PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Qi D, Harlim J. 2022 Machine learning-based statistical closure models for turbulent dynamical systems. *Phil. Trans. R. Soc. A* **380**: 20210205.
https://doi.org/10.1098/rsta.2021.0205

Received: 1 September 2021
Accepted: 13 December 2021

One contribution of 16 to a theme issue 'Data-driven prediction in dynamical systems'.

Subject Areas:
applied mathematics, mathematical modelling, computational mathematics

Keywords:
reduced-order model, non-Markovian closure, long-time statistical prediction, long-short-term-memory network

Author for correspondence:
Di Qi
e-mail: qidi@purdue.edu

Machine learning-based statistical closure models for turbulent dynamical systems

Di Qi[1] and John Harlim[2,3]

[1]Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
[2]Department of Mathematics, and [3]Department of Meteorology and Atmospheric Science, Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA

DQ, 0000-0002-9865-2327

We propose a machine learning (ML) non-Markovian closure modelling framework for accurate predictions of statistical responses of turbulent dynamical systems subjected to external forcings. One of the difficulties in this statistical closure problem is the lack of training data, which is a configuration that is not desirable in supervised learning with neural network models. In this study with the 40-dimensional Lorenz-96 model, the shortage of data is due to the stationarity of the statistics beyond the decorrelation time. Thus, the only informative content in the training data is from the short-time transient statistics. We adopt a unified closure framework on various truncation regimes, including and excluding the detailed dynamical equations for the variances. The closure framework employs a Long-Short-Term-Memory architecture to represent the higher-order unresolved statistical feedbacks with a choice of ansatz that accounts for the intrinsic instability yet produces stable long-time predictions. We found that this unified agnostic ML approach performs well under various truncation scenarios. Numerically, it is shown that the ML closure model can accurately predict the long-time statistical responses subjected to various time-dependent external forces that have larger maximum forcing amplitudes and are not in the training dataset.

This article is part of the theme issue 'Data-driven prediction in dynamical systems'.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

The closure problem in nonlinear dynamical systems is one of the most challenging tasks in computational statistics, see e.g. [1–6]. In the context of turbulent fluid flows, the closure problem has been studied for over a century dating back to Boussinesq's eddy viscosity hypothesis [7], where the goal is to describe the Reynold stress term (which is effectively a second-order statistic) as a function of the mean flow. In a nutshell, the underlying closure problem is to find a closed system that can describe the evolution of observable (such as low-order statistics), and by 'closure', the goal is to specify a map that allows one to untangle the dependence on unresolved variables (such as higher-order statistics). In the context of the low-order statistical closure problem, which is the primary interest in this work, predicting the time evolution of mean statistics is useful for point estimation, while predicting the time evolution of the covariance statistics has a wide range of applications, including uncertainty quantification [8–10] and data assimilation [11–13].

As machine learning (ML) becomes popular, finding such a 'closure' map can be formulated as a supervised learning task. With ML algorithms, one approximates the closure system by solving a regression problem on an appropriate hypothesis space, replacing the traditional approach of finding an analytical expression that can be very difficult in general. In the context of turbulent fluid flows, numerous neural network-based ML closure systems have been proposed (e.g. [6,14,15]). While the success of the estimation depends crucially on the choice of neural network architectures, a natural hypothesis space for modelling time series is the RNNs architecture. In the closure modelling applications, the Long-Short-Term-Memory (LSTM) [16], a special class of RNNs, has been shown to produce state-of-the-art accuracies in the prediction of high-dimensional time series [17–20].

Building on these empirical successes, we consider the LSTM-based neural network architecture for statistical closure modelling of turbulent dynamical systems. In this paper, we will examine the effectiveness of ML in uncovering the non-Markovian statistical model. In previous works [20,21], a closure model for predicting the trajectory of the observed state variables is constructed using a long time series of the corresponding observable. Despite the similarity to the closure modelling framework formulated in [20,21], the proposed statistical closure problem in this article is more challenging. In the present work, a unified model framework is proposed aiming to directly predict the leading-order statistical moments subjected to general external perturbations, with limited training data. Particularly, we will consider short-time transient statistical sequences for training. This consideration is motivated by practical issues (e.g. stiff numerical solver and large storage) in obtaining longer time series when the full-order model is multiscale and high-dimensional. Even in moderately low-dimensional problems, as we shall see in this paper when the perturbed dynamics correlation statistics are decaying, we only have short-time series of transient statistics that are informative for training. While the lack of training data makes the closure problems in this paper a stringent test for the ML algorithm, we are not only concerned to predict the evolution of the low-order statistics of the underlying unperturbed system. Our ultimate goal is to capture statistical responses subjected to unseen external forces, extending previous works [20,21] which only examined the accuracy of the unperturbed dynamical system on new initial conditions.

To achieve this goal, we assume that one can numerically simulate the full-order model in a short time window (as in many reduced-order modelling configurations, e.g. [9,22–25]) to generate a training dataset under pre-selected simple constant forcing functions and initial conditions. We will simulate this training dataset by a Monte–Carlo simulation. While this task can be expensive depending on the choice of integration scheme for solving the underlying full-order model, the length of the time integration to reach correlation time scales, and the sample size needed to achieve a robust statistical estimation, it only needs to be performed once for pre-selected constant external forcings. Subsequently, we validate the closure model by examining how well it can extrapolate beyond the training data to predict the statistical responses subjected to various new time-dependent forcing functions and initial conditions.

Numerically, we examine the ML closure on a simple test model, the Lorenz '96 (L-96) system, that was first introduced by Lorenz [26] as an idealization of atmospheric waves in midlatitude. While the model is simple, it carries some properties of realistic turbulent complex systems [27,28] such as the energy preserving advection-like nonlinear term, and a wide spectrum of unstable modes through the nonlinear coupling between states. Beyond the simplicity, which allows us to carry the numerical verification with moderate computational costs, our choice to investigate this case is largely motivated by the fact that closure models for a coupled system of the mean and covariance statistics have been well developed and improved in [4,9,28]. These parametric closure models, developed based on clever physical intuition, have demonstrated accurate statistical predictions. In such a configuration, we found that the ML-based model can produce accurate statistical responses (comparable to the parametric model) on moderate to large forcing amplitudes. Despite the effective prediction with parametric closure models in [9,28], the cost of calibrating the statistical modes throughout the entire spectrum can become very high as the dimensionality of the problem increases [5,10]. In addition to this practical problem, a more fundamental issue with parametric modelling is that the design of accurate closure models crucially depends on knowing the physics well enough, such as self-similarity or some structure of the underlying dynamics. As an example that illustrates this issue, we will compare parametric and ML closure models for only the mean statistics (no dynamical models for the variance are involved) in the simple L-96 example. In this scenario, we find that the agnostic ML framework can produce more accurate predictions, beating the parametric-based approach. This simple test suggests that the agnostic approach is easily portable for any truncation scenario. On the other hand, while the parametric modelling assumption [9,28] works well on the coupled system of the mean of covariance statistics, different parametric assumptions need to be considered for accurate closure of only the mean statistics.

The remainder of this paper is organized as follows. In §2, we discuss the general statistical closure modelling framework of turbulent dynamical systems using L-96 as a prototypical example and provide a hierarchy of low-order closure models. In §3, we provide details on the ML algorithm used to estimate the non-Markovian dynamical components. In §4, we present numerical results on the hierarchy of closure models introduced in §2. In §5, we close the paper with a summary.

2. Statistical closure of complex nonlinear systems

The general formulation of the turbulent dynamical systems [9,27] can be described by the canonical equations for the state variable $\mathbf{u} \in \mathbb{R}^N$ as

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = (\mathcal{L} + \mathcal{D})\mathbf{u} + B(\mathbf{u}, \mathbf{u}) + \mathbf{F}(t). \tag{2.1}$$

On the right-hand side of the above 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; and $\mathcal{D} < 0$ is a negative definite operator. The nonlinear effect in the dynamical system is introduced through a quadratic form, $B(\mathbf{u}, \mathbf{u})$, that satisfies the conservation law, $\mathbf{u} \cdot B(\mathbf{u}, \mathbf{u}) = 0$, and the Liouville property, $\mathrm{div}_u B(\mathbf{u}, \mathbf{u}) = 0$ [27].

Following [9,29], the dynamics of the statistical moments are constructed by representing the state space $\mathbf{u}$ as

$$\mathbf{u}(t) = \bar{\mathbf{u}}(t) + \sum_{i=1}^{N} Z_i(t)\mathbf{e}_i \quad \text{and} \quad R_{ij} = \left\langle Z_i Z_j^* \right\rangle \tag{2.2}$$

where $\bar{\mathbf{u}}(t) = \langle \mathbf{u}(t) \rangle$ represents the mean statistics, and the coefficients $\{Z_i(t)\}$ are fluctuation terms along the coordinates $\mathbf{e}_i$. In the above description, the notation $\langle \cdot \rangle$ is to denote the canonical statistical ensemble average that approximates the integral over the phase space at the limit of large ensemble size, following the standard notion in statistical mechanics [30]. Inserting the

representation in (2.2) to (2.1), one obtains a system of dynamical moments equations, where the first two moments satisfy

$$\frac{\mathrm{d}\bar{\mathbf{u}}}{\mathrm{d}t} = (\mathcal{L} + \mathcal{D})\bar{\mathbf{u}} + B(\bar{\mathbf{u}}, \bar{\mathbf{u}}) + \sum_{i,j} R_{ij} B(\mathbf{e}_i, \mathbf{e}_j) + \mathbf{F} \tag{2.3a}$$

and

$$\frac{\mathrm{d}R}{\mathrm{d}t} = L_v(\bar{\mathbf{u}})R + RL_v^*(\bar{\mathbf{u}}) + \theta, \tag{2.3b}$$

with

$$(L_v)_{ij} = [(\mathcal{L} + \mathcal{D})\mathbf{e}_j + B(\bar{\mathbf{u}}, \mathbf{e}_j) + B(\mathbf{e}_j, \bar{\mathbf{u}})] \cdot \mathbf{e}_i \tag{2.4a}$$

and

$$(\theta)_{ij} = \sum_{m,n} \langle Z_m Z_n Z_j \rangle B(\mathbf{e}_m, \mathbf{e}_n) \cdot \mathbf{e}_i + \langle Z_m Z_n Z_i \rangle B(\mathbf{e}_m, \mathbf{e}_n) \cdot \mathbf{e}_j. \tag{2.4b}$$

Here, the energy flux θ expresses nonlinear energy exchanges between different fluctuation modes due to the nonlinearity of the dynamics modelled through third-order moments. In general, such a representation gives rise to a non-closed system (possibly infinite-dimensional ODEs) as each moment equation is coupled to higher-order moments.

Despite the fact that the exact equations for the statistical mean (2.3a) and the covariance fluctuations (2.3b) are not a closed system, the total energy in the mean plus the total variance defined as $E = \frac{1}{2}\bar{\mathbf{u}} \cdot \bar{\mathbf{u}} + \frac{1}{2}\mathrm{tr}(R)$ satisfies the following scalar dynamical equation [9]

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \bar{\mathbf{u}} \cdot \mathcal{D}\bar{\mathbf{u}} + \mathrm{tr}(\mathcal{D}R) + \bar{\mathbf{u}} \cdot \mathbf{F}, \tag{2.5}$$

where $\bar{\mathbf{u}}$ and R are the exact solutions from the statistical equations. While the mean and covariance dynamics in (2.3a) and (2.3b) are not explicitly written in terms of E, we found that by allowing the unresolved components in (2.3b) to depend on E, one can achieve an effective non-Markovian closure model, especially for reduced-order model building upon the coupled system (2.3a), (2.3b), (2.5) in leading modes [28].

For the convenience of notation in the following discussion, we consider a discrete dynamical system induced by numerical integration of the coupled system (2.3a), (2.3b), (2.5), and a non-Markovian equation for the energy flux θ,

$$\left.\begin{aligned} &\bar{\mathbf{u}}_{i+1} = \mathcal{F}_1(\bar{\mathbf{u}}_i, R_i, \mathbf{F}_{i+1}), \\ &R_{i+1} = \mathcal{F}_2(\bar{\mathbf{u}}_i, R_i, \theta_i) \\ &E_{i+1} = \mathcal{F}_3(\bar{\mathbf{u}}_i, R_i, \mathbf{F}_{i+1}) \\ \text{and} \quad &\theta_{i+1} = \mathcal{G}(\bar{\mathbf{u}}_i, \ldots, \bar{\mathbf{u}}_{i-m+1}; R_i, \ldots, R_{i-m+1}; E_i \ldots, E_{i-m+1}; \theta_i, \ldots, \theta_{i-m+1}). \end{aligned}\right\} \tag{2.6}$$

Here, we have defined $\bar{\mathbf{u}}_i := \bar{\mathbf{u}}(t_i)$, $R_i := R(t_i)$, $E_i := E(t_i)$, $\theta_i := \theta(t_i)$, and $\{\mathcal{F}_j\}$ to denote the corresponding operators associated with the numerical integration of (2.3a), (2.3b), (2.5) for a suitable time step $\Delta t := t_{i+1} - t_i$. The operator $\mathcal{G}$ denotes a hidden non-Markovian model that maps the delay coordinates of variables $\{\bar{\mathbf{u}}, R, E, \theta\}$ to the energy flux θ at the next time step. We should point out that in the absence of external forces, the non-Markovian system in (2.6) is an exact representation (no approximation) of the corresponding temporal discretization of the full dynamical system in (2.1) in terms of $\{\bar{\mathbf{u}}, R, E, \theta\}$ that satisfies mild conditions of the delay embedding theorem [31,32]. To see this, one can employ the discrete Mori–Zwanzig formulation [20] to the full system (2.1) with a projection operator defined as the conditional expectation of the delay embedding coordinates of these observables, $\mathbb{E}[\mathbf{X}_{i+1}|\mathbf{x}_i, \ldots, \mathbf{x}_{i-m+1}]$ for some $m > 1$, where $\mathbf{X}_i$ denotes the random variable associated with the dynamical process $\mathbf{x}_i := (\mathbf{u}_i, R_i, E_i, \theta_i)$ (see §3 of [33] for such a derivation).

While one can, in principle, deduce the hidden dynamics $\mathcal{G}$, such a mathematical derivation is far from trivial even if the structure of the full dynamics in (2.1) is known. Following the idea in [20,21], we will use ML to approximate the hidden map $\mathcal{G}$ in an efficient way. Theoretically,

under the assumption that $\mathcal{F}_j, \mathcal{G}$ are uniformly Lipschitz, one can guarantee accurate solutions (in a strong sense) up to a finite time with an error bound that depends linearly on the total error of learning $\mathcal{G}$ (see theorem 3 in [20]). Numerically, we will consider a specific type of RNN known as the LSTM model for the estimation of $\mathcal{G}$, motivated by the robust numerical results on other closure problems reported in [20]. In fact, using the approximation theory of a two-layer neural network, the work in [34] shows that there exists an RNN closure model that gives the desired consistency up to a finite time.

To illustrate the approach, we focus on the Lorenz'96 (L-96) model [26] that fits into the general structure of (2.1). The L-96 model is a 40-dimensional ODE system with state variables $\mathbf{u} = (u_0, u_1, \ldots, u_{J-1})^\top$

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} = (u_{j+1} - u_{j-2})u_{j-1} - d(t)u_j + F_j(t), \quad j = 0, \ldots, J-1 = 39, \tag{2.7}$$

with a periodic boundary condition, $u_J = u_0$, mimicking geophysical waves in the mid-latitude atmosphere. While the model is rather simple, it carries representative properties of realistic complex systems with the energy preserving advection-like nonlinear terms, and the exchanges between the damping and forcing terms.

To compare with the abstract form (2.1), we can write the linear and quadratic operators for the L-96 system as

$$\mathcal{L} = 0, \quad \mathcal{D}(t) = \mathrm{diag}(-d_0(t), \ldots, -d_{J-1}(t)) \quad \text{and} \quad B(\mathbf{u}, \mathbf{v}) = \{u^*_{i-1}(v_{i+1} - v_{i-2})\}_{i=0}^{J-1}$$

and define the state variables in (2.2) with $\mathbf{e}_k := \{\mathrm{e}^{2\pi ik(j/J)}\}_j$ for $j = 0, \ldots, J-1$.

For simplicity, we consider uniform damping and forcing terms, $d(t)$ and $F(t)$ respectively, that are only functions of time and identical for any grid points $j = 0, \ldots J-1$. For an extensive test of the model prediction skill, we will consider several forcing functions with distinctive features (figure 1*b*). With this assumption, the first two moments can be further simplified to a uniform mean state, $\bar{\mathbf{u}}(t) = \bar{u}(t)(1, \ldots, 1)^{\mathrm{T}}$, and a diagonal covariance matrix, $R(t) = \mathrm{diag}(r_0(t), \ldots, r_{J/2}(t))$. The corresponding moment equations are given as

$$\frac{\mathrm{d}\bar{u}(t)}{\mathrm{d}t} = -d(t)\bar{u}(t) + \phi(t) + F(t), \tag{2.8a}$$

$$\frac{\mathrm{d}r_k(t)}{\mathrm{d}t} = -2[\Gamma_k\bar{u}(t) + d(t)]r_k(t) + \theta_k(t), \quad k = 0, 1, \ldots, \frac{J}{2} \tag{2.8b}$$

and

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = -2d(t)E(t) + F(t)\bar{u}(t), \tag{2.8c}$$

where we have defined the coupling coefficients $\Gamma_k = 1/J(\cos(4\pi k/J) - \cos(2\pi k/J))$, $r_{-k} = \langle Z_{-k}Z^*_{-k}\rangle = \langle Z_k Z^*_k\rangle = r_k$, the variance feedback ϕ to the mean equation, and the nonlinear flux θ_k in the variance equations

$$\phi = \sum_k r_k\Gamma_k \quad \text{and} \quad \theta_k = 2\sum_m \mathfrak{Re}\left\{\langle Z_m Z^*_{m+k} Z_k\rangle\,(\mathrm{e}^{-2\pi i(2m+k/J)} - \mathrm{e}^{2\pi i((m+2k)/J)})\right\},$$

respectively, with statistical energy conservation $\mathrm{tr}(\theta_k) = 0$. See appendix A in [28] for a detail derivation of these terms. A numerical discretization of the right hand sides of (2.8a)–(2.8c) gives an explicit example for the abstract operators $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ in (2.6).

(a) Mean-covariance closure model

Here, we will specify the closure model for the discrete dynamical system in the form (2.6) induced by the time discretization of (2.8a)–(2.8c). In §4, we will numerically validate the

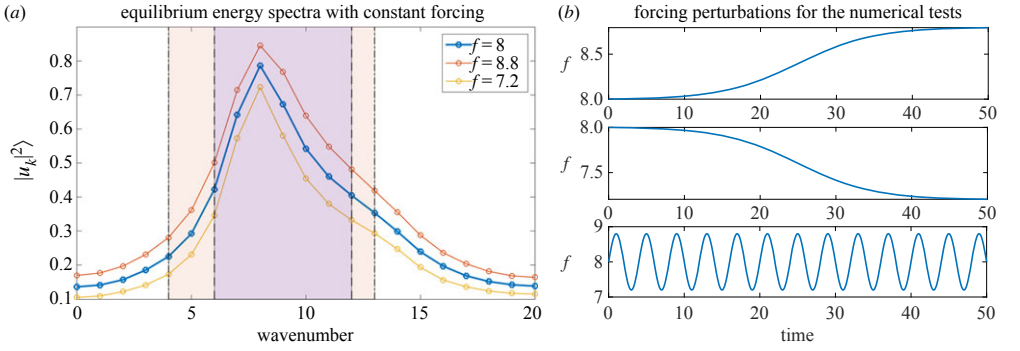

Figure 1. Direct Monte–Carlo simulation solutions of the 40-mode L-96 system as the standard test model. (*a*) The equilibrium energy spectra and the inner shaded area includes the resolved modes in the reduced-order model (2.10). Unstable modes span in a wider range $4 \leq k \leq 13$ than the resolved state in the reduced-order model. (*b*) Several external forces we will consider for testing the prediction skill. (Online version in colour.)

effectiveness of the ML strategy in recovering the dynamical maps $\mathcal{G}_k$ that model the evolution of the nonlinear flux θ_k that are missing in this formulation

$$\left.\begin{aligned} \bar{u}_{i+1} &= \mathcal{F}_1(\bar{u}_i, \{r_{k,i}\}_{k=0,\ldots,J/2}, F_{i+1}),\\ r_{k,i+1} &= \mathcal{F}_{2,k}(\bar{u}_i, r_{k,i}, \theta_{k,i}),\\ E_{i+1} &= \mathcal{F}_3(\bar{u}_i, E_i, F_{i+1})\\ \text{and}\qquad \theta_{k,i+1} &= \mathcal{G}_k(\bar{u}_i, \ldots, \bar{u}_{i-m+1}; \{\theta_{k,i}, \ldots, \theta_{k,i-m+1}\}_{k=0,\ldots,J/2}; E_i, \ldots, E_{i-m+1}) \end{aligned}\right\} \tag{2.9}$$

for $k = 0, \ldots, J/2$. Here, we should point out that $\bar{u}_i, r_{k,i}, E_i, \theta_{k,i}$ are all real-valued scalar variables, where we used subscript-i to denote the discrete time index. Here, we have adopted another simplification by ignoring the explicit dependence of $\mathcal{G}_k$ on $\{r_k\}$. This simplification is partly motivated by the implicit dependence of the variance information through E. Numerically, this simplification avoids the complexity in training the neural network model in approximating $\mathcal{G} := (\mathcal{G}_0, \ldots, \mathcal{G}_{J/2})$. In such a case, we should point out that $\{\mathcal{G}_k : \mathbb{R}^{(3+J/2)m} \to \mathbb{R}\}_{k=0,\ldots,J/2}$ is already high-dimensional when m is large, even without the explicit dependence on $\{r_k\}$.

(b) Reduced-order mean-covariance closure model

Next, we will consider a reduced-order model by truncating the summation term in (2.8a) to only account for leading modes $\mathcal{K} := \{k \in \mathbb{Z} : k_{\min} \leq k \leq k_{\max}\}$ that carry large variances (see the variances as functions of modes in figure 1*a* for various constant forcings). In our numerical experiment, we consider $k_{\min} = 6$ and $k_{\max} = 12$ such that the resolved subset $\mathcal{K}$ includes only the most unstable modes.

Specifically, the reduced-order model is given by a coupled system consisting of

$$\frac{\mathrm{d}\bar{u}(t)}{\mathrm{d}t} = -d(t)\bar{u}(t) + \sum_{k\in\mathcal{K}} r_k(t)\Gamma_k + \tilde{\phi}(t) + F(t),$$

and the dynamics of $\{r_k : k \in \mathcal{K}\}$ in (2.8b) and the dynamics of E in (2.8c). The key idea here is to consider a non-Markovian model (to be learned via appropriate ML algorithm) for the evolution

of the unresolved total variance feedback $\tilde{\phi} := \sum_{k\in\{0,\dots,J/2\}\setminus\mathcal{K}} r_k(t)\Gamma_k$. In discrete form, our task will be to learn the dynamical maps $\mathcal{G}_{k,1}$ and $\mathcal{G}_2$ of,

$$\left.\begin{aligned} \bar{u}_{i+1} &= \mathcal{F}_1(\bar{u}_i, \{r_{k,i}\}_{k\in\mathcal{K}}, F_{i+1}) + \tilde{\phi}_i, \\ r_{k,i+1} &= \mathcal{F}_{k,2}(\bar{u}_i, r_{k,i}, \theta_{k,i}), \\ E_{i+1} &= \mathcal{F}_3(\bar{u}_i, E_i, F_{i+1}), \\ \theta_{k,i+1} &= \mathcal{G}_{1,k}(\bar{u}_i, \dots, \bar{u}_{i-m+1}; \{\theta_{k,i}, \dots, \theta_{k,i-m+1}\}_{k\in\mathcal{K}}; E_i, \dots, E_{i-m+1}) \\ \text{and} \quad \tilde{\phi}_{i+1} &= \mathcal{G}_2(\bar{u}_i, \dots, \bar{u}_{i-m+1}; \tilde{\phi}_i, \dots, \tilde{\phi}_{i-m+1}; E_i, \dots, E_{i-m+1}), \end{aligned}\right\} \tag{2.10}$$

for $k \in \mathcal{K}$ using appropriate ML algorithms. With the reduced-order model (2.10), we only need to learn $\{\mathcal{G}_1 : \mathbb{R}^{(2+|\mathcal{K}|)m} \to \mathbb{R}^{|\mathcal{K}|}, \mathcal{G}_2 : \mathbb{R}^{3m} \to \mathbb{R}\}_{k\in\mathcal{K}}$, where $\mathcal{G}_1 := (\mathcal{G}_{1,k})_{k\in\mathcal{K}}$ and we have denoted the number of modes in $\mathcal{K}$ by $|\mathcal{K}| = k_{\max} - k_{\min} + 1$. Compared to the full-order closure model in (2.9), $\mathcal{G}_1$ is a lower-dimensional map which makes the computational cost less expensive when $|\mathcal{K}| < J/2$. In our numerical simulation, we will consider $\mathcal{K} = \{6, \dots, 12\}$ such that $|\mathcal{K}| = 7 < 21$ and the closure map to be recovered, $\mathcal{G}_1 : \mathbb{R}^{9m} \to \mathbb{R}^7$, is a much smaller dimensional map relative to that in the full-order model where $\mathcal{G}_1 : \mathbb{R}^{24m} \to \mathbb{R}^{21}$.

(c) Mean closure model

Finally, we will consider a closure model that ignores the detail evolution of the covariance terms r_k. In such a severe truncation scenario, we will introduce a non-Markovian closure for $\phi := \sum_{k=-J/2+1}^{J/2} r_k(t)\Gamma_k$ to account for the combined contribution of the truncated covariance terms in (2.8a). The corresponding discrete form is given as follows,

$$\left.\begin{aligned} \bar{u}_{i+1} &= \mathcal{F}_1(\bar{u}_i, F_{i+1}) + \phi_i, \\ E_{i+1} &= \mathcal{F}_3(\bar{u}_i, E_i, F_{i+1}) \\ \text{and} \quad \phi_{i+1} &= \mathcal{G}(\bar{u}_i, \dots, \bar{u}_{i-m+1}; \phi_i, \dots, \phi_{i-m+1}; E_i, \dots, E_{i-m+1}). \end{aligned}\right\} \tag{2.11}$$

Computationally, we only need to learn one map $\mathcal{G} : \mathbb{R}^{3m} \to \mathbb{R}$, which is a significant reduction compared to the previous models in (2.9) and (2.10).

(d) An important strategy for modelling unstable dynamics

In the full-order and reduced-order covariance models (2.9) and (2.10), neural network models will be constructed to update the variances r_k. One major challenge is the inclusion of strong inherent instability that is common among turbulent dynamical systems. For example, in the L-96 system, the covariance equation (2.8b) for r_k contains positive unstable modes with positive Lyapunov exponents if $-\Gamma_k\bar{u} > 0$. A lack of careful consideration in the detailed balance in unstable variance dynamics will lead to unbounded model divergence in the numerical verification. Particularly, an empirically trained neural-network model for the map $\mathcal{G}$ in (2.9) (or $\mathcal{G}_1$ in (2.10)) may not produce marginally stable dynamics that maintain accurate long-term stable forecasts.

To address this issue, we consider a more structural modelling, adopting the ideas in [9] by including an explicit nonlinear coupling term in the variance equation. To illustrate this, we modify the dynamical equation for $\mathcal{G}$ in (2.9) and (2.10) as follows: we decompose the higher-order nonlinear flux θ_k containing all the third moments in a (nonlinear) effective damping $d^M_{k,i+1}$

and noise $\sigma^M_{k,i+1}$ such that

$$\left.\begin{aligned} \theta_{k,i+1} &= -d^M_{k,i+1} r_{k,i} + \sigma^M_{k,i+1}, \\ Q^M_{k,i+1} &= \mathcal{G}_k(\bar{u}_i, \ldots, \bar{u}_{i-m+1}; \{\theta_{k,i}, \ldots, \theta_{k,i-m+1}\}_{k=0,\ldots,J/2}; E_i, \ldots, E_{i-m+1}) \\ d^M_{k,i+1} &= -\frac{\min\{Q^M_{k,i+1}, 0\}}{r_{k,\mathrm{eq}}} \\ \text{and} \qquad \sigma^M_{k,i+1} &= \max\{Q^M_{k,i+1}, 0\}. \end{aligned}\right\} \tag{2.12}$$

Here, the map $\mathcal{G}_k$ models the full nonlinear flux at each time instant and we employ an LSTM network in the next section to approximate $\mathcal{G} := (\mathcal{G}_0, \ldots, \mathcal{G}_{J/2})$. However, instead of directly setting $\theta_{k,i+1} = Q^M_{k,i+1}$, which gives the last equation in (2.9), we split the model output into two positive effective damping $d^M_k > 0$ and effective noise $\sigma^M_k > 0$. The effective damping is recovered from the unperturbed equilibrium statistics $r_{k,\mathrm{eq}}$. In this way, the unstable directions in the system are stabilized by the effective damping modelling the nonlinear transfer of energy without altering the detailed statistical balance in the equilibrium. One can see that if $Q^M_{k,i+1}$ is positive (that is, the mode is stable), then $\theta_{k,i+1} = Q^M_{k,i+1}$ and we retain the original model in (2.9).

3. Machine learning of the missing non-Markovian components

In this section, we briefly discuss how to employ the LSTM [16], a recurrent neural network, to learn the hidden non-Markovian maps in the proposed closure statistical models in (2.9)–(2.11). To simplify the discussion, let us identify the input variable (or covariate) with a sequence of correlated state variables $\{\mathbf{x}_j\}^i_{j=i-m+1}$ measured at m time instants ahead of the prediction time $i+1$ and the output (response) variable at discrete time index-$i+1$ as $\mathbf{y}_{i+1}$. In the case of (2.9), the input variable is $\mathbf{x}_j = \{\bar{u}_j, \theta_{0,j}, \ldots, \theta_{J/2,j}, E_j\}$ and the output variable is $\mathbf{y}_{i+1} = \{\theta_{0,i+1}, \ldots, \theta_{J/2,i+1}\}$. For (2.10), the input variable is $\mathbf{x}_j = \{\bar{u}_j, \{\theta_{k,j}\}_{k\in\mathcal{K}}, \tilde{\phi}_j, E_j\}$ and the output variable is $\mathbf{y}_{i+1} = \{\{\theta_{k,i+1}\}_{k\in\mathcal{K}}, \tilde{\phi}_{i+1}\}$. For (2.11), the input variable is $\mathbf{x}_j = \{\bar{u}_j, \phi_j, E_j\}$ and the output variable is $\mathbf{y}_{i+1} = \{\phi_{i+1}\}$.

Recurrent neural networks offer the desirable structure to incorporate temporal processes of sequential data with long temporal correlations and keep tracking of hidden processes. The LSTM network is designed to avoid the problem of vanishing gradients. The building block of LSTM is to consider the following model, which is known as an LSTM cell

$$\left.\begin{aligned} \mathbf{f}_i &= \sigma_g(W_f \mathbf{x}_i + U_f \mathbf{h}_{i-1} + V_f \mathbf{c}_{i-1} + \mathbf{b}_f), \\ \mathbf{I}_i &= \sigma_g(W_i \mathbf{x}_i + U_i \mathbf{h}_{i-1} + V_i \mathbf{c}_{i-1} + \mathbf{b}_i), \\ \mathbf{c}_i &= \mathbf{f}_i \otimes \mathbf{c}_{i-1} + \mathbf{I}_i \otimes \tanh(W_c \mathbf{x}_i + U_c \mathbf{h}_{i-1} + \mathbf{b}_c), \\ \mathbf{o}_i &= \sigma_g(W_o \mathbf{x}_i + U_o \mathbf{h}_{i-1} + V_o \mathbf{c}_i + \mathbf{b}_o) \\ \text{and} \qquad \mathbf{h}_i &= \mathbf{o}_i \otimes \tanh(\mathbf{c}_i). \end{aligned}\right\} \tag{3.1}$$

In (3.1), $\sigma_g = 1/1 + \mathrm{e}^{-x}$ is the sigmoid activation function, and $\otimes$ represents the element-wise product. The model cell includes forget, input and output gates $\mathbf{f}_i, \mathbf{I}_i, \mathbf{o}_i$, and the cell state $\mathbf{c}_i$. The hidden process $\{\mathbf{h}_{i-m+1}, \ldots, \mathbf{h}_{i-1}, \mathbf{h}_i\}$ represents the time-series of the unresolved process. In a compact form, let us denote the LSTM cell in (3.1) as $\mathbf{h}_{i+1} = \mathrm{Lc}(\mathbf{x}_i, \mathbf{h}_i)$, where we have suppressed the dependence on the parameters for simplicity.

The LSTM network is constructed from m LSTM cells Lc with the same structure and parameters $\mathbf{W}$. The cells are connected by the intermediate hidden state $\mathbf{h}_i \in \mathbb{R}^h$. Every LSTM cell takes in the input data $\mathbf{x}_i$ at the ith step and the output $\mathbf{h}_i$ from the previous adjacent cell, and gives out the inner hidden state $\mathbf{h}_{i+1}$ to be used for prediction of the next state. The full LSTM

chain is connected through m sequential cell structures, that is,

$$\mathbf{h}_m = \mathrm{Lc}^{(m)}\{\mathbf{h}_0; \mathbf{x}_{i-m+1}, \ldots, \mathbf{x}_i\} \equiv \mathrm{Lc}(\mathbf{x}_i) \circ \cdots \circ \mathrm{Lc}(\mathbf{x}_{i-m+1})(\mathbf{h}_0), \tag{3.2}$$

where the composition operator is defined with respect to the hidden state $\mathbf{h}_i$. In (3.2), the data at different time instance, $\mathbf{x}_i$, is fed into the corresponding LSTM cell, and the hidden state $\mathbf{h}_i$ is the output of the previous cell and input for the next cell. For simplicity, the initial value of the hidden state is often set as zero, $\mathbf{h}_0 = 0$. The final output $\mathbf{h}_m$ from the last step of the LSTM chain goes through a final single layer fully connected linear model given as

$$\hat{\mathbf{y}}_{i+1} = A\mathbf{h}_m + \mathbf{b}, \tag{3.3}$$

where $A \in \mathbb{R}^{d_y \times h}$, $\mathbf{b} \in \mathbb{R}^{d_y}$ are the model coefficients in the final layer and $d_y = \dim(\mathbf{y})$ denotes the dimension of the output variables. In our numerical implementation, for the reduced-order model in (2.10), we consider two LSTM networks, one for estimating $\mathcal{G}_1$ and another one for estimating $\mathcal{G}_2$. One can also consider separate LSTM networks for each component $\mathcal{G}_k$ in (2.9) (or $\mathcal{G}_{1,k}$ in (2.10)), which we do not pursue in our numerical experiments.

(a) Empirical loss functions

The neural network parameters $\mathbf{W} := \{A, \mathbf{b}, W_f, U_f, V_f, \mathbf{b}_f, W_i, U_i, V_i, \mathbf{b}_i, \ldots\}$ are obtained by solving a nonlinear non-convex optimization problem to minimize the difference between the training output data $\{\mathbf{y}_j^\ell\}_{\ell,j=1}^{n,M}$ and the LSTM output data $\{\hat{\mathbf{y}}_j^\ell\}_{\ell,j=1}^{n,M}$, subjected to the same input data $\{\mathbf{x}_j^\ell\}_{j=i-m+1,\ell=1}^{i,n}$, where n denotes the total number of training samples. There are many ways to design loss functions. Denote the true output data as $\mathbf{y}_j^\ell = \{\phi_j^\ell, \theta_{k,j}^\ell\}$ and the LSTM model output data as $\hat{\mathbf{y}}_j^\ell = \{\hat{\phi}_j^\ell, \hat{\theta}_{k,j}^\ell\}$, then we can, for example, consider the following empirical loss function:

$$\sum_{j=1}^{M}\left[\alpha\sum_{\ell=1}^{n}\left(\phi_j^\ell - \hat{\phi}_j^\ell(\mathbf{W}^\phi)\right)^2 + \sum_k \beta_k \sum_{\ell=1}^{n}\left|\theta_{k,j}^\ell - \hat{\theta}_{k,j}^\ell(\mathbf{W}^\theta)\right|\right], \tag{3.4}$$

where we have defined $\mathbf{W}^\phi$ and $\mathbf{W}^\theta$ to distinguish the parameters of the two network models. For the full-order mean-covariance model in (2.9), we set $\alpha = 0$ and minimize (3.4) for $\mathbf{W}^\theta$. For the reduced-order mean-covariance model in (2.10), we set $\alpha = 1$ and $\beta_k > 0$ such that they have comparable scales and minimize (3.4) for both $\mathbf{W}^\phi$ and $\mathbf{W}^\theta$. For the mean closure model in (2.11), we set $\alpha = 1$ and $\beta_k = 0, \forall k$, and minimize (3.4) for $\mathbf{W}^\phi$. While other choices exist, such as to include the error in the mean and variance components, we found the improvement is not significant. We should also point out that the empirical loss function in (3.4) is defined over a path of length-M, the model parameters are obtained by one minimization problem. In practice, we found that with $M = 10$, the resulting estimate yields more stable long-time predictions, compared to just setting $M = 1$ (for which one can solve separate minimization problems to obtain independent LSTM networks for $\mathcal{G}_1$ and $\mathcal{G}_2$, by fitting to one-step forecast data as employed in [20]).

(b) Small training dataset

While the general unperturbed underlying non-Markovian dynamics in (2.6) is an example of the missing dynamical model formulated in [20,21], the statistical configuration here is more challenging due to the shortage of informative training data that reflect the key features of the underlying dynamical process. In our numerical test problem (the L-96 model), the statistics are homogeneous such that the statistics of each of the solutions forced by a constant forcing will decay to a constant value in a short time (figure 2). To compensate for the lack of observed statistical data, in practice, we generate the training data by a direct short-time Monte-Carlo simulation following these steps:

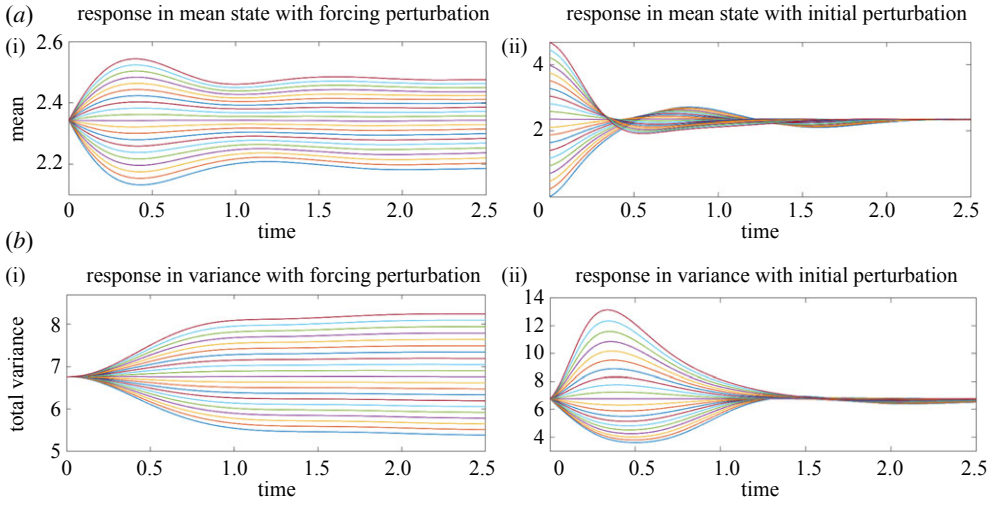

Figure 2. Statistical responses correspond to perturbations in external forcing (*a*(i),*b*(i)) and in the initial mean state (*a*(ii),*b*(ii)). Responses of the statistical mean and total variance with different perturbation amplitudes are shown. (*a*) Response in $\bar{u}$ (*b*) Response in tr(R). (Online version in colour.)

(i) Generate an ensemble of unperturbed equilibrium statistical solutions with the reference forcing $F = F_{\text{eq}}$. Each ensemble member solves an initial value problem corresponding to a randomly drawn initial condition from the standard Gaussian distribution.

(ii) Simulate an ensemble of solutions to the statistical steady state subjected to various constant external perturbations $F = F_{\text{eq}} + \delta f$. The ensemble of solutions at the final time from [i.] is used as initial conditions. The empirical mean and variance of these initial conditions correspond to the mean state $\bar{u} = \bar{u}_{\text{eq}} \approx 2.35$, and total variance, $\text{tr}(R) = \text{tr}(R_{\text{eq}}) \approx 6.8$, respectively (figure 2*a*(i),*b*(i)).

(iii) Simulate an ensemble of solutions to the statistical steady state corresponding to unperturbed constant external perturbation $F = F_{\text{eq}}$ and perturbed initial conditions $\bar{u} \to \bar{u} + \delta\bar{u}$. We perturb each ensemble member of the initial condition by adding a constant value $\delta\bar{u}$ to each ensemble member at the final time from [i.]. In figure 2*a*(ii),*b*(ii), one can see that the ensemble mean states at the initial time vary while the total variances at the initial time stay the same.

Typical statistical trajectories with different initial and forcing perturbation amplitudes are depicted in figure 2. The statistics of the perturbed states, which exhibit strong nonlinear coupling effects, decay to new (or original unperturbed) equilibrium states beyond the decorrelation time. Notice that for this problem, the decaying behaviour of these trajectories yields a small training dataset (in terms of temporal length). Beyond the transient time, the time series saturates and, thus, is not informative.

In general, even when long time statistics are informative (e.g. for non-trivial time-dependent external forces or non-stationary statistical dynamics), from a practical standpoint, attaining longer time series is computationally infeasible, especially when the dimension of the underlying state space is high and/or the system requires a stiff numerical solver. Thus, the configuration that we consider (training with an ensemble of short time series) in the present paper can be used for a wide class of high-dimensional systems when long time series are not accessible.

4. Numerical results

We now examine the effectiveness of the statistical model schemes discussed above with detailed numerical tests. We start with the full-order mean-covariance model (2.9) to learn the unresolved

high-order nonlinear flux directly from data. Second, the reduced-order mean-covariance model (2.10) is proposed for efficient computation with the most energetic leading modes. Finally, we show an even more efficient computation of the mean statistical prediction using only the mean closure model (2.11), focusing on the mean responses subjected to various forcing perturbations.

(a) Model configuration for training and prediction

In the training stage, the training data are generated from 41 response solutions (shown in figure 2) with either perturbed initial states $\delta\bar{u} \in [-\bar{u}_{\rm eq}, \bar{u}_{\rm eq}]$ or constant forcing perturbations $\delta f \in [-0.1F_{\rm eq}, 0.1F_{\rm eq}]$ from direct Monte-Carlo solutions of the L-96 system with an ensemble size 10 000. The true equation (2.7) is integrated with a fourth-order Runge–Kutta scheme with a small time step $\delta t = 0.001$, while the data are sampled at every 10 steps. Thus, we have the data sampling step $\Delta t = 0.01$. The training model is updated $M = 10$ times to account for the integrated error along the time integration. Notice that this choice of larger measurement step size leads to numerical discretization errors in computing the time integration and recovering the parameters of ϕ and θ_k. The total number of samples is $n = 1640$, and they are obtained by collecting non-overlapping time interval $M\Delta t = 0.1$ units from the statistical response trajectories. With such a small sample size, the learning problem is rather challenging as the neural-network model has a large number of parameters. While one can, of course, generate more data by additional perturbations and initial conditions, we will not pursue this direction since our goal is to understand the effectiveness of the agnostic ML model in such a stringent configuration with a small training dataset.

In the prediction stage, we verify the model performance by considering the long-time statistical prediction under a variety of time-dependent forcing scenarios that are not observed in the training dataset. For long-time prediction, the model output in the previous step is reiterated as an input in the next forecast stage, thus model errors accumulate in time. Therefore, it requires the closure models to be numerically stable in resistance to the accumulated model errors in the neural network model. In our numerical tests, we consider the ramp-type forcings and the periodic forcing as standard test examples where the large external perturbation is introduced to its equilibrium state forcing, $F_{\rm eq} = 8$ (see figure 1 for changes in the energy spectra for different forcing perturbations). In application, such testing configurations can be used to simulate the climate change scenario where the original state is driven away from its previous equilibrium state due to external perturbations [10,28] and other uncertainty quantification tasks.

In addition, a residual structure is adopted in the neural network for the closure models (2.9)–(2.11)

$$\theta_{i+1} = \mathcal{G} = \theta_i + \tilde{\mathcal{G}}, \tag{4.1}$$

where $\tilde{\mathcal{G}}$ denotes the LSTM network (3.2) to update the increment of the unresolved higher-order component. The LSTM chain contains $m = 100$ repeating cells with the same structure, taking a time sequence of time length $T = 1$ which is still shorter than the correlation time of the system (figure 2). The dimensions of the hidden states in LSTM are taken as $h_v = 50$ for the variance equation and $h_m = 10$ for the mean equation. The optimization for the loss (3.4) is carried out by the ADAM scheme. A total of 100 epochs is repeated during training, starting from the learning rate $\mathrm{lr} = 5 \times 10^{-4}$ which is reduced three times to half of its original value at the epoch numbers 25, 50 and 75.

(b) Prediction skill of the mean-covariance model

In this section, we numerically verify the prediction skill of the full statistical mean-covariance model (2.9). For clarity, we split the discussion into two subsections. First, we state the concrete discrete closure model corresponding to this example. Subsequently, we report the detailed prediction skill.

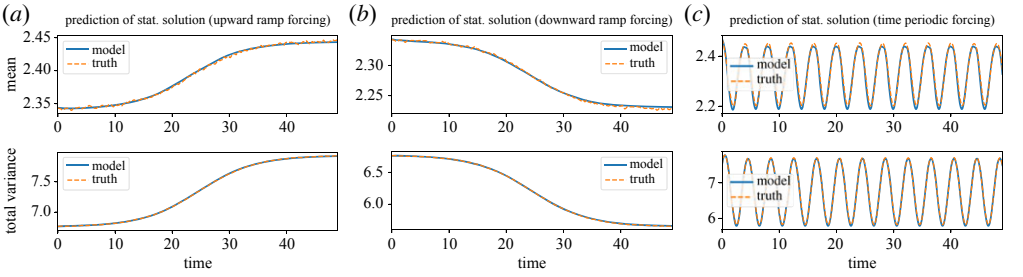

Figure 3. Long-time model prediction with the full mean-covariance closure model. Predictions of the statistical mean and total variance under three different external forcing scenarios are compared. (*a*) Upward ramp forcing. (*b*) downward ramp forcing. (*c*) periodic forcing. (Online version in colour.)

(i) Training model to learn the unresolved nonlinear flux

In the full mean-covariance model in (2.8a)–(2.8c), the dynamical equations for the mean state $\bar{u}$, the total energy E, and variance r_k are given explicitly and Markovian. In our numerical experiment, the discrete form in (2.9) is obtained by adopting the mid-point implicit scheme on (2.8a)–(2.8c) to ensure a more robust numerical performance with the larger time step $\Delta t = 10\delta t$. Together with the structural form in (2.12) to avoid instabilities and the residual network architecture in (4.1), the overall dynamical closure model adopted here is given as follows:

$$\left.\begin{aligned}
\bar{u}_{i+1} - \bar{u}_i &= \Delta t\left[-\frac{d}{2}(\bar{u}_i + \bar{u}_{i+1}) + \sum_k \frac{\Gamma_k}{2}(r_{k,i} + r_{k,i+1}) + \frac{1}{2}(F_i + F_{i+1})\right],\\
E_{i+1} - E_i &= \Delta t\left[-d(E_i + E_{i+1}) + \frac{1}{2}(\bar{u}_i F_i + \bar{u}_{i+1}F_{i+1})\right],\\
r_{k,i+1} - r_{k,i} &= \Delta t[-\Gamma_k(\bar{u}_i r_{k,i} + \bar{u}_{i+1} r_{k,i+1}) - d(r_{k,i} + r_{k,i+1}) + \theta_{k,i+1}],\\
\theta_{k,i+1} - \theta_{k,i} &= \frac{\min\{Q^M_{k,i+1}, 0\}}{r_{k,\mathrm{eq}} r_{k,i}} + \max\{Q^m_{k,i+1}, 0\},\\
\text{and}\qquad Q^M_{k,i+1} &= \mathcal{G}_k(\bar{u}_i, \ldots, \bar{u}_{i-m+1}, \{\theta_{k,i}, \ldots, \theta_{k,i-m+1}\}_{k=0,\ldots,J/2}, E_i, \ldots, E_{i-m+1}).
\end{aligned}\right\} \tag{4.2}$$

In (4.2), the states are discretized at time intervals $t_{i+1} - t_i = \Delta t$. The exact dynamical equations for the mean $\bar{u}$, the total statistical energy E, and r_k are adopted, while dynamics of θ_k are learned from data, with $\mathcal{G}_k$ modelled by an LSTM architecture.

(ii) Numerical results for detailed mean and variance prediction

Figure 3 shows the model prediction performance under the three forcing scenarios. The numerical model (4.2) is trained with a very short time dataset under constant forcings (in figure 2), while the prediction performance is tested on time-dependent forcing perturbations (in figure 1*a*). It is shown that for the long-time prediction (up to $T = 50$), the trained neural network model is stable and generates accurate predictions of both the statistical mean and variance throughout the time interval among all three test cases. For a more detailed comparison of the variance response on individual mode, figure 4 compares the predictions of the variances of the first three leading modes. Again, we observe robust accurate prediction of variances in all the modes under the tested forcing cases containing different statistical features.

In addition, we confirm the importance of adopting the strategy discussed in §(d) to guarantee long-time numerical stability. The variance dynamics include a large number of unstable directions that will amplify even small errors. Considering this, the neural network approximation for θ_k adopts the decomposed structure (2.12) so that the marginally stable modes

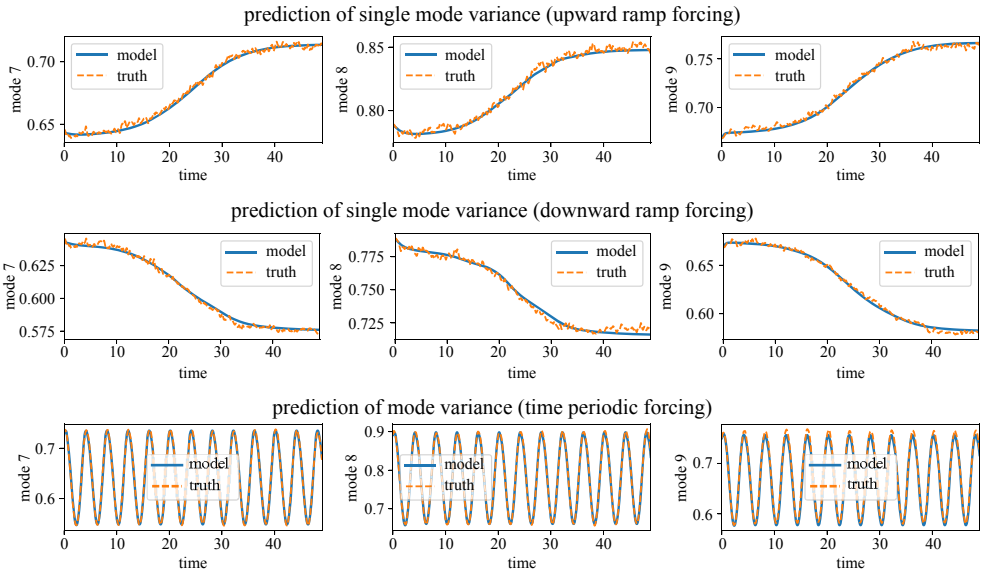

Figure 4. Detailed prediction of the variances of the first three most energetic modes using the full mean-covariance model. (Online version in colour.)

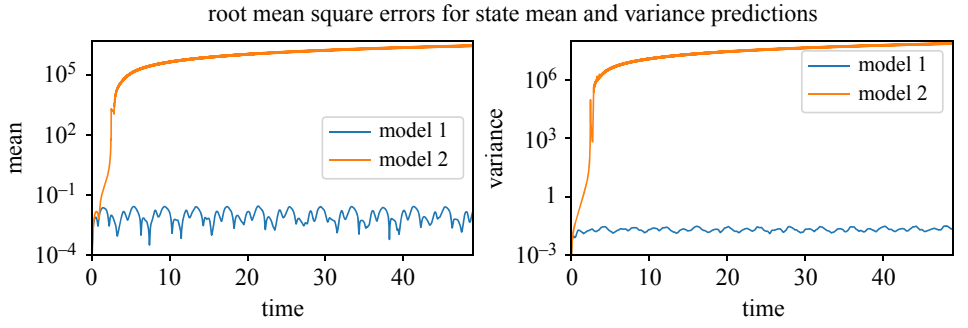

Figure 5. Prediction RMSEs using the trained models with different model structures. Model 1: the full-order model with decomposed effective damping and noise in (2.12); Model 2: the direct model without using the proper nonlinear flux decomposition. (Online version in colour.)

are balanced. Otherwise, if a neural network is applied directly to the model structure θ_k without proper consideration of the physical mechanism, severe numerical instability may occur due to the insufficient modelling of the unstable dynamics. Numerically, we compare the root mean square errors (RMSEs) in mean and total variance prediction in figure 5. Indeed, we see that the optimal model with decomposed damping and noise structure (model 1) maintains high accuracy for the long prediction period. By contrast, if the nonlinear flux θ_k is directly learned from the neural network (model 2), the predicted solution diverges after a short time due to the strong inherent persistent instability in the system.

(c) Prediction skill of the reduced-order mean-covariance model

Next, we consider the reduced-order mean-covariance model for efficient computation of only the most energetic modes $k_{\min} \le k \le k_{\max}$ in the variance equation. The total contribution of the less energetic unresolved modes is accounted with another neural network model for $\tilde{\phi}$. Thus,

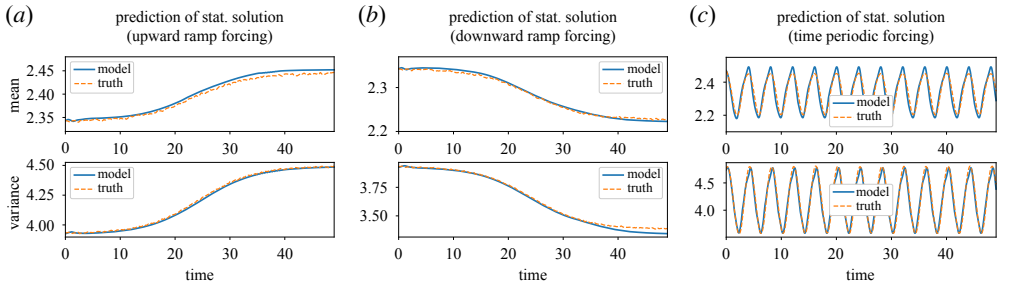

Figure 6. Prediction of the statistical mean and variance on the resolved modes using the reduced-order mean-variance model. The same neural network model is applied to different types of external forcing forms. Only the most energetic modes $6 \leq k \leq 12$ are computed in the model. (*a*) upward ramp forcing. (*b*) downward ramp forcing. (*c*) periodic forcing. (Online version in colour.)

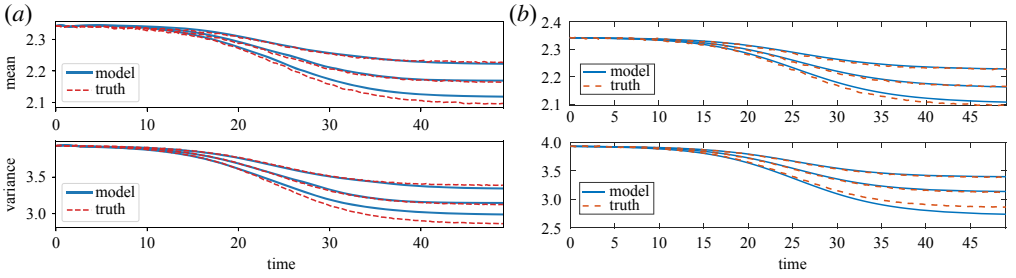

Figure 7. Comparison of reduced-order model predictions under different forcing perturbation amplitudes. The downward ramp forcing case is shown here as a typical example. Panels (*a*) show the predictions of ML model and (*b*) show the prediction of the parametric closure model proposed in [28]. (Online version in colour.)

the computational scheme follows the discretized mean and variance equations (2.10), with the dynamical equation for θ being modified as in (2.12) to avoid instability of the flux with residual network structure in (4.1).

In figure 6, we compare the reduced-order model prediction for different forcing perturbations. Again, we attain accurate predictions on both the statistical mean state and variances in the resolved subspace among the different kinds of forcing cases. In comparison to the full-order model prediction in figure 3, a slightly larger error occurs here, especially in the mean state. This reflects the additional model error due to the model approximation for the many unresolved modes. However, the computation cost is significantly reduced since we only compute a small portion of the full system (7 out of the total 21 modes).

Furthermore, to check the robustness of the model, we verify the model prediction skill with even stronger forcing perturbation amplitudes. Figure 7 shows the downward forcing case with stronger maximum forcing perturbations, $\delta f = -0.1F_{\text{eq}}, -0.15F_{\text{eq}}, -0.2F_{\text{eq}}$ (beyond the maximum forcing $|\delta f| = 0.1F_{\text{eq}}$ in the training data). Notice that the long-time prediction skill remains accurate for $-0.15F_{\text{eq}}$ and starts to deteriorate for larger forcing amplitude, $-0.2F_{\text{eq}}$. This somewhat negative result for larger perturbation is not so surprising as it displays the difficulty of the ML model in extrapolating beyond the information contents in the training data.

As a benchmark, we also show the corresponding prediction skill of the parametric closure model [28] in figure 7*b*. By visual comparison, one can see that the prediction skill is very similar to the ML-based closure model; for the largest forcing amplitude, the parametric closure gives a slightly better prediction. While the prediction performance is comparable, this parametric model requires a complicated calibration strategy that involves an expensive brute-force minimization

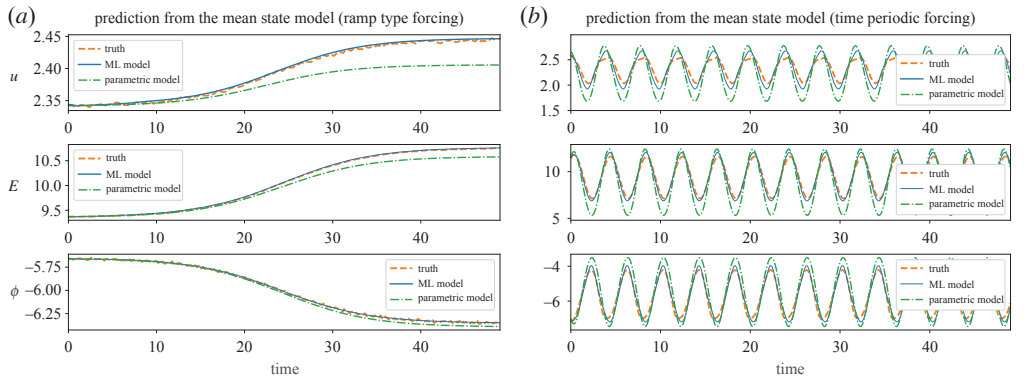

Figure 8. Predictions for the mean state $\bar{u}$, total statistical energy E, and the mean dynamical feedback ϕ with the mean statistical model under different forcing scenarios. The optimized ML model is compared with the parametric closure model. (Online version in colour.)

of a loss function that depends on long time statistics (linear response statistics). Particularly, the evaluation of the loss function involves an integration of the reduced-order model for a long time for each choice of parameter. Beyond this step that can be expensive for high-dimensional problems (as the dimension of the parameter space increases), a more fundamental issue is that it requires a physical insight for choosing the parametric model for the flux term. On the other hand, the more agnostic neural-network model can capture the changes in the statistics without specifying some detailed nonlinear flux structure beyond (2.12) that overcome instability.

(d) Prediction skill of the mean closure model

Finally, we test the prediction skill of the mean closure model (2.11), where we adopt the implicit midpoint rule for the discretization of (2.8a) and (2.8c), and use the standard LSTM network for the unresolved high-order feedback. We test the performance of the neural network model for long-term mean state prediction under the ramp down and periodic forcings in figure 1. In figure 8, we plot the predicted mean state $\bar{u}$, total statistical energy E, as well as the variance feedback in the mean ϕ. It shows that the ML model successfully captures the changes in the mean state under this extreme model set-up (with severely truncated dynamics) without including the explicit dynamical equations of the second-order moments. With forcing to a non-Gaussian regime (downward ramp forcing) or a periodic forcing with larger amplitude, the prediction becomes less accurate compared to the closure models that include more detailed variance dynamics (e.g. figures 3 and 6). Still, the closure model maintains high prediction skills under the unseen forcings. Again, if we compare the ML model results with the parametric closure model in [28], the ML framework produces more accurate predictions with cheaper computational costs. This shows the robustness of the agnostic neural network-based approach on various truncated configurations. On the other hand, the less accurate parametric model is due to the difficulty in modelling the truncated flux terms with a simple parametric equation.

5. Summary

In this paper, we developed non-Markovian statistical modelling strategies with ML. In the construction of the statistical closure models, we considered learning the complicated dynamical structure of the high-order nonlinear flux terms directly from data by imposing the LSTM neural-network architecture to uncover the non-Markovianity induced by partially observed components. Three statistical mean and covariance models were considered, with different emphases on the prediction of full variance spectrum, most energetic leading modes, and only

the mean state. With limited training data due to the stationarity of the statistics beyond the correlation time, we enriched the training dataset by simulating the transient behaviour of the statistics under various constant forcings and perturbed initial conditions.

The performance of the hierarchical ML models was verified on the L-96 system with homogeneous statistics. Uniformly accurate long-time predictions are observed using the resulting ML model under different forcing perturbation functions and strong perturbation amplitudes beyond the data in the training set. In addition, the true nonlinear physical energy transfer mechanism was considered in the model construction to guarantee numerical stability in long-term numerical integration. The ML model displays strong resistance to accumulated model errors with a long-time stable prediction despite the inherent instability in a wide spectrum of modes in the L-96 system. We found that the ML-based model prediction is comparable to that of the existing parametric model, which requires a more detailed calibration strategy, in two scenarios: learning the full-order and reduced-order coupled systems of mean-covariance statistics. On the other hand, the ML model is more accurate than the parametric approach in the severely truncated regime, learning the closure of only the mean statistics.

From this study, we conclude that the agnostic ML model is portable on various truncation scenarios since the strategy does not require physical knowledge of the high-order flux terms (as in parametric modelling) beyond avoiding instabilities. Numerically, the proposed scheme benefits from the advancement in the optimization of neural-network modelling, which allows us to carry the supervised learning task conveniently under one caveat (a reasonable neural-network architecture, in our case LSTM, and various tuning parameters). In addition, it is found from our numerical tests that the model performance is insensitive to different choices of neural-network hyperparameters such as the input chain length and hidden state size, implying robust prediction skills of the model framework.

While this result is encouraging, we only view this work as a first step. Particularly, this work only focuses on the L-96 system with homogeneous statistical dynamics. This assumption simplifies the closure model as the covariance matrix naturally becomes diagonal, so we only need to close the diagonal variances and their reduction. A more important and challenging direction is to extend the proposed ML approach to non-homogeneous statistical dynamics, which involve non-trivial off-diagonal covariance components. Besides the curse of dimension problem (the dynamical equation of the covariance matrix has N^2-terms), a direct closure on the covariance statistics may not preserve positive definite-ness under the ML prediction. To overcome these two issues, one possibly needs to consider closing the dynamical equation for the fluctuation components Z_i in (2.2) directly, extending the idea from [35] with the ML model, which is part of our future work.

Data accessibility. This article has no additional data.

Authors' contributions. D.Q.: conceptualization, investigation, methodology, software, validation, writing—original draft, writing—review and editing; J.H.: conceptualization, investigation, methodology, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. The research of J.H. was partially supported under the NSF grant no. DMS-1854299 and the ONR grant N00014-22-1-2193. This research of D.Q. is partially supported by the Office of Naval Research N00014-19-1-2286 and the start-up funds provided by Purdue University.

References

1. Lesieur M. 1987 *Turbulence in fluids: stochastic and numerical modelling*, vol. **488**. Boston, MA: Nijhoff.
2. Kwasniok F. 2012 Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling. *Phil. Trans. R. Soc. A* **370**, 1061–1086. (doi:10.1098/rsta.2011.0384)

3. Lu F, Lin K, Chorin A. 2016 Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems. *Commun. Appl. Math. Comput. Sci.* **11**, 187–216. (doi:10.2140/camcos.2016.11.187)
4. Sapsis TP, Majda AJ. 2013 Statistically accurate low-order models for uncertainty quantification in turbulent dynamical systems. *Proc. Natl Acad. Sci. USA* **110**, 13 705–13 710. (doi:10.1073/pnas.1313065110)
5. Qi D, Majda AJ. 2016 Low-dimensional reduced-order models for statistical response and uncertainty quantification: two-layer baroclinic turbulence. *J. Atmos. Sci.* **73**, 4609–4639. (doi:10.1175/JAS-D-16-0192.1)
6. Maulik R, San O, Rasheed A, Vedula P. 2019 Subgrid modelling for two-dimensional turbulence using neural networks. *J. Fluid Mech.* **858**, 122–144. (doi:10.1017/jfm.2018.770)
7. Schmitt FG. 2007 About Boussinesq's turbulent viscosity hypothesis: historical remarks and a direct evaluation of its validity. *C. R. Mec.* **335**, 617–627. (doi:10.1016/j.crme.2007.08.004)
8. Leith CE. 1975 Climate response and fluctuation dissipation. *J. Atmos. Sci.* **32**, 2022–2026. (doi:10.1175/1520-0469(1975)032<2022:CRAFD>2.0.CO;2)
9. Majda AJ, Qi D. 2018 Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems. *SIAM Rev.* **60**, 491–549. (doi:10.1137/16M1104664)
10. Majda AJ, Qi D. 2019 Linear and nonlinear statistical response theories with prototype applications to sensitivity analysis and statistical control of complex turbulent dynamical systems. *Chaos* **29**, 103131. (doi:10.1063/1.5118690)
11. Berry T, Harlim J. 2014 Linear theory for filtering nonlinear multiscale systems with model error. *Proc. R. Soc. A* **470**, 20140168. (doi:10.1098/rspa.2014.0168)
12. Harlim J. 2017 *Model error in data assimilation*. Cambridge, UK: Cambridge University Press.
13. Zhen Y, Harlim J. 2015 Adaptive error covariance estimation methods for ensemble Kalman filtering. *J. Comput. Phys.* **294**, 619–638. (doi:10.1016/j.jcp.2015.03.061)
14. Gamahara M, Hattori Y. 2017 Searching for turbulence models by artificial neural network. *Phys. Rev. Fluids* **2**, 054604. (doi:10.1103/PhysRevFluids.2.054604)
15. Singh AP, Medida S, Duraisamy K. 2017 Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils. *AIAA J.* **55**, 2215–2227. (doi:10.2514/1.J055595)
16. Hochreiter S, Schmidhuber J. 1997 Long short-term memory. *Neural Comput.* **9**, 1735–1780. (doi:10.1162/neco.1997.9.8.1735)
17. Ma C, Wang J, Weinan E. 2018 Model reduction with memory and the machine learning of dynamical systems. *Commun. Comput. Phys.* **25**, 947–962.
18. Vlachas PR, Byeon W, Wan ZY, Sapsis TP, Koumoutsakos P. 2018 Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proc. R. Soc. A* **474**, 20170844. (doi:10.1098/rspa.2017.0844)
19. Maulik R, Mohan A, Lusch B, Madireddy S, Balaprakash P, Livescu D. 2020 Time-series learning of latent-space dynamics for reduced-order model closure. *Physica D* **405**, 132368. (doi:10.1016/j.physd.2020.132368)
20. Harlim J, Jiang SW, Liang S, Yang H. 2020 Machine learning for prediction with missing dynamics. *J. Comput. Phys.* **428**, 109922. (doi:10.1016/j.jcp.2020.109922)
21. Jiang SW, Harlim J. 2020 Modeling of missing dynamical systems: deriving parametric models using a nonparametric framework. *Res. Math. Sci.* **7**, 1–25. (doi:10.1007/s40687-020-00217-4)
22. Givon D, Kupferman R, Stuart A. 2004 Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity* **17**, R55–R127. (doi:10.1088/0951-7715/17/6/R01)
23. Weinan E, Enguist B. 2003 The heterogeneous multi-scale methods. *Commun. Math. Sci.* **1**, 87–133. (doi:10.4310/CMS.2003.v1.n1.a8)
24. Chorin AJ, Hald OH, Kupferman R. 2002 Optimal prediction with memory. *Physica D* **166**, 239–257. (doi:10.1016/S0167-2789(02)00446-3)
25. Gouasmi A, Parish EJ, Duraisamy K. 2017 A priori estimation of memory effects in reduced-order models of nonlinear systems using the Mori–Zwanzig formalism. *Proc. R. Soc. A* **473**, 20170385. (doi:10.1098/rspa.2017.0385)
26. Lorenz EN. 1996 Predictability: a problem partly solved. In *Proc. seminar on predictability*, vol. 1, pp. 1–18. Reading, UK: ECMWF.
27. Majda AJ. 2016 *Introduction to turbulent dynamical systems in complex systems*. Berlin, Germany: Springer.

28. Majda AJ, Qi D. 2016 Improving prediction skill of imperfect turbulent models through statistical response and information theory. *J. Nonlinear Sci.* **26**, 233–285. (doi:10.1007/s00332-015-9274-5)
29. Sapsis TP, Majda AJ. 2013 A statistically accurate modified quasilinear Gaussian closure for uncertainty quantification in turbulent dynamical systems. *Physica D* **252**, 34–45. (doi:10.1016/j.physd.2013.02.009)
30. Zwanzig R. 2001 *Nonequilibrium statistical mechanics*. Oxford, UK: Oxford University Press.
31. Takens F. 2010 Chapter 7—reconstruction theory and nonlinear time series analysis. In *Handbook of dynamical systems*, vol. 3 (eds H Broer, B Hasselblatt, F Takens), pp. 345–377. Amsterdam, The Netherlands: Elsevier Science.
32. Sauer T, Yorke JA, Casdagli M. 1991 Embedology. *J. Stat. Phys.* **65**, 579–616. (doi:10.1007/BF01053745)
33. Gilani F, Giannakis D, Harlim J. 2021 Kernel-based prediction of non-Markovian time series. *Physica D* **418**, 132829. (doi:10.1016/j.physd.2020.132829)
34. Levine ME, Stuart AM. 2021 A framework for machine learning of model error in dynamical systems. (arXiv preprint arXiv:2107.06658)
35. Sapsis TP, Majda AJ. 2013 Blending modified Gaussian closure and non-Gaussian reduced subspace methods for turbulent dynamical systems. *J. Nonlinear Sci.* **23**, 1039–1071. (doi:10.1007/s00332-013-9178-1)