cdot\partial_{x_i} w_z, \tag{19a}$$

$$\partial_t\tilde{w}_z = \tilde{\mathcal{L}}_z^{\mathcal{I}^m}\tilde{w}_z = \sum_{i=1}^N\left[V(x_i) + \sum_{k,l} c_{kl} I_i^m(k) I_i^m(l) z_k z_l B_{kl} + F\right]\cdot\partial_{x_i}\tilde{w}_z. \tag{19b}$$

Above, the RBM generator $\tilde{\mathcal{L}}_z^{\mathcal{I}m}$ is also conditional on the random batch partition of modes $\mathcal{I}m = \{\mathcal{I}_i^m\}$ in each time interval $t \in (t_{m-1}, t_m]$. To formulate the equations under comparable terms, we introduce the index function during the time interval $t_{m-1} < t \le t_m$ as

$$I_i^m(k) = \begin{cases} 1 & \text{if } k \in \mathcal{I}_i^m,\\ 0 & \text{otherwise.}\end{cases} \tag{20}$$

The index function I_i^m is fixed during the time interval and will be modified at the beginning of each time updating step t_m subject to the random partition (we neglect the subscript "m" for time step t_m in the rest part of this section for brevity of notations).

The proof of Theorem 2 follows the method of weak convergence in Ref. 22. The main difference here is that we have to deal with the nonlinear coupling from the fluctuation modes Z. We need the following lemmas to compare the difference between the statistical functions w and $\tilde{w}$. The proofs of these lemmas can be found in Appendix A.

First, Lemma 3 indicates the values of the new coupling coefficients c_{kl} in (11) of the RBM model according to expectations on the random batch partition.

Lemma 3: *Let $I_i(k)$ be the index function (20), indicating that the mode k belongs to the batch $\mathcal{I}_i$ with $k = 1,\dots,Np = K$ (K is the total number of modes partitioned into N random batches). Then, we have the expectations about the partition $\mathcal{I}$ for any k*

$$\mathbb{E}I_i^2(k) = \mathbb{E}I_i(k) = \frac{p}{K}$$

and for any $k \ne l$

$$\mathbb{E}I_i(k) I_i(l) = \frac{p}{K}\frac{p-1}{K-1}.$$

Next, the following lemma describes the estimates on the moments of entire fluctuation modes. The supremum bound is guaranteed by the crucial stability assumption (14c) with uniform negative damping in the dynamical equations of all the modes Z_k.

Lemma 4: *Under the stable dynamics (14c), the fluctuation modes Z_k and $\tilde{Z}_k$ in (12) and (13) satisfy*

$$\sup_{t\le T}\mathbb{E}\|Z_t\|^{2q} < C_q, \quad \sup_{t\le T}\mathbb{E}\left\|\tilde{Z}_t\right\|^{2q} < C_q$$

for any integer $q \ge 1$ and $\|Z_t\|^2 = \sum_{k=1}^K |Z_k(t)|^2$. The constant C_q is independent of T.

The last lemma shows the regularity in the function $w(x,t)$ that is uniformly bounded up to second-order differentiations.

Lemma 5: *For $t \le T$, we have the uniform bound for the characteristic function (17) at least up to second-order derivatives*

$$\left\|\nabla_x^2 w(\cdot,t)\right\|_\infty = \sum_{i,j=1}^N\left\|\partial_{x_i x_j}^2 w(\cdot,t)\right\|_\infty < C_\varphi(T),$$

with the constant $C_\varphi(T) > 0$ independent of N.

With the above lemmas, we can give the proof of our main theorem. First, we estimate the one-step error between $w_z(x,t_{m+1})$ and $\tilde{\mathcal{S}}_z w(x,t)$, and then the final estimate follows by taking the expectation about Z.

Proof of Theorem 2. In the RBM model, the batches of modes $\{\tilde{Z}_k\}$ are regrouped before each time step update at $t_m = m\tau$. Thus, we first focus on the one-step update during the time internal $t_m < t \le t_{m+1}$ for all $m \le n$ in the backward equations for the full model (19a) and the RBM model (19b).

For the one-step time update of the RBM model and given $z(s), t_m < s \le t_{m+1}$, the discrete semigroup operator $\tilde{\mathcal{S}}_z^{\mathcal{I}}$ for a fixed batch partition $\mathcal{I}$ applied on the function $w(x,t_m)$ according to (19b) has the expanded expression

$$\begin{aligned}\tilde{\mathcal{S}}_z^{\mathcal{I}} w(x,t_m) = e^{\tau\tilde{\mathcal{L}}_z^{\mathcal{I}}} w(x,t_m) &= w(x,t_m) + \tau\tilde{\mathcal{L}}_z^{\mathcal{I}} w(x,t_m)\\ &+ \int_0^\tau(\tau-s)\left(\tilde{\mathcal{L}}_z^{\mathcal{I}}\right)^2 e^{s\tilde{\mathcal{L}}_z^{\mathcal{I}}} w(x,t_m)\,ds.\end{aligned}$$

Correspondingly, for the full model (19a), we have the continuous generator applying on the same function at $t = t_{m+1}$

$$\begin{aligned}w_z(x,t_{m+1}) = e^{\tau\mathcal{L}_z} w(x,t_m) &= w(x,t_m) + \tau\mathcal{L}_z w(x,t_m)\\ &+ \int_0^\tau(\tau-s)\,\mathcal{L}_z^2 e^{s\mathcal{L}_z} w(x,t_m)\,ds.\end{aligned}$$

Notice that above, we consider the one-step update from $w(x,t_m)$ conditional on the realization of fluctuation modes z during the updating interval $(t_m, t_{m+1}]$.

Next, we take the expectation on the random batch partition $\mathcal{I}$ at step t_m on the RBM generator $\tilde{\mathcal{L}}_z^{\mathcal{I}}$. Then, the RBM generator becomes consistent with the full model generator under the proper choice of the coefficients c_{kl}

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

This is guaranteed by Lemma 3, and the coefficients c_{kl} in (11) appear naturally from the lemma. With this, we have from the definition in (18)

$$\begin{aligned}
&\tilde{\mathcal{S}}w(x,t_m) - w(x,t_{m+1}) \\
&\quad = \mathbb{E}^Z\mathbb{E}^{\mathcal{I}}\tilde{\mathcal{S}}_Z^{\mathcal{I}}w(x,t_m) - \mathbb{E}^Z w_Z(x,t_{m+1}) \\
&\quad = \int_0^\tau (\tau - s)\left[\mathbb{E}^{Z,\mathcal{I}}\left(\tilde{\mathcal{L}}_Z^{\mathcal{I}}\right)^2 e^{s\tilde{\mathcal{L}}_Z^{\mathcal{I}}} - \mathbb{E}^Z\mathcal{L}_Z^2 e^{s\mathcal{L}_Z}\right] w(x,t_m)\, ds.
\end{aligned}$$

Then, we show that the residual terms in the above integrants are uniformly bounded with the constant C independent of N, K,

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

In fact, using the fact that $\mathcal{L}_z^2$ and $e^{s\mathcal{L}_z}$ are commutative and $e^{s\mathcal{L}_z}$ is a contraction under L^∞, we have the estimation

$$\begin{aligned}
&\left\|\mathbb{E}^Z\mathcal{L}_Z^2 e^{s\mathcal{L}_Z}w(\cdot,t)\right\|_\infty \\
&\quad \le \mathbb{E}^Z\left\|e^{s\mathcal{L}_Z}\mathcal{L}_Z^2 w(\cdot,t)\right\|_\infty \le \mathbb{E}^Z\left\|\mathcal{L}_Z^2 w(\cdot,t)\right\|_\infty \\
&\quad \le \left(C_0 + \frac{C_1}{K}\sum_{k,l}\mathbb{E}\left|Z_k Z_l\right| + \frac{C_2}{K^2}\sum_{k,l,m,n}\mathbb{E}\left|Z_k Z_l Z_m Z_n\right|\right) \\
&\qquad \times \left\|\nabla_x^2 w(\cdot,t)\right\|_\infty.
\end{aligned}$$

In the last inequality above, we expand that the operator $\mathcal{L}_Z^2$ and C_0, C_1, C_2 are three positive constants independent of N, K, and $\left\|\nabla_x^2 w\right\|_\infty = \sum_{i,j}\left\|\partial_{x_i x_j}^2 w\right\|_\infty$. The uniform bounds for the total moments $\frac{1}{K}\sum_{k,l}\mathbb{E}\left|Z_k Z_l\right| \le \mathbb{E}\left\|Z\right\|^2$ and $\frac{1}{K^2}\sum_{k,l,m,n}\mathbb{E}\left|Z_k Z_l Z_m Z_n\right| \le \mathbb{E}\left\|Z\right\|^4$ are guaranteed by Lemma 4. We have the uniform bound for $\left\|\nabla_x^2 w\right\|_\infty$ by Lemma 5. The same result can be achieved for $\tilde{\mathcal{L}}_Z^{\mathcal{I}}$ with a similar argument.

Therefore, during the time updating from t_{m-1} to t_m, we have the one-step error between the RBM solution $\tilde{\mathcal{S}}w(x,t_m)$ and the full model $w(x,t_{m+1})$ from the time integration of the residual term inside the time interval

$$\begin{aligned}
\left\|\tilde{\mathcal{S}}w(\cdot,t_m) - w(\cdot,t_{m+1})\right\|_\infty &= \left\|\mathbb{E}^Z\left[\tilde{\mathcal{S}}_Z w(\cdot,t_m) - w_Z(\cdot,t_{m+1})\right]\right\|_\infty \\
&\le C\tau^2.
\end{aligned}$$

Finally, applying $\tilde{\mathcal{S}}^{(n)}$ on the initial function φ and using $\tilde{\mathcal{S}}\tilde{w}(x,t_m) = \tilde{w}(x,t_{m+1})$ by recurrently applying the semigroup and using the L^∞ contraction property for $\tilde{\mathcal{S}}$, the total error at $t = t_n = n\tau$ can be computed as

$$\begin{aligned}
\left\|\tilde{\mathcal{S}}^{(n)}\varphi - w(\cdot,t_n)\right\|_\infty &\le \left\|\tilde{\mathcal{S}}\left[\tilde{\mathcal{S}}^{(n-1)}\varphi - w(\cdot,t_{n-1})\right]\right\|_\infty \\
&\quad + \left\|\tilde{\mathcal{S}}w(\cdot,t_{n-1}) - w(\cdot,t_n)\right\|_\infty \\
&\le \left\|\tilde{\mathcal{S}}^{(n-1)}\varphi - w(\cdot,t_{n-1})\right\|_\infty \\
&\quad + \left\|\tilde{\mathcal{S}}w(\cdot,t_{n-1}) - w(\cdot,t_n)\right\|_\infty \\
&\le \sum_{m=0}^{n-1}\left\|\tilde{\mathcal{S}}w(\cdot,t_m) - w(\cdot,t_{m+1})\right\|_\infty \le C(t_n)\,\tau.
\end{aligned}$$

This gives the final error estimate by maximizing among the initial samples x_i. □

V. NUMERICAL TESTS OF THE RANDOM BATCH ALGORITHM ON TURBULENT MODELS

Now, we evaluate the numerical performance of the general RBM model in Algorithm 1 using turbulent models containing a wide spectrum of fluctuation modes. In particular, we test the algorithm on two prototype benchmark models, that is, the conceptual turbulent model and the topographic barotropic model, which are shown to generate various representative phenomena in multiscale turbulence. The RBM algorithm displays uniformly high skill in capturing the key statistical behaviors in the dominant large-scale states displaying highly non-Gaussian PDFs and intermittent extreme events with much lower computational cost.

A. Random batch algorithm for the conceptual turbulent model

One particular concrete example accepting the general model framework (3) is the *conceptual dynamical model for turbulence* developed in Ref. 43,

$$\begin{aligned}
\frac{d\bar{u}}{dt} &= -\bar{d}\bar{u} + \frac{\gamma}{K}\sum_{k=1}^K v_k^2 - \bar{\alpha}\bar{u}^3 + \bar{F}, \\
\frac{dv_k}{dt} &= -\gamma\bar{u}v_k - d_k v_k + \sigma_k\dot{W}_k, \quad 1 \le k \le K.
\end{aligned} \tag{21}$$

Above, the state variables $(\bar{u}, v_k) \in \mathbb{R}^{1+K}$ constitute a $(K+1)$-dimensional system. The scalar large-scale mean state $\bar{u}$ is coupled with each small-scale mode v_k through the nonlinear interaction coefficient $\gamma > 0$, while the large number of small-scale fluctuation modes v_k impact the large-scale mean state $\bar{u}$ together through a quadratic coupling term. It is easy to check that the nonlinear coupling term conserves the total energy $E = \frac{1}{2}\left(\bar{u}^2 + \frac{1}{K}\sum_k v_k^2\right)$ so that the structural property in (1) is satisfied. The model (21) gives a typical characterization for the anisotropic turbulence in which fluctuating energy flows intermittently from a wide range of small scales to affect the largest-scale mean flow. Besides, unstable dynamics are also induced in the leading fluctuation modes v_k through coupling with the mean state when $-d_k - \gamma\bar{u} > 0$, where strong intermittency and extreme events are

TABLE I. Parameter values for the conceptual turbulent model.

K	$\bar{d}$	$\bar{\alpha}$	$\bar{F}$	γ	d_k	σ_k	E_0	K_1	p	N (full MC)	N_1 (RBM)
100	-0.1	0.05	-0.055	1.5	$1+0.02k^2$	$\sqrt{2E_k d_k}$	0.004	4	5	10 000	100

triggered with non-Gaussian statistics through the chaotic fluctuations. This characterizes another key observation in turbulence, which is captured in this conceptual model.

The model parameters $\left(\bar{d}, \bar{\alpha}, \bar{F}, \gamma, d_k, \sigma_k\right)$ for a strongly unstable regime are listed in Table I. Strong instability in the large-scale dynamics for $\bar{u}$ is imposed through the linear anti-damping term $\bar{d} < 0$, which needs to be balanced by the nonlinear feedbacks from both small and large scales. A wide spectrum of modes $K = 100$ is included for multiscale fluctuations. d_k and σ_k are used to describe the turbulent dissipation and white noise forcing in these fluctuation modes, respectively. For the convenience of the numerical test, we assign the Kolmogorov spectrum $E_k = \frac{\sigma_k^2}{2d_k} = E_0 k^{-5/3}$ for the leading modes $k \le K_1 = 4$, while all the other smaller-scale modes have equipartition of energy $E_k = E_0 K_1^{-5/3}$. Under this model setup, it can be shown that an ergodic invariant measure[40] will be reached, while instability on both small and large scales will create intermittent behavior and non-Gaussian PDFs during the model evolution.

For direct ensemble simulation in a large phase space with dimension $1 + K = 101$, we take a large ensemble size $N = 1 \times 10^4$ for accurate MC simulation to get the reference true distribution. Next, to apply the RBM model to recover the probabilistic solutions in leading states, only the mean state $\bar{u}$ together with the first four leading fluctuation modes $v_k, k \le K_1 = 4$ is sampled. This is considering their unstable dynamics to satisfy the assumption in (14c) so that only stable small-scale fluctuation modes are partitioned in the random batches of a small size $p = 5$. Thus, a much smaller sample size $N_1 = 100$ is sufficient to model the $1 + K_1 = 5$ dimensional subspace. The algorithm can be developed directly according to the steps in Algorithm 1 with a simple generalization of also resolving the leading fluctuation modes. The joint PDFs of the resolved states can be approximated by sample histograms as the empirical representations in (7) and (8). We summarize the detailed numerical RBM equations for the conceptual model (21) in Appendix B 1.

To illustrate the basic statistical features, Fig. 1 plots the equilibrium energy spectrum $\mathbb{E}^{\text{eq}} |v_k|^2$ and the decorrelation time $\int_0^\infty \mathbb{E}^{\text{eq}} |v_k(t)\, v_k(0)|\, dt$ of the fluctuation modes (with $\mathbb{E}^{\text{eq}}$ denoting the average about the equilibrium measure). This displays a typical example of the common features in turbulent flows: the first few leading modes accumulate most of the energy and a relatively long decorrelation time, while there exists a long extended spectrum of small-scale fast-mixing fluctuating modes containing small energy in each mode but having a non-negligible combined contribution to the large-scale mean flow. This is shown more clearly in the right column of Fig. 1 for the time evolution of energy $\sum_k |v_k|^2$ in the leading modes $k \le 4$ and all the rest small-scale modes $k > 4$. The leading modes show intermittent bursts of large energy due to the destabilizing coupling with the mean flow, while the large number of small-scale fluctuating modes account for a major amount of energy during the quiescent regime for most of the time during the evolution. This confirms that the contributions from the many small-scale modes play an important role of driving the large scales to the final equilibrium and cannot be simply ignored in simulations.

First, we check the recovery of the solution trajectories using the RBM model. Figure 2 plots the typical time series of the mean

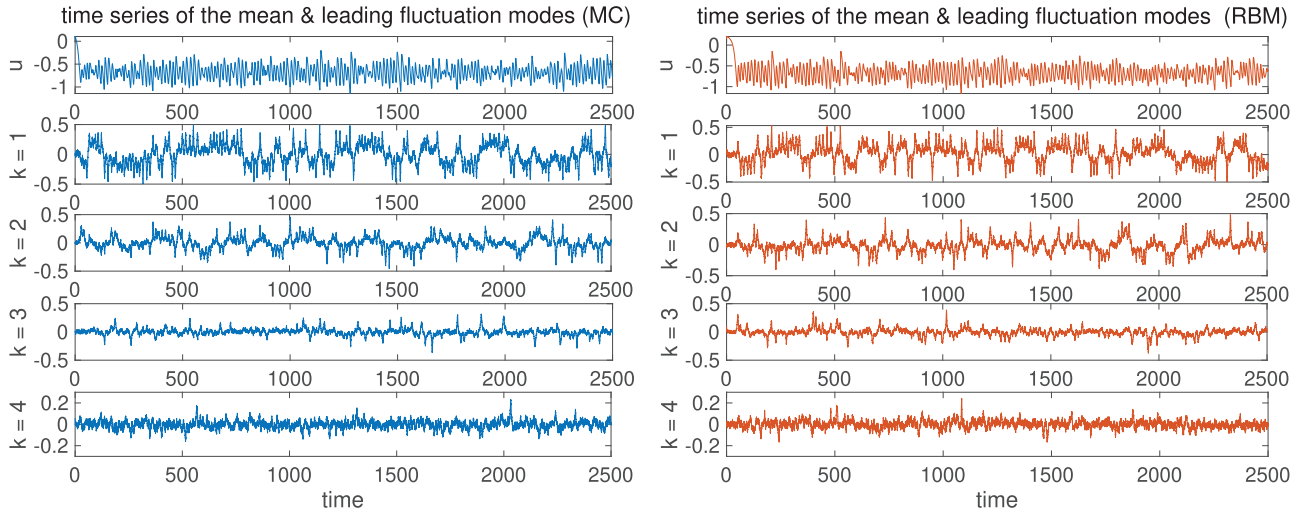
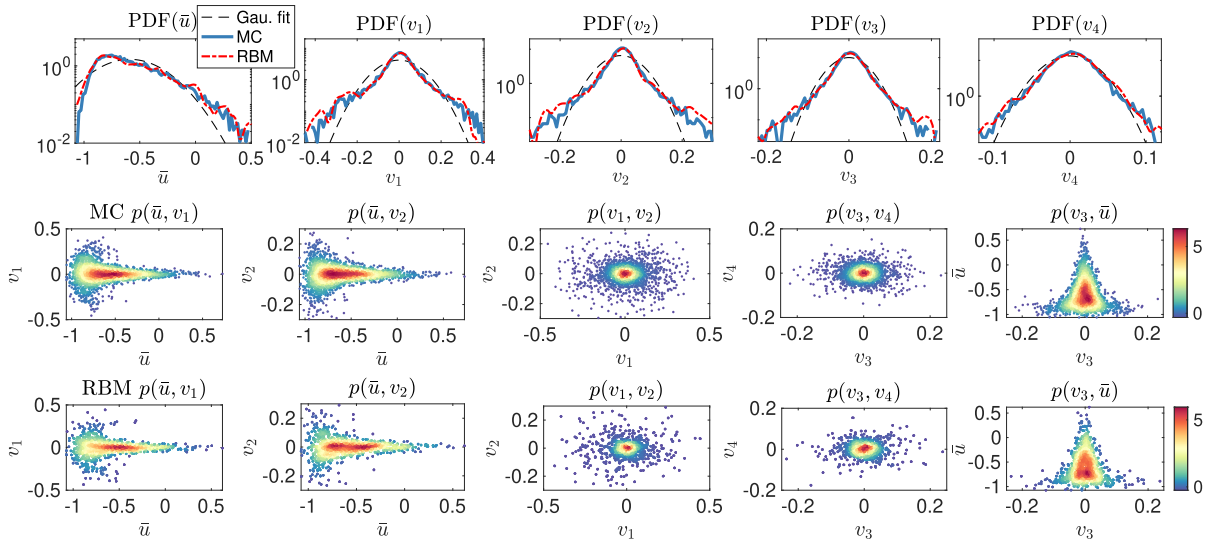
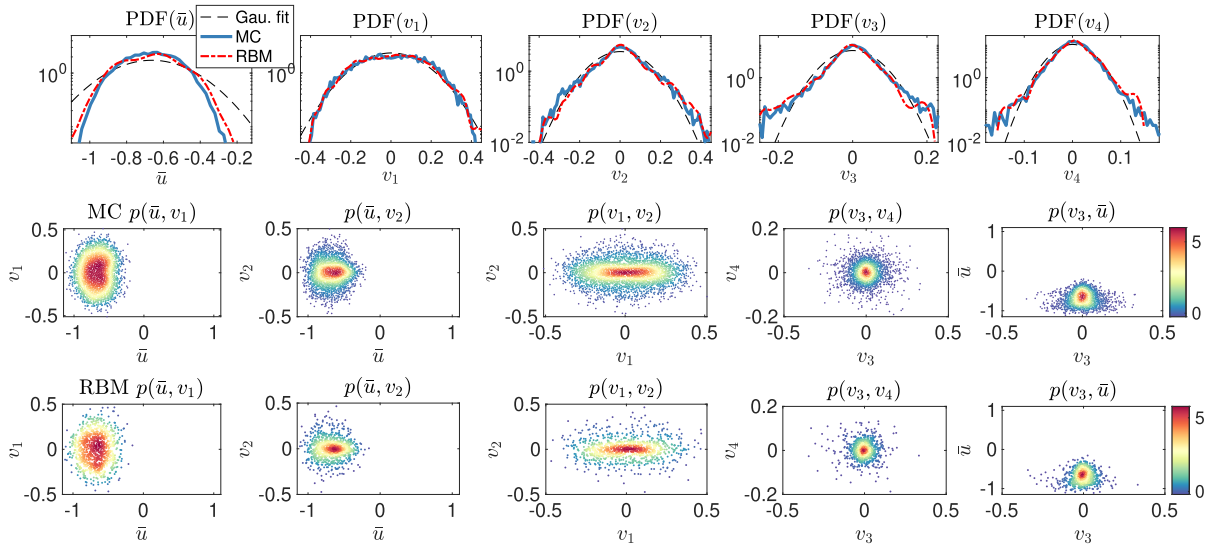


FIG. 2. Time trajectories of the mean state $\bar{u}$ and the first four leading mode v_k from one typical realization of the direct simulation (left) and the random batch method (right).



(a)



(b)

FIG. 3. Predicted PDFs from the direct MC simulation with $N = 10\,000$ samples and from the RBM model using $N_1 = 100$ samples for the conceptual model. The 1D and 2D marginal PDFs of the leading modes $(\bar{u}, v_1, v_2, v_3, v_4)$ are compared. In the 1D PDFs, the Gaussian fit with the same mean and variance is plotted in a dashed line. The 2D joint PDFs are shown by scatterplots with colors indicating sample density. The PDFs in the starting transient state (upper) and the final equilibrium state (lower) are compared.

state $\bar{u}$ and the first four leading modes v_k from direct simulation and the random batch approximation. Highly non-Gaussian features with intermittent bursts of extreme events can already be observed in the time series of the leading modes. The intermittent flow structure is a key feature to model that is closely related to the largest-scale mean state. Qualitatively, we observe that the structures in the most energetic modes are precisely captured from the RBM approach especially with the random intermittent bursts and extreme events. Notice that in each time step update of the model, only $p = 5$ modes are included in updating the mean equation

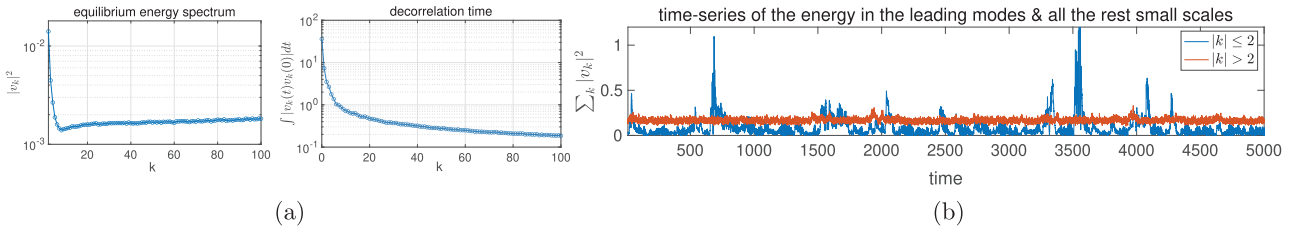


FIG. 4. Statistics of the topographic barotropic model (22). Left: equilibrium energy spectrum and decorrelation time in the fluctuation modes. Right: time series of the energy in the first two leading modes and energy in all the rest fluctuation modes.

randomly picked from the total $K = 100$ small-scale modes. It is shown that the batch size can be further reduced to even $p = 2$ in Fig. 6. Thus, the exhausting computational cost to resolve the long spectrum of all fluctuation modes in each ensemble member at each time step can be effectively avoided.

Next, we test the performance of the RBM model in efficient ensemble prediction of the probability distributions. In particular, the prediction skill of both the final equilibrium probability distribution and the transient probability distributions before equilibrium is considered. In the tests, the initial samples at the starting time are drawn from independent Gaussian distributions for each mode with mean zero and a small variance. Thus, the PDFs will go through a statistical transition from the Gaussian initial distribution to the non-Gaussian final equilibrium. Figure 3 compares the prediction of marginal and joint PDFs in both the starting transient state and the final statistical equilibrium. First, from the sample distributions, we observe the drastic deviation from Gaussian in the PDFs. Strongly non-Gaussian structures appear in both the starting transient state and the final statistical equilibrium. The fat-tailed PDFs refer to the intermittent extreme flow structures and are of particular interest in the study of turbulent flows. To capture such extreme features, a very large sample size is required in the direct MC simulation to sufficiently characterize the outliers that constitute the PDF tails. In contrast, the RBM model focuses on the mean and the first four dominant modes and recovers the joint PDFs of these most important modes. It is able to capture both the transient and equilibrium non-Gaussian PDFs, especially the non-Gaussian PDF tails to a large extent accurately, while saving the computational cost to a large degree requiring a much smaller number of samples.

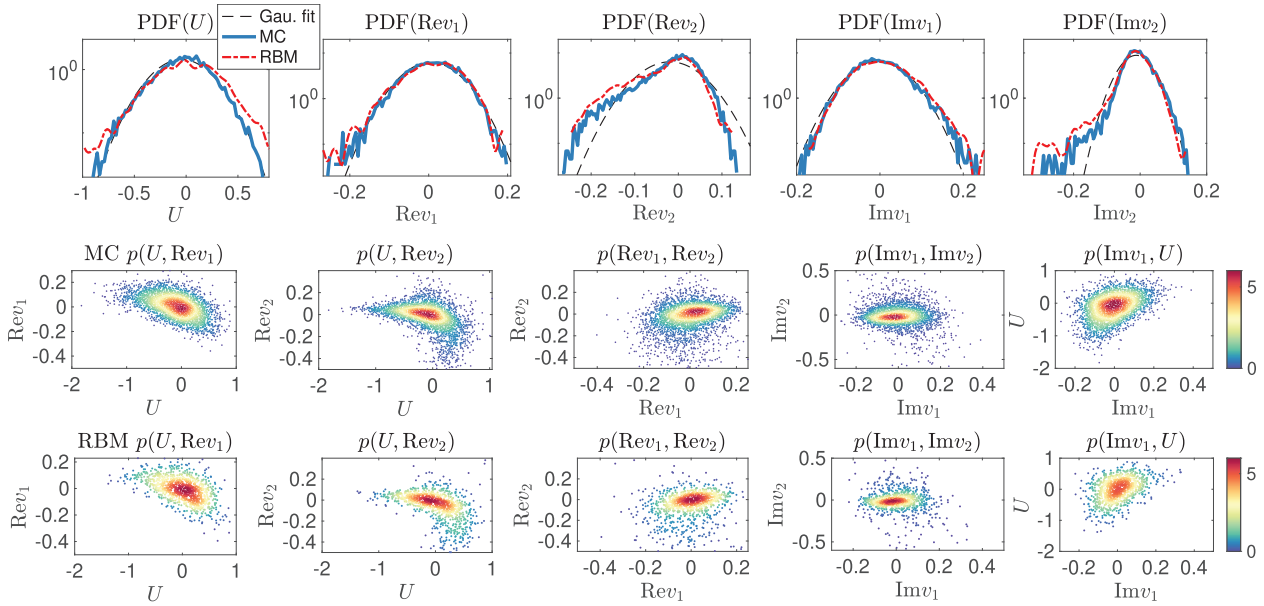


FIG. 5. Predicted PDFs from the direct MC simulation with $N = 10\,000$ samples and from the RBM model using $N_1 = 100$ samples for the topographic model. The 1D and 2D marginal PDFs of the leading modes (U, Rev_1, Rev_2, Imv_1, Imv_2) are compared. In the 1D PDFs, the Gaussian fit with the same mean and variance is plotted in a dashed line. The 2D joint PDFs are shown by scatterplots with colors indicating sample density.

B. Random batch algorithm for the topographic barotropic model

The topographic barotropic flow is a paradigm model in geophysical turbulence,[44] which generates many representative features found in real atmosphere and ocean. It identifies the multiscale interactions between a large-scale mean flow and fluctuations through interaction with topography. Under projection along one characteristic wavenumber direction, the *topographic barotropic model of layered topography* can be expressed in terms of a large-scale mean flow U and a wide spectrum of fluctuation modes v_k in complex values as

$$\begin{aligned} \frac{\mathrm{d}U}{\mathrm{d}t} &= \frac{1}{K}\sum_{k=-K}^{K} h_k^* v_k - d_0 U + \sigma_0 \dot{W}_0, \\ \frac{\mathrm{d}v_k}{\mathrm{d}t} &= i\left(k^{-1}\beta - kU\right) v_k - U h_k - d_k v_k + \sigma_k \dot{W}_k, \quad |k| \le K. \end{aligned} \tag{22}$$

Above, the complex flow modes satisfying $v_{-k} = v_k^*$ are derived from the Fourier expansion of the fluctuation flow field. The model (22) constitutes a system of dimension $1 + 2K$. The fixed complex modes h_k represent the topographic effect, and $\beta, d_0, d_k, \sigma_0, \sigma_k$ are other model parameters representing rotation, damping, and unresolved forcing in large and small scales. The topographic model eliminates the nonlinear interactions between the fluctuation modes and focuses on the coupling effect between the large and small scales. Equation (22) becomes consistent with the general multiscale model framework (3) by introducing the auxiliary dynamics $\frac{\mathrm{d}h_k}{\mathrm{d}t} \equiv 0$, which represents the constant topographic structure. Therefore, Algorithm 1 still applies to implementing the RBM model for this system. The topographic model also generates many key features of interest in turbulent flows, including skewed non-Gaussian PDFs and the related extreme events from a new generating mechanism. Different from the previous conceptual model (21), the small-scale feedback to the mean U is coupled through a combined interaction with the topographic mode h_k, and instability in the leading fluctuation modes is introduced from the coupling with topographic stress. A detailed derivation from the two-dimensional barotropic model and many desirable properties, such as conservation of energy, can be found in Refs. 28 and 45.

The model parameters in (22) are listed in Table II. The topographic modes h_k are defined by the Fourier modes of the spatial topography structure as a combination of two major large scales and multiple small-scale random perturbations

$$h = H(\sin x + \cos x) + \frac{H}{2}(\sin 2x + \cos 2x) + \sum_{3\le|k|\le K} e^{i(kx+\theta_k)},$$

with θ_k being the random phase shift drawn independently from a uniform distribution in $[0, 2\pi)$. A white noise forcing with small amplitude $\sigma_0 = \frac{1}{4\sqrt{2}}$ is also added in large scale to induce stronger chaotic dynamics in the mean state. This can also fit into the model framework with simple generalization. In the small scales, $\beta = 2$ represents the rotational effect of the flow. Again, we adopt the Kolmogorov spectrum $E_k = \frac{\sigma_k^2}{2d_k} = E_0 k^{-5/3}$ in the first two complex modes $|k| \le K_1 = 2$ and with equipartition of energy for all the small scales, $\frac{\sigma_k^2}{2d_k} = E_0 K_1^{-5/3}, |k| > K_1$. These choices of parameter values are following the non-dimensionalization of the real physics measurements of the characteristic scales.[28] We first demonstrate the model statistical features in Fig. 4. The equilibrium energy spectrum and the decorrelation time in the fluctuation modes display again the decaying energy and the fast mixing rate in small scales typical in turbulent flows. The small-scale fluctuation modes contain small

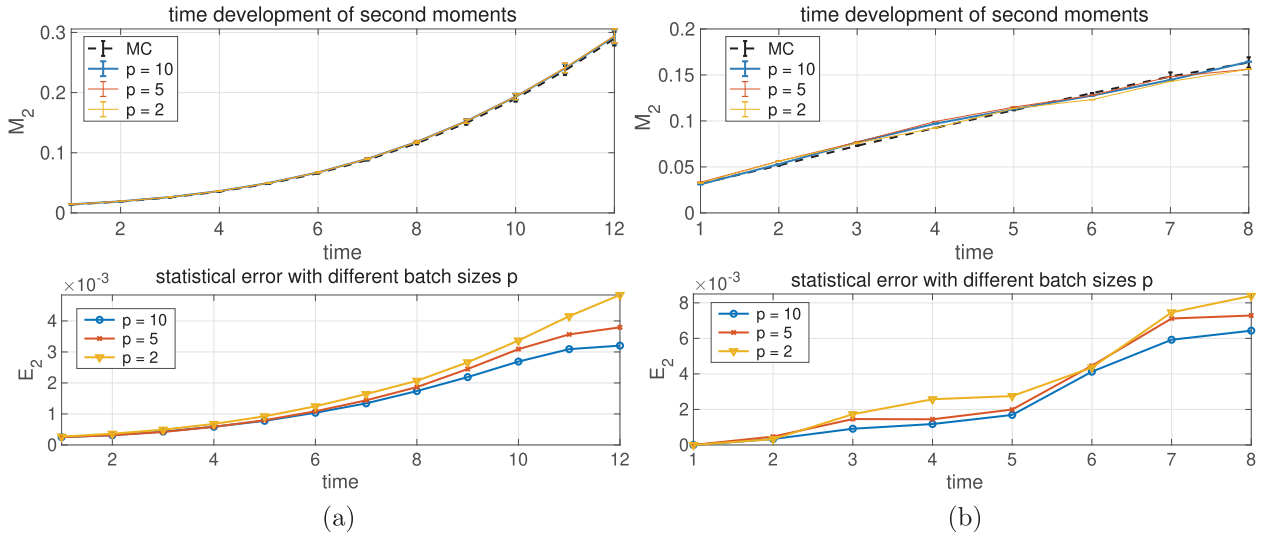


FIG. 6. Statistical moments $M_2 = \frac{1}{N}\sum_i |u^{(i)}|^2$ and errors $E_2 = \left|\frac{1}{N}\sum_i |u^{(i)}|^2 - \frac{1}{N}\sum_i |\tilde{u}^{(i)}|^2\right|$ of the RBM model in the time development of the conceptual model (left) and the topographic model (right). The results with different batch sizes $p = 2, 5, 10$ are compared with the direct MC approach.

TABLE II. Parameter values for the topographic barotropic model.

K	H	β	d_0	σ_0	d_k	σ_k	E_0	K_1	p	N (full MC)	N_1 (RBM)
100	1	2	0.0125	$\frac{1}{4\sqrt{2}}$	$0.0125k$	$\sqrt{2E_k d_k}$	0.02	2	5	10 000	100

energy and fast decaying autocorrelation in each single mode, while they constitute a major contribution in their combined feedback to the mean state. The bursts in the first two leading modes imply the occurrence of extreme events excited by such a multiscale interaction. Thus, the small scales play a non-negligible role and are suitable for the RBM approximation.

In the test for ensemble forecast, we aim to capture the PDFs in the mean-flow state U together with the first two leading modes v_1, v_2. Since the fluctuation modes are in complex values, it forms a five-dimensional subspace compared with the full model dimension $1 + 2K = 201$. A large ensemble size $N = 1 \times 10^4$ is again needed to sufficiently sample the high-dimensional phase space of the system in the full MC model, while in contrast, the RBM model uses a much smaller ensemble of $N_1 = 100$ samples. We put the detailed RBM model formulation of the topographic model (22) following Algorithm 1 in Appendix 2. In this topographic model, one typical feature is the intermittent bursts of extreme events in both the mean flow U and the leading fluctuation modes v_1, v_2, reflected by the skewed fat-tails in the resulting PDFs. This makes an even more challenging case for accurate ensemble forecast since it usually requires a much larger ensemble size to capture the extreme events in the asymmetric PDF tails with accuracy. The RBM prediction for the marginal PDFs and the joint PDFs in the resolved leading modes is shown in Fig. 5 compared with the direct MC results. As shown in both the marginal PDFs of the leading modes and the joint distributions, complicated non-Gaussian structures are generated during the evolution of the states. Again, the RBM model maintains a high skill to capture the skewed PDF structures while greatly reducing the computational cost.

Finally, we offer a quantitative calibration for the prediction errors by measuring the empirical statistics in the first two moments as $E_2 = \left| \frac{1}{N}\sum_i \left|u^{(i)}\right|^2 - \frac{1}{N}\sum_i \left|\tilde{u}^{(i)}\right|^2 \right|$ with u being the mean state of the full MC model and $\tilde{u}$ being the RBM solution. In comparison, for the error amplitudes, we also plot the time evolution of the absolute moments $M_2 = \frac{1}{N}\sum_i \left|u^{(i)}\right|^2$ together with the error bars estimated from several repeated simulations. A large sample size $N = 1 \times 10^4$ is used to reduce the error from the empirical sample average approximation. In Fig. 6, we plot the evolution of errors for the two test models. Notice that there exist errors from the ensemble approximation of the expectation due to the finite sample size N. Still, it shows that the error stays small in the ensemble approximations during the entire time evolution of the second moments M_2. It is observed that the RBM model maintains accurate prediction skill of the statistics with small errors during the model evolution. The errors gradually grow in time and will saturate at a low level when the system reaches a statistical equilibrium. As a further comparison, we also compare the errors under different batch sizes p. It shows that we can even push the batch sizes to an extreme $p = 2$, and the model still provides accurate prediction with just a slightly larger error. Overall, this confirms the robust performance of the RBM model subject to different turbulent dynamical features and for different statistical regime in both transient and final equilibrium state.

VI. SUMMARY

We developed a new efficient ensemble prediction strategy to model and forecast the time evolution of probability distributions and the associated statistical features in the dominant large-scale states of complex turbulent systems with a coupled multiscale structure. Standard Monte Carlo simulation of a high-dimensional turbulent system suffers the curse-of-dimensionality and, thus, requires an unaffordable ensemble size to even maintain a low-order approximation. The proposed RBM model circumvents the inherent difficulty by just sampling a low-dimensional subspace, which contains the dominant large-scale states, while the contributions from all the small-scale fluctuation modes are fully considered through a random batch decomposition. The wide spectrum of the K small-scale fluctuation modes is randomly divided into small batches of size p at the start of each time updating step, and each of the K/p batches is associated with one of the N large-scale state samples to update the coupled nonlinear feedback term computed inside the batch. The modes in each batch serve as the different samples to compute the small-scale feedback in the ensemble update of the large-scale state. The computational cost is then greatly reduced based on the random batch decomposition, which avoids the expensive ensemble simulation of a large number of small-scale fluctuation modes. The true multiscale dynamics is recovered in the efficient algorithm due to the frequent resampling of the batches at each time updating step and the fast mixing rate of the ergodic small-scale fluctuation modes. The resulting algorithm is also very easy to implement for a general group of multiscale turbulent models capable of creating a wide variety of realistic complex phenomena. Therefore, the efficient RBM model developed here provides a useful tool for improving the understanding of various turbulent features observed in natural and engineering systems and the further development of effective methods in uncertainty quantification and data assimilation[4,46] of complex turbulent systems.

In the analysis of the RBM model for the coupled large- and small-scale turbulent systems, the approximation errors under statistical expectation of the empirical ensemble average are derived by comparing the semigroups generated by the backward equation of the original model and the RBM approximation. The error is shown to be only related to the numerical time step and independent of the sample size and the full dimension of the system. The RBM model is then applied to two representative turbulent models with a close link to several realistic phenomena, such as extreme events

and intermittent instability. One central issue in practical forecast of turbulent systems concerns the accurate characterization of extreme events represented in the long extended PDF tails and the deviation from the Gaussian distribution. The RBM model is show to have a uniformly high skill in predicting various different structures in the PDFs during the time evolution in both test models driven by different types of coupling mechanisms. Only a very small ensemble size $N = 100$ is needed to achieve sufficient accuracy for models with a high-dimensional fluctuation state dimension $K = 100$ and $K = 200$. In contrast, the direct MC simulation requires at least $N = 1 \times 10^4$ samples to reach a relatively high accuracy. The RBM model here is designed for the large–small scale interaction model (3) with a wide spectrum of fast mixing and rapid decaying small-scale fluctuation modes. In the next stage development of the method, a more complete approach is needed to include both the large–small scale interaction and the self-coupling between the small scales in the fully coupled model (5) and to apply to more realistic applications with coupled multiscale effects. The RBM model here shows potential to deal with the different levels of multiscale coupling effects combining all stochastic modes and to overcome the curse-of-dimensionality for a wider group of practical problems involving prediction and data assimilation of fully turbulent high-dimensional flows.

ACKNOWLEDGMENTS

The research of J.-G. L. was partially supported under the NSF grant (No. DMS-2106988). We thank the many suggestions and comments from the reviewers that help to improve this manuscript.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Di Qi: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Writing – original draft (equal). **Jian-Guo Liu:** Conceptualization (equal); Formal analysis (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: PROOFS OF THE LEMMAS

Proof of Lemma 3. This is the direct conclusion by counting the number of ordered combinations of the n random batches. First, we define the total number of ways of listing np distinguishable objects into n ordered batches of size p as

$$n!M(n) = \frac{(np)!}{(p!)^n}.$$

We use $M(n)$ to denote the number of combinations to put np objects to p groups without order. Next, for the kth mode falling in the batch i, we determine the ith batch to contain k and select the other $p-1$ objects in this batch from the remaining $np-1$ objects and then order the rest $n-1$ batches. This gives

$$\mathbb{E}I_i(k) = \frac{\binom{np-1}{p-1}(n-1)!M(n-1)}{n!M(n)} = \frac{1}{n} = \frac{p}{K}.$$

Similarly, for two modes $k \neq l$ falling in the same batch i, we put these two objects together with other $p-2$ objects and then still order the rest $n-1$ batches, which gives

$$\mathbb{E}I_i(k)\, I_i(l) = \frac{\binom{np-2}{p-2}(n-1)!M(n-1)}{n!M(n)} = \frac{1}{n}\frac{p-1}{np-1} = \frac{p}{K}\frac{p-1}{K-1}.$$

□

Proof of Lemma 4. Applying Ito's lemma for $f(z) = \|z\|^{2q} = \left\{\left(\sum_k |z_k|^2\right)\right\} q$ according to the SDE of Z_k in (12), we have the conditional expectation $\mathbb{E}_u$ with a fixed u as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_u\|Z\|^{2q} = 2q\sum_k \left(\gamma_k(u) - d_k\right)\mathbb{E}_u Z_k^2\|Z\|^{2(q-1)} + q\mathbb{E}_u\|Z\|^{2(q-1)}\sum_k \sigma_k^2 + 2q(q-1) \times \sum_k \sigma_k^2 \mathbb{E}_u Z_k^2\|Z\|^{2(q-2)}.$$

From (14c) in Assumption 1, the coefficients are uniformly bounded, $d_k - \gamma_k(u) \geq r > 0$ and $\sum_k \sigma_k^2 \leq C$. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_u\|Z\|^{2q} \leq -2qr\mathbb{E}_u\|Z\|^{2q} + q^2C'\mathbb{E}_u\|Z\|^{2(q-1)}.$$

First, for $q = 1$, the last term on the above inequality becomes a constant; thus, $\mathbb{E}\|Z_t\|^2 = \mathbb{E}^U\mathbb{E}\|Z(t \mid U)\|^2 \leq \frac{C'}{2r} \equiv C_1$. Next, we have by induction $\mathbb{E}\|Z_t\|^{2q} \leq C_q$ for any integer $q \geq 1$. In the same way, we have $\mathbb{E}\left\|\tilde{Z}_t\right\|^{2q} < C_q$ for any $t > 0$. □

Proof of Lemma 5. By using the backward Kolmogorov equation (19) for $w_z(x,t)$ and taking its derivative about the ith coordinate $x_i \in \mathbb{R}^d$ with $i = 1, \ldots, N$, it yields

$$\partial_t \partial_{x_i} w_z = \mathcal{L}_z \partial_{x_i} w_z + \nabla V(x_i) \cdot \partial_{x_i} w_z.$$

Above, we define the vector function $\partial_{x_i} w_z = \{\partial_{x_i^j} w_z\}_{j=1}^d \in \mathbb{R}^d$ with $x_i = \{x_i^j\}_j$ and $\nabla V(x) = \{\partial_{x^j} V_j\}_j : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. From the definition of the generator, the only term that is dependent on x in $\mathcal{L}_z$ is $V(x_i)$. This leads to the formal solution

$$\partial_{x_i} w_z(x,t) = e^{t\mathcal{L}_z}\partial_{x_i} w_z(x,0) + \int_0^t e^{(t-s)\mathcal{L}_z}\nabla V(x_i) \cdot \partial_{x_i} w_z(x,s)\,\mathrm{d}s.$$

By the contraction of the semigroup $e^{t\mathcal{L}_z}$, we have

$$\left\|e^{t\mathcal{L}_z} w_z(x,0)\right\|_\infty \leq \|w_z(x,0)\|_\infty$$

and

$$\left\| e^{(t-s)\mathcal{L}_z} \nabla V(x_i) \cdot \partial_{x_i} w_z(x,s) \right\|_\infty \le \left\| \nabla V(x_i) \cdot \partial_{x_i} w_z(x,s) \right\|_\infty .$$

Therefore, using the uniform boundedness of V in (14b),

$$\begin{aligned} \left\| \partial_{x_i} w_z(x,t) \right\|_\infty &\le \left\| \partial_{x_i} w_z(x,0) \right\|_\infty + \int_0^t \left\| \nabla V(x_i) \cdot \partial_{x_i} w_z(x,s) \right\|_\infty \mathrm{d}s \\ &\le \left\| \partial_{x_i} w(x,0) \right\|_\infty + C \int_0^t \left\| \partial_{x_i} w_z(x,s) \right\|_\infty \mathrm{d}s. \end{aligned}$$

Using the integral form of Grönwall's inequality, we get

$$\begin{aligned} \left\| \partial_{x_i} w(x,t) \right\|_\infty \le \mathbb{E}_Z \left\| \partial_{x_i} w_Z(x,t) \right\|_\infty &\le C(t) \left\| \partial_{x_i} w(x,0) \right\|_\infty \\ &\le \frac{C(t,\varphi)}{N}. \end{aligned}$$

Above, in the last inequality, by definition, $w(x,0) = \frac{1}{N}\sum_{i=1}^N \varphi(x_i)$, then $\partial_{x_i} w(x,0) = \frac{1}{N}\nabla\varphi(x_i)$ is bounded since $\varphi \in C_b^2$.

Next, by taking a second derivative about x_j on the backward equation, we have

$$\begin{aligned} \partial_t \partial^2_{x_i x_j} w_z = \mathcal{L}_z \partial^2_{x_i x_j} w_z &+ \left[\nabla V(x_i) + \nabla V(x_j) \right] \cdot \partial^2_{x_i x_j} w_z \\ &+ \delta_{ij} \nabla^2 V(x_i) \cdot \partial_{x_i} w_z. \end{aligned}$$

Above, the last term on the right hand side only appears when $i = j$. Following the same argument as before and using the boundedness of the initial condition $\partial^2_{x_i x_j} w(x,0) = \frac{1}{N}\nabla^2\varphi(x_i)\,\delta_{ij}$ and $\nabla^2 V$, we first get the bound for a second derivative about i, j,

$$\left\| \partial^2_{x_i x_j} w(x,t) \right\|_\infty \le \frac{C}{N}.$$

Then, under similar estimation of the second-order derivation equation and taking summation among all the samples i, $j = 1, \ldots, N$,

$$\begin{aligned} \sum_{i,j} \left\| \mathbb{E}^Z \partial^2_{x_i x_j} w_Z(x,t) \right\|_\infty &\le \sum_{i,j} \left\| \partial^2_{x_i x_j} w(x,0) \right\|_\infty \\ &\quad + C_0 \int_0^t \sum_i \left\| \mathbb{E}^Z \partial_{x_i} w_Z(x,s) \right\|_\infty \mathrm{d}s \\ &\quad + C_1 \int_0^t \sum_{i,j} \left\| \mathbb{E}^Z \partial^2_{x_i x_j} w_Z(x,s) \right\|_\infty \mathrm{d}s \\ &\le C(T) + C_1 \int_0^t \sum_{i,j} \left\| \mathbb{E}^Z \partial^2_{x_i x_j} w_Z(x,s) \right\|_\infty \mathrm{d}s. \end{aligned}$$

In the second row, we use the initial condition $w(x,0) = \frac{1}{N}\sum_{i=1}^N \varphi(x_i)$ and the estimation on the first derivative so that the bounds on the right are both of order $O(1)$ and independent of N

$$\sum_{i,j} \left\| \partial^2_{x_i x_j} w(\cdot,0) \right\|_\infty = \sum_i \frac{1}{N} \left\| \nabla^2 \varphi \right\|_\infty,$$

$$\sum_i \left\| \partial_{x_i} w(\cdot,s) \right\|_\infty \le \sum_i \frac{C}{N}.$$

Letting $f(t) = \sum_{i,j} \left\| \mathbb{E}^Z \partial^2_{x_i x_j} w_Z(\cdot,t) \right\|_\infty$, then

$$f(t) \le C + C_1 \int_0^t f(s)\,\mathrm{d}s.$$

Grönwall's inequality gives

$$\sum_{i,j} \left\| \partial^2_{x_i x_j} w \right\|_\infty \le C(t,\varphi),$$

with the constant C on the right only dependent on the time t and the test function φ. □

APPENDIX B: DETAILED RBM FORMULATION FOR THE TEST MODELS

1. RBM equations for the conceptual turbulent model

Here, we show the detailed equations of the RBM model for the ensemble prediction of the conceptual turbulent model (21). First, we decompose the small-scale states $v = \{v_{1,k}, v_{2,k}\}$ further into the unstable leading modes $v_{1,k}, k \le K_1$ and the rest fluctuating smaller scales, $v_{2,k}, K_1 < k \le K$, with stable dynamics. This decomposition is used to also resolve the leading fluctuation modes containing intermittent unstable growth when $-(d_k + \gamma\bar{u}) > 0$ from the coupling with the mean. In this way, we still only need to sample a much lower dimensional subspace containing the most energetic leading modes together with the mean state, that is, $\{\bar{u}, v_{1,k}\}_{k\le K_1}$, while the much less energetic stable smaller-scale fluctuation modes are modeled by the random batches. The ensemble approximation of the marginal PDF with N_1 samples becomes

$$p_{\mathrm{RBM}}(\bar{u}, v_1) = \frac{1}{N_1}\sum_{i=1}^{N_1} \delta\left(\bar{u} - \bar{u}^{(i)}\right) \otimes \prod_{k=1}^{K_1} \delta\left(v_{1,k} - v_{1,k}^{(i)}\right).$$

The above empirical approximation requires a much smaller ensemble size N_1 and is independent of the full dimension $K (\gg K_1)$ of the system. Next, the random batch partition is applied to the large group of small-scale fluctuation modes $\{v_{2,k}\}_{k>K}$ exploiting their ergodicity with a fast mixing rate. In particular, the large number of fluctuation modes $\{v_{2,k}\}$ are divided into small batches of size p each. In the RBM model, the sample size N_1 in the ensemble simulation is associated with the same number of batches. For the ith sample in the large-scale ensemble $\{\bar{u}^{(i)}, v_{1,k}^{(i)}\}$, the large-scale mean equation only includes p randomly picked small-scale modes in one batch, $\{v_{2,k} : k \in \mathcal{I}_i\}$. The RBM model with samples $i = 1, \ldots, N_1$ at time step $t = t_n$ becomes

$$\begin{aligned} \frac{\mathrm{d}\bar{u}^{(i)}}{\mathrm{d}t} &= -\bar{d}\bar{u}^{(i)} + \frac{\gamma}{K_1}\sum_{k=1}^{K_1}\left(v_{1,k}^{(i)}\right)^2 + \frac{\gamma}{p}\sum_{k\in\mathcal{I}_i}\left(v_{2,k}\right)^2 - \bar{\alpha}\left(\bar{u}^{(i)}\right)^3 + \bar{F}, \\ \frac{\mathrm{d}v_{1,k}^{(i)}}{\mathrm{d}t} &= -d_k v_{1,k}^{(i)} - \gamma\bar{u}^{(i)} v_{1,k}^{(i)} + \sigma_k \dot{W}_k^{(i)}, \quad 1 \le k \le K_1, \qquad \text{(B1)} \\ \frac{\mathrm{d}v_{2,k}}{\mathrm{d}t} &= -d_k v_{2,k} - \gamma\bar{u}^{(i)} v_{2,k} + \sigma_k \dot{W}_k, \qquad k \in \mathcal{I}_i, \quad i = 1, \ldots, N_1. \end{aligned}$$

The small-scale modes $v_{2,k}$ are segmented into small batches with $\cup_i \mathcal{I}_i = \{k : K_1 < k \le K\}$. In addition, we add a small ensemble $n_2 (= 5)$ for the small-scale modes. Therefore, there are

$N_1 = \left\lceil n_2 \frac{K-K_1}{p} \right\rceil$ samples to sufficiently sample the resolved subspace. Only a very small ensemble is needed for the high-dimensional subspace for $\{v_{2,k}\}$ satisfying

$$n_2 = \frac{N_1 p}{K - K_1}.$$

We can estimate the computational cost of (B1) as $O\left(N_1\left(1+K_1\right)p\right) \sim O\left(n_2\left(1+K_1\right)K\right)$. Notice that the cost will not have the exponential growth depending on the full dimension $1+K$ of the system by avoiding sampling the full high-dimensional space.

2. RBM equations for the topographic barotropic model

In a similar way, we display the detailed equations for the implementation of the RBM model for the topographic barotropic model (22). Again, we focus on the ensemble approximation of the dominant mean state and the first K_1 leading modes, $\left(U, v_{1,k}\right)$ with $|k| \le K_1$

$$p_{\mathrm{RBM}}\left(U, v_1\right) = \frac{1}{N_1}\sum_{i=1}^{N_1} \delta\left(U - U^{(i)}\right) \otimes \prod_{k=1}^{K_1} \delta\left(v_{1,k} - v_{1,k}^{(i)}\right),$$

while all the other small-scale modes $\{v_{2,k}\}, K_1 < |k| \le K$ are modeled in the random batches $\cup_i \mathcal{I}_i = \{k : K_1 < \lceil k \rceil \le K\}$ with the size of the batches $p = |\mathcal{I}_i|$. The RBM model for samples $i = 1, \ldots, N_1$ at time step $t = t_n$ becomes

$$\frac{\mathrm{d}U^{(i)}}{\mathrm{d}t} = \frac{1}{K_1}\sum_{|k|\le K_1} h_k^* v_{1,k}^{(i)} + \frac{1}{p}\frac{\tilde{K}-1}{p-1}\sum_{k\in\mathcal{I}_i} h_k^* v_{2,k} - d_0 U^{(i)} + \sigma_0 \dot{W}_0^{(i)},$$

$$\frac{\mathrm{d}v_{1,k}^{(i)}}{\mathrm{d}t} = i\left(k^{-1}\beta - kU^{(i)}\right)v_{1,k}^{(i)} - U^{(i)}h_k - d_k v_{1,k}^{(i)} + \sigma_k \dot{W}_k^{(i)}, \quad |k| \le K_1, \tag{B2}$$

$$\frac{\mathrm{d}v_{2,k}}{\mathrm{d}t} = i\left(k^{-1}\beta - kU^{(i)}\right)v_{2,k} - U^{(i)}h_k - d_k v_{2,k} + \sigma_k \dot{W}_k, \qquad k \in \mathcal{I}_i.$$

Notice that we have the fixed-in-time auxiliary variable h_k; thus, we need to take the new quadratic coupling coefficients $c_{kl} = \frac{1}{p}\frac{\tilde{K}-1}{p-1}$ in (11) between the two different modes $v_{2,k}, h_k$. In this case, the $\tilde{K} = K - K_1$ small-scale modes $v_{2,k}$ are partitioned into N_1 small batches, and only the modes in the ith batch are used to update the ith mean state sample $U^{(i)}$. Still, the first K_1 leading modes $v_{1,k}^{(i)}$ are resolved explicitly by the ensemble and acting on the corresponding mean state $U^{(i)}$ considering their important role in generating the correct dynamics.

REFERENCES

[1] D. R. Nicholson and D. R. Nicholson, *Introduction to Plasma Theory* (Wiley, New York, 1983), Vol. 582.
[2] U. Frisch, *Turbulence: The Legacy of AN Kolmogorov* (Cambridge University Press, 1995).
[3] R. Salmon, *Lectures on Geophysical Fluid Dynamics* (Oxford University Press, 1998).
[4] A. J. Majda, *Introduction to Turbulent Dynamical Systems in Complex Systems* (Springer, 2016).
[5] W.-K. Tao, J.-D. Chern, R. Atlas, D. Randall, M. Khairoutdinov, J.-L. Li, D. E. Waliser, A. Hou, X. Lin, C. Peters-Lidard *et al.*, Bull. Am. Meteorol. Soc. **90**, 515 (2009).
[6] L.-S. Young, J. Stat. Phys. **108**, 733 (2002).
[7] J. D. Neelin, B. R. Lintner, B. Tian, Q. Li, L. Zhang, P. K. Patra, M. T. Chahine, and S. N. Stechmann, Geophys. Res. Lett. **37**, L05804, https://doi.org/10.1029/2009GL041726 (2010).
[8] K. Srinivasan and W. Young, J. Atmos. Sci. **69**, 1633 (2012).
[9] D. C. Wilcox, AIAA J. **26**, 1311 (1988).
[10] M. Leutbecher and T. N. Palmer, J. Comput. Phys. **227**, 3515 (2008).
[11] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo Statistical Methods* (Springer, 1999), Vol. 2.
[12] J.-G. Liu and R. Yang, Math. Comput. **86**, 725 (2017).
[13] D. Qi and E. Vanden-Eijnden, AIP Adv. **12**, 025016 (2022).
[14] Y. Gao, T. Li, X. Li, and J.-G. Liu, "Transition path theory for Langevin dynamics on manifolds: Optimal control and data-driven solver," Multiscale Model. Simul. **21**, 1–33 (2023).
[15] N. Chen and A. J. Majda, Chaos **30**, 033101 (2020).
[16] T. Gneiting and A. E. Raftery, Science **310**, 248 (2005).
[17] B. Giggins and G. A. Gottwald, Q. J. R. Meteorol. Soc. **145**, 642 (2019).
[18] B. Giggins and G. A. Gottwald, Q. J. R. Meteorol. Soc. **146**, 4038 (2020).
[19] A. J. Majda and D. Qi, SIAM Rev. **60**, 491 (2018).
[20] A. J. Majda and D. Qi, Chaos **29**, 103131 (2019).
[21] S. Jin, L. Li, and J.-G. Liu, J. Comput. Phys. **400**, 108877 (2020).
[22] S. Jin, L. Li, and J.-G. Liu, SIAM J. Numer. Anal. **59**, 746 (2021).
[23] Y. Gao, J.-G. Liu, and N. Wu, Appl. Comput. Harmon. Anal. **62**, 261 (2023).
[24] H. Robbins and S. Monro, Ann. Math. Stat. **22**, 400 (1951).
[25] S. Bubeck, Found. Trends Mach. Learn. **8**, 231 (2015).
[26] W. Hu, C. J. Li, L. Li, and J.-G. Liu, Ann. Math. Sci. Appl. **4**, 3 (2019).
[27] S. Jin and X. Li, "Random batch algorithms for quantum Monte Carlo simulations," arXiv:2008.12990 (2020).
[28] A. Majda and X. Wang, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows* (Cambridge University Press, 2006).
[29] G. K. Vallis, *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation* (Cambridge University Press, 2006).
[30] D. Olbers, in *Stochastic Climate Models* (Springer, 2001), pp. 3–63.
[31] D. Qi, A. J. Majda, and A. J. Cerfon, Phys. Plasmas **27**, 102304 (2020).
[32] S. S. Varadhan, *Stochastic Processes* (American Mathematical Society, 2007), Vol. 16.
[33] P. H. Diamond, S. Itoh, K. Itoh, and T. Hahm, Plasma Phys. Control. Fusion **47**, R35 (2005).
[34] R. H. Kraichnan, J. Math. Phys. **2**, 124 (1961).
[35] A. S. Lanotte, R. Benzi, S. K. Malapaka, F. Toschi, and L. Biferale, Phys. Rev. Lett. **115**, 264502 (2015).
[36] A. J. Majda and N. Chen, Entropy **20**, 644 (2018).
[37] F. Daum and J. Huang, in *2003 IEEE Aerospace Conference Proceedings (Cat. No. 03TH8652)* (IEEE, 2003), Vol. 4, pp. 4_1979–4_1993.
[38] J. H. Friedman, Data Min. Knowl. Discov. **1**, 55 (1997).
[39] E. Weinan, J. C. Mattingly, and Y. Sinai, "Gibbsian dynamics and ergodicity for the stochastically forced Navier-Stokes equation," Commun. Math. Phys. **224**(1), 83–106 (2001).
[40] J. C. Mattingly, A. M. Stuart, and D. J. Higham, Stoch. Process. Their Appl. **101**, 185 (2002).
[41] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications* (Springer Science & Business Media, 2013).
[42] Y. Feng, L. Li, and J.-G. Liu, Commun. Math. Sci. **16**, 777 (2018).
[43] A. J. Majda and Y. Lee, Proc. Natl. Acad. Sci. U.S.A. **111**, 6548 (2014).
[44] G. F. Carnevale and J. S. Frederiksen, J. Fluid Mech. **175**, 157 (1987).
[45] A. J. Majda and P. R. Kramer, Phys. Rep. **314**, 237 (1999).
[46] S. Reich and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation* (Cambridge University Press, 2015).