A Data-Driven Statistical-Stochastic Model for Effective Ensemble Forecast of Complex Turbulent Systems

Di Qi\textsuperscript{a} and John Harlim\textsuperscript{b}

August 24, 2022

\textsuperscript{a} Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA. Email: qidi@purdue.edu

\textsuperscript{b} Department of Mathematics, Department of Meteorology and Atmospheric Science, Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802. Email: jharlim@psu.edu

Abstract

In this paper, we propose a system of statistical-stochastic equations to predict the response of the mean and variance statistics under perturbations of the initial condition and external forcing. 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 theory, 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 turbulent 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.

1 Introduction and background

Moment closure of nonlinear dynamical systems is a long-standing problem that has been widely studied in many fields of science and engineering [8, 20, 18]. In the context of low-order statistical closure problem, which is the primary interest in this work, predicting the time evolution of response statistics under external perturbations commands a wide range of applications in uncertainty quantification [7, 14, 15] and data assimilation [13, 19, 3]. For turbulent dynamical systems, the main challenge is to resolve the moment interaction induced by the bilinear quadratic form that represents the nonlinear advection of fluid flows. To be more precise, the system of 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 variable, 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 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 proposed for the unresolved subgrid scale terms [1, 21, 16]). In fact, for the second-order moment closure problem, machine learning with
LSTM architecture on homogeneous statistics has been proposed [17] to emulate the feedback from the third-order moments. In the homogeneous case, the dimension of the state variable is \((1 + N)\)-dimensional consisting of the one-dimensional mean variable and \(N\)-dimensional variance components of the diagonal covariance matrix. While the purely statistical closure models for the homogeneous case [17] produce accurate response statistical predictions under different forces and initial conditions, there are several issues emerging when the statistics are inhomogeneous. Beyond the practical issue of resolving the \(N \times N\) covariance matrix that is non-diagonal, constructing a machine learning model for the feedback from the third-order moment that preserves the positive-definite covariance remains a challenging problem, especially when the covariance dynamical equation is conditionally unstable.

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 \(N_1\), where \(N_1 < N\), the dynamical equation of the proposed statistical-stochastic equation is \(N_1(1 + M)\) dimensional, accounting for the \(N_1\)-dimensional vector for the mean and the \(M\) ensemble members of \(N_1\)-dimensional fluctuation equations. While the system is high-dimensional, especially since \(M \gg N_1\), the unresolved processes to be modeled in this formulation are only \(2N_1\)-dimensional, accounting for the modeling error in the \(N_1\)-dimensional reduced-order mean dynamics, and the unresolved fluxes in the \(N_1\)-dimensional fluctuation dynamics.

For a scalable and accurate statistical prediction on this systematic modeling framework, two issues need to be addressed. First, the internal instability in the fluctuation dynamics often leads to an unstable numerical scheme and fast divergence of the solution. This becomes a major obstacle for accurate long-term prediction since the unavoidable small errors from the data-driven approximation will be amplified in time rapidly. To overcome this issue, we will consider a semi-parametric framework with consistent variances, embedding the neural-network model on a parametric modeling framework proposed in [17, 14]. Second, while the ensemble structure in the proposed statistical-stochastic system is natural for ensemble prediction and data assimilation, the dimensionality of this fully coupled system \(N_1(1 + M)\) is often too high for accurate learning. In our numerical example, the state-space dimension is of order 1000 for a \(N_1 = 14\) dimensional reduced-order state variable. In this case, the standard learning procedure of fitting an empirical loss function that compares the trajectory of these states [4] may not be numerically feasible as it may lead to an enormously expensive computational cost since it requires a very large training data set. 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 a loss function, deduced from an empirical learning theory, 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.

Numerically, we will examine the stochastic-statistical formulation on the Lorenz-96 model [10] which admits the characteristic of the canonical equation for the turbulence 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 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 the 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 turbulent systems with uncertainty

In this section, we give a quick overview of a general moment closure formulation for the complex turbulent systems common in natural and engineering problems and develop efficient machine learning reduced-order models. 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 $u \in \mathbb{R}^N$ in a high-dimensional phase space,

$$\frac{du}{dt} = (L + D) u + B (u, u) + F + \sigma W.$$  \hspace{1cm} (1)

On the right hand side of the equation (1), the first two components, $(L + D) u$, represent linear dispersion and dissipation effects, where $L^* = -L$ is an energy-conserving skew-symmetric operator for dispersive effects; and $D < 0$ is a negative definite operator for dissipations. The nonlinear effect in the dynamical system is introduced through a quadratic form, $B (u, u)$, that satisfies the conservation law $u \cdot B (u, u) = \sum_{j=1}^N u_j B_j (u, u) := 0$ [14]. Besides, the system is usually subject to time-dependent external forcing effects that are decomposed into a deterministic component, $F(t)$, and a stochastic component represented by a Gaussian random process, $\sigma(t) W(t; \omega)$. It needs to be emphasized that in many situations $F$ might be inhomogeneous, and thus, introduce anisotropic structures into the system.

One way to characterize the effect of internal instabilities and the uncertainties from initial state and forcing in the turbulent system (1) is through a statistical description for the time evolution of the moment of the state variable $u$. 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 $p(u, t)$), it remains challenging to computationally solve such a PDE, especially when state space dimension, $N$, is large. The ensemble approach [9, 22], which uses an ensemble of solutions of (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 running the original model (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 multiscale 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 $u$, 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 $u$ 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 $\{ e_i \}_{i=1}^N$ as,

$$u(t, \omega) = \bar{u}(t) + u'(t; \omega) = \bar{u}(t) + \sum_{i=1}^N Z_i (t; \omega) e_i,$$  \hspace{1cm} (2)

where $\bar{u}(t) = \langle u(t) \rangle$ (here and after, we use $\langle \cdot \rangle$ to denote the statistical expectation about the PDF $p(u(t))$), represents the statistical expectation of the model state, i.e. the mean field; and $\{ Z_i (t; \omega) \}$ are mean-zero stochastic coefficients measuring the uncertainty in fluctuation processes $u'$ along each eigenmode direction $e_i$. 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 (1) and using the mean-fluctuation
decomposition (2), the evolution equation of the statistical mean state $\bar{u}$ is given by the following dynamical equation,

$$\frac{d\bar{u}}{dt} = (L + D) \bar{u} + B(\bar{u}, \bar{u}) + \sum_{i,j=1}^{N} R_{ij} B(e_i, e_j) + F,$$  

(3)

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

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

(4)

where $A_{ij}(\bar{u}) = [(L + D) e_j + B(\bar{u}, e_j) + B(e_j, \bar{u})] \cdot e_i$ characterizes the quasilinear coupling between the mean state $\bar{u}$ and the fluctuations $u' = \sum_i Z_i e_i$. The interactions between the fluctuation modes in different scales are summarized in the second term on the right hand side of (4) with the coupling coefficient $\gamma_{imn} = B(e_m, e_n) \cdot e_i$. Alternatively, from the stochastic equation (4) we directly obtain the exact evolution equation of the covariance matrix $R$,

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

(5)

where $A(\bar{u})$ is the same quasilinear operator from (4). $Q_F$ is the nonlinear energy flux that includes all the third moments $\langle Z_m Z_n Z_i \rangle$ for the higher-order feedbacks to the covariance dynamics, and $Q_\sigma$ is the contribution from the unresolved white noise forcing. Detailed expressions for the equation (5) can be found in [14].

As a further remark on this mean-fluctuation formulation of the original system, we could use either the stochastic equation (4) or the equivalent statistical covariance equation (5) to model the uncertainty in each fluctuation mode $e_i$. In fact, a data-driven statistical closure model combining (3) and (5) has been developed in [17] 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 (3) and (4) enjoys the advantage of more flexibility to run ensemble forecasts for both uncertainty quantification and data assimilation. In addition, this statistical-stochastic model can naturally estimate inhomogeneous statistics and avoids the main issue with the purely statistical formulation in (3), (5) in preserving the positive-definite covariance estimation.

### 2.2 A generic statistical-stochastic closure models 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 $\mathcal{I}$ are resolved, that is,

$$u^{\mathcal{I}} = \bar{u}^{\mathcal{I}} + \sum_{i \in \mathcal{I}} Z_i e_i,$$

(6)

where $\bar{u}^{\mathcal{I}} = \text{Pr}_{\mathcal{I}} \bar{u} = \sum_{i \in \mathcal{I}} \bar{u} e_i$ 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 $|I| = M \ll N$. Inspecting the coupling terms in the true dynamics (3) and (4), several difficulties will emerge in accurate modeling of the detailed coupling mechanisms in the constrained reduced-order representation (6). First, the high-order nonlinear coupling terms in the mean and fluctuation equation consist of the 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 $\bar{u}$ through $A_{ij}(\bar{u})$ poses a challenge for the marginally stable scheme in equilibrium to accurately model 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 challenging.
2.2.1 Effective closure models for the mean and fluctuations

First, we introduce the reduced statistical mean equation by projecting the full equation (3) to the resolved low-dimensional subspace

$$\frac{d\bar{u}^Z}{dt} = (L + D) \bar{u}^Z + \text{Pr}^Z B (\bar{u}^Z, \bar{u}^Z) + \sum_{i,j \in I} R^Z_{ij} \text{Pr}^Z (e_i, e_j) + F^Z + \Theta^m. \quad (7)$$

In the above equation, only the projected dynamics in the constrained subspace are resolved. We introduce the additional unresolved process $\Theta^m$ to model the unresolved mean feedback from the nonlinear mean interactions $B (\bar{u}, \bar{u})$ and the combined feedback from the unresolved covariance $R_{ij}$. Various statistical closure strategies have been developed [14, 12] 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 (4). Again we concentrate on modes in the subset $I$ as in (6). Let $Z^I = \text{Pr}^Z Z = \{Z_i\}_{i \in I}$ be the vector of the resolved fluctuation modes. Similar to the statistical mean closure (7), we propose to construct projected dynamics for the resolved modes with the unresolved feedback learned from a properly designed data-driven model. Therefore, the resulting reduced-order fluctuation equation for the stochastic coefficients $Z$ becomes

$$\frac{dZ^I}{dt} = A (\bar{u}^Z) Z^I + \sigma \hat{W} \cdot e_i + \Theta^v. \quad (8)$$

Notice that the quasilinear coefficient $A_{ij} (\bar{u}) = [(L + D) e_j + B (\bar{u}, e_j) + B (e_j, \bar{u})] \cdot e_i$ for $i, j \in I$ includes the mean-fluctuation interaction leading to inherent internal instability for turbulent dynamics (that is, with positive eigenvalues in $A (\bar{u})$). To have a stable scheme for the stochastic model, we introduce a more detailed parameterization for the unresolved process as

$$\Theta^v = -DZ^I + \Sigma \hat{W}, \quad (9)$$

where the white noises $\hat{W}$ are independent to $\hat{W}$. In (9), the parameter $D$ is introduced to suppress the instability induced by the high-order nonlinear flux, while the parameter $\Sigma$ is to account for the energy source from the nonlinear exchange of energy. The effective decomposition in (9) generalizes the idea in the statistical closure model for nonlinear energy mechanism [17, 14]. Here, effective parameters $D, \Sigma$ will be constructed by fitting to the consistent covariance statistics. Especially, according to (5), which is derived using Itô’s Lemma, the corresponding covariance equation for the stochastic coefficients (8) is given by

$$\frac{dR^I}{dt} = A (\bar{u}^Z) R^I + R^I A^* (\bar{u}^Z) + Q^I_\sigma + Q^I_F,$$

where $R^I = \langle Z^I (Z^I)^* \rangle$ is the covariance matrix of the resolved fluctuation modes, and $Q^I_F$ is the nonlinear flux induced by the coupling from different stochastic coefficients and the truncation error. For the choice of parameterization in (9), the covariance consistency requires

$$Q^I_F = -DR^I - R^I D^* + \Sigma \Sigma^*, \quad (10)$$

to be satisfied. While such a choice of parameterization is ideal, it is difficult to numerically find $D$ and $R$ that satisfies (10) as the covariance $R^I$ is a time-dependent variable. To avoid such a practical issue, we consider fitting to the equilibrium statistics,

$$Q^I_F \approx -DR^I_{eq} - R^I_{eq} D^* + \Sigma \Sigma^*, \quad (11)$$

where $R^I_{eq}$ denotes the stationary covariance statistics of the resolved modes in $I$ that can be empirically estimated. To achieve further consistency in the time-dependent state away from equilibrium, we decompose the nonlinear flux into a positive and a negative definite component $Q^I_F = (Q^I_F)^+ - (Q^I_F)^-$. Then the effective damping and noise matrix can be approximated accordingly by fitting the negative and positive define components, respectively, as

$$D \approx D^M := \frac{1}{2}(Q^I_F)^- (R^I_{eq})^{-1}, \quad \Sigma \Sigma^* \approx \Sigma^M (\Sigma^M)^* := (Q^I_F)^+. \quad (12)$$

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, (7)-(9) with the approximate coefficients $D^M$ and $\Sigma^M$ in (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 (7) and the stochastic equation (8) for the fluctuation modes. The resolved mean state $\bar{u}^T$ enters the fluctuation equation (8) through the quasilinear term $A(\bar{u}^T)$. Especially, it induces positive growth rate among the unstable modes. Inversely, the statistical mean equation (7) depends on the covariance feedback from the resolved modes $R^T$, which will be empirically estimated by a Monte-Carlo average over an ensemble of solutions of the fluctuation equation (8). Denoting the ensemble solutions of the fluctuation equation (8) as $\{Z^{M,(i)}\}_{i=1,...,M}$ the second-order moment can be estimated empirically as,

$$R^T = \langle Z^T (Z^T)^* \rangle \approx R^M := \frac{1}{M-1} \sum_{i=1}^{M} Z^{M,(i)} (Z^{M,(i)})^*.$$  

With this empirical estimation, the general reduced-order statistical-stochastic closure model is given as,

$$\frac{d\bar{u}^M}{dt} = (L + D) \bar{u}^M + Pr_ZB(\bar{u}^M, \bar{u}^M) + \sum_{i,j \in I} R_{ij}^M Pr_ZB(e_i, e_j) + F^T + \Theta^m$$

$$\frac{dZ^M}{dt} = A(\bar{u}^M) Z^M + \sigma \tilde{W} \cdot e_i - D^M Z^M + \Sigma^M \tilde{\tilde{W}},$$  

where $D^M$ and $\Sigma^M$ will be parameterized following the approximation in (12) such that $Q_f^M$ in (10) is approximated by the flux model,

$$Q^M := -D^M R^T - R^T (D^M)^* + \Sigma^M (\Sigma^M)^*.$$  

Subsequently, the data-driven closure is adopted by fitting the standard Long Short Term Memory (LSTM) models to learn the unresolved terms $\Theta^m$ and $Q^M$. In Section 3, we will provide specific examples of (14) induced by the moment closure of homogeneous and inhomogeneous turbulence dynamics.

### 2.2.3 Empirical loss functions based on information theory

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. In fact, we demonstrate numerically in Appendix A that fitting the stochastic components of (14) directly to the pathwise trajectory of (8) given the true mean $\bar{u}^T$ often leads to overfitting, thus does not provide accurate statistical prediction when testing on new inputs. However, fitting to the mean and variance statistical responses corresponding to the same trajectory solutions (that lead to overfitted model) produces closure models with accurate statistical predictions on new inputs.

Particularly, we consider the following practical metric based on Kullback-Leibler divergence [6, 11] of two empirical measures induced by the response mean and variance statistics of the underlying dynamics in (1) and the reduced-order model in (14), respectively. Let $\delta \bar{u} := \bar{u}_s - \bar{u}_{eq}$ and $\delta R := R_\delta - R_{eq}$ be the response mean and covariance statistics of the underlying dynamics in (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, $\bar{u}_{eq}$ 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 (1). Analogously, we also define $\delta \tilde{u}^M := \tilde{u}_s^M - \tilde{u}_{eq}^M$ and $\delta R^M := R_\delta^M - R_{eq}^M$ as the corresponding statistical responses of the reduced-order model in (14), where $D$ and $F^T$ are perturbed by additional damping and forcing of amplitude $\delta$. Assuming that the perturbed distributions vary smoothly under parameter $\delta$ and denoting $\text{diag}(C)$ as the diagonal matrix whose diagonal entries are $C_{i,i}$, the KL-divergence between Gaussian measures $\pi_\delta = \mathcal{N}(\bar{u}_{eq} + \delta \bar{u}, \text{diag}(R_{eq} + \delta R))$ and $\pi_\delta^M = \mathcal{N}(\bar{u}_{eq} + \delta \tilde{u}^M, \text{diag}(R_{eq} + \delta R^M))$ can be written as,

$$\text{KL}(\pi_\delta, \pi_\delta^M) = \text{KL}(\pi_{eq}, \pi_{eq}^M) + \frac{1}{2} \sum_{k \in I} \sum_{i,k} R_{eq,k}^{-1} (\delta \tilde{u}_k - \delta \tilde{u}^M_k)^2 + \frac{1}{2} \sum_{k \in I} \sum_{i,k} R_{eq,k}^{-2} (\delta R_k - \delta R^M_k)^2 + O(\delta^3),$$  

where $\delta \bar{u}_i$ and $\delta \tilde{u}_i^M$ denote the $i$-th component of the mean responses $\delta \bar{u}$ and $\delta \tilde{u}^M$, respectively, and $\text{diag}(R_{eq,i}, \delta R_{eq,i})$ and $\text{diag}(R_{eq,i}^M, \delta R_{eq,i}^M)$ denote the $i$th component of the matrices $R_{eq}, \delta R$, and $\delta R^M$, respectively. The choice of fitting only the
diagonal entries of the response statistics is motivated by the practical difficulties in the empirical estimation of the non-diagonal entries of $R$, especially if the covariance values to be estimated is small and the number of samples for an empirical estimation is small.

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

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

(16)

which measures the signal and dispersion contribution at discrete time indices $\{t_j = j\Delta t\}_{j=0,\ldots,T}$ 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 $\theta$, denoting parameters in the class of models used to approximate, $\Theta^m$ and $Q^M$. When a neural-network model is used, then $\theta$ corresponds to the neural-network model parameters.

From the classical supervised learning perspective, the loss function (16) compares the label,

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

(17)

the statistical responses of the underlying dynamics in (1), to the predicted label,

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

(18)

the statistical response induced by the reduced-order model in (14). For convenience of the discussion, let us define the operator $M$ as $\hat{y}^M = M(x)$, where $x$ denotes the initial conditions of (14) (which will include the appropriate inputs for $\Theta^m$ and $Q^M$). We will specify the input variable $x$ in Section 3.4 corresponding to a specific reduced-order model accounting for the inputs of $\Theta^m$ and $Q^M$. With this notation, we write the loss function $L(\theta) := L(\theta, y, \hat{y}^M)$ to emphasize its dependence on the label (17) and predicted label (18). The supervised machine learning training corresponds to minimizing the following empirical risk function,

$$R_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} L(\theta, y_i, \hat{y}^M_i),$$

(19)

which is an empirical average of the loss function over $n$ training data $(x_i, y_i)_{i=1,\ldots,n}$. Here, we should emphasize that $\hat{y}^M_i = M(x_i)$ is the predicted label correspond to the input $x_i$. In Section 3.4, we will specify the input variable $x$ of a specific example of (14) and provide a pseudo-algorithm to evaluate the operator $M$. 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 models to capture leading statistics, we display the detailed construction of the general modeling framework 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 perturbations. Especially in modeling systems with strongly turbulent dynamics, a crucial issue is to construct a stable numerical scheme that can deal with 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 turbulent dynamics. Particularly, we consider the 40-dimensional Lorenz ’96 (L-96) system [10] of state variables $u = (u_1, u_2, \ldots, u_N)^T$ with general inhomogeneous damping and forcing,

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

(20)
This ODE system is defined with periodic boundary condition mimicking geophysical waves in the mid-latitude atmosphere. Various representative statistical features can be generated in the model (20) comparable with the real data from observations. Especially, inhomogeneous processes are introduced by the spatially varying damping and forcing effects $d_j$ and $f_j$. 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 (1) above, we can write the linear and quadratic operators for L-96 system as

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

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

We aim to deduce moment closure equations for (20) that include inhomogeneous structures in the statistical mean and stochastic fluctuation modes. In order to achieve this, we project the general inhomogeneous forcing/damping and model state onto each spectral mode such that

$$f_j = \hat{f}_0 + \sum_{k \neq 0} \hat{f}_k e^{i2\pi k \frac{m}{N}}, \quad d_j = \hat{d}_0 + \sum_{k \neq 0} \hat{d}_k e^{i2\pi k \frac{m}{N}}$$

$$u_j(t, \omega) = \bar{u}_j(t) + \sum_{k = -N/2+1}^{N/2} Z_k(t; \omega) e^{i2\pi k \frac{m}{N}}. \quad (21)$$

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 in the decomposition in (21), 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,

$$\bar{u}_j(t) = \bar{u}_0(t) + \sum_{k = -N/2+1}^{N/2} \hat{u}_k(t) e^{i2\pi k \frac{m}{N}},$$

where $\hat{u}_k(t)$ corresponds to the $k$-th Fourier mode of $\bar{u}_j$. If $\bar{u}_j = \bar{u}$ is spatially homogeneous, then the zeroth mode $\bar{u}_0(t)$ 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 (20) to different spectral modes, we obtain the statistical mean equation for the homogeneous and inhomogeneous components

$$\frac{d\bar{u}_0}{dt} = -\bar{d}\bar{u}_0 - \sum_{k \neq 0} d_{0,k} \bar{u}_k + \hat{f}_0 + \sum_{|k| \leq N/2} (|\bar{u}_k|^2 + \langle |Z_k|^2 \rangle) \gamma_k, \quad (22a)$$

$$\frac{d\hat{u}_k}{dt} = -\sum_{|m| \leq N/2} d_{k,m} \hat{u}_m + \hat{f}_k + \sum_{|m| \leq N/2} (\langle \hat{u}_m \hat{u}_{k-m} + \langle Z_m \hat{Z}_{k-m} \rangle \rangle \gamma^*_m e^{-i2\pi \frac{m}{N}}), \quad (22b)$$

with the uniform damping rate $\bar{d} = \frac{1}{N} \sum_j d_j$, and the damping coefficients for each inhomogeneous mode $d_{k,m} = \frac{1}{N} \sum_j d_j e^{i2\pi (m-k) \frac{m}{N}} = \hat{d}_{k-m}$. The nonlinear coupling between different scales is connected by the coefficient $\gamma_k = e^{-i\pi \frac{k}{N}} - e^{i2\pi \frac{k}{N}}$. Notice that the first equation (22a) only contains homogeneous dynamics (no cross-correlation between different wavenumbers $k$). In addition to the homogeneous mean mode $\bar{u}_0$, 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 (22) from the original equation (20) and subsequently projecting it to each spectral mode. Following these steps, we have the
The governing equation for the stochastic coefficients $Z_k$ as,
\[
\frac{dZ_k}{dt} = -\sum_{|m| \leq N/2} d_{k,m} Z_m + \sum_{|m| \leq N/2} \mu_{k,m} \hat{u}_{k-m} Z_m + \sum_{|m| \leq N/2} \left( Z_m Z_{k-m} - \langle Z_m Z_{k-m} \rangle \right) \gamma^*_m e^{i2\pi \frac{m}{N}},
\]
with the coupling coefficient $\mu_{k,m} = e^{i2\pi \frac{m-k}{N}} + e^{i2\pi \frac{2m-N-k}{N}} - e^{i2\pi \frac{m-N-k}{N}} - e^{i2\pi \frac{-k-m}{N}}$. In the right-hand-side of (23), 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 $R_{kl} = \langle Z_k Z_l^\ast \rangle$ according to (5)
\[
\frac{dR_{kl}}{dt} = -2dR_{km} - \left( \gamma_{-k} + \gamma^*_{-k} \right) \bar{u}_0 R_{kl} - \sum_{m \neq k} \left( d_{k,m} R_{ml} + d_{l,m}^\ast R_{km} \right) + \sum_{m \neq k} \left( \mu_{k,m} \bar{u}_{k-m} R_{ml} + \mu_{l,m}^\ast \hat{u}_{l-m}^\ast R_{km} \right) + \sum_{m \neq 0} \left( \langle Z_m Z_{k-m} Z_l^\ast \rangle \gamma^*_m e^{i2\pi \frac{m}{N}} + \langle Z_m Z_{l-m} Z_k^\ast \rangle \gamma_m e^{i2\pi \frac{k}{N}} \right).
\]

The homogeneous effects due to damping and mean interaction are summarized in the first row of (24). The inhomogeneous damping and mean interactions are shown in the second row of (24). 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 (24).

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. The same idea can be directly applied to the general abstract formulation described in (3) and (4) in comparison with the particular case (22) and (23).

### 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 (21) together with homogeneous initial perturbations $u_{0,j} := u_0$. In this case, the mean and fluctuation equations in (22) and (23) can be simplified as
\[
\begin{align*}
\frac{d\bar{u}}{dt} &= -\gamma \bar{u} + \sum_k \langle Z_k Z_k^\ast \rangle \gamma_k + f, \\
\frac{dZ_k}{dt} &= -\left( \gamma + \gamma_k \bar{u} \right) Z_k + \sum_{m \neq 0} Z_m Z_{k-m} \gamma^*_m e^{i2\pi \frac{m}{N}},
\end{align*}
\]
with $\gamma_k = e^{-i\frac{2\pi k}{N}} - e^{i\frac{2\pi k}{N}}, \bar{u} = \bar{u}_0, \bar{d} = \gamma$, and $R_k := R_{kk} = \langle Z_k Z_k^\ast \rangle$ denotes the variance of the stochastic coefficient $Z_k(t; \omega)$. Under the homogeneous statistics, the statistical mean state becomes a scalar and the covariance matrix becomes diagonal, that is, 
\[
\bar{u}_j = \bar{u} = \bar{u}_0, \quad \bar{u}_k := 0 \neq 0, \quad \text{and} \quad R_{kl} = R_{k0} \delta_{kl}.
\]

Thus we do not need to consider the inhomogeneous mean equation (22b) involving $\hat{u}_k$ and the cross-correlations between different spectral mode $\langle Z_k Z_l^\ast \rangle, k \neq l$. On the other hand, nonlinear dynamics and non-Gaussian statistics still play a central role due to the strongly coupled feedbacks in equations (25). 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 interactions. For example, in (22b) strong positive growth rate will be induced when $\bar{u}_0 = \bar{u} > 0$ for modes with $\Re \gamma_k < 0$. To illustrate this, we plot in Figure 3.1 the quasilinear growth rate $-\left( \gamma + \gamma_k \bar{u} \right)$ of each spectral mode in the L96 model, subjected to different initial perturbations (that we will described 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 external perturbations, occur intermittently, and thus, create a practical challenge for learning a stable and accurate model.
3.2.1 Direct modeling of the unresolved feedbacks

Using the hybrid statistical-stochastic model (25), it requires the computation of the statistical expectation \( R_k = \langle Z_k^M \rangle \) for the mean equation (25a) from the solution of the stochastic equation (25b). In practice, this can be achieved by ensemble simulation of the stochastic coefficients \( Z_k \). Here, we propose a reduced-order closure model for homogeneous statistics with unresolved higher-order feedback modeled by processes \( \Theta^m, \Theta_k^v \) to be learned directly from data

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

In (26), \( Z_k^{M(i)} = (Z_k^{M(i)}, k \in I, i = 1, \cdots, M \) is each (independent) ensemble member in the ensemble simulation of the fluctuation equations. Its statistical feedback in the mean equation is approximated by the empirical ensemble average in (13).

In the above model, we consider \( I \) to include modes with large variances (modes 7-13, see Figure 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 (26) with the full exact model (25), the process \( \Theta^m \) models the variance feedback to the mean dynamics among all the unresolved small-scale modes \( k \in \{-N/2 + 1, \ldots, N/2\} \). In addition, \( \Theta^m \) 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 \( \Theta_k^v \) approximates the total contribution from the nonlinear coupling among all the fluctuation modes.

3.2.2 Stable numerical scheme with effective damping and noise

First, to overcome the instabilities occur 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 calibration of each stochastic trajectory of \( Z_k^{(i)} \), we choose to only measure the error in the ensemble empirical statistics as in Section 2.2.2, so that the strong uncertainty in training for stochastic trajectories can be effectively avoided. Notice that for the homogeneous case, the dynamical equation for the covariance matrix \( R \) in (5) can be simplified. Particularly, \( R \) only has nontrivial diagonal components \( R_k \) that satisfied,

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

Figure 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.
where $Q_F$ is a diagonal matrix for the high-order statistical nonlinear fluxes. Following the decomposition in (11) on each diagonal component $Q_{F,k}$, we have,

$$Q_{F,k} = Q_{F,k}^+ - Q_{F,k}^- \approx \Sigma_k \Sigma_k^T - D_k R_k,$$  

(28)

Therefore, the following closure system of equations with explicit effective damping and noise are proposed:

$$\frac{d\bar{u}_k^M}{dt} = -\gamma \bar{u}_k^M + \sum_{i=1}^{M} \left( \frac{1}{M-1} \sum_{i=1}^{M} Z_k^{M,(i)} (Z_k^{M,(i)})^* \right) \gamma_k + \Theta^m + f,$$

$$\frac{dZ_k^{M,(i)}}{dt} = - \left( \gamma + \gamma_k \bar{u}_k^M \right) Z_k^{M,(i)} - D_k^M Z_k^{M,(i)} + \Sigma_k^M W_k^{(i)}, \quad k \in \mathcal{I}, \quad i = 1, \cdots, M,$$

(29)

which is an example of (14). One way to parameterize $D_k$ and $\Sigma_k$ such that they satisfy the flux consistency in (28) is to fit a time-dependent model, $Q_k^M$, for the time series of the flux,

$$Q_{F,k} \approx Q_k^M,$$

$$D_k^M = - \min \left\{ Q_k^M, 0 \right\} / R_{eq,k},$$

$$\Sigma_k^M = \sqrt{\max \left\{ Q_k^M, 0 \right\}}.$$  

(30)

We should point out that the component-wise decomposition (30) is only possible since the statistics is homogeneous, and thus, avoiding an expensive matrix decomposition to identify the positive and negative definite components, $(Q_F^T)^+$ and $(Q_F^T)^-$, respectively, that satisfy $Q_F^T := (Q_F^T)^+ - (Q_F^T)^-$ for non-diagonal matrix $Q_F^T$.

In section 3.4, we will specify the class of machine learning models to identify $\Theta^m_k$ 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 will 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 when inhomogeneous forcing and perturbation are introduced. For this case, we have the additional observations for the inhomogeneous equations (22) and (23):

- The homogeneous mean state $\bar{u} = \bar{u}_0$ 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 $|\bar{u}_k|^2$;

- The inhomogeneous mean modes $\hat{u}_k$ are subject to the cross interactions between the mean states $\bar{u}_0 \hat{u}_k$ and the cross-covariances $(Z_k Z_k^T)$;

- 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 Figure 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 $\bar{u}_0$ 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 result in 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 Figure 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 (22b).
Given these statistical features, we consider a diagonal closure model, measuring only the statistics in the inhomogeneous mean and diagonal variances. Particularly, we consider the dynamical closure equations for the homogeneous mean \( \bar{u}_0 \), inhomogeneous mean \( \hat{u}_k \), and the fluctuation modes \( Z_k \) corresponding to the resolved subset \( k \in \mathcal{I} \) as follows:

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

which is an example of (14). Here \( d_{k,m} \) are defined as in (22a)-(22b). The variance feedback \( R_k^M \) in the mean equation is defined as the diagonal component of (13) attained with \( M \) samples. Also, the feedbacks from the non-diagonal covariance entries are not computed explicitly in the second equation in (31) due to their relatively small amplitudes. Following the homogeneous case, we introduce \( \Theta_0^m, \Theta_k^m \) to account for 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 \( \Sigma_k^M \) are constructed exactly as in (30), except that the model \( Q_k^M \) in (30) is fitted to the higher-order statistical flux \( Q_{F,k} \) induced by the inhomogeneous dynamics for the variance component,

\[
\frac{dR_k}{dt} = - \sum_{m \in \mathcal{I}} (d_{k,m} R_{mk} + d_{k,m}^e R_{km}) - (\gamma_k + \gamma_k^e) \bar{u} R_k + Q_{F,k},
\]

replacing (27) of the homogeneous case. In this diagonal closure modeling, we should point out that the element-wise decomposition (30) can still be performed and thus avoiding the matrix decomposition \( Q_F^T := (Q_F^T)^+ - (Q_F^T)^- \) 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 \( R_{kk}, k \neq \ell \) that are not calibrated 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.

Figure 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.
3.4 Neural network model for the unresolved processes

To parameterize \( \{ \Theta^m, Q^M \} \) in (29) or \( \{ \Theta^m_0, \Theta^m_k, Q^M \} \) in (31), we consider using a non-Markovian closure model to learn these terms, following our previous works in modeling variance closure [17] and trajectory of partially observed discrete-time ergodic Markov chain [4]. 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 [4, 2]), the dimension of the theoretically justifiable observable in the current application is too high for a tractably implementable. Particularly, for the statistical-stochastic models in (29) whose state variable is \( M|I| + 1 \) dimensional, accounting \( M \) ensemble members, identifying a model that depends on \( L \)-time delay corresponds to estimating an \( (M|I| + 1)L \) dimensional map, which is a very high-dimensional estimation problem since \( M \gg 1 \) is needed for reasonably accurate ensemble estimations. Instead of following such an agnostic approach, we identify the input of the non-Markovian map to reflect some explicit expression of \( \{ \Theta^m_0, \Theta^m_k, Q^M \} \) as reported in (22). 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 [17], we will consider the class of residual network with vanilla LSTM architecture [5]. For the homogeneous model, the correction terms in both the mean equation (25a) and the fluctuation equations (25b) are approximated by a neural network with residual structure,

\[
\begin{align*}
\Theta^m(t_{i+1}) &= \Theta^m(t_i) + \text{LSTM}^m(\bar{u}(t_{i-L}; i), \{ R_k(t_{i-L}; i) \}, \Theta^m(t_{i-L}; i) ; \theta), \\
Q^M_k (t_{i+1}) &= Q^M_k(t_i) + \text{LSTM}^k(\bar{u}(t_{i-L}; i), \{ R_k(t_{i-L}; i) \}, \{ Q^M_k(t_{i-L}; i) \} ; \theta),
\end{align*}
\]

(33)

where we have used the notation \( a(t_{i-L}; i) := (a(t_{i-L}), a(t_{i-L+1}), a(t_i)) \) for any dependent variable \( a \) and \( \theta \) to denote the parameters in the LSTM network. Notice that the right-hand-side of both equations in (33) are \( (2|I| + 1)L \) dimensional maps, independent of the ensemble size \( M \). For the inhomogeneous case, the unresolved model parameters \( \{ \Theta^m_0, \Theta^m_k, Q^M_k \} \) can also be directly learned by fitting the LSTM neural networks with analogous residual structure, that is,

\[
\begin{align*}
\Theta^m_0(t_{i+1}) &= \Theta^m_0(t_i) + \text{LSTM}^m_0(\bar{u}(t_{i-L}; i), \{ \bar{u}_k(t_{i-L}; i) \}, \{ R_k(t_{i-L}; i) \}, \Theta^m_0(t_{i-L}; i) ; \theta), \\
\Theta^m_k(t_{i+1}) &= \Theta^m_k(t_i) + \text{LSTM}^k(\bar{u}(t_{i-L}; i), \{ \bar{u}_k(t_{i-L}; i) \}, \{ R_k(t_{i-L}; i) \}, \Theta^m_k(t_{i-L}; i) ; \theta), \\
Q^M_k(t_{i+1}) &= Q^M_k(t_i) + \text{LSTM}^k(\bar{u}(t_{i-L}; i), \{ \bar{u}_k(t_{i-L}; i) \}, \{ R_k(t_{i-L}; i) \}, \{ Q^M_k(t_{i-L}; i) \} ; \theta).
\end{align*}
\]

(34)

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 \( (3|I| + 1)L \)-dimensional time delay embedding maps.

The LSTM model parameters, \( \theta \), are attained by minimizing the empirical risk in (19) 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 \( y_i^M = M(x_i) \). In the following pseudo-algorithm, we provide the computational steps for evaluating \( y_i^M = M(x_i) \), where the input \( x \) corresponding to the statistical-stochastic reduced-order model in (31) is also stated precisely.

We remark that similar pseudo-algorithm is used for the homogeneous case, where the reduced-order model in (31) is replaced with (29) and the LSTM closure models in (34) is replaced with (33). In the homogeneous case, the input \( x \) does not have \( \{ \bar{u}_k \}_{k \in I} \).

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 turbulent systems with strong instability and non-Gaussian statistics. The inhomogeneous regime serves as a more challenging problem induced by nonlinear multiscale 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.
we sample the transient state statistics from only an initial perturbation of the ensemble samples in the following
\[ \bar{u} \] if efficient numerical integration with larger time step is used. Subsequently, we compute the empirical mean and
\[ \Delta \] the data sampling step.

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

Evaluating the predicted label
\[ y^M \] consists of \( \delta u_0^M, \delta u_k^M, \delta R_k^M \) for all \( k \in \mathcal{I} \) at times \( t_1, \ldots, t_T \).

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
(33) or (34) 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 \( t_\ell \) using the mean models (the first two equations
in (31)). Subsequently, we attain the response mean statistics \( \delta \hat{u}_0^M(t_\ell) := \hat{u}_0^M(t_\ell) - \hat{u}_{eq,0} \) and \( \delta \hat{u}_k^M(t_\ell) := \hat{u}_{\delta,k}^M(t_\ell) - \hat{u}_{eq,k} \), where \( \hat{u}_{eq,0}, \hat{u}_{eq,k} \) are the reference equilibrium mean;
- Update the effective damping and noise \( D_k^M, \Sigma_k^M \) at time \( t_\ell \) using the decomposition in (30) 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 (31) for each ensemble member
with the effective damping and noise corrections.
- Compute the empirical variances of the perturbed coefficients, \( R_k^{M,i}(t_\ell) = \frac{1}{M-1} \sum_{i=2}^{M} Z_{\delta,k}^{M,i}(t_\ell)(Z_{\delta,k}^{M,i}(t_\ell))^* \). Subsequently, we compute the response variance \( \delta R_k^M(t_\ell) := R_k^M(t_\ell) - R_{eq,k} \), where \( R_{eq,k} \) denotes the
equilibrium variance of the unperturbed system;
- Update \( \ell = \ell + 1; \)
end while

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

To generate the training data (or the label \( y \) in (17)), we first integrate the L-96 system under homogeneous
reference forcing \( F_{ref} = 8 \) and damping \( d_{ref} = 1 \). 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 \( \bar{u}_{eq} \) and \( R_{eq,k} \), over these discrete time realizations.

In order 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 in the following
form
\[ u^{(i)} := \alpha \bar{u}_{eq} + \sqrt{\beta} \left( u_{ref}^{(i)} - \bar{u}_{eq} \right), \quad i = 1, \ldots, M = 500, \] (35)

where \( \{ u_{ref}^{(i)} \}_{i=1}^{M} \) denotes a set of \( M = 500 \) random samples of the discrete solution of the L-96 system under
the reference damping and forcing \( F_{ref}, d_{ref} \). New initial ensembles are generated by perturbing the mean through
the parameter \( \alpha \) and the variance through the parameter \( \beta \). Figure 4.1 plots several realizations of the statistical
responses for the mean and variance subject to different perturbation parameters \( \alpha, \beta \). The converging trajectories
of the mean and variance also illustrate the decorrelation time that characterizes the mixing rate of the states. The
solutions will rapidly converge to the unperturbed equilibrium within the decorrelation time around \( T = 1.5 \). For
training, we will consider 6 different values for each \( \alpha \in [0.5, 1.5] \) and \( \beta \in [0.5, 1.5] \), 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 Figure 4.1). To increase the number of training data in the homogeneous case, we
consider 4 additional external constant forcings in addition to the reference forcing, \( F_{\delta} = F_{ref} + \delta F \), 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, \( \delta u \), 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
Effectively, these inhomogeneous forcings and dampings (see Figure 4.2) exerted a single Fourier mode data points of \( \{ \delta u(t) \} \) at each of times 1.1 units: \( \{ 0.1, 0.1, 0.2, \ldots, 0.9, 2 \} \). 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 Algorithm 1, we label these statistical timeseries as the quantities at \( t = -L, 0, 1, \ldots, 10 \), with \( L = 100 \). We will use the first \( L + 1 = 101 \) data points of \( \{ \bar{u}(t) \} \) to construct the input data \( x \). Particularly, we take FFT on \( \bar{u}(t) \) to attain \( \bar{u}(t) \) and \( \{ \hat{u}_k(t) \} : k \in I \) to attain an ensemble perturbation \( \{ Z_k^{(i)}(t) \} : k \in I, i = 1, \ldots, M \}. \) This completes the construction of an input \( x \) for each partition. Finally, we take the last 10 data points of \( \delta \bar{u}(t) \) and \( \{ \delta R_k(t) \} \) at \( t = 1, \ldots, 10 \) 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 \( (x_i, y_i) \) 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 \( \{ \delta F = 1.5 \sin(kx) \} \) and the damping is one of the four cases: reference damping \( \delta d = 0 \) or \( \{ \delta d = 0.5 \sin(kx) \} \). Effectively, these inhomogeneous forcings and dampings (see Figure 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 hyper-parameters of the LSTM network are summarized in Table 1. The empirical risk minimization task of (19) is performed using the stochastic gradient descent 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.

4.2 Training and prediction of the homogeneous statistical regime

First, we consider the homogeneous perturbation case using uniform perturbations in the forcing \( F = F_{eq} + \delta F e_0 \). Only the most energetic leading modes, \( I = \{ k : 6 \leq |k| \leq 12 \} \), 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 \( \bar{u} = \bar{u} e_0 \) and the diagonal variance in resolved mode \( R_k = \frac{1}{M} \sum_i Z_k^{(i)}(Z_k^{(i)})^* \) based on the ensemble solutions.
| 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 | \( \Delta t \) 0.01 |
| LSTM sequence length | \( L \) 100 |
| forward prediction steps in training | \( T \) 10 |
| LSTM hidden state size | \( h \) 100 |

Table 1: Hyper-parameters for training the standard LSTM neural network model

![Training loss and MSEs for leading statistics](image1)

(a) Errors in mean and total variance during training iterations

![Training MSEs for unresolved fluxes](image2)

(b) Errors in the unresolved fluxes during training iterations

Figure 4.3: Training errors using the reduced-order model (29) 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 \( \Theta^m \) and \( Q^M_k \).

We first show the evolution of the errors during the training iterations in Figure 4.3(a). The first row plots the values of the empirical risk function (19), where the training errors in the predicted mean, \( \bar{u}^M \), and variance, \( R^M_k \), 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 MSEs of the homogeneous mean \( \bar{u} \) and the total variance of resolved modes \( \text{tr} R^M = \sum_{k \in I} R^M_k \) can be both effectively minimized to very small values. 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 \( \Theta^m \) and \( Q^M_k \) in (33) that are not directly compared in the loss function. In this case, we found that the error in \( \Theta^m \), which is not measured directly in the loss function, decays. On the other hand, the discrepancy between the model constructed flux, \( Q^M_k \), and the statistical flux, \( Q_{F,k} \), actually increases. This is not a surprise since we used the decomposition in (28) to determine \( D^M_k \) and \( \Sigma^M_k \). Recall that this parameterization uses the equilibrium variance \( R_{eq,k} \) to avoid the elaborate computational cost induced by fitting to the more ideal time-dependent variance, \( R_k \), as suggested in (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 $i$th step, 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 Figure 3.1. The prediction errors in the mean state and variance from the empirical ensemble average are plotted in the first row of Figure 4.4. The errors in 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 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 numerical scheme 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 Figure 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. 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 Figure 4.5. Consistent with the prediction errors in Figure 4.4, the evolution of the mean and variances starting from three pair 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
Figure 4.5: Prediction of the statistical mean and variance using the trained reduced-order statistical-stochastic model for homogeneous statistics. The upper panel plots the predictions of the mean and total variance with different ensemble sizes $M = 500, 100, 50$. 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. The detailed comparison for the variance in each resolved mode is plotted in the lower panel.
to these wave numbers as shown in Figure 4.2.

As in the previous section, we first display the training results using the reduced-order model (31) 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 $0 \leq |k| \leq 12$. The evolution of the errors during training iterations is displayed in Figure 4.6. Similar to the homogeneous case, the errors can be effectively minimized within the 100 training epochs in both the homogeneous mean $u_0 = \bar{u}$ 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 discrepancy is due to the use of coefficients $D_k^M$ and $\Sigma_k^M$ attained by equilibrium fitting in (30) as a way to realize the decomposition in (28).

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 prediction 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 Figure 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 Figure 4.2, for new perturbations on the initial mean and sample variance as in (35). 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 Figure 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 Figure 4.8 for the three different perturbation cases.

While all previous numerical experiments were focused to capture statistical response 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 $\delta F = 0.5 + 1.5 \sin(x)$ or the inhomogeneous leading mean modes $\delta F = f_1 + f_2 + f_3$, with $f_k = \sin(kx) + \cos(kx)$, where these additional constant mean forces are not in the training data set. Figure 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 the dynamics of covariance in (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. 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 relatively long time before the error starts to accumulate in time.
MSEs for long time prediction of statistical inhomogeneous mean & total variance

(a) prediction MSEs in different test cases

prediction of statistical inhomogeneous mean & total variance

(b) prediction of the inhomogeneous statistics in 3 initial perturbation cases (in 3 columns)

Figure 4.7: Upper panel: Prediction errors in the homogeneous mean $\bar{u}_0^M$, resolved inhomogeneous mean $\bar{u}_k^M = \sum_{k \in I} \hat{u}_k^M$ and total variance $trR^M = \sum_{k \in I} r_k^M$ 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.
Figure 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.

Figure 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.
5 Summary

In this paper, we proposed a generic statistical-stochastic closure model framework for effective ensemble prediction of leading order statistics in turbulent systems containing strong instability and multiscale interactions. The mean dynamic is modeled with a set of statistical equations that represents the homogeneous and inhomogeneous components subject to external perturbations. On the other hand, the fluctuation dynamics, which characterize the uncertainty among the multiscale modes, are modeled with a stochastic formulation. The mean dynamics are exerted by 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 ensemble simulation. Subsequently, machine learning tools are adopted to identify non-Markovian models for the nonlinear flux feedback from the unresolved processes and imperfect model errors. The coupled statistical and stochastic equations provide a flexible general framework for accurate prediction of the statistical responses in leading order mean and variance subject to various initial and external perturbation forms.

To overcome the inherent instability of the reduced-order fluctuation modes, we adopted the effective damping and noise parameterization to model the unresolved statistical flux that was originally proposed in [17, 14]. In this paper, we extended this parameterization with an LSTM neural-network model that is fitted to the unresolved flux such that the variance statistics are consistent. We numerically found that this approach is effective in stabilizing the numerical integration with accurate statistical prediction. To avoid the overfitting problem that often arises in the fitting of the high-dimensional fluctuation states with a limited training dataset, we train the neural network model by solving an empirical risk minimization of a loss function that compares the multistep outputs of the response mean and variances.

The skill of the reduced-order statistical-stochastic model is tested on several different statistical regimes of the L-96 system, including the homogeneous and inhomogeneous statistics produced by exerting different types of forcing and damping perturbations. The trained model displays uniformly high skill in capturing mean and variance responses among different statistical regimes and remains stable for a long time prediction even beyond the decorrelation time of the states. The proposed general modeling framework can be adopted for uncertainty quantification and data assimilation in high dimensional systems which usually requires a large ensemble size.

Acknowledgments

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.

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 trajectories of the reduced-order model. 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 (26)) can be numerically demanding, especially when \( M \) is large. In the following, we will conduct an experiment fitting only the stochastic component of (26) 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 (28) is not used.

Specifically, we parameterize \( \Theta_k^v \) in (26) by minimizing an empirical risk defined with the following loss function,

\[
L(\theta, Z^I, Z^M) := \sum_{k \in I} |Z_k - Z^M_k|^2, \tag{A1}
\]
where,
\[
\frac{dZ_k^M}{dt} = -\left(\gamma + \gamma_k \bar{u}\right) Z_k^M + \Theta^v_k,
\]
\[
(\Theta^v_k(t_{i+1}) = \Theta^v_k(T_i) + \text{LSTM}^m(\bar{u}(t_{i-L:}) , \{ R_k(t_{i-L:}) \} , \Theta^v_k(t_{i-L:}) ; \theta). \tag{A2}
\]

In this numerical experiment, the input data is
\[
x = \left( Z_k^M(t_i), \bar{u}(t_{i-L:}), R_k(t_{i-L:}), \Theta^v_k(t_{i-L:}) \right) \in \mathbb{R}^{(2|\mathcal{I}|+1)L + |\mathcal{I}|},
\]
and output is \( y = Z_k \in \mathbb{R}^{|\mathcal{I}|} \). In this homogeneous statistics case, we set \( \mathcal{I} = \{ k : 6 \leq |k| \leq 12 \} \) as in Section 4.2, which results in \( N_1 = |\mathcal{I}| = 14 \). Setting \( L = 100 \) as in Table 1, this learning problem is to find a \( (2|\mathcal{I}|+1)L + |\mathcal{I}| = 2914 \) 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 Figure A1. The loss and mean square errors (MSEs) in the coefficients \( Z_k \) and unresolved flux term \( \Theta^v_k \) during training iterations are shown in the first row of Figure A1. 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 Figure A1 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 \( T_{\text{decorr}} \). 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 (A2) 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 (A2) is used. In addition to these issues, the Monte-Carlo error induced by the empirical average in (13) will amplify difficulties which translates into a numerically expensive training procedure when the true \( \bar{u} \) is not available in which one has to also learn the unresolved term \( \Theta^m \) in (26).

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] Masataka Gamahara and Yuji Hattori. Searching for turbulence models by artificial neural network. Physical Review Fluids, 2(5):054604, 2017.

[2] Faheem Gilani, Dimitrios Giannakis, and John Harlim. Kernel-based prediction of non-Markovian time series. Physica D: Nonlinear Phenomena, 418:132829, 2021.

[3] John Harlim. Model error in data assimilation. In C. Franzke and T. O’Kane, editors, Nonlinear and Stochastic Climate Dynamics. Cambridge University Press, 2017.
Figure A1: 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.
[4] John Harlim, Shixiao W. Jiang, Senwei Liang, and Haizhao Yang. Machine learning for prediction with missing dynamics. *Journal of Computational Physics*, page 109922, 2020.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[6] Solomon Kullback. Letter to the editor: The kullback-leibler distance. *AMERICAN STATISTICIAN*, 1987.

[7] Cecil E Leith. Climate response and fluctuation dissipation. *Journal of Atmospheric Sciences*, 32(10):2022–2026, 1975.

[8] Marcel Lesieur. *Turbulence in fluids: stochastic and numerical modelling*, volume 488. Nijhoff Boston, MA, 1987.

[9] Martin Leutbecher and Tim N Palmer. Ensemble forecasting. *Journal of computational physics*, 227(7):3515–3539, 2008.

[10] Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

[11] Andrew Majda, Rafail V Abramov, and Marcus J Grote. *Information theory and stochastics for multiscale nonlinear systems*, volume 25. American Mathematical Soc., 2005.

[12] Andrew J Majda. *Introduction to turbulent dynamical systems in complex systems*. Springer, 2016.

[13] Andrew J Majda and John Harlim. *Filtering complex turbulent systems*. Cambridge University Press, 2012.

[14] Andrew J Majda and Di Qi. Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems. *SIAM Review*, 60(3):491–549, 2018.

[15] Andrew J Majda and Di Qi. Linear and nonlinear statistical response theories with prototype applications to sensitivity analysis and statistical control of complex turbulent dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(10):103131, 2019.

[16] Romit Maulik, Omer San, Adil Rasheed, and Prakash Vedula. Subgrid modelling for two-dimensional turbulence using neural networks. *Journal of Fluid Mechanics*, 858:122–144, 2019.

[17] Di Qi and John Harlim. Machine learning-based statistical closure models for turbulent dynamical systems. *Philosophical Transactions of the Royal Society A*, 380(2229):20210205, 2022.

[18] Di Qi and Andrew J Majda. Low-dimensional reduced-order models for statistical response and uncertainty quantification: Two-layer baroclinic turbulence. *Journal of the Atmospheric Sciences*, 73(12):4609–4639, 2016.

[19] Sebastian Reich and Colin Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

[20] Themistoklis P. Sapsis and Andrew J. Majda. Statistically accurate low-order models for uncertainty quantification in turbulent dynamical systems. *Proceedings of the National Academy of Sciences*, 110(34):13705–13710, 2013.

[21] Anand Pratap Singh, Shivaji Medida, and Karthik Duraisamy. Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils. *AIAA journal*, 55(7):2215–2227, 2017.

[22] Zoltan Toth and Eugenia Kalnay. Ensemble forecasting at ncep and the breeding method. *Monthly Weather Review*, 125(12):3297–3319, 1997.