

A data-driven statistical-stochastic surrogate modeling strategy for complex nonlinear non-stationary dynamics

Di Qi^{a,*}, John Harlim^b

^a Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA

^b Department of Mathematics, Department of Meteorology and Atmospheric Science, Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:

Received 27 August 2022

Received in revised form 1 March 2023

Accepted 19 March 2023

Available online 27 March 2023

Dataset link: <https://github.com/qidigit/non-Markovian-closure-LSTM>

Keywords:

Statistical response

Stochastic closure model

Inhomogeneous dynamics with instability

Data-driven parameterization

Information metric

ABSTRACT

We propose a statistical-stochastic surrogate modeling approach to predict the response of the mean and variance statistics under various initial conditions and external forcing perturbations. The proposed modeling framework extends the purely statistical modeling approach that is practically limited to the homogeneous statistical regime for high-dimensional state variables. The new closure system allows one to overcome several practical issues that emerge in the non-homogeneous statistical regimes. First, the proposed ensemble modeling that couples the mean statistics and stochastic fluctuations naturally produces positive-definite covariance matrix estimation, which is a challenging issue that hampers the purely statistical modeling approaches. Second, the proposed closure model, which embeds a non-Markovian neural-network model for the unresolved fluxes such that the variance of the dynamics is consistent, overcomes the inherent instability of the stochastic fluctuation dynamics. Effectively, the proposed framework extends the classical stochastic parametric modeling paradigm for the unresolved dynamics to a semi-parametric parameterization with a residual Long-Short-Term-Memory neural network architecture. Third, based on empirical information metric, we provide an efficient and effective training procedure by fitting a loss function that measures the differences between response statistics. Supporting numerical examples are provided with the Lorenz-96 model, a system of ODEs that admits the characteristic of chaotic dynamics with both homogeneous and inhomogeneous statistical regimes. In the latter case, we will see the effectiveness of the statistical prediction even though the resolved Fourier modes corresponding to the leading mean energy and variance spectra do not coincide.

© 2023 Elsevier Inc. All rights reserved.

1. Introduction and background

One of the key challenges in uncertainty quantification of dynamical systems [1–3] and data assimilation [4–6] is to construct a surrogate model that allows one to accurately and efficiently predict the evolution of the low-order statistics under perturbation of model parameters (e.g., additional forces) and initial conditions. Computationally, how uncertainty propagates in dynamical systems is usually characterized by understanding how the second-order statistics change when the system's initial conditions and/or parameters are perturbed. Under mild perturbations and appropriate mathematical

* Corresponding author.

E-mail addresses: qidigit@purdue.edu (D. Qi), jharlim@psu.edu (J. Harlim).

conditions, this problem has also been studied in non-equilibrium statistical mechanics (see e.g. Chapter 7 of [7]). The time evolution of mean and covariance statistics is also essential in data assimilation. In this application, most algorithms (such as Kalman filtering and its variant) construct conditional mean and covariance statistics by a Bayesian formula that updates these low-order statistics to account for the newly measured observations. Subsequently, the resulting conditional statistics are fed into the dynamical model as initial conditions for predicting the state with the mean and uncertainty characterized by the covariance statistics.

An expensive method to compute these statistics is to employ a Monte-Carlo simulation. This approach involves solving an ensemble of solutions for the dynamical systems and using these solutions to empirically estimate the statistics of interest. In the data assimilation context, such an idea has been realized by the well-known Ensemble Kalman Filter algorithm [8]. With the Monte-Carlo approach, the computational cost is determined by the complexity of integrating the dynamical system multiplying the ensemble size. One known issue with such an approach is that the ensemble size required to maintain the desired accuracy will grow exponentially as a function of the state-spaced dimension. This issue poses major computational challenge when online statistical predictions under new initial conditions and forces are needed in uncertainty quantification and data assimilation applications. The approach adopted in this paper is to construct a reduced-order model to accurately predict the evolution of the low-order statistics, where an offline machine learning algorithm is employed to emulate the feedback from higher-order statistical moments.

While the proposed formulation here is applicable to any complex spatially extended nonlinear dynamical systems, it is also closely related to a long-standing moment closure problem of turbulent dynamical systems that has been widely studied in many fields of science and engineering [9–11]. This serendipity motivates us to present the formulation on a specific class of nonlinear dynamical system, where the moment interactions are induced by a bilinear quadratic form that is typically inherited from a discretization (or spectral projection) of nonlinear advection in fluid dynamical models, such as in the Navier-Stokes and Burger's equations, and the spatiotemporally chaotic Kuramoto-Sivashinsky equation. In any nonlinear systems, the moment dynamics is not closed in the sense that the dynamical equation for each moment depends on the higher-order moments in addition to the lower-order statistics. This inherent hierarchical structure poses some practical issues especially if one is interested to resolve at least the first- and second-order moments. Particularly, for a system with N -dimensional state space variables, while the evolution of the first-order moment is represented by a system of N -dimensional differential equations, the evolution of the second-order moment is represented by an $N \times N$ matrix-valued differential equations that further depend on components of the unresolved third-order moments of size $N \times N \times N$.

For the first-order moment closure of turbulent dynamics, machine learning has been used to approximate the unresolved subgrid scale terms [12–14]). In fact, for the second-order moment closure problem, machine learning with Long-Short-Term-Memory (LSTM) architecture has been proposed [15] to emulate the feedback from the third-order moments. The approach in [15], unfortunately, is restricted to spatially homogeneous statistics. When the statistics are spatially homogeneous, the mean and variance dynamics can be represented by a $(1 + N)$ -dimensional system, consisting of the one-dimensional mean variable and N -dimensional variance components since the non-diagonal entries in the covariance matrix are all zeros. Such a reduced representation, unfortunately, is invalid in non-homogeneous systems. The work in this paper is to extend the machine learning approach in [15] to non-homogeneous statistics. In this regime, several issues emerge. Beyond the practical issue of resolving the $N \times N$ non-diagonal covariance matrix, constructing a machine learning model for the feedback from the third-order moments that preserves the positive-definite covariance remains a challenging task, especially when the covariance dynamical equation is conditionally unstable due to the quasilinear coupling with the mean state.

To overcome this issue, we propose a reduced-order system of differential equations for the leading-order mean state and stochastic fluctuations. In this fully coupled statistical-stochastic system, the reduced-order mean dynamics depend on the covariance matrix that is empirically estimated using the ensemble prediction of the fluctuation terms and the resolved mean state determines the linear stability of the stochastic fluctuation dynamics. Denoting the resolved state space dimension by K , where $K < N$, the dynamical equation of the proposed statistical-stochastic model has the complexity of order $K(1 + M)$, accounting for the K -dimensional vector for the mean and the M ensemble members of K -dimensional fluctuation equations. While the system can be moderately high-dimensional, especially if the ensemble size $M > K$, the unresolved processes to be modeled in this formulation are only K -dimensional, accounting for the modeling error in the K -dimensional reduced-order mean dynamics and the unresolved fluxes in the K -dimensional fluctuation dynamics.

For a scalable and accurate statistical prediction on this systematic modeling framework, two issues need to be addressed. First, internal instability as a common feature in chaotic dynamics often leads to unstable dynamic and fast divergence of the solution when the data-driven surrogate model is chosen from an arbitrary hypothesis space (such as the neural-network models) without any dynamical constraints. To overcome this issue, we will impose dynamical constraints through a semi-parametric framework with consistent variances, embedding the neural-network model on a parametric modeling framework proposed in [15,2]. Second, while the ensemble structure in the proposed statistical-stochastic system is natural for ensemble prediction and data assimilation, the computational cost of this fully coupled system $K(1 + M)$ is often too high for accurate learning. In our numerical example, the state-space dimension is of order 1000 for a $K = 14$ dimensional reduced-order state variable. In this case, the standard learning procedure of fitting an empirical loss function that compares the trajectories of these states [16] may not be numerically feasible since it requires a very large training data set resulting in an enormously expensive computational cost. In the Appendix of this paper, we documented that applying such a learning

procedure with a generic neural-network model to identify the unresolved fluxes in the fluctuation dynamics leads to overfitting. To overcome this issue, we will consider an empirical information metric as a loss function [17] that compares the response mean and variance statistics. We will show that fitting to the statistical responses corresponding to the same trajectories that led to overfitting in the standard procedure produces accurate response statistical prediction subject to new initial conditions and external forces, both in homogeneous and inhomogeneous statistical regimes.

As we already mentioned above, while the approach can be implemented on any spatially extended chaotic dynamical systems, since the approach is closely related to the moment closure problem, we will present our approach on nonlinear systems with a bilinear form. Specifically, we will examine the stochastic-statistical formulation of the Lorenz-96 model [18] which admits the characteristic of the chaotic dynamical systems and can be adjusted to generate non-trivial inhomogeneous statistics. First, we should point out that the variance spectrum in this system decays slowly (relative to the Kolmogorov decay in classical turbulence theory) and the corresponding Fourier modes with large variance spectrum are all unstable. When spatially large-scale disturbances are injected into the system, while they excite the entire mean energy spectrum and variance spectrum, the changes in the mean energy spectrum are significantly noticeable in the Fourier modes corresponding to the lower variance spectrum. The mismatch between modes that have a large mean energy spectrum and those that have a large variance spectrum makes this system ideal for testing the proposed reduced-order statistical-stochastic framework. Particularly, this test model would allow us to understand to which extent the reduced ordering can be employed and whether the internal instability in this system can be overcome with the proposed modeling framework.

The remainder of this paper is organized as follows. In Section 2, we discuss the general statistical-stochastic closure modeling framework of turbulent dynamical systems. In Section 3, we discuss the proposed machine learning strategy on a concrete example, the Lorenz-96 model. In Section 4, we discuss the training configuration and present numerical results for the proposed closure framework, both on homogeneous and inhomogeneous statistical regimes. In Section 5, we close the paper with a summary. As mentioned in the above discussion, we include an example demonstrating the difficulty of attaining accurate trajectory prediction on the test dataset using the standard machine learning procedure in Appendix A, which motivates this work.

2. General mathematical formulation for complex systems with uncertainty

In this section, we give a quick overview of a general moment closure formulation for a class of complex nonlinear systems that is common in natural and engineering problems and formulate an efficient machine learning reduced-order model. One representative feature that makes the moment closure problem challenging in such complex systems is the nonlinear energy-conserving interaction that transports energy across scales. The general formulation of the turbulent dynamical systems can be characterized by the canonical equations of the state variable $\mathbf{u} \in \mathbb{R}^N$ in a high-dimensional phase space,

$$\frac{d\mathbf{u}}{dt} = (\mathcal{L} + \mathcal{D})\mathbf{u} + B(\mathbf{u}, \mathbf{u}) + \mathbf{F} + \sigma \dot{\mathbf{W}}. \quad (2.1)$$

On the right hand side of the equation (2.1), the first two components, $(\mathcal{L} + \mathcal{D})\mathbf{u}$, represent linear dispersion and dissipation effects, where $\mathcal{L}^* = -\mathcal{L}$ is an energy-conserving skew-symmetric operator for dispersive effects; and $\mathcal{D} < 0$ is a negative definite operator for dissipations. The nonlinear effect in the dynamical system is introduced through a quadratic form, $B(\mathbf{u}, \mathbf{u})$ [2] that arise in a discretization of the nonlinear advection in fluid mechanics. Besides, the system is usually subject to time-dependent external forcing effects that are decomposed into a deterministic component, $\mathbf{F}(t)$, and a stochastic component represented by a Gaussian random process, $\sigma(t)\dot{\mathbf{W}}(t; \omega)$. It needs to be emphasized that in many situations \mathbf{F} might be spatially inhomogeneous, and thus, introduce anisotropic structures into the system.

One way to characterize the effect of internal instabilities and the uncertainties from the initial state and forcing in the turbulent system (2.1) is through a statistical description for the time evolution of the moment of the state variable \mathbf{u} . While in principle the dynamical equations of the statistical moments follow the backward-Kolmogorov PDE (which is the L^2 adjoint of the Fokker-Planck equation that characterizes the evolution of the density function $p(\mathbf{u}, t)$), it remains challenging to computationally solve such a PDE, especially when state space dimension, N , is large. The Monte-Carlo approach [19,20], which uses an ensemble of solutions of (2.1) subjected to initial and forcing perturbations, provides an alternative means to quantify the essential statistics that quantify the uncertainties through empirical ensemble averages.

2.1. The exact formulation for statistical mean and stochastic fluctuation interactions

Despite its simplicity, a direct ensemble forecast obtained from integrating the original model (2.1) has several difficulties in accurately recovering the key model statistics in a high dimensional space. First, the ensemble size required to maintain the accuracy will grow exponentially in direct ensemble simulation of the full model as the dimension of the system increases. This requirement may not be computationally desirable, especially when online predictions under new initial conditions and forces are needed, and subsequently motivates the need for reduced-order modeling. On the other hand, turbulent systems often contain strong internal instability and mixed spatio-temporal structures. These features pose some computational challenges for developing effective reduced-order models directly under the original model formulation, especially when the reference dynamical system is nonlinear and non-Gaussian.

To address these difficulties, we introduce a statistical-stochastic decomposition of the model state \mathbf{u} , so that the mean-fluctuation interactions can be identified. Efficient model reduction strategies will be proposed where data-driven components can be introduced naturally to account for the unresolved fluctuation interactions. To achieve this, we view the model state \mathbf{u} as a random field and project it onto the composition of a statistical mean and stochastic fluctuations in a finite-dimensional representation under a suitable orthonormal basis $\{\mathbf{e}_i\}_{i=1}^N$ as,

$$\mathbf{u}(t; \omega) = \bar{\mathbf{u}}(t) + \mathbf{u}'(t; \omega) = \bar{\mathbf{u}}(t) + \sum_{i=1}^N Z_i(t; \omega) \mathbf{e}_i, \quad (2.2)$$

where $\bar{\mathbf{u}}(t) = \langle \mathbf{u}(t) \rangle$ (here and after, we use $\langle \cdot \rangle$ to denote the statistical expectation about the PDF $p(\mathbf{u}, t)$), represents the statistical expectation of the model state, i.e. the mean field; and $\{Z_i(t; \omega)\}$ as the mean-zero stochastic coefficients measuring the uncertainty in fluctuation processes \mathbf{u}' along each direction \mathbf{e}_i . The statistical uncertainty among the fluctuation modes can be characterized by the covariance between the stochastic modes.

By taking the statistical (ensemble) average over the original equation (2.1) and using the mean-fluctuation decomposition (2.2), the evolution equation of the statistical mean state $\bar{\mathbf{u}}$ is given by the following dynamical equation,

$$\frac{d\bar{\mathbf{u}}}{dt} = (\mathcal{L} + \mathcal{D})\bar{\mathbf{u}} + B(\bar{\mathbf{u}}, \bar{\mathbf{u}}) + \sum_{i,j=1}^N R_{ij} B(\mathbf{e}_i, \mathbf{e}_j) + \mathbf{F}, \quad (2.3)$$

where $R := \langle \mathbf{Z}\mathbf{Z}^* \rangle$ denotes the second-order covariance matrix of the stochastic coefficients $\mathbf{Z} = \{Z_i\}_{i=1}^N$. The term $B(\bar{\mathbf{u}}, \bar{\mathbf{u}})$ represents the nonlinear interactions between the mean state, and $R_{ij} B(\mathbf{e}_i, \mathbf{e}_j)$ is the higher-order feedback from the fluctuation modes to the mean state dynamics. Next, by projecting the above equation (2.1) to each orthonormal basis element \mathbf{e}_i we obtain the evolution equation for the stochastic fluctuation coefficients,

$$\frac{dZ_i}{dt} = \sum_{j=1}^N A_{ij}(\bar{\mathbf{u}}) Z_j + \sum_{m,n=1}^N \gamma_{imn} (Z_m Z_n^* - R_{mn}) + \sigma \dot{\mathbf{W}} \cdot \mathbf{e}_i, \quad (2.4)$$

where $A_{ij}(\bar{\mathbf{u}}) = [(\mathcal{L} + \mathcal{D})\mathbf{e}_j + B(\bar{\mathbf{u}}, \mathbf{e}_j) + B(\mathbf{e}_j, \bar{\mathbf{u}})] \cdot \mathbf{e}_i$ characterizes the quasilinear coupling between the mean state $\bar{\mathbf{u}}$ and the fluctuations $\mathbf{u}' = \sum_i Z_i \mathbf{e}_i$. The interactions between the fluctuation modes of different scales are summarized in the second term on the right hand side of (2.4) with the coupling coefficient $\gamma_{imn} = B(\mathbf{e}_m, \mathbf{e}_n) \cdot \mathbf{e}_i$. Alternatively, from the stochastic equation (2.4) we directly obtain the exact evolution equation of the covariance matrix R ,

$$\frac{dR}{dt} = A(\bar{\mathbf{u}})R + RA^*(\bar{\mathbf{u}}) + Q_F + Q_\sigma, \quad (2.5)$$

where $A(\bar{\mathbf{u}})$ is the same quasilinear operator from (2.4) containing instability represented by its positive eigenvalues, while Q_F is the nonlinear energy flux, which includes all the third moments $\langle Z_m Z_n Z_i \rangle$ feedback to balance the unstable linear growth. The term $Q_{\sigma,kl} = \sum_m (\mathbf{e}_k \cdot \sigma_m)(\sigma_m \cdot \mathbf{e}_l)$ is the contribution from the unresolved white noise forcing. Detailed expression for the equation (2.5) can be found in [2].

As a further remark on this mean-fluctuation formulation of the original system, we could use either the stochastic equation (2.4) or the equivalent statistical covariance equation (2.5) to model the uncertainty in each fluctuation mode \mathbf{e}_i . In fact, a data-driven statistical closure model combining (2.3) and (2.5) has been developed in [15] to effectively capture the leading-order statistical responses in mean and variance of homogeneous turbulent dynamics. On the other hand, the statistical-stochastic formulation using (2.3) and (2.4) enjoys the advantage of more flexibility to run ensemble forecasts for both uncertainty quantification and data assimilation, and compute statistical quantities other than the covariance. In addition, this statistical-stochastic model can naturally estimate inhomogeneous statistics and avoids the main issue with the purely statistical formulation in (2.3), (2.5) in preserving the positive-definite covariance estimation.

2.2. A generic statistical-stochastic closure model for mean and variance statistics

Now, we present the main idea in the efficient combined statistical-stochastic model to effectively capture the central statistical features. To effectively reduce the computational cost in finding the solution of high dimensional phase space, we introduce a proper low wavenumber truncation so that only the most important leading modes in the subset \mathcal{I} (for example, the subset \mathcal{I} can be taken to include the most energetic modes \mathbf{e}_i in the projection (2.2) with the largest mean energy and/or variances) are resolved, that is,

$$\mathbf{u}^\mathcal{I} = \bar{\mathbf{u}}^\mathcal{I} + \sum_{i \in \mathcal{I}} Z_i \mathbf{e}_i, \quad (2.6)$$

where $\bar{\mathbf{u}}^\mathcal{I} = \text{Pr}_\mathcal{I} \bar{\mathbf{u}} = \sum_{i \in \mathcal{I}} \bar{u}_i \mathbf{e}_i$ is a low-dimensional representation of the mean state and Z_i denotes the stochastic coefficients corresponding to the low-dimensional subset of the full state space $|\mathcal{I}| = K \ll N$. Inspecting the coupling terms in the

true dynamics (2.3) and (2.4), several difficulties will emerge for accurate modeling of the detailed coupling mechanisms in the constrained reduced-order representation (2.6). First, the high-order nonlinear coupling terms in the mean and fluctuation equations consist of a multiscale interaction of modes along the entire spectrum, while we only have access to a subset \mathcal{I} of the resolved mean and fluctuation modes. Second, inherent instability in the fluctuation modes Z_i due to the quasilinear coupling with $\bar{\mathbf{u}}$ through $A_{ij}(\bar{\mathbf{u}})$ poses challenge in constructing a stable dynamical model that can accurately predict the statistical responses to various perturbations. Third, the fluctuation components Z_i are stochastic processes coupled to the statistical mean equation, making direct modeling of the random trajectories very expensive.

2.2.1. Effective closure equations for the mean and fluctuations

First, we introduce the *reduced statistical mean equation* by projecting the full equation (2.3) to the resolved low-dimensional subspace

$$\frac{d\bar{\mathbf{u}}^{\mathcal{I}}}{dt} = (\mathcal{L} + \mathcal{D})\bar{\mathbf{u}}^{\mathcal{I}} + \text{Pr}_{\mathcal{I}}B(\bar{\mathbf{u}}^{\mathcal{I}}, \bar{\mathbf{u}}^{\mathcal{I}}) + \sum_{i,j \in \mathcal{I}} R_{ij} \text{Pr}_{\mathcal{I}}B(\mathbf{e}_i, \mathbf{e}_j) + \mathbf{F}^{\mathcal{I}} + \Theta^m. \quad (2.7)$$

In the above equation, only the projected dynamics in the reduced subspace are resolved. Here, the unresolved mean feedback that we denoted as Θ^m accounts for the residual (or truncation error) induced by the projection, namely the difference between the right hand side of the full model in (2.3) and the first three resolved components in the right-hand-side of (2.7). Various statistical closure strategies have been developed [2,21] using the parametric approximation of the unresolved structures. In this paper, we aim to design a machine learning scheme to identify this unresolved process directly from data.

Second, we consider the stochastic closure for the fluctuation equation (2.4). Again we concentrate on modes in the subset \mathcal{I} as in (2.6). Let $\mathbf{Z}^{\mathcal{I}} = \text{Pr}_{\mathcal{I}}\mathbf{Z} = \{Z_i\}_{i \in \mathcal{I}}$ be the vector of the resolved fluctuation modes. Similar to the statistical mean closure (2.7), we propose to construct a projected dynamical model for the resolved modes and learn the unresolved feedback with a properly designed data-driven model. Specifically, the resulting *reduced-order fluctuation equation* for the stochastic coefficients \mathbf{Z} becomes

$$\frac{d\mathbf{Z}^{\mathcal{I}}}{dt} = A(\bar{\mathbf{u}}^{\mathcal{I}})\mathbf{Z}^{\mathcal{I}} + \sigma \dot{\mathbf{W}}^{\mathcal{I}} + \Theta^v, \quad (2.8)$$

where Θ^v denotes the residual induced by the projection, namely the difference between the right hand side of the full model in (2.5) and the first two resolved components in the right-hand-side of (2.8).

Notice that the quasilinear coefficient $A_{ij}(\bar{\mathbf{u}}) = [(\mathcal{L} + \mathcal{D})\mathbf{e}_j + B(\bar{\mathbf{u}}, \mathbf{e}_j) + B(\mathbf{e}_j, \bar{\mathbf{u}})] \cdot \mathbf{e}_i$ for $i, j \in \mathcal{I}$ includes the mean-fluctuation interaction leading to inherent internal instability for turbulent dynamics (that is, with positive eigenvalues in $A(\bar{\mathbf{u}})$). While this linear instability is suppressed in the full model by the second term in (2.4), or equivalently by Q_F in (2.5), dynamical instability can occur in the reduced-order model, especially when the term Θ^v is numerically approximated with an arbitrary class of hypothesis models. To construct a stable approximate dynamical equation that can suppress instability induced by the marginally stable dynamics in (2.8), we introduce a more detailed parameterization for the unresolved process as

$$\Theta^v = -D\mathbf{Z}^{\mathcal{I}} + \Sigma\tilde{\mathbf{W}}, \quad (2.9)$$

where the white noises $\tilde{\mathbf{W}}$ are independent to $\dot{\mathbf{W}}$. Here, D and Σ are coefficient matrices to be approximated (see (2.12) below). In (2.9), the parameters D and Σ are introduced to play the equivalent role as the nonlinear flux term Q_F in the corresponding statistical equation (2.5): the parameter D is introduced to act as the equivalent damping suppressing the unstable positive growth rate, while the parameter Σ is to account for the energy source from the nonlinear exchange of energy. The effective decomposition in (2.9) generalizes the idea in the statistical closure model for nonlinear energy mechanism [15,2]. Here, effective parameters D, Σ will be constructed by fitting to the consistent covariance statistics. Applying Itô's lemma to $f(\mathbf{Z}^{\mathcal{I}}) = \frac{1}{2}\mathbf{Z}^{\mathcal{I}}(\mathbf{Z}^{\mathcal{I}})^*$ and taking expectation, we obtain,

$$\frac{dR^{\mathcal{I}}}{dt} = A(\bar{\mathbf{u}}^{\mathcal{I}})R^{\mathcal{I}} + R^{\mathcal{I}}A^*(\bar{\mathbf{u}}^{\mathcal{I}}) + Q_{\sigma}^{\mathcal{I}} + Q_F^{\mathcal{I}},$$

where $R^{\mathcal{I}} = \langle \mathbf{Z}^{\mathcal{I}}(\mathbf{Z}^{\mathcal{I}})^* \rangle$ is the covariance matrix of the resolved fluctuation modes, and $Q_F^{\mathcal{I}}$ is the nonlinear flux induced by the coupling from different stochastic coefficients and the truncation error. When Θ^v is parameterized by (2.9), the Itô's lemma for the flux parameterization is given by,

$$Q_F^{\mathcal{I}} = -DR^{\mathcal{I}} - R^{\mathcal{I}}D^* + \Sigma\Sigma^*. \quad (2.10)$$

While such a choice of parameterization is ideal, it is difficult to numerically find D and Σ that satisfy (2.10) as the covariance $R^{\mathcal{I}}$ is a time-dependent variable. To avoid such a practical issue, we consider fitting to the stationary (equilibrium) statistics,

$$Q_F^{\mathcal{I}} \approx -\tilde{D}R_{\text{eq}}^{\mathcal{I}} - R_{\text{eq}}^{\mathcal{I}}\tilde{D}^* + \tilde{\Sigma}\tilde{\Sigma}^* := \tilde{Q}, \quad (2.11)$$

where $R_{\text{eq}}^{\mathcal{I}}$ denotes the stationary covariance statistics of the resolved modes in \mathcal{I} that can be empirically estimated. In the above approximation, we introduced the notations \tilde{Q} , \tilde{D} , and $\tilde{\Sigma}$ to denote the approximate model (or parameterization). The above approximation in (2.11) is to first enforce equilibrium consistency in the unperturbed case and then assume the identity is still valid for small perturbations. The approximation in (2.11) becomes exact equality at the long-term limit guaranteeing the final convergence to the equilibrium covariance statistics of the unperturbed dynamics. With this approximation, we can decompose the approximate model into a positive and a negative definite component $\tilde{Q} = (\tilde{Q})^+ - (\tilde{Q})^-$. Then the effective damping and noise matrix can be approximated accordingly by fitting the negative and positive-definite components, respectively, as

$$D \approx \tilde{D} := \frac{1}{2}(\tilde{Q})^-(R_{\text{eq}}^{\mathcal{I}})^{-1}, \quad \Sigma\Sigma^* \approx \tilde{\Sigma}(\tilde{\Sigma})^* := (\tilde{Q})^+. \quad (2.12)$$

The above approximation is based on the equivalent roles of the effective damping and noise as discussed above. We note that the equilibrium covariance matrix $R_{\text{eq}}^{\mathcal{I}}$ is nonsingular, and a well-conditioned matrix when the variance is not too small. This motivates the choice of modes in \mathcal{I} with large variance energy spectra. When the covariance matrix is diagonally dominant, we found that a further simplification can be made to avoid the computational complexity of realizing the matrix factorization above. The general framework above, (2.7)-(2.9) with the approximate coefficients \tilde{D} and $\tilde{\Sigma}$ in (2.12), will be implemented on an explicit model next in Section 3 as a concrete example of the idea.

2.2.2. An ensemble-based statistical and stochastic model

Finally, we need to couple the statistical mean equation (2.7) and the stochastic equation (2.8) for the fluctuation modes. The resolved mean state $\bar{\mathbf{u}}^{\mathcal{I}}$ enters the fluctuation equation (2.8) through the quasilinear term $A(\bar{\mathbf{u}}^{\mathcal{I}})$. Especially, it induces positive growth rate among the unstable modes. Inversely, the statistical mean equation (2.7) depends on the covariance feedback from the resolved modes $R^{\mathcal{I}}$, which will be empirically estimated by a Monte-Carlo average over an ensemble of solutions of the fluctuation equation (2.8). Denoting the ensemble solutions of the fluctuation equation (2.8) as $\{\mathbf{Z}^{M,(i)}\}_{i=1,\dots,M}$ the second-order moment can be estimated empirically as,

$$R^{\mathcal{I}} = \langle \mathbf{Z}^{\mathcal{I}}(\mathbf{Z}^{\mathcal{I}})^* \rangle \approx R^M := \frac{1}{M-1} \sum_{i=1}^M \mathbf{Z}^{M,(i)}(\mathbf{Z}^{M,(i)})^*. \quad (2.13)$$

With this empirical estimation, we have the complete *general reduced-order statistical-stochastic closure model* as,

$$\begin{aligned} \frac{d\bar{\mathbf{u}}^M}{dt} &= (\mathcal{L} + \mathcal{D})\bar{\mathbf{u}}^M + \text{Pr}_{\mathcal{I}B}(\bar{\mathbf{u}}^M, \bar{\mathbf{u}}^M) + \sum_{i,j \in \mathcal{I}} R_{ij}^M \text{Pr}_{\mathcal{I}B}(\mathbf{e}_i, \mathbf{e}_j) + \mathbf{F}^{\mathcal{I}} + \Theta^m \\ \frac{d\mathbf{Z}^M}{dt} &= A(\bar{\mathbf{u}}^M)\mathbf{Z}^M + \sigma\dot{\mathbf{W}}^{\mathcal{I}} - D^M\mathbf{Z}^M + \Sigma^M\dot{\tilde{\mathbf{W}}}^{\mathcal{I}}, \end{aligned} \quad (2.14)$$

where Θ^m is the nonlinear mean feedback defined in (2.7). The coefficients D^M and Σ^M will be parameterized following the approximation in (2.12),

$$D^M := \frac{1}{2}(Q^M)^-(R_{\text{eq}}^{\mathcal{I}})^{-1}, \quad (\Sigma^M)(\Sigma^M)^* := (Q^M)^+, \quad (2.15)$$

where the model $Q^M = (Q^M)^+ - (Q^M)^-$ approximates $Q_F^{\mathcal{I}}$ is defined to follow (2.11),

$$Q^M := -D^M R^{\mathcal{I}} - R^{\mathcal{I}}(D^M)^* + \Sigma^M(\Sigma^M)^*.$$

The closure model (2.14) with parameterization (2.15) provides a new formulation for the leading order statistics by combining the statistical mean equation with the stochastic fluctuation dynamics. Subsequently, the data-driven closure is adopted by fitting the standard Long Short Term Memory (LSTM) models to learn the unresolved terms Θ^m and Q^M . In Section 3, we will provide specific examples of (2.14) induced by the moment closure of homogeneous and inhomogeneous turbulence dynamics.

2.2.3. Empirical loss functions based on an information metric

It is important to design a suitable criterion for the loss function that reflects the appropriate quantity of interests, which are the response mean and variance statistics rather than the individual trajectory of the stochastic fluctuations [22,23]. In fact, we demonstrate numerically in Appendix A that fitting the stochastic components of (2.14) directly to the pathwise trajectory of (2.8) given the true mean $\bar{\mathbf{u}}^{\mathcal{I}}$ often leads to overfitting, thus does not provide accurate statistical prediction when testing on new inputs. In this context new inputs correspond to new initial values and forcing perturbations as in the numerical tests. However, fitting to the mean and variance statistical responses corresponding to the same trajectory

solutions (that lead to an overfitted model when trajectory is fitted) produces a closure model with accurate statistical predictions on new inputs.

A natural way to fit the statistics is to consider the information distance as it allows one to measure the errors between the probability distributions achieved from the empirical average of the ensemble simulations. Particularly, we consider the following practical metric based on Kullback-Leibler (KL) divergence [24,25] of two empirical measures induced by the response mean and variance statistics of the underlying dynamics in (2.1) and the reduced-order model in (2.14), respectively. The KL divergence offers a balanced calibration between the statistical errors in the mean and variance. Let $\delta\bar{\mathbf{u}} := \bar{\mathbf{u}}_\delta - \bar{\mathbf{u}}_{\text{eq}}$ and $\delta R := R_\delta - R_{\text{eq}}$ be the response mean and covariance statistics of the underlying dynamics in (2.1) subject to additional damping and forcing of small $0 < \delta \ll 1$ perturbation amplitudes in addition to the reference damping and forcing parameters. Here, the response statistics are defined as the differences between the time-dependent statistics subject to the additional damping and forces and the equilibrium statistics, $\bar{\mathbf{u}}_{\text{eq}}$ and R_{eq} , corresponding to the solutions with reference damping and forcing that can be empirically estimated offline by an ensemble simulation of the underlying system in (2.1). Analogously, we also define $\delta\bar{\mathbf{u}}^M := \bar{\mathbf{u}}_\delta^M - \bar{\mathbf{u}}_{\text{eq}}$ and $\delta R^M := R_\delta^M - R_{\text{eq}}$ as the corresponding statistical responses of the reduced-order model in (2.14), where \mathcal{D} and \mathbf{F}^T are perturbed by additional damping and forcing of amplitude δ . Assuming that the perturbed distributions vary smoothly under parameter δ and denoting $\text{diag}(R)$ as the diagonal matrix whose diagonal entries are R_k , the KL-divergence between Gaussian measures $\pi_\delta = \mathcal{N}(\bar{\mathbf{u}}_{\text{eq}} + \delta\bar{\mathbf{u}}, \text{diag}(R_{\text{eq}} + \delta R))$ and $\pi_\delta^M = \mathcal{N}(\bar{\mathbf{u}}_{\text{eq}} + \delta\bar{\mathbf{u}}^M, \text{diag}(R_{\text{eq}} + \delta R^M))$ can be written as,

$$\text{KL}(\pi_\delta, \pi_\delta^M) = \text{KL}(\pi_{\text{eq}}, \pi_{\text{eq}}^M) + \frac{1}{2} \sum_{k \in \mathcal{I}} R_{\text{eq},k}^{-1} (\delta\bar{\mathbf{u}}_k - \delta\bar{\mathbf{u}}_k^M)^2 + \frac{1}{4} \sum_{k \in \mathcal{I}} R_{\text{eq},k}^{-2} (\delta R_k - \delta R_k^M)^2 + O(\delta^3), \quad (2.16)$$

where $\delta\bar{\mathbf{u}}_k$ and $\delta\bar{\mathbf{u}}_k^M$ denote the k -th component of the mean responses $\delta\bar{\mathbf{u}}$ and $\delta\bar{\mathbf{u}}^M$, respectively, and $R_{\text{eq},k}$, δR_k , and δR_k^M denote the k -th diagonal component of the matrices R_{eq} , δR , and δR^M , respectively. The choice of fitting only the diagonal entries of the response statistics is reasonable when the covariance is diagonally dominant with small non-diagonal entries. In practice, an accurate proxy of the non-diagonal entries of R_{eq} may not be available since such accurate training data may require a simulation with a very large ensemble size especially if the full model is high-dimensional. So, fitting to inaccurate non-diagonal entries in R_{eq} may introduce additional errors. In Section 4.3, we show that fitting to only the diagonal entries of R_{eq} with the following loss function still produces an accurate estimation for the variance response.

Since $\text{KL}(\pi_{\text{eq}}, \pi_{\text{eq}}^M) = 0$ by equilibrium consistency, we propose the following loss function,

$$\text{L}(\theta) = \sum_{j=1}^T \left(\frac{1}{2} \sum_{k \in \mathcal{I}} R_{\text{eq},k}^{-1} (\delta\bar{\mathbf{u}}_k(t_j) - \delta\bar{\mathbf{u}}_k^M(t_j; \theta))^2 + \frac{1}{4} \sum_{k \in \mathcal{I}} R_{\text{eq},k}^{-2} (\delta R_k(t_j) - \delta R_k^M(t_j; \theta))^2 \right), \quad (2.17)$$

which measures the signal and dispersion contribution at discrete time indices $\{t_j = j\Delta t\}_{j=0,\dots,T}$ to the discrepancy of the response mean and variance statistics between the underlying dynamics and the reduced-order model to be fitted. We specify the loss function to depend on θ , denoting parameters in the class of models used to approximate, Θ^m and Q^M . When a neural-network model is used, then θ corresponds to the neural-network model parameters.

From the classical supervised learning perspective [26], the loss function (2.17) for the regression problem compares the real-valued labels,

$$\mathbf{y} := (\delta\bar{\mathbf{u}}_k(t_j), \delta R_k(t_j) : \forall j \in \{1, \dots, T\}, k \in \mathcal{I}), \quad (2.18)$$

the statistical responses of the underlying dynamics in (2.1), to the predicted labels,

$$\mathbf{y}^M := (\delta\bar{\mathbf{u}}_k^M(t_j), \delta R_k^M(t_j) : \forall j \in \{1, \dots, T\}, k \in \mathcal{I}) \quad (2.19)$$

the statistical response induced by the reduced-order model in (2.14). For convenience of the discussion, let us define the operator \mathcal{M} as $\mathbf{y}^M = \mathcal{M}(\mathbf{x})$, where \mathbf{x} denotes the initial conditions of (2.14) (which will include the appropriate inputs for Θ^m and Q^M). We will specify the input variable \mathbf{x} in Section 3.4 corresponding to a specific reduced-order model accounting for the inputs of Θ^m and Q^M . With this notation, we write the loss function $\text{L}(\theta) := \text{L}(\theta, \mathbf{y}, \mathbf{y}^M)$ to emphasize its dependence on the label (2.18) and predicted label (2.19). The supervised machine learning training corresponds to minimizing the following empirical risk function,

$$R_n(\theta) := \frac{1}{n} \sum_{i=1}^n \text{L}(\theta, \mathbf{y}_i, \mathbf{y}_i^M), \quad (2.20)$$

which is an empirical average of the loss function over n training data $(\mathbf{x}_i, \mathbf{y}_i)_{i=1,\dots,n}$. Here, we should emphasize that $\mathbf{y}_i^M = \mathcal{M}(\mathbf{x}_i)$ is the predicted response statistics (real-valued label) corresponds to the input \mathbf{x}_i . In Section 3.4, we will specify the input variable \mathbf{x} of a specific example of (2.14) and provide a pseudo-algorithm to evaluate the operator \mathcal{M} . In Section 4.1, we will provide more detailed discussion on the generation of training data.

3. Machine learning strategies for modeling unresolved structures with strong instability

To illustrate the key idea in the data-driven modeling framework to capture leading statistics, we display the detailed construction of the general model described in Section 2 in a step-by-step fashion on the L-96 system as one representative example. First, we start with a simpler case only including homogeneous statistics. Then, the inhomogeneous model is developed by adding additional structures subject to the inhomogeneous damping and forcing effects. Especially in modeling systems with chaotic dynamics, a crucial issue is to construct stable approximate dynamical equations that can avoid the inherent instability in the system.

3.1. Lorenz '96 system as a representative test model

In this section, we realize the closure modeling approach described in Section 2 on a simple prototypical example that exhibits a range of statistical features that arose in chaotic dynamics. Particularly, we consider the 40-dimensional Lorenz '96 (L-96) system [18] of state variables $\mathbf{u} = (u_1, u_2, \dots, u_N)^\top$ with general spatially inhomogeneous damping and forcing,

$$\frac{du_j}{dt} = (u_{j+1} - u_{j-2})u_{j-1} - d_j(t)u_j + f_j(t), \quad j = 1, \dots, N = 40. \quad (3.1)$$

This ODE system is defined with a periodic boundary condition mimicking geophysical weather dynamics on a midlatitude belt of roughly 32,000 km. The choice of $N = 40$ grid points corresponds to a spatial discretization of about 800 km which is the length scale of Rossby radius observed in nature. Various statistical features that reflect the real observations in nature can be generated by the simple model (3.1). Especially, inhomogeneous processes are introduced by the spatially varying damping and forcing effects d_j and f_j as a generalization to the standard L-96 model configuration with uniform damping $d_j \equiv d_{eq} = 1$ and forcing $f_j \equiv F_{eq} = 8$ (some explicit forms of the inhomogeneous forcing and damping that we use in our numerical are illustrated in Fig. 4.2 in Section 4). This will lead to more complicated inhomogeneous statistics in the mean modes as well as the non-zero off-diagonal covariances. To compare with the abstract form (2.1), we can write the linear and quadratic operators for L-96 system as

$$\mathcal{L} + \mathcal{D} = \text{diag}(-d_1, \dots, -d_N), \quad B(\mathbf{u}, \mathbf{v}) = \{u_{i-1}^*(v_{i+1} - v_{i-2})\}_{i=1}^N$$

and project the state variables onto the Fourier basis $\mathbf{e}_k = \{e^{i2\pi k \frac{j}{N}}\}_{j=1}^N$ considering the periodic boundary condition.

We aim to deduce moment closure equations for (3.1) that include inhomogeneous structures in the statistical mean and stochastic fluctuation modes. In order to achieve this, we project the general inhomogeneous forcing and damping as well as the model state onto each spectral mode such that

$$\begin{aligned} f_j &= \hat{f}_0 + \sum_{k \neq 0} \hat{f}_k e^{i2\pi k \frac{j}{N}}, & d_j &= \hat{d}_0 + \sum_{k \neq 0} \hat{d}_k e^{i2\pi k \frac{j}{N}}, \\ u_j(t, \omega) &= \bar{u}_j(t) + \sum_{|k| \leq N/2} Z_k(t; \omega) e^{i2\pi k \frac{j}{N}}. \end{aligned} \quad (3.2)$$

Above, we denote the homogeneous components of the forcing and damping as \hat{f}_0 and \hat{d}_0 , respectively, corresponding to the Fourier mode $k = 0$. Notice that in the decomposition in (3.2), the state variable $u_j = \bar{u}_j + u'_j$ is decomposed into the statistical mean \bar{u}_j and the fluctuations u'_j , which is then written as a linear combination of the fluctuation modes Z_k in Fourier coordinates. We will further decompose the mean state into the contributions of the homogeneous and inhomogeneous terms as,

$$\bar{u}_j(t) = \hat{u}_0(t) + \sum_{|k| \leq N/2} \hat{u}_k(t) e^{i2\pi k \frac{j}{N}},$$

where $\hat{u}_k(t)$ corresponds to the k -th Fourier mode of the mean \bar{u}_j . If $\bar{u}_j = \bar{u}$ is spatially homogeneous, then the zeroth mode $\hat{u}_0(t)$ is precisely the homogeneous mean $\bar{u}(t)$. This observation implies that the non-zero Fourier modes characterize the inhomogeneity of the dynamical processes.

3.1.1. Statistical mean dynamics

Projecting Equation (3.1) to different spectral modes, we obtain the statistical mean equation for the homogeneous and inhomogeneous components

$$\frac{d\hat{u}_0}{dt} = -\bar{d}\hat{u}_0 - \sum_{k \neq 0} d_{0,k} \hat{u}_k + \hat{f}_0 + \sum_{|k| \leq N/2} (|\hat{u}_k|^2 + |Z_k|^2) \gamma_k, \quad (3.3a)$$

$$\frac{d\hat{u}_k}{dt} = - \sum_{|m| \leq N/2} d_{k,m} \hat{u}_m + \hat{f}_k + \sum_{|m| \leq N/2} (\hat{u}_m \hat{u}_{k-m} + \langle Z_m Z_{k-m} \rangle) \gamma_m^* e^{-i2\pi \frac{k}{N}}, \quad (3.3b)$$

with the uniform damping rate $\bar{d} = \frac{1}{N} \sum_j d_j$, and the damping coefficients for each inhomogeneous mode $d_{k,m} = \frac{1}{N} \sum_j d_j e^{i2\pi(m-k)\frac{j}{N}} = \hat{d}_{k-m}$. The nonlinear coupling between different scales is connected by the coefficient $\gamma_k = e^{-i\frac{4\pi k}{N}} - e^{i\frac{2\pi k}{N}}$. Notice that the first equation (3.3a) only contains homogeneous dynamics (no cross-correlation between different wavenumbers k). In addition to the homogeneous mean mode \hat{u}_0 , we also need to compute the inhomogeneous mean modes \hat{u}_k if inhomogeneous forcing and damping effects are included.

3.1.2. Stochastic coefficient dynamics

The dynamical equation for the stochastic coefficients can be attained by subtracting the mean dynamics (3.3) from the original equation (3.1) and subsequently projecting it to each spectral mode. Following these steps, we have the governing equation for the stochastic coefficients Z_k as,

$$\frac{dZ_k}{dt} = - \sum_{|m| \leq N/2} d_{k,m} Z_m + \sum_{|m| \leq N/2} \mu_{k,m} \hat{u}_{k-m} Z_m + \sum_{|m| \leq N/2} (Z_m Z_{k-m} - \langle Z_m Z_{k-m} \rangle) \gamma_m^* e^{i2\pi \frac{-k}{N}}, \quad (3.4)$$

with the coupling coefficient $\mu_{k,m} = e^{i2\pi \frac{k-2m}{N}} + e^{i2\pi \frac{2m-k}{N}} - e^{i2\pi \frac{m-2k}{N}} - e^{i2\pi \frac{-k-m}{N}}$. On the right-hand-side of (3.4), the first term denotes linear damping, the second term characterizes the coupling through the homogeneous and inhomogeneous means, and the third term characterizes the nonlinear coupling between the fluctuation modes between different scales.

For a complete investigation of the energy transferring mechanism subject to linear and nonlinear interactions, we can also derive the corresponding dynamical equation for the covariance $R_{kl} = \langle Z_k Z_l^* \rangle$ according to (2.5)

$$\begin{aligned} \frac{dR_{kl}}{dt} = & -2\bar{d}R_{km} - (\gamma_{-k} + \gamma_{-k}^*) \hat{u}_0 R_{kl} \\ & - \sum_{m \neq k} (d_{k,m} R_{ml} + d_{l,m}^* R_{km}) + \sum_{m \neq k} (\mu_{k,m} \hat{u}_{k-m} R_{ml} + \mu_{l,m}^* \hat{u}_{l-m} R_{km}) \\ & + \sum_{m \neq 0} \langle Z_m Z_{k-m} Z_l^* \rangle \gamma_m^* e^{i2\pi \frac{-k}{N}} + \langle Z_m Z_{l-m}^* Z_k \rangle \gamma_m e^{i2\pi \frac{l}{N}}. \end{aligned} \quad (3.5)$$

The homogeneous effects due to damping and mean interaction are summarized in the first row of (3.5). The inhomogeneous damping and mean interactions are shown in the second row of (3.5). Higher-order feedbacks from the third-order moments with non-Gaussian statistics among all the spectral modes enter the covariance equation in the third row of (3.5).

In the following, we describe the step-by-step construction of the data-driven reduced-order model on the L-96 system as a canonical example, following the same reduced modeling approach that is stated for the more abstract model in (2.3) and (2.4) on the particular case (3.3) and (3.4).

3.2. Hybrid statistical-stochastic model for homogeneous statistics

We start with the simple model set up with homogeneous damping and forcing, $d_j := \gamma$, $f_j := f$, in (3.2) together with homogeneous initial perturbations $u_{0,j} := u_0$. In this case, the mean and fluctuation equations in (3.3) and (3.4) can be simplified as

$$\frac{d\bar{u}}{dt} = -\gamma \bar{u} + \sum_k R_k \gamma_k + f, \quad (3.6a)$$

$$\frac{dZ_k}{dt} = -(\gamma + \gamma_k \bar{u}) Z_k + \sum_{m \neq 0} Z_m Z_{k-m} \gamma_m^* e^{i2\pi \frac{-k}{N}}, \quad (3.6b)$$

with $\gamma_k = e^{-i\frac{4\pi k}{N}} - e^{i\frac{2\pi k}{N}}$, $\bar{u} = \hat{u}_0$, $\bar{d} = \gamma$, and $R_k := R_{kk} = \langle Z_k Z_k^* \rangle$ denotes the variance of the stochastic coefficient $\hat{Z}_k(t; \omega)$. Under the homogeneous statistics, the statistical mean state becomes a scalar and the covariance matrix becomes diagonal, that is,

$$\bar{u}_j = \bar{u} = \hat{u}_0, \quad \hat{u}_k := 0, \quad k \neq 0, \quad \text{and} \quad R_{kl} = R_k \delta_{kl}.$$

Thus we do not need to consider the inhomogeneous mean equation (3.3b) involving \hat{u}_k and the cross-correlations between different spectral mode $\langle Z_k Z_l^* \rangle$, $k \neq l$. On the other hand, nonlinear dynamics and non-Gaussian statistics still play a central role due to the strongly coupled feedbacks in equations (3.6). Different scales are mixed in the feedbacks with summations over all the wavenumbers. In particular, the system may contain strong internal instability through the mean-fluctuation

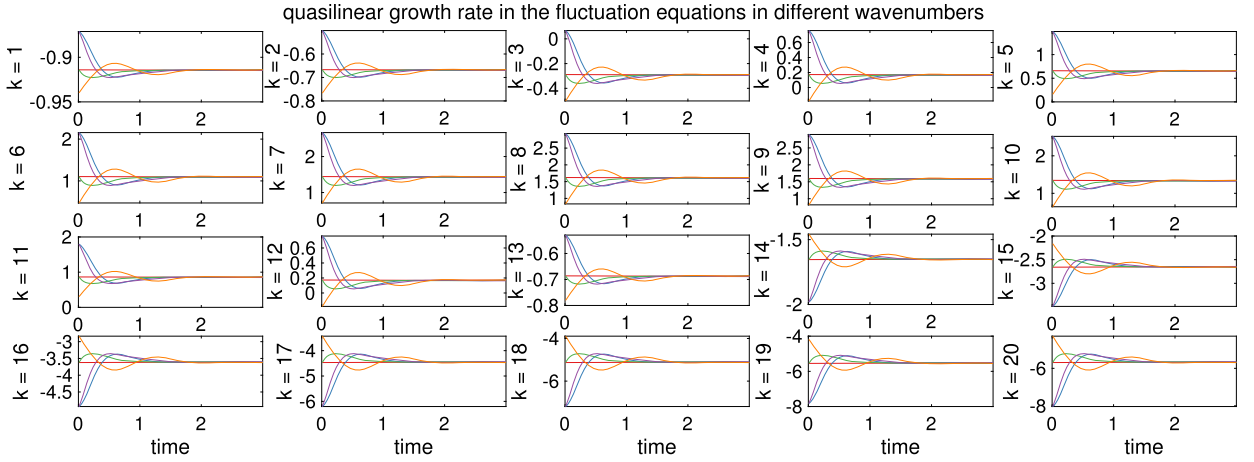


Fig. 3.1. The time evolution of the quasilinear growth rate computed for each spectral mode k of the L-96 model. Different lines are subject to the initial perturbations described in Section 4.1.

interactions. For example, in (3.3b) strong positive growth rate will occur when $\hat{u}_0 = \bar{u} > 0$ for modes with $\Re \gamma_k < 0$. To illustrate this, we plot in Fig. 3.1 the quasilinear growth rate $-(\gamma + \gamma_k \bar{u})$ of each spectral mode in the L-96 model, subject to different initial perturbations (that we will describe in Section 4.1). Positive value implies instability of the mode. We notice that the instabilities occur on a wide range of modes depending on the initial value perturbations and their intermittent occurrences create a practical challenge for learning a dynamically stable model and accurate prediction of model statistics.

3.2.1. Direct modeling of the unresolved feedbacks

Using the hybrid statistical-stochastic model (3.6), it requires the computation of the statistical expectation $R_k = \langle Z_k Z_k^* \rangle$ for the mean equation (3.6a) from the solution of the stochastic equation (3.6b). In practice, this can be achieved by ensemble simulation of the stochastic coefficients Z_i . With the ensemble approximation in (2.13), the analog of the closure model in (2.14) in this example for the homogeneous case is given by:

$$\begin{aligned} \frac{d\bar{u}^M}{dt} &= -\gamma \bar{u}^M + \sum_{k \in \mathcal{I}} \left(\frac{1}{M-1} \sum_{i=1}^M Z_k^{M,(i)} (Z_k^{M,(i)})^* \right) \gamma_k + \Theta^m + f, \\ \frac{dZ_k^{M,(i)}}{dt} &= -(\gamma + \gamma_k \bar{u}^M) Z_k^{M,(i)} + \Theta_k^v, \quad k \in \mathcal{I}, i = 1, \dots, M. \end{aligned} \quad (3.7)$$

In (3.7), $\mathbf{Z}^{M,(i)} = (Z_k^{M,(i)}), k \in \mathcal{I}, i = 1, \dots, M$ is each (independent) ensemble member in the ensemble simulation of the fluctuation equations, \mathcal{I} is the set containing the resolved fluctuation modes corresponding to the largest variances. Its statistical feedback in the mean equation is approximated by the empirical ensemble average in (2.13).

In the above model, we consider \mathcal{I} to include modes with largest variances but still only accounting for a small portion of the total energy (that is, $k = 7 - 13$, see Fig. 3.2(a)) so that the leading order dynamics can be replicated. The resolved fluctuations provide the explicit variance feedback and nonlinear coupling in the mean and fluctuation equations, whereas the data-driven component behaves as a higher-order correction. Comparing the reduced-order model (3.7) with the full exact model (3.6), the process Θ^m models the variance feedback to the mean dynamics among all the unresolved small-scale modes $k \in \{-N/2 + 1, \dots, N/2\} \setminus \mathcal{I}$. In addition, Θ^m also accounts for the approximation error in the resolved variances from the empirical ensemble average from a finite sample size, M . In the stochastic equations for Z_k^M , the model Θ_k^v approximates the total contribution from the nonlinear coupling among all the fluctuation modes.

3.2.2. Stable dynamical equations with effective damping and noise

First, to overcome the instabilities occurring on unstable modes, where $\gamma + \gamma_k \bar{u} < 0$ (see Fig. 3.1), we follow the strategy discussed in Section 2.2.1. Next, instead of an extensive pointwise calibration of each stochastic trajectory of $Z_k^{(i)}$, we propose to only measure the error in the ensemble statistics discussed in Section 2.2.2, so that the high computational cost in training is effectively avoided while the statistical accuracy is also maintained. Notice that for the homogeneous case, the dynamical equation for the covariance matrix R in (2.5) can be simplified. Particularly, R only has nontrivial diagonal components R_k that satisfied,

$$\frac{dR_k}{dt} = -2\gamma R_k - (\gamma_k + \gamma_k^*) \bar{u} R_k + Q_{F,k}, \quad (3.8)$$

where Q_F is a diagonal matrix for the high-order statistical nonlinear fluxes.

In this homogeneous case, the proposed statistical-stochastic model in (2.14) is simplified to,

$$\begin{aligned} \frac{d\bar{u}^M}{dt} &= -\gamma \bar{u}^M + \sum_{k \in \mathcal{I}} \left(\frac{1}{M-1} \sum_{i=1}^M Z_k^{M,(i)} (Z_k^{M,(i)})^* \right) \gamma_k + \Theta^m + f, \\ \frac{dZ_k^{M,(i)}}{dt} &= -\left(\gamma + \gamma_k \bar{u}^M \right) Z_k^{M,(i)} - D_k^M Z_k^{M,(i)} + \Sigma_k^M \dot{W}_k^{(i)}, \quad k \in \mathcal{I}, i = 1, \dots, M, \end{aligned} \quad (3.9)$$

where D_k and Σ_k are parameterized as in (2.15).

$$\begin{aligned} D_k^M &= -\frac{\min\{Q_k^M, 0\}}{2R_{\text{eq},k}}, \\ \Sigma_k^M &= \sqrt{\max\{Q_k^M, 0\}}. \end{aligned} \quad (3.10)$$

We should point out that the component-wise decomposition (3.10) is only possible since the statistics is homogeneous, and thus, avoiding an expensive matrix decomposition to identify the positive and negative definite components, $(Q_M)^+$ and $(Q_M)^-$, respectively, that satisfy $Q_M := (Q_M)^+ - (Q_M)^-$ for non-diagonal matrix Q_M .

In section 3.4, we will specify the class of machine learning models to identify Θ_k^m and Q_k^M in terms of time delay embedding of these variables, respectively, in addition to the time delay embedding of the mean and variance. Before discussing this, we consider a slight modification to the closure model above to accommodate for inhomogeneous statistics in the next section.

3.3. The statistical-stochastic model for inhomogeneous statistics

Next, we consider to predict the mean and variance responses under a more general case with inhomogeneous statistics introduced by spatially inhomogeneous forcing and initial perturbations. For this case, we have the additional observations for the inhomogeneous equations (3.3) and (3.4):

- The homogeneous mean state $\bar{u} = \hat{u}_0$ is subject to feedbacks from not only the variances R_k (as in the homogeneous case), but also the energy in the inhomogeneous mean states $|\hat{u}_k|^2$;
- The inhomogeneous mean modes \hat{u}_k are subject to the cross interactions between the mean states $\hat{u}_0 \hat{u}_k$ and the cross-covariances $\langle Z_k Z_0 \rangle$;
- The stochastic coefficients Z_k are subject to the cross interactions between the mean and the fluctuation modes as well as the nonlinear coupling between different wavenumber modes.

In the inhomogeneous case, it is expensive to resolve the cross interaction terms between the entire spectrum. In Fig. 3.2, we plot the typical spectra for the mean and variance under several different inhomogeneous perturbations. First, we should point out that the homogeneous mean \hat{u}_0 and variance R_k for $7 \leq k \leq 13$ are still dominant under various inhomogeneous forces. While including these non-trivial modes in \mathcal{I} is sufficient for homogeneous modeling, excluding other modes (such as $1 \leq k \leq 6$) whose mean energy spectra are significantly increased under inhomogeneous forces will produce a poor statistical recovery. In general, the reduced model should include modes that are significantly excited by the inhomogeneous perturbations, which makes the modeling choice slightly more complicated than that of the homogeneous case. From Fig. 3.2(b), we also notice that the covariance matrices of the perturbed dynamics are diagonally banded with the detailed structure depending crucially on the perturbations. While the non-diagonal components are non-negligible, they are much smaller compared to the diagonal components. This scale separation poses an additional computational challenge for an accurate estimation of the non-diagonal covariance components, which is crucial for stable modeling of the inhomogeneous components as shown in (3.3b).

Given these statistical features, we consider a diagonal closure model for the feedback from the higher-order moments, measuring only the statistics in the inhomogeneous mean and diagonal variances. Particularly, we consider the dynamical closure equations for the homogeneous mean \hat{u}_0 , inhomogeneous mean \hat{u}_k , and the fluctuation modes Z_k corresponding to the resolved subset $k \in \mathcal{I}$ as follows:

$$\begin{aligned} \frac{d\hat{u}_0^M}{dt} &= -\sum_{j \in \mathcal{I}} d_{0,j} \hat{u}_j^M + \bar{f} + \sum_{k \in \mathcal{I}} \left(|\hat{u}_k^M|^2 + R_k^M \right) \gamma_k + \Theta_0^m, \\ \frac{d\hat{u}_k^M}{dt} &= -\sum_{\ell \in \mathcal{I}} d_{k,\ell} \hat{u}_\ell^M + \hat{f}_k - \gamma_k \hat{u}_0^M \hat{u}_k^M + \Theta_k^m, \\ \frac{dZ_k^{M,(i)}}{dt} &= -\sum_{\ell \in \mathcal{I}} d_{k,\ell} Z_\ell^{M,(i)} - \gamma_k \hat{u}_0^M Z_k^{M,(i)} - D_k^M Z_k^{M,(i)} + \Sigma_k^M \dot{W}_k^{(i)}, \end{aligned} \quad (3.11)$$

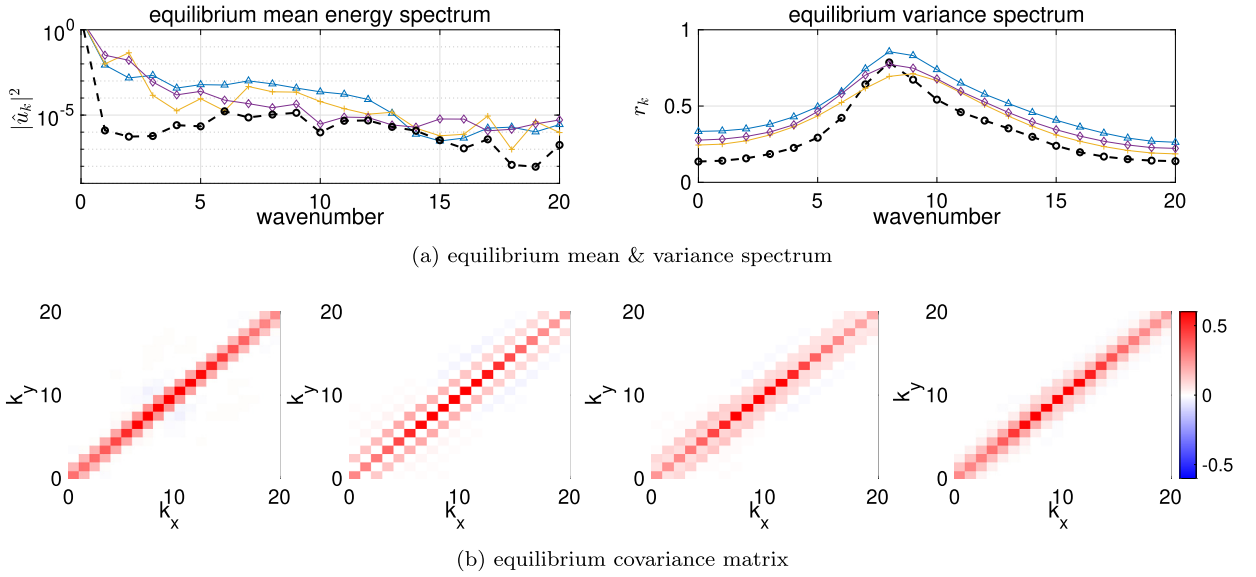


Fig. 3.2. Equilibrium statistics of the L-96 model with inhomogeneous perturbations. First row: equilibrium spectra of energy in the mean and variance covariance (unperturbed homogeneous case in dashed line). Second row: the equilibrium covariance under several inhomogeneous forcing and damping effects. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

which is an example of (2.14). Here $d_{k,m}$ are defined as in (3.3a)-(3.3b). The variance feedback R_k^M in the mean equation is defined as the diagonal component of (2.13) attained with M samples. Also, the feedbacks from the non-diagonal covariance entries are not computed explicitly in the second equation in (3.11) due to their relatively small amplitudes. Following the homogeneous case, we introduce Θ_0^m, Θ_k^m to account for the truncation error and the feedbacks in the homogeneous and inhomogeneous mean state from unresolved small-scale processes. The stabilizing decomposition for effective damping D_k^M and noise Σ_k^M is constructed exactly as in (3.10), except that the model Q_k^M in (3.10) is fitted to the higher-order statistical flux $Q_{F,k}$ induced by the inhomogeneous dynamics for the variance component,

$$\frac{dR_k}{dt} = - \sum_{m \in \mathcal{I}} \left(d_{k,m} R_{mk} + d_{k,m}^* R_{km} \right) - (\gamma_k + \gamma_k^*) \bar{u} R_k + Q_{F,k}, \quad (3.12)$$

replacing (3.8) of the homogeneous case. In this diagonal closure modeling, we should point out that the element-wise decomposition (3.10) can still be performed and thus avoiding the matrix decomposition $Q_M := (Q_M)^+ - (Q_M)^-$ for non-diagonal case.

In this inhomogeneous model, we make a final remark that the proposed closure Q_k^M is to account for modeling error induced by: i) the higher-order moment feedback to the variance; ii) the variances of the unresolved modes $k \in \mathcal{I}^c$; and iii) the neglected cross-covariances $R_{k\ell}, k \neq \ell$ that are not explicitly computed in the model. In Section 4, we will empirically show that the proposed diagonal reduced-order model, which is numerically efficient, does not introduce significant error to the prediction of the mean and variance statistics.

An extension from the previous work. In the previous work where only modeling homogeneous dynamics is considered [15], in addition to not having non-homogeneous mean modes, $\{\hat{u}_k\}_{k \in \mathcal{I}, k \neq 0}$, we also employed a closure to the diagonal model for R_k in (3.8). In that work, we employed the same fitting as in (3.10) to $Q_{F,k} \approx Q_k^M$, and considered the following closure model for the variance dynamics,

$$\frac{dR_k^M}{dt} = -2\gamma R_k^M - (\gamma_k + \gamma_k^*) \bar{u} R_k^M - 2D_k^M R_k^M + (\Sigma_k^M)^2. \quad (3.13)$$

Compare to the dynamics of $Z_k^{M,(i)}$ in (3.11), the diagonal model in (3.13) does not capture the feedback from different modes through $\sum_{\ell \in \mathcal{I}} d_{k,\ell} Z_\ell^{M,(i)}$ in (3.11). The current approach with (3.11) is analogous to extending the previous approach in (3.13) to include the non-diagonal second-order statistical interactions. Indeed, one can see that in the limit of large ensemble size, the covariance dynamics of (3.11) is identical to that achieved by imposing the diagonal closure in (3.10) to approximate the third-order moments in (3.12),

$$\begin{aligned}
\frac{dR_{kl}}{dt} = & -2\bar{d}R_{km} - (\gamma_{-k} + \gamma_{-k}^*) \hat{u}_0 R_{kl} \\
& - \sum_{m \neq k} \left(d_{k,m} R_{ml} + d_{l,m}^* R_{km} \right) + \sum_{m \neq k} \left(\mu_{k,m} \hat{u}_{k-m} R_{ml} + \mu_{l,m}^* \hat{u}_{l-m}^* R_{km} \right) \\
& + (-2D_k^M R_k^M + (\Sigma_k^M)^2) \delta_{kl}.
\end{aligned} \tag{3.14}$$

While one can use this closure model for R_{kl} , we choose to consider the fluctuation dynamics for Z_k^M for the following reasons. First, the closure model for the fluctuation components allows one to estimate statistics other than covariance $R^{\mathcal{I}}$, whenever improved parameterization of the unresolved feedback from higher-order moments becomes available. Second, using the ensemble approach, the empirical estimate of the covariance is always symmetric and positive. While this may not pose any serious issue with the current diagonal closure in (3.14) provided the $K \times K$ covariance dynamics is numerically integrated with an adequate ODE solver and time step, preserving the symmetry can be challenging when more non-diagonal parameterization model Q_M becomes available. Third, we note that the fluctuation dynamics in (3.11) consists of an ensemble of independent K -dimensional model for $\{Z_k^{M,(i)}\}_{k \in \mathcal{I}}$. This independent structure allows one to employ a parallel computation for each ensemble member in the model integration to reduce the time complexity when $M > K$, and possibly be faster than directly integrating the fully coupled $K \times K$ system in (3.14). Particularly, parallel computing will be natural when the neural-network model for Q_k^M is employed on a GPU cluster.

3.4. Neural network model for the unresolved processes

To parameterize $\{\Theta^m, Q_k^M\}$ in (3.9) or $\{\Theta_0^m, \Theta_k^m, Q_k^M\}$ in (3.11), we consider using a non-Markovian closure model to learn these terms, following our previous works in modeling variance closure [15] and trajectory of partially observed discrete-time ergodic Markov chain [16]. While such a general formulation can be theoretically justified in the context of predicting time-evolution of state variables (using the discrete Mori-Zwanzig representation and the time delay Taken's embedding theory [16,27]), the dimension of the theoretically justifiable observable in the current application is too high for a numerically tractable implementation. Particularly, for the statistical-stochastic models in (3.9) accounting M ensemble members, identifying a model of $K = |\mathcal{I}|$ variables that depends on L -time delay corresponds to estimating a map with $(MK + 1)L$ variables, which is a very high-dimensional estimation problem since $M \gg 1$ is needed for reasonably accurate ensemble estimations. Rather than learning the entire complicated dynamical processes, we identify the input of the non-Markovian map to reflect some explicit expression of $\{\Theta_0^m, \Theta_k^m, Q_k^M\}$ as reported in (3.3) and design the neural network only to learn these unresolved processes. Particularly, since these variables are ultimately functions of the statistical quantities, we will identify them as time delay mappings of the mean and variances of the resolved components in addition to the time-delay of the variable of interest, neglecting their dependence on the smaller non-diagonal covariance components.

Following the work in [15], we will consider the class of residual network with vanilla LSTM architecture [28]. For the homogeneous model, the correction terms in both the mean equation (3.6a) and the fluctuation equations (3.6b) are approximated by a neural network with residual structure,

$$\begin{aligned}
\Theta^m(t_{\ell+1}) &= \Theta^m(t_\ell) + \text{LSTM}^m(\bar{u}(t_{\ell-L:\ell}), \{R_k(t_{\ell-L:\ell})\}, \Theta^m(t_{\ell-L:\ell}); \theta), \\
Q_k^M(t_{\ell+1}) &= Q_k^M(t_\ell) + \text{LSTM}_k^v(\bar{u}(t_{\ell-L:\ell}), \{R_k(t_{\ell-L:\ell})\}, \{Q_k^M(t_{\ell-L:\ell})\}; \theta),
\end{aligned} \tag{3.15}$$

where we have used the notation $a(t_{\ell-L:\ell}) := (a(t_{\ell-L}), a(t_{\ell-L+1}), \dots, a(t_\ell))$ for any dependent variable a and θ to denote the parameters in the LSTM network. Notice that the right-hand-sides of both equations in (3.15) are $(2K + 1)L$ dimensional maps, independent of the ensemble size M . For the inhomogeneous case, the unresolved model parameters $\{\Theta_0^m, \Theta_k^m, Q_k^M\}$ can also be learned directly by fitting the LSTM neural networks with analogous residual structure, that is,

$$\begin{aligned}
\Theta_0^m(t_{\ell+1}) &= \Theta_0^m(t_\ell) + \text{LSTM}_0^m(\bar{u}(t_{\ell-L:\ell}), \{\hat{u}_k(t_{\ell-L:\ell})\}, \{R_k(t_{\ell-L:\ell})\}, \Theta_0^m(t_{\ell-L:\ell}); \theta), \\
\Theta_k^m(t_{\ell+1}) &= \Theta_k^m(t_\ell) + \text{LSTM}_k^m(\bar{u}(t_{\ell-L:\ell}), \{\hat{u}_k(t_{\ell-L:\ell})\}, \{R_k(t_{\ell-L:\ell})\}, \{\Theta_k^m(t_{\ell-L:\ell})\}; \theta), \\
Q_k^M(t_{\ell+1}) &= Q_k^M(t_\ell) + \text{LSTM}_k^v(\bar{u}(t_{\ell-L:\ell}), \{\hat{u}_k(t_{\ell-L:\ell})\}, \{R_k(t_{\ell-L:\ell})\}, \{Q_k^M(t_{\ell-L:\ell})\}; \theta).
\end{aligned} \tag{3.16}$$

We should point that since the non-diagonal terms in the covariance are small relative to the diagonal components, we only include the variance components as inputs, and thus, arrive at a problem of estimating time delay embedding maps with $(3K + 1)L$ variables.

The LSTM model parameters, θ , are attained by minimizing the empirical risk in (2.20) defined by averaging the loss function $L(\theta, y_i, y_i^M)$ on n training samples of $(x_i, y_i)_{i=1}^n$, where x_i takes in the sequences of input data \bar{u}, \hat{u}_k, R_k and $y_i^M = \mathcal{M}(x_i)$ gives the LSTM output. In the following pseudo-code, we provide the computational steps for evaluating $y^M = \mathcal{M}(x)$, where the input x corresponding to the statistical-stochastic reduced-order model in (3.11) is also stated precisely. We remark that similar pseudo-code is used for the homogeneous case, where the reduced-order model in (3.11) is replaced with (3.9) and the LSTM closure models in (3.16) is replaced with (3.15). In the homogeneous case, the input x does not have $\{\hat{u}_k\}_{k \in \mathcal{I}}$.

Pseudo-code 1 Evaluating the label $y^M = \mathcal{M}(x, \theta)$ corresponds to the reduced-order statistical-stochastic model.

Input: x consists of $\hat{u}_0, \hat{u}_k, R_k$ at time t_{-L}, \dots, t_0 and $Z_k^{(i)}$ at time t_0 , where $\Delta t = t_\ell - t_{\ell-1}$ for all $k \in \mathcal{I}$ and $i = 1, \dots, M$. (θ denotes the parameters in the LSTM model.)
Output: y^M consists of $\delta \hat{u}_0^M, \delta \hat{u}_k^M, \delta R_k^M$ for all $k \in \mathcal{I}$ at times t_1, \dots, t_T .
Require: $\ell = 1, T > 0$
while $\ell < T$ **do**
 • Compute the unresolved fluxes $\Theta_0^M, \Theta_k^M, Q_k^M$ at $t = t_\ell$ for the mean and variance using the LSTM model in (3.15) or (3.16), evaluated at the input parameter value θ , with the time-delay inputs from the previous L time steps;
 • Evaluate the perturbed mean states $\hat{u}_{\delta,0}^M, \hat{u}_{\delta,k}^M$ at time t_ℓ using the mean models (the first two equations in (3.11)). Subsequently, we attain the response mean statistics $\delta \hat{u}_0^M(t_\ell) := \hat{u}_{\delta,0}^M(t_\ell) - \hat{u}_{\text{eq},0}$ and $\delta \hat{u}_k^M(t_\ell) := \hat{u}_{\delta,k}^M(t_\ell) - \hat{u}_{\text{eq},k}$, where $\hat{u}_{\text{eq},0}, \hat{u}_{\text{eq},k}$ are the reference equilibrium mean;
 • Update the effective damping and noise D_k^M, Σ_k^M at time t_ℓ using the decomposition in (3.10) of the statistical flux model $Q_k^M(t_\ell)$;
 • Update the stochastic coefficients $Z_{\delta,k}^{M,(i)}(t_\ell)$ by solving the third equation in (3.11) for each ensemble member with the effective damping and noise corrections.
 • Compute the empirical variances of the perturbed coefficients, $R_{\delta,k}^M(t_\ell) = \frac{1}{M-1} \sum_{i=1}^M Z_{\delta,k}^{M,(i)}(t_\ell) (Z_{\delta,k}^{M,(i)}(t_\ell))^*$. Subsequently, compute the response variance $\delta R_k^M(t_\ell) := R_{\delta,k}^M(t_\ell) - R_{\text{eq},k}$, where $R_{\text{eq},k}$ denotes the equilibrium variance of the unperturbed system;
 • Update $\ell = \ell + 1$;
end while

4. Predicting leading-order statistics of the L-96 system

In this section, we numerically validate the prediction skill of the proposed reduced-order statistical-stochastic models to recover the leading-order statistics on different statistical structures in the L-96 system. In particular, we consider two representative regimes, generating homogeneous and inhomogeneous statistics. The homogeneous regime provides a simpler test case for validating the proposed algorithm on chaotic complex systems with strong instability and non-Gaussian statistics. The inhomogeneous regime serves as a more challenging problem induced by nonlinear spatio-temporal interactions and non-zero cross-correlations. We organize the section as follows: First, we report the experiment configuration and the training data generation in Section 4.1. Then, we report the results for the homogeneous and inhomogeneous cases in Sections 4.2 and 4.3, respectively.

4.1. Model configuration and training dataset for the L-96 system

To generate the training data (or the label y in (2.18)), we first integrate the L-96 system under homogeneous reference forcing $F_{\text{ref}} = 8$ and damping $d_{\text{ref}} = 1$. The equation is integrated using the 4th-order Runge-Kutta scheme with a small time step $dt = 0.001$, and the data is subsequently sampled at every 10 steps. Thus we have the data sampling step $\Delta t = 0.01$. The use of larger sampling step size, while introduce additional numerical discretization, is to reflect the practical situation when frequent measurements are not often available, especially if efficient numerical integration with larger time step is used. Subsequently, we compute the empirical mean and variance $\bar{\mathbf{u}}_{\text{eq}}$ and $R_{\text{eq},k}$, over these discrete time realizations.

To produce a unified training data set independent of the particular forcing and damping perturbations, we sample the transient state statistics from only an initial perturbation of the ensemble samples of the following form

$$\mathbf{u}^{(i)} := \alpha \bar{\mathbf{u}}_{\text{eq}} + \sqrt{\beta} \left(\mathbf{u}_{\text{ref}}^{(i)} - \bar{\mathbf{u}}_{\text{eq}} \right), \quad i = 1, \dots, M = 500, \quad (4.1)$$

where $\{\mathbf{u}_{\text{ref}}^{(i)}\}_{i=1,\dots,M}$ denotes a set of $M = 500$ samples randomly drawn from a long discrete trajectory of solutions of the L-96 system corresponding to the reference damping and forcing $F_{\text{ref}}, d_{\text{ref}}$. New initial ensembles are generated by perturbing the mean through the parameter α and the variance through the parameter β . Fig. 4.1 plots several realizations of the statistical responses for the mean and variance subject to different perturbation parameters α, β . The converging trajectories of the mean and variance also illustrate the decorrelation time that characterizes the mixing rate of the states. The solutions will finally converge to the unperturbed equilibrium within the decorrelation time around $T = 1.5$, which is a time scale that we expect the prediction skill to be accurate over the testing data. For training, we will consider 6 different values for each $\alpha, \beta \in \{0.5, 0.7, 0.9, \dots, 1.5\}$, resulting to $6 \times 6 = 36$ different initial perturbation cases.

Based on these initial conditions, we have 36 trajectories of transient statistics for the reference systems (see some of these trajectories in Fig. 4.1). To increase the number of training data in the homogeneous case, we consider 4 additional external constant forcings, $F_\delta = F_{\text{ref}} + \delta F$, where $\delta F \in \{-1, -0.5, 0.5, 1\}$ in addition to reference forcing with $\delta F = 0$. With these additional perturbations, we have $36 \times 5 = 180$ trajectories of the response mean, $\delta \hat{\mathbf{u}}$, and variance statistics, $\{\delta R_k\}_{k \in \mathcal{I}}$, at discrete time $t_\ell = \ell \Delta t \in [0, 2]$. For training data, we ignore the solutions beyond 2 time units since most of the statistical quantities at these times are constant. We partition the statistics on time interval $[0, 2]$ into 10 overlapping sub-intervals, each of time length 1.1 units: $[0, 1.1], [0.1, 1.2], \dots, [0.9, 2]$. On each sub-interval, since the discrete time step is $\Delta t = 0.01$, we have statistics at 111 data points. Following the notation in Pseudo-code 1, we label these statistical timeseries as the quantities at $t_\ell = -L, \dots, 0, 1, \dots, 10$, with $L = 100$. We will use the first $L + 1 = 101$ data points of $\{\mathbf{u}^{(i)}\}_{i=1,\dots,M}, \bar{\mathbf{u}}$ and $\{R_k\}_{k \in \mathcal{I}}$ to construct the input data x . Particularly, we take FFT on $\bar{\mathbf{u}}(t_\ell) \in \mathbb{R}^N$ to attain $\hat{u}_0(t_\ell)$ and $\{\hat{u}_k(t_\ell) : k \in \mathcal{I}\}$ for $\ell = -100, \dots, 0$. For the homogeneous case, it is clear that $\hat{u}_0 = \bar{u}$ and $\hat{u}_k = 0$ when $k \neq 0$. We also use the decomposition in (3.2) on $\{\mathbf{u}^{(i)}(t_0)\}_{i=1,\dots,M}$ to attain an ensemble perturbation $\{Z_k^{(i)}(t_0) : k \in \mathcal{I}, i = 1, \dots, M\}$. This completes the construction

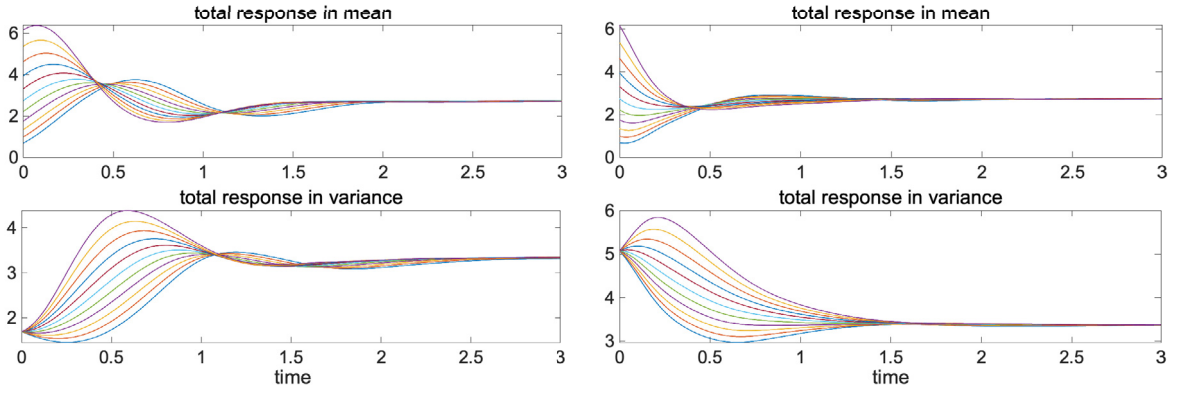


Fig. 4.1. Statistical responses of the total energy in the mean and the total variance of the L-96 system. Different lines represent the different amplitudes of mean perturbation $\alpha \in [0.5, 1.5]$. Left: $\beta = 0.5$; right: $\beta = 1.5$.

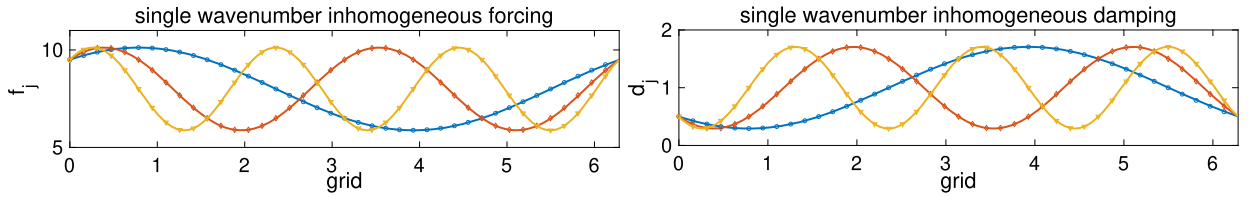


Fig. 4.2. Inhomogeneous forcing and damping perturbations on top of the equilibrium state $F_{eq} = 8$ and $d_{eq} = 1$. Perturbations are added to wavenumbers $k = 1, 2, 3$.

Table 1

Hyper-parameters for training the standard Long-Short-Term-Memory (LSTM) neural network model using stochastic gradient descent (SGD).

total training epochs	100
ensemble size	500
SGD batch size	1
initial learning rate	0.001
learning rate reduction at iteration step	25, 50, 75
time step size between two measurements Δt	0.01
LSTM sequence length L	100
forward prediction steps in training T	10
LSTM hidden state size h	100

of an input x for each partition. Finally, we take the last 10 data points of $\delta \bar{\mathbf{u}}(t_\ell)$ and $\{\delta R_k(t_\ell)\}_{k \in \mathcal{I}}$ at $\ell = 1, \dots, 10$ as the label data y on each partition.

Accounting for the number of sub-interval from the partition, we have a total of $n = 36 \times 5 \times 10 = 1800$ training data $(x_i, y_i)_{i=1, \dots, n}$ for the homogeneous case. For the inhomogeneous case, we consider 16 different forcings and dampings, where in each case, the forcing is chosen to be one of the four cases: reference case $\delta f = 0$ or $\{\delta f_j = 1.5 \sin(\frac{2\pi k j}{N})\}_{k=1,2,3}$ and the damping is one of the four cases: reference damping $\delta d = 0$ or $\{\delta d_j = 0.5 \sin(\frac{2\pi k j}{N})\}_{k=1,2,3}$. Effectively, these inhomogeneous forcings and dampings (see Fig. 4.2) exerted a single Fourier mode $k = 1, 2$ or 3 . Applying the same temporal partitioning as in the homogeneous case on each trajectory of response statistics, we have a total of $n = 36 \times 16 \times 10 = 5760$ training data for the inhomogeneous case.

The training procedure is to solve the empirical risk minimization task of (2.20), where the average is over the training pairs $(x_i, y_i)_{i=1, \dots, n}$ discussed above. In the empirical risk function in (2.20), the predicted label, $y_i^M = \mathcal{M}(x_i, \theta)$ is computed using Pseudocode 1. Computationally, the optimization problem is to find the parameters θ in the LSTM models in (3.15) or (3.16) that is used in the first step of Pseudo-code 1. The hyper-parameters of the LSTM network are summarized in Table 1. We solve this optimization problem using the stochastic gradient descent (SGD) algorithm with batch size 1 to minimize the computational cost (as we do not find any advantage of using larger batch sizes). The learning rate is reduced by 50% at iteration steps 25, 50, and 75. Once the model is trained, denoting θ^* as the parameter obtained from the SGD algorithm, we use Pseudocode 1 to evaluate the response statistics, $\mathcal{M}(x_i^{\text{new}}, \theta^*)$, corresponding to new input x_i^{new} that is not in the training data set.

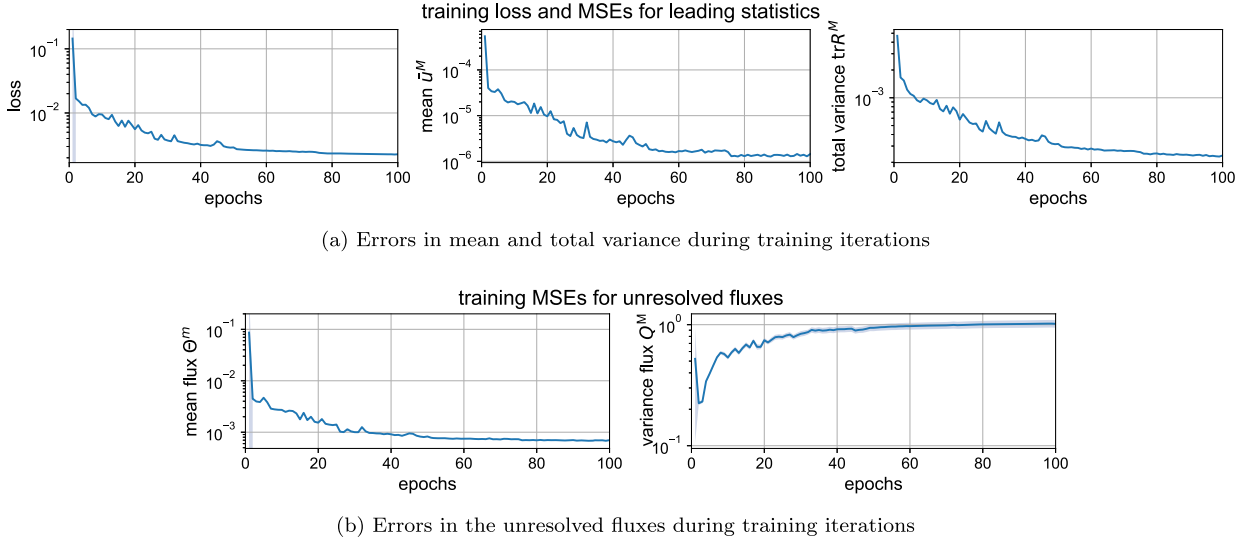


Fig. 4.3. Training errors using the reduced-order model (3.9) for the homogeneous statistical regime. The first row shows the evolution of the loss function and MSEs in the predicted mean and variance. The second row compares the difference between the true flux and the neural network outputs in mean Θ^m and Q_k^M .

4.2. Training and prediction of the homogeneous statistical regime

First, we consider the homogeneous perturbation case using uniform perturbations in the forcing $\mathbf{F} = F_{\text{eq}} + \delta f \mathbf{e}_0$. Only the most energetic leading modes, $\mathcal{I} = \{k : 6 \leq |k| \leq 12\}$, are resolved in the fluctuation equation for Z_k^M (compared to total 40 modes). Here, the target is to predict the homogeneous mean state $\bar{\mathbf{u}} = \bar{u} \mathbf{e}_0$ and the diagonal variance in resolved mode $R_k = \frac{1}{M-1} \sum_i Z_k^{(i)} (Z_k^{(i)})^*$ based on the ensemble solutions.

We first show the evolution of the errors during the training iterations in Fig. 4.3(a). The first row plots the values of the empirical risk function (2.20), where the training errors in the predicted mean, \bar{u}^M , and variance, R_k^M , are computed based on the empirical average of the ensemble of solutions with training inputs, $\{x_i\}_{i=1}^n$. The neural network model is trained with 100 repeating epochs and a small number of forwarding steps $T = 10$. The result shows that the loss function can be minimized to small values after a much smaller number of iterations (around 40 epochs). Correspondingly, the mean square errors ($\text{MSE}(f, f^M) = \frac{1}{n} \sum_{i=1}^n |f_i - f_i^M|^2$) of the homogeneous mean \bar{u} and the total variance of resolved modes $\text{tr}R^M = \sum_{k \in \mathcal{I}} R_k^M$ can be both effectively minimized to very small values. This indicates the accurate fitting of both the mean and statistics in the reduced order model. The variance of the sample errors is also plotted by the shaded area around the lines, which is reduced to negligibly small values. The decay of training error demonstrates the effectiveness of the training process in reducing the model errors uniformly among all the training samples. For more detailed comparisons of the training performance, we also plot the training errors in the neural network outputs of the unresolved flux terms Θ^m and Q_k^M in (3.15) that are not directly compared in the loss function, decays. In this case, we found that the error in Θ^m , which is not measured directly in the loss function, decays. On the other hand, the discrepancy between the model constructed flux, Q_k^M , and the statistical flux, $Q_{F,k}$, actually increases. This is not a surprise since we used the decomposition in (2.11) to determine D_k^M and Σ_k^M . Recall that this parameterization uses the equilibrium variance $R_{\text{eq},k}$ to avoid the elaborate computational cost induced by fitting to the more ideal time-dependent variance, R_k , as suggested in (2.11).

In the forecast stage, the trained model is applied to predict the key statistics under perturbations of initial conditions that are different from the training input data. We should point out that the model output in the forecast stage at time t_i is the result of iterating the model i steps, thus the model errors are accumulated through these iterations, and attaining accurate prediction becomes challenging for the unstable modes with a positive growth rate as illustrated in Fig. 3.1. The prediction errors in the mean state and variance from the empirical ensemble average are plotted in the first row of Fig. 4.4. The errors in the test solutions with different perturbations of initial conditions are compared during the time evolution up to a long time $T = 2.5$ with 250 iterations (beyond the decorrelation time $T_{\text{decorr}} \approx 1.5$ of the state). The result suggests that the trained model produces accurate statistical predictions under various tested perturbations of initial conditions. Particularly, the errors in both mean and variance in resolved fluctuation modes remain small during the prediction time interval shown. This confirms the stable model dynamics using the effective damping and forcing introduced in this reduced model.

We also report the prediction skill for the smaller number of samples in the second and third row of Fig. 4.4. Specifically, the prediction errors using smaller ensemble sizes $M = 100, 50$ are compared using the same model that is trained with an ensemble of size $M = 500$. Compared with the larger ensemble case $M = 500$ in the first row, the errors begin to grow as the ensemble size decreases. This is expected since the estimated statistics via ensemble average become less accurate.

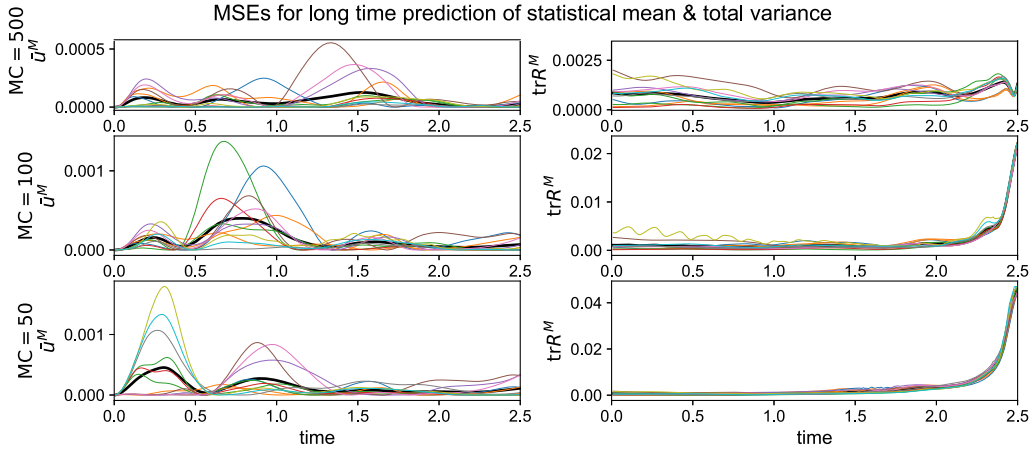


Fig. 4.4. Prediction errors using the trained reduced-order model (3.9) for the homogeneous statistical regime. MSEs of initial conditions that are different than the training input data (in thin colored lines) in the mean state and ensemble variance are compared together with the overall average of all the test cases (in thick black line). Errors using different ensemble sizes $M = 500, 100, 50$ to recover the statistics are also compared.

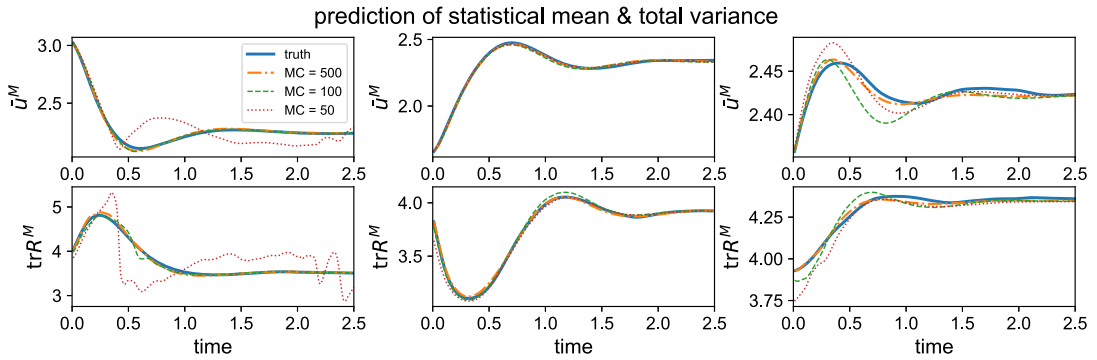


Fig. 4.5. Prediction of the statistical mean and variance using the trained reduced-order statistical-stochastic model for homogeneous statistics. The predictions of the mean and total variance with different ensemble sizes $M = 500, 100, 50$ are compared. The first, second and third columns show predictions starting from an initial condition that does not belong to the training input data under three forcing perturbations $\delta F = -1, 0, 1$ that were used to generate the training data set, respectively.

However, we can still see that the prediction is accurate in most of the test cases. A more detailed comparison between the statistical predictions of the mean, the trace of the variance, and the variance of each resolved mode under three forcing perturbations are shown in Fig. 4.5. Consistent with the prediction errors in Fig. 4.4, the evolution of the mean and variances starting from three pairs of initial conditions and forcings is captured accurately when $M = 500$. When the ensemble size is reduced to $M = 100$, there is a slight increase in errors, however, the overall qualitative transient behavior in each resolved mode is still accurately predicted. When the ensemble size is reduced to $M = 50$, the performance accuracy varies wildly. Under large forcing perturbations $\delta F = \pm 1$ (see the first and third columns), the prediction accuracy significantly deteriorates. Under the reference perturbation with $\delta F = 0$ in the second column, the statistics are accurately predicted.

4.3. Training and prediction of the inhomogeneous statistical regime

Finally, we consider the model prediction skill in the challenging case induced by inhomogeneous statistics. As we have discussed in Section 4.1, we train the model using data generated by applying inhomogeneous damping and forcing on the first three leading modes, which corresponds to spatially periodic forcing and damping corresponding to these wave numbers as shown in Fig. 4.2.

As in the previous section, we first display the training results using the reduced-order model (3.11) to learn and recover the inhomogeneous statistics of the perturbed L-96 system. In this inhomogeneous case, the resolved mean modes include the inhomogeneously forced and damped wavenumbers $k = 1, 2, 3$, and the resolved fluctuation modes include still the most energetic ones $6 \leq |k| \leq 12$. The evolution of the errors during training iterations is displayed in Fig. 4.6. Similar to the homogeneous case, the errors can be effectively minimized within the 100 training epochs in both the homogeneous mean $\hat{u}_0 = \bar{u}$ and variance R_k as well as all the inhomogeneous mean Fourier coefficients \hat{u}_k . Again, it is useful to notice that the error in the variance feedback Q_k^M actually increases during the training process. As in the homogeneous case, this

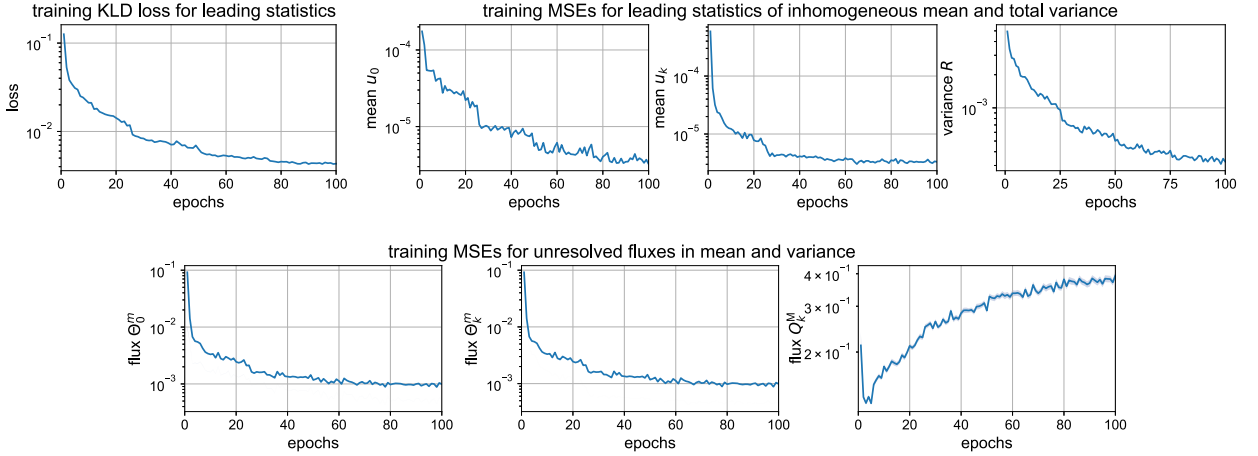


Fig. 4.6. Training relative entropy loss and MSEs during training iterations of the reduced stochastic model.

discrepancy is due to the use of coefficients D_k^M and Σ_k^M attained by equilibrium fitting in (3.10) as a way to realize the decomposition in (2.11).

Next, we check the long-term prediction of the trained model for capturing the inhomogeneous mean and variance, starting from new initial conditions that are different from the training input data. As in the homogeneous case, instead of iterating the model in a small number of steps (10 forward steps) in the training stage, the prediction stage iterates the optimized model in 175 steps to achieve prediction up to 1.75 model unit time. The first row of Fig. 4.7 shows the prediction MSEs in the homogeneous and inhomogeneous mean state and resolved variance under three inhomogeneous forcing and damping cases on wavenumbers $k = 1, 2, 3$ as in Fig. 4.2, for new perturbations on the initial mean and sample variance as in (4.1). This is to check the model responses in leading order statistics subject to the inhomogeneous statistical structures induced by the forcing and damping. It shows that the trained reduced-order model produces accurate prediction skill on both the homogeneous and inhomogeneous components among all the different test cases. The next three rows of Fig. 4.7 display the detailed comparison of the true and predicted statistics corresponding to the same initial condition for three different damping and forcing perturbations (that are imposed to obtain the MSE in the first row). The predicted inhomogeneous mean state in the first three modes and responses in leading variance mode are also compared in Fig. 4.8 for the three different perturbation cases.

While all previous numerical experiments were focused to capture statistical responses under new initial state perturbations, we further test the model prediction for mean and variance responses to different external forcing perturbations. The additional forcing perturbations are exerted on either the homogeneous mean state $\delta f = 0.5 + 1.5 \sin(\frac{2\pi j}{N})$ or the inhomogeneous leading mean modes $\delta f = f_1 + f_2 + f_3$, with $f_k = \sin(\frac{2\pi k j}{N}) + \cos(\frac{2\pi k j}{N})$, where these additional constant mean forces are not in the training data set. Fig. 4.9 shows that the closure model predicts the statistical responses accurately under these different forcing perturbations.

We should point out that for inhomogeneous cases, the inhomogeneous mean and cross-covariances play an important role. Thus the dynamics of covariance in (2.5) is fully coupled with the inhomogeneous mean modes. This fact makes accurate dynamical modeling of even just the variance components nontrivial since one has to account for the interaction with all other modes. In addition, the inherent instability in the dynamical model will amplify the unavoidable small errors in the neural network output at each time iteration step and accumulate them in time. Despite these challenging issues, we found that the proposed reduced-order model can accurately predict the response mean and variance statistics under different initial and forcing perturbations for a long prediction time beyond the decorrelation time of the states before the error starts to accumulate in time.

5. Summary

In this paper, we proposed a generic statistical-stochastic closure modeling framework for effective ensemble prediction of leading order statistics in complex systems containing strong instability and interactions among different spatio-temporal scales. The mean dynamics are modeled with a set of statistical equations that represents the homogeneous and inhomogeneous components subject to external perturbations. The fluctuation dynamics, which characterize the uncertainty among the multiscale modes, are modeled with a stochastic formulation. The mean equations are coupled with the covariance matrix that is empirically estimated using the ensemble prediction of the fluctuation terms. On the other hand, the stability of the fluctuation dynamics depends on the mean state. Such a formulation guarantees the positive definiteness of the covariance matrix. A reduced-order closure strategy is formulated to resolve the most energetic mean and variance modes for efficient computation with an ensemble simulation. Subsequently, machine learning tools are adopted to identify non-Markovian models for the nonlinear feedback from unresolved processes and imperfect model errors.

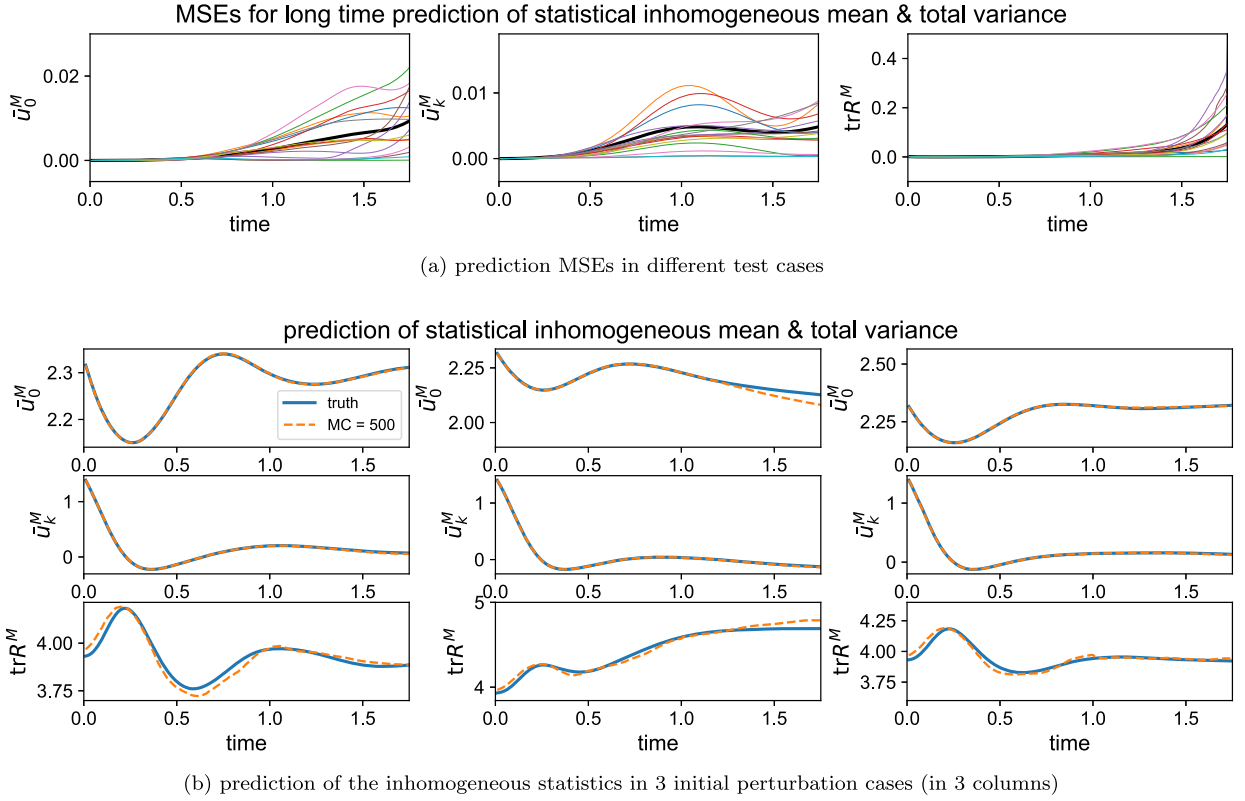


Fig. 4.7. Upper panel: Prediction errors in the homogeneous mean \bar{u}_0^M , resolved inhomogeneous mean $\bar{u}_k^M = \sum_{k \in \mathcal{I}} \hat{u}_k^M$ and total variance $\text{tr}R^M = \sum_{k \in \mathcal{I}} r_k^M$ with inhomogeneous forcing and damping. Errors with different initial perturbations are plotted in thin colored lines and the total averaged error is plotted in thick black lines. Lower panel: Prediction of the homogeneous and resolved inhomogeneous mean, and total resolved variance for the same initial state perturbation that is not in training data set. Each column corresponds to a specific choice of damping and forcing that is used to generate the training data set. The true solution is in solid blue line and the model prediction is in dashed orange line.

To combat instability and allow for a scalable training procedure, we considered a closure model with an effective damping and noise parameterization of the form (2.9). Here, the damping coefficients and noise amplitudes are identified with a non-Markovian model that is designed to be consistent with the unperturbed equilibrium covariance statistics in the long term. Further diagonal approximation is employed in our numerical examples where the covariance statistics are diagonally dominant. Finally, efficient training of the neural network model under limited training dataset is achieved by measuring only the statistical output using an information metric-based loss function, fitting to the available training response mean and variance statistics. With such a training procedure, we effectively avoid the usually exhausting process of fitting of a large number of high-dimensional stochastic trajectories that often leads to overfitting as illustrated in Appendix A.

We numerically found that the proposed approach is effective in identifying stable dynamics with accurate statistical prediction. The dimension of the reduced-order model is of order $K(1 + M)$, where the resolved state dimension $K \ll N$ is much smaller than the dimension of the full state, N . Thus, a smaller ensemble size M is needed to sample the lower dimensional resolved subspace. The skill of the reduced-order statistical-stochastic model is tested on several statistical regimes of the L-96 system, including the homogeneous and inhomogeneous statistics produced by exerting different types of forcing and damping perturbations. In the simpler homogeneous case, the trained model showed uniformly high skill in dealing with different perturbations. In this case, the reduced-order model only resolves the first seven most energetic complex-valued leading Fourier modes. In the inhomogeneous case, the situation becomes more challenging. Inhomogeneous perturbations exert the mean energy of some Fourier modes with low variance energy and the covariance becomes non-diagonal although it is still diagonally dominant. Including three additional Fourier modes corresponding to the largest mean energy in the reduced-order model, we achieve accurate predictions of the leading mean and variance states on various nontrivial inhomogeneous statistical regimes and obtain a model that remains stable for at least up to the decorrelation time of the states. However, we should also point out that there is no guarantee that the present model does not blow up if we keep iterating for longer times since the closure parameterization ignores the non-diagonal component of Q_F which may suppress conditional linear instability.

While the numerical results are encouraging, constructing an improved parameterization that accounts for non-diagonal components of the unresolved feedback Q_F that can guarantee stable dynamics for a long time under various perturbations induced by the learning and the temporal discretization errors remains an open issue. While finding a stable parameteriza-

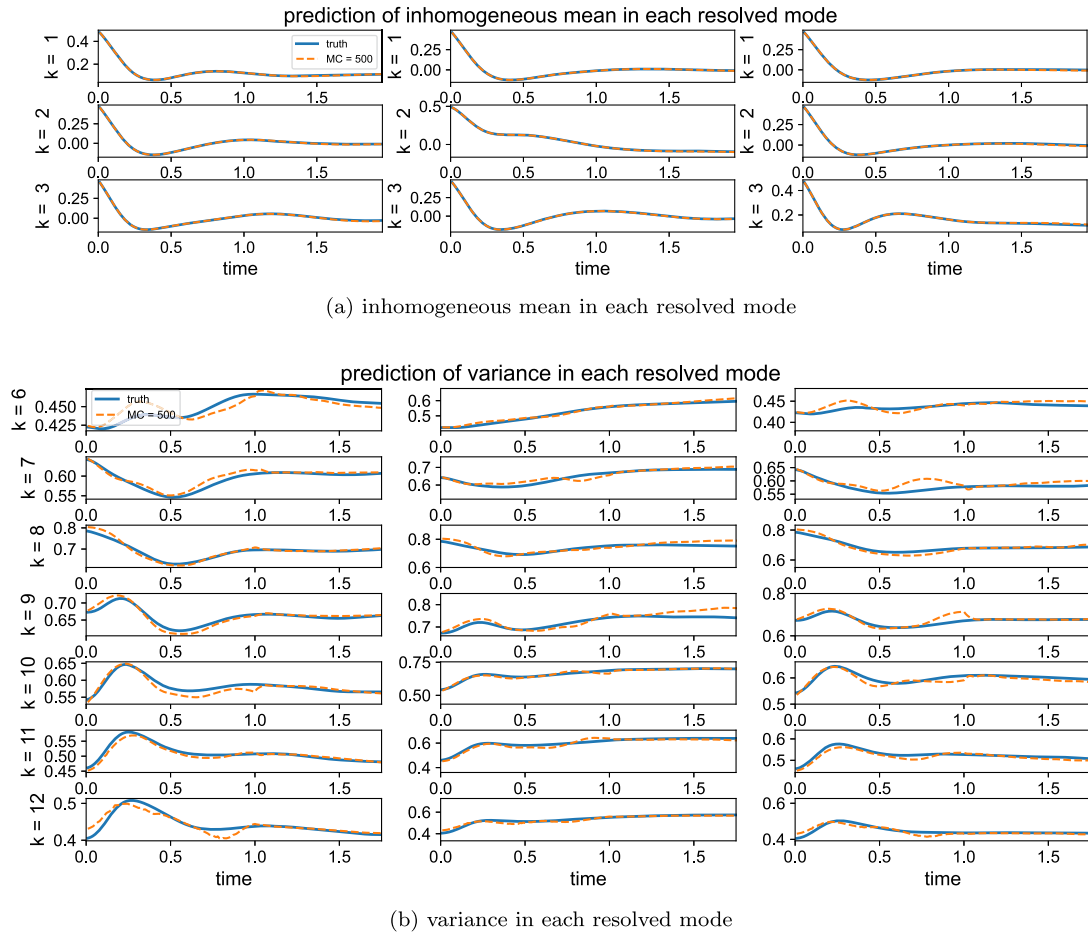


Fig. 4.8. Detailed prediction results of the statistical mean state in the first three inhomogeneous modes, and the predicted variance in each resolved mode. The three columns represent 3 typical test cases with different perturbations.

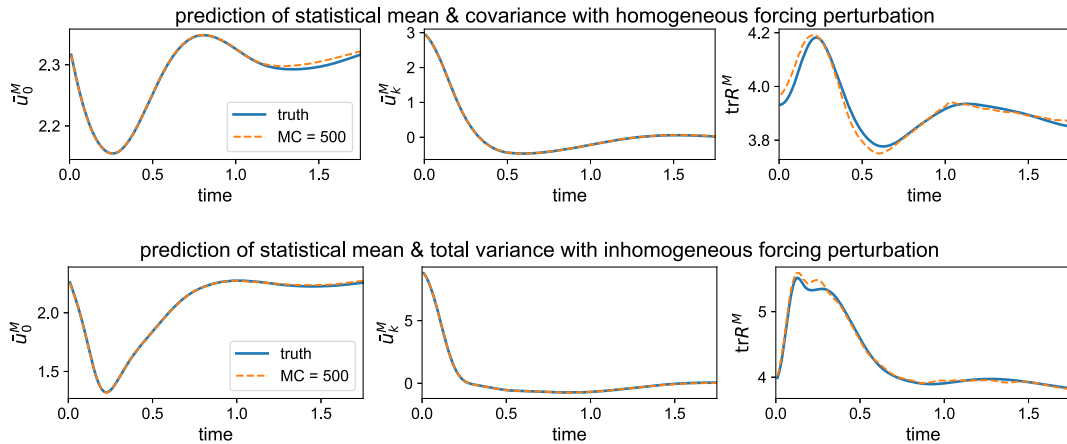


Fig. 4.9. Prediction of the mean and total resolved variance with external forcing perturbation on the mean $\delta f = 0.5 + 1.5 \sin(x)$ (upper) and on the first 3 inhomogeneous modes $\delta f = f_1 + f_2 + f_3$ (lower) with inhomogeneous statistics. The true solution is in solid blue line and the model prediction is in dashed orange line.

tion for Q_F is a general problem, we also believe that this issue is related to an open problem in linear response theory for chaotic dynamical systems. The study of the validity of linear response in [29] is crucially important to predicting the response at the steady state of the perturbed dynamical system. Particularly, the main goal of such a study is to check whether the perturbed system has an invariant measure that is smooth (under appropriate topology) as a function of the parameters that reflect the perturbations. If such a condition is invalid, then small perturbations induced by the learning error of the approximate closure model will generate a drastic change in the dynamical behavior, and thus, prohibit us to emulate accurate long-time statistics.

CRedit authorship contribution statement

The authors have equal contributions in designing the research, performing the research, validating the results, and writing the paper.

Di Qi: Conceptualization, Methodology, Validation, Writing – original draft and review & editing.

John Harlim: Conceptualization, Methodology, Validation, Writing – original draft and review & editing.

Declaration of competing interest

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

Data availability

The codes are written in Python and are available on GitHub (<https://github.com/qidigit/non-Markovian-closure-LSTM>).

Acknowledgements

The research of D.Q. was partially supported by the start-up funds and the PCCRC Seed Funding provided by Purdue University. The research of J.H. was partially supported by the NSF grants DMS-1854299, DMS-2207328, DMS-2229435, and the ONR grant N00014-22-1-2193.

Appendix A. Trajectory training and prediction using the direct stochastic model

In this Appendix, we demonstrate the limitation of the standard learning procedure with a loss function that measures the discrepancies between stochastic trajectories. Since the reduced-order model couples a statistical quantity that depends on empirical variance of a stochastic fluctuation, fitting trajectories of an entire statistical-stochastic system (such as (3.7)) can be numerically demanding, especially when M is large. In the following, we will conduct an experiment fitting only the stochastic component of (3.7) by assuming that the time series of the underlying mean \bar{u} is always available for us, and thus, ignoring the error induced by finite ensemble size M in the mean dynamics. While this scenario is not useful for real-time prediction, we will demonstrate that the standard machine learning procedure may not produce an effective learning even in such a simple case when the proposed damping and forcing parameterization in (2.11) is not used.

Specifically, we parameterize Θ_k^v in (3.7) by minimizing an empirical risk defined with the following loss function,

$$L(\theta, \mathbf{Z}^I, \mathbf{Z}^M) := \sum_{k \in \mathcal{I}} |Z_k - Z_k^M|^2, \quad (\text{A.1})$$

where,

$$\begin{aligned} \frac{dZ_k^M}{dt} &= -(\gamma + \gamma_k \bar{u}) Z_k^M + \Theta_k^v, \\ \Theta_k^v(t_{i+1}) &= \Theta_k^v(t_i) + \text{LSTM}^m(\bar{u}(t_{i-L:i}), \{R_k(t_{i-L:i})\}, \Theta_k^v(t_{i-L:i}); \theta). \end{aligned} \quad (\text{A.2})$$

In this numerical experiment, the input data is

$$\mathbf{x} = \left(Z_k^M(t_i), \bar{u}(t_{i-L:i}), R_k(t_{i-L:i}), \Theta_k^v(t_{i-L:i}) \right) \in \mathbb{R}^{(2K+1)L+K},$$

and output is $\mathbf{y} = \mathbf{Z} \in \mathbb{R}^K$. In this homogeneous statistics case, we set $\mathcal{I} = \{k : 6 \leq |k| \leq 12\}$ as in Section 4.2, which results in $K = |\mathcal{I}| = 14$. Setting $L = 100$ as in Table 1, this learning problem is to find a $(2K + 1)L + K = 2914$ dimensional map, which is quite high-dimensional. We fit this into the fluctuation coefficients $\{Z_k : k \in \mathcal{I}\}$ corresponding to the statistical training data for homogeneous case discussed in Section 4.1. Since we are fitting each realization of the fluctuation, the size of training data set is $n = 1800 \times 500$, accounting $M = 500$ ensemble members.

The training and prediction performance of the direct stochastic model is shown in Fig. A.1. The loss and mean square errors (MSEs) in the coefficients Z_k and unresolved flux term Θ_k^v during training iterations are shown in the first row of

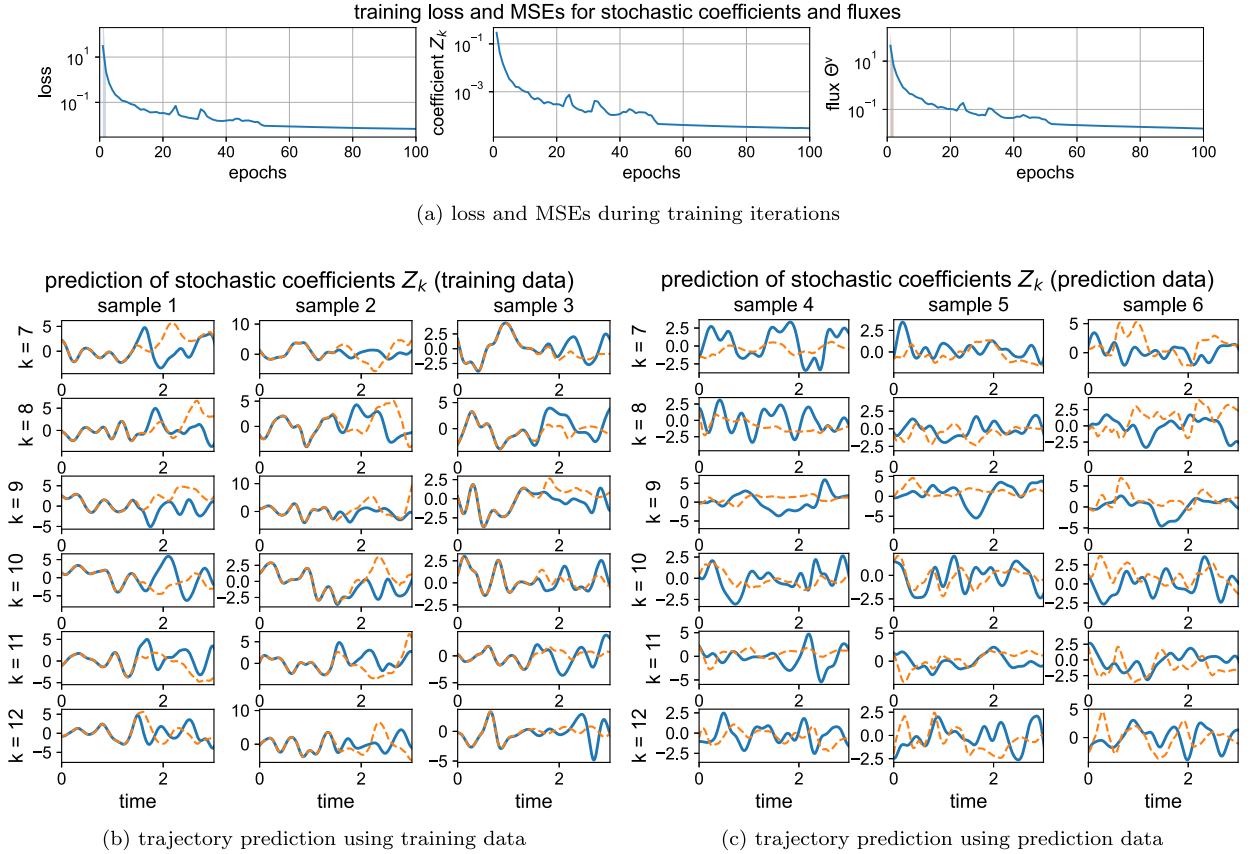


Fig. A.1. Training and prediction using the direct stochastic model. The first row shows the training iterations of errors. The second row shows predicted time trajectories of stochastic coefficients Z_k in the most energetic modes $k = 7, 8, 9, 10, 11, 12$. Several different sample trajectories are compared: the 3 samples on the left using the training data set and the 3 samples on the right using the new prediction data set. The truth is in solid blue lines while the model prediction is in dashed orange lines.

Fig. A.1. It appears that the training is effective with the pointwise errors minimized among all the trained samples. Then the trained model is tested on both the previous training data and the new prediction data away from the training set. The predicted trajectories are recurrently updated in time for a large number of iterations up to a long time $T = 3$ (300 iterations compared with only 10 iterations in training). The first 6 most energetic stochastic modes Z_k are plotted in the second row of Fig. A.1 in several samples. Testing on trajectories in the same training set, the predicted solution stays accurate for a while before the solution begins to diverge in time. Referring to the converging rate in Sec. 4.1, the predicted solutions always begin to diverge around the decorrelation time $T_{\text{decorr}} \sim 1.5$ when the autocorrelation decays to zero. This implies the inherent barrier in training the individual stochastic trajectories beyond the decorrelation time for a turbulent system containing instability.

More importantly, the trained model fails to predict stochastic trajectories away from the training set. Using new trajectories that are not included in the training data, the prediction diverges immediately and shows no skill in capturing the true trajectory. This shows a typical example of overfitting in training a neural network model. In this specific example, we suspect that the failure can be attributed to combinations of several issues. First, we suspect that the required amount of data to capture the large degrees of uncertainties in this high-dimensional problem is much larger than what we used in this experiment. Second, the stochastic components of the dynamical equations (A.2) are all conditionally unstable modes. With this inherent stability, identification of a stable neural-network modeling becomes a challenging issue, especially if no additional structures are imposed as in our experiments where the standard LSTM model with the residual structure in (A.2) is used. In addition to these issues, the Monte-Carlo error induced by the empirical average in (2.13) will amplify difficulties which translates into a numerically expensive training procedure when the true \mathbf{u} is not available in which one has to also learn the unresolved term Θ^m in (3.7).

This practical issue motivates the idea of fitting response statistics discussed in Section 2.2.3, especially when we are mostly interested in the statistical prediction of moments generated by the ensemble averages. As for the instability issue, we consider the use of damping and forcing parameterization discussed in Section 2.2.1 on the neural-network models.

References

- [1] C.E. Leith, Climate response and fluctuation dissipation, *J. Atmos. Sci.* 32 (10) (1975) 2022–2026.
- [2] A.J. Majda, D. Qi, Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems, *SIAM Rev.* 60 (3) (2018) 491–549.
- [3] A.J. Majda, D. Qi, Linear and nonlinear statistical response theories with prototype applications to sensitivity analysis and statistical control of complex turbulent dynamical systems, *Chaos, Interdiscip. J. Nonlinear Sci.* 29 (10) (2019) 103131.
- [4] A.J. Majda, J. Harlim, *Filtering Complex Turbulent Systems*, Cambridge University Press, 2012.
- [5] S. Reich, C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, 2015.
- [6] J. Harlim, Model error in data assimilation, in: C. Franzke, T. O’Kane (Eds.), *Nonlinear and Stochastic Climate Dynamics*, Cambridge University Press, 2017.
- [7] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, 2001.
- [8] G. Evensen, The ensemble Kalman filter: theoretical formulation and practical implementation, *Ocean Dyn.* 53 (2003) 343–367.
- [9] M. Lesieur, *Turbulence in Fluids: Stochastic and Numerical Modelling*, vol. 488, Nijhoff, Boston, MA, 1987.
- [10] T.P. Sapsis, A.J. Majda, Statistically accurate low-order models for uncertainty quantification in turbulent dynamical systems, *Proc. Natl. Acad. Sci.* 110 (34) (2013) 13705–13710, <https://doi.org/10.1073/pnas.1313065110>.
- [11] D. Qi, A.J. Majda, Low-dimensional reduced-order models for statistical response and uncertainty quantification: two-layer baroclinic turbulence, *J. Atmos. Sci.* 73 (12) (2016) 4609–4639.
- [12] M. Gamahara, Y. Hattori, Searching for turbulence models by artificial neural network, *Phys. Rev. Fluids* 2 (5) (2017) 054604.
- [13] A.P. Singh, S. Medida, K. Duraisamy, Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils, *AIAA J.* 55 (7) (2017) 2215–2227.
- [14] R. Maulik, O. San, A. Rasheed, P. Vedula, Subgrid modelling for two-dimensional turbulence using neural networks, *J. Fluid Mech.* 858 (2019) 122–144.
- [15] D. Qi, J. Harlim, Machine learning-based statistical closure models for turbulent dynamical systems, *Philos. Trans. R. Soc. A, Math. Phys. Eng. Sci.* 380 (2229) (2022) 20210205.
- [16] J. Harlim, S.W. Jiang, S. Liang, H. Yang, Machine learning for prediction with missing dynamics, *J. Comput. Phys.* (2020) 109922, <https://doi.org/10.1016/j.jcp.2020.109922>.
- [17] N. Chen, D. Qi, A physics-informed data-driven algorithm for ensemble forecast of complex turbulent systems, arXiv preprint arXiv:2204.08547, 2022.
- [18] E.N. Lorenz, Predictability: a problem partly solved, in: *Proc. Seminar on Predictability*, vol. 1, 1996.
- [19] M. Leutbecher, T.N. Palmer, Ensemble forecasting, *J. Comput. Phys.* 227 (7) (2008) 3515–3539.
- [20] Z. Toth, E. Kalnay, Ensemble forecasting at NCEP and the breeding method, *Mon. Weather Rev.* 125 (12) (1997) 3297–3319.
- [21] A.J. Majda, *Introduction to Turbulent Dynamical Systems in Complex Systems*, Springer, 2016.
- [22] C.K. Williams, C.E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2, MIT Press, Cambridge, MA, 2006.
- [23] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.
- [24] S. Kullback, Letter to the editor: the Kullback-Leibler distance, *Am. Stat.* (1987).
- [25] A. Majda, R.V. Abramov, M.J. Grote, *Information Theory and Stochastics for Multiscale Nonlinear Systems*, vol. 25, American Mathematical Soc., 2005.
- [26] M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of Machine Learning*, MIT Press, 2018.
- [27] F. Gilani, D. Giannakis, J. Harlim, Kernel-based prediction of non-Markovian time series, *Phys. D: Nonlinear Phenom.* 418 (2021) 132829.
- [28] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [29] V. Baladi, Linear response, or else, arXiv preprint arXiv:1408.2937, 2014.