Machine Learning for Prediction with Missing Dynamics

John Harlim
Department of Mathematics
Department of Meteorology and Atmospheric Science
Institute for Computational and Data Sciences
The Pennsylvania State University, PA 16802, USA
jharlim@psu.edu

Shixiao W. Jiang
Department of Mathematics
The Pennsylvania State University, PA 16802, USA
suj235@psu.edu

Senwei Liang
Department of Mathematics
Purdue University[*], IN 47907, USA
Department of Mathematics
National University of Singapore[†], Singapore
liang339@purdue.edu

Haizhao Yang
Department of Mathematics
Purdue University, IN 47907, USA
haizhao@purdue.edu

July 8, 2020

Abstract

This article presents a general framework for recovering missing dynamical systems using available data and machine learning techniques. The proposed framework reformulates the prediction problem as a supervised learning problem to approximate a map that takes the memories of the resolved and identifiable unresolved variables to the missing components in the resolved dynamics. We demonstrate the effectiveness of the proposed framework with a strong convergence error bound of the resolved variables up to finite time and numerical tests on prototypical models in various scientific domains. These include the 57-mode barotropic stress models with multiscale interactions that mimic the blocked and unblocked patterns observed in the atmosphere, the nonlinear Schrödinger equation which found many applications in physics such as optics and Bose-Einstein-Condense, the Kuramoto-Sivashinsky equation which spatiotemporal chaotic pattern formation models trapped ion mode in plasma and phase dynamics in reaction-diffusion systems. While many machine learning techniques can be used to validate the proposed framework, we found that recurrent neural networks outperform kernel regression methods in terms of recovering the trajectory of the resolved components and the equilibrium one-point and two-point statistics. This superb performance suggests that a recurrent neural network is an effective tool for recovering the missing dynamics that involves approximation of high-dimensional functions.

Keywords. Closure Modeling, Missing dynamics, Machine Learning, Long Short Term Memory
AMS subject classifications: 44A55, 65R10 and 65T50.

[*]Current institute.
[†]Part of the work was done at the National University of Singapore

1 Introduction

The problem of missing dynamics is ubiquitous in any scientific domain that concerns with prediction through computational models. This long-standing problem has been posted under various names, including model error, sub-grid scale parameterization, closure modeling [16, 32, 6, 27, 36, 39, 41, 45, 73]. Another relevant topic of broad interest is the reduced-order modeling whose ultimate goal is to systematically deduce a computationally efficient model to predict the evolution of the resolved variables when the full underlying model is too expensive to solve [22, 48, 49, 71, 13, 14, 24, 26, 33]. In our context, the proposed framework adopted here does not require any knowledge of the full equations that govern the underlying dynamical systems. The proposed approach that we consider assumes that only the dynamical components corresponding to the resolved variables are given. As in [30], the missing components will be learned from a historical time series of the resolved variables and the identifiable unresolved variables, where the latter serves as feedback from the interaction of the resolved and unresolved scales. In essence, the proposed approach is to learn a dynamical model for the identifiable unresolved variables that depend on both the resolved and unresolved variables.

The success of deep learning as a supervised learning algorithm has drawn tremendous interest on reduced-order modeling applications. A closely related approach to the modeling framework in this paper is presented in [57]. They proposed a Feedforward Neural Network (FNN) as a representation of the dynamics of the irrelevant variables. These authors also provide a linear control theory perspective to justify the identifiability of their dynamical representation on a class of nonlinear systems with a dual linear closure. In this article, we consider the closure modeling framework with a nonparametric formulation and provide a strong convergence error bound of the estimation of the resolved variables for the first time. For discrete dynamics obtained from a temporal discretization of differential equations, we found that when the unresolved variables are fully identifiable, the error rate is $O(T^2\Delta^2\epsilon)$, under mild assumptions. Here, $T > 0$ denotes the prediction time index, Δ denotes the discrete time step, and $\epsilon > 0$ denotes the approximation error of the missing dynamics. Recalling the theory of nonparametric regression [68], if the missing dynamics is a function of a Sobolev class, H^β, where $\beta > 0$ denotes the regularity parameter, the learning rate ϵ of any nonparametric regression algorithm with i.i.d data has an optimal global error rate of an order $\epsilon = O(N^{-\frac{\beta}{2\beta+d}})$, where N denotes the length of training data and d denotes the dimension of the domain of the function. Hence, nonparametric regression algorithms suffer from the curse of dimensionality unless when the regularity parameter $\beta = O(d)$. However, even for the case of $\beta = O(d)$, there are no efficient tools to carry out the computation for high-dimensional problems.

Fortunately, recent advances in the theoretical analysis of deep neural networks show that they can avoid the curse of dimensionality in terms of approximation error in both the case of sufficiently smooth functions [5, 51, 53, 52, 42], and even for continuous functions [64]. Also, there is no curse of dimensionality of deep neural networks in terms of generalization error when the target functions admit sufficient smoothness [18], when the data are sampled on a low-dimensional manifold [56], or in the case of classification functions [9]. While the generalization error for deep neural networks on general functions is an open problem, empirical numerical evidence has indicated that deep neural networks together with their stochastic training algorithms (e.g., batch-based stochastic gradient descent) are automatic tools that can identify the "low complexity" of the underlying systems, e.g., the smoothness or the low-dimensional domain that leads to no curse of dimensionality [5, 52, 53, 63, 42, 18]. In particular, recurrent neural networks as a special case of deep neural networks has the potential to avoid the curse of dimension when learning and predicting discrete dynamical systems with low complexity structures.

While the closure modeling framework can be numerically realized using any approximation/regression methods, we will consider a special type of recurrent neural networks called the Long-Short-Term-Memory (LSTM) [29]. We will show that this approach can overcome the curse of dimension suffered by the standard

nonparametric regression method such as the kernel mean embedding approximation used in [30]. Our choice of using the LSTM is also encouraged by the success of it in recent closure modeling applications as proposed in [44, 70, 50]. We should stress that these existing approaches [44, 70, 50] share a similarity, that is, they specify the closure model as a function of only the memory of the resolved variables and motivate their framework using a heuristic connection with the Mori-Zwanzig formalism [54, 75]. In contrast, we will demonstrate that it is critical for the closure model to also depend on the memory of the identifiable unresolved variables in addition to the resolved components. We will demonstrate the effectiveness of our framework on several tough prototype complex systems that arise in geophysical fluid dynamics, optics, quantum fluid such as Bose-Einstein-Condensate, and plasma physics, in addition to theoretical justification.

The rest of the paper is organized as follows. In Section 2, we formulate the problem, provide a simple example to elucidate the proposed formulation, and discuss the theoretical aspect of the proposed approach. In Section 3, we provide a short discussion on the numerical aspect of LSTM as a special class of RNN. In Section 4, we report the numerical results on three prototypical examples of different types of dynamical systems. In Section 5, we close the paper with a summary. The technical proofs of the theoretical result will be reported in Appendices.

2 Data-Driven Modeling for Missing Dynamics

Throughout this paper, we will describe the closure modeling approach in the context of discrete maps that naturally arise from numerical discretization of partial or ordinary differential equations. Stochastic differential equation will be discussed in the following sections. Let the resolved, $x_t \in \mathcal{X}$, and unresolved, $y_t \in \mathcal{Y}$, variables be the solution of the following deterministic discrete dynamical systems,

$$x_{t+1} = \mathcal{F}(x_t, y_t), \quad y_{t+1} = \mathcal{G}(x_t, y_t), \tag{1}$$

given initial conditions x_0, y_0.

Assumption 1. *Furthermore, we assume that the full system is ergodic with a unique invariant measure μ. For the measure space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}, \mu)$, where $\mathcal{B}$ denotes the σ-algebra of set $\mathcal{X} \times \mathcal{Y}$, the map defined by $\Phi := (\mathcal{F}, \mathcal{G})$ is invariant under measure μ. That is, $\mu(\Phi^{-1}(B)) = \mu(B)$, for $B \in \mathcal{B}$.*

Under this assumption, the missing dynamics problem is to predict $\{x_t : t \in \mathbb{N}\}$ and its statistics, such as, the mean, covariance, and auto-correlation functions, given only partial dynamics, $\mathcal{F}$. Basically, the absence of $\mathcal{G}$ means that we are missing the unresolved dynamics for y. Our goal is to reconstruct the missing dynamics in (1) from the given historical time series, $\{x_t, \theta_t\}_{t=1,\dots,N}$, where $\theta_t := \Theta(x_t, y_t)$ is the identifiable unresolved variable. Here, the observables $X : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and $\Theta : \mathcal{X} \times \mathcal{Y} \to \mathcal{W}$ are random variables defined as $X(x_t, y_t) = x_t$ and $\Theta(x_t, y_t) = \theta_t$, respectively.

In particular, θ_t is the component of the unresolved variables that can be identified from $\mathcal{F}(x, y) := \mathcal{F}(x, \Theta(x, y))$ in (1) and observations $\{x_t\}$. With this definition, we abuse the notation $\mathcal{F}$ to emphasize its dependence on θ (and suppress its dependence on y), so that in general, $\theta \neq y$ (see e.g, (2)). This restriction is motivated by practical constraints where only the resolved variables are observed. For example, any $\mathcal{F}$ can be decomposed into,

$$\mathcal{F}(x_t, y_t) = \bar{\mathcal{F}}(x_t) + \Theta(x_t, y_t) = \bar{\mathcal{F}}(x_t) + \theta_t, \tag{2}$$

for some $\bar{\mathcal{F}}$ that involves only the resolved variables and the remainder term is the identifiable unresolved variable. Given $\{x_t\}$, one can extract a time series of $\theta_t = x_{t+1} - \bar{\mathcal{F}}(x_t)$ by a direct subtraction. In the case when the observed x_t is noisy, one can also use appropriate filtering methods [27, 7]. We should point out that our formulation below also holds even if θ depends only on the unresolved variables y, so long as the time series

of θ is available. In this case, the identifiability of θ will be related to the notion of reachability/observability in the control theory (e.g. see Chapters 3 and 4 of [15]). In the numerical simulations shown below, we assume that a historical timeseries of $\{x_t, \theta_t\}_{t=1,\ldots,N}$ is available to us.

Our goal is to predict $\{x_t : t \in \mathbb{N}\}$ and its statistics, such as the mean, covariance, and auto-correlation functions, with the above constraints. We also aim to reconstruct the missing dynamics in (1). Define $\boldsymbol{z}_{t,m} := (\boldsymbol{x}_{t-m:t}, \boldsymbol{\theta}_{t-m:t}) \in \mathcal{Z}$ with $\boldsymbol{x}_{t-m:t} := (x_{t-m}, x_{t-m+1}, \ldots, x_t)$ and $\boldsymbol{\theta}_{t-m:t} := (\theta_{t-m}, \theta_{t-m+1}, \ldots, \theta_t)$ for some integer $m \geq 0$ which characterizes the memory length. We consider a general closure model of the following form,

$$\hat{x}_{t+1} = \mathcal{F}(\hat{x}_t, \hat{\theta}_t), \quad \hat{\theta}_{t+1} = \mathbb{E}^\epsilon[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}] + \xi_{t+1}, \tag{3}$$

where $\hat{\cdot}$ is used to denote the numerical approximation of the corresponding variable in the closure model. In (3), the notation $\mathbb{E}^\epsilon[\Theta_{t+1}|\cdot] : \mathcal{Z} \to \mathcal{W}$ is to denote an estimator to the target function $\mathbb{E}[\Theta_{t+1}|\cdot] : \mathcal{Z} \to \mathcal{W}$ with error $\epsilon > 0$ in appropriate sense. Here, the random variable $\Theta_{t+1} := S^{t+1} \circ \Theta$, where $S : L^2(\mu, \mathcal{W}) \to L^2(\mu, \mathcal{W})$ denotes the associated Koopman operator that is defined as, $Sf := f \circ \Phi$, for all function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{W}$ of the Hilbert space $L^2(\mu, \mathcal{W})$, equipped with the inner product $\langle f, g \rangle_{L^2(\mu)} = \int_{\mathcal{X}\times\mathcal{Y}} \langle f(x,y), g(x,y) \rangle_{\mathcal{W}} d\mu(x,y)$. Here, the map $\Phi := (\mathcal{F}, \mathcal{G})$ denotes the full dynamics. With this definition, one can verify that evaluating Θ_{t+1} on initial condition (x_0, y_0) produces $\Theta_{t+1}(x_0, y_0) = S^{t+1}\Theta(x_0, y_0) = \Theta \circ \Phi^{t+1}(x_0, y_0) = \Theta(x_{t+1}, y_{t+1}) = \theta_{t+1}$.

While the proposed framework suggests that one can use any supervised learning method to construct an estimator $\mathbb{E}^\epsilon[\Theta_{t+1}|\cdot]$ (that can be in the form of parametric or nonparametric as we shall discuss in subsection 2.2), to guarantee an accurate estimation of the path x_t, one should consider a consistent estimator, that is, $\mathbb{E}^\epsilon[\Theta_{t+1}|\cdot] \to \mathbb{E}[\Theta_{t+1}|\cdot]$ as $\epsilon \to 0$ in L^2 sense, as we shall discussed below in subsections 2.1 and 2.3. Another question that will be clarified in these two subsections is to which target function does the proposed estimator converge to. In other words, what is the target function $\mathbb{E}[\Theta_{t+1}|\cdot]$? As we shall see later, this depends on the observable Θ.

In (3), the noise ξ_t is added to account for the residual due to misspecification of hypothesis space for the target function $\mathbb{E}[\Theta_{t+1}|\cdot]$. In fact, we shall see from our numerical experiments below that this additional noise is not needed for the deterministic problems when LSTM is used as the estimator $\mathbb{E}^\epsilon[\Theta_{t+1}|\cdot]$. For simplicity, we only consider $\xi_t \sim \Xi$ to be Gaussian with variance,

$$\mathbb{E}[\Xi^2] := \mathbb{E}\Big[(\Theta_{t+1} - \mathbb{E}^\epsilon[\Theta_{t+1}|Z_{t,m}])^2\Big], \tag{4}$$

where we used the notation $Z_{t,m}$ to denote the random variables associated to the realizations $\boldsymbol{z}_{t,m}$. Here, the variance will also be estimated by Monte-Carlo approximation to (4) using the historical solutions of the ergodic system in (1).

To summarize, the closure model is reformulated as a supervised learning problem to learn the map $\hat{\boldsymbol{z}}_{t,m} \mapsto \mathbb{E}[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}]$ and the variance $\mathbb{E}[\Xi^2]$ of the residual. In the next subsection, we discuss a simple case for which $\Theta(x, y) = y$. Subsequently, we discuss the notion of parametric and nonparametric estimators, $\mathbb{E}^\epsilon[\Theta_{t+1}|\cdot]$. We finally close this section with a study on the general case of Θ, which constitutes a more complicated target function, $\mathbb{E}[\Theta_{t+1}|\cdot]$, obtained using a discrete Dyson formula.

2.1 Fully identifiable unresolved variables

To give an intuition, suppose that the entire unresolved variables can be identified, that is, $\Theta(x, y) = y$. Since $\theta_t = y_t$, let us replace the notation Θ_{t+1} with $Y_{t+1} := S^{t+1} \circ Y$ defined with the random variable $Y : \mathcal{X}\times\mathcal{Y} \to \mathcal{Y}$ in $L^2(\mu, \mathcal{Y})$ with $Y(x_t, y_t) = y_t$. In this case, it is clear that the target function is nothing but the full missing dynamics, namely, $\mathbb{E}[\Theta_{t+1}|\boldsymbol{z}_{t,0}] = \mathbb{E}[Y_{t+1}|x_t, y_t] = \mathcal{G}(x_t, y_t)$ such that one can rewrite (1) as,

$$x_{t+1} = \mathcal{F}(x_t, y_t), \quad y_{t+1} = \mathbb{E}[Y_{t+1}|x_t, y_t]. \tag{5}$$

If $\mathbb{E}^\epsilon[Y_{t+1}|\cdot]$ is a consistent estimator of $\mathbb{E}[Y_{t+1}|\cdot]$ with variance error rate ϵ^2, it is clear from (4) that $\mathbb{E}[\Xi^2] = O(\epsilon^2)$. In this case, we can show that

Theorem 1. *Let $\mathcal{F}$ and $\mathcal{G}$ be Lipschitz in x and y. Let x_{t+1} be the solutions of* (5) *and $\hat{x}_{t+1}$ be the solutions of,*

$$\hat{x}_{t+1} = \mathcal{F}(\hat{x}_t, \hat{y}_t), \quad \hat{y}_{t+1} = \mathbb{E}^\epsilon[Y_{t+1}|\hat{x}_t, \hat{y}_t] + \xi_{t+1}, \tag{6}$$

under the same initial conditions $x_0 = \hat{x}_0, y_0 = \hat{y}_0$. Under the Assumption 1, if $\mathbb{E}^\epsilon[Y_{t+1}|\cdot] \to \mathbb{E}[Y_{t+1}|\cdot]$ as $\epsilon \to 0$ with variance error of order ϵ^2, then

$$\mathbb{E}\Big[\max_{t\in\{0,\dots,T\}} |\hat{x}_t - x_t|\Big] = O(a^T\epsilon). \tag{7}$$

for some constant $a > 1$ that is independent of T and ϵ.

Proof. See Appendix A. □

This rather pessimistic error bound (exponential on T) is not surprising since the assumption on $\mathcal{F}$ and $\mathcal{G}$ is mild, Lipschitz continuous. This error bound is basically an extension of a classical result in dynamical system theory, the continuous dependence of the solutions of uniformly perturbed vector field (e.g., Chapter 17.5 in [28]). To obtain an improved result (such as polynomial on T), one should consider the structure on $\mathcal{F}, \mathcal{G}$. For example, when the discrete dynamical system is a result of the Euler-Maruyama discretization on a system of stochastic differential equations,

$$dx = f(x,y)\,dt + \sigma_x dW_{x,t}, \qquad dy = g(x,y)\,dt + \sigma_y dW_{y,t},$$

where $dW_{x,t}$ and $dW_{y,t}$ denote independent Gaussian white noises, we have:

$$x_{t+1} = x_t + f(x_t, y_t)\Delta + \sigma_x\Delta^{1/2}\xi_{x,t+1}, \qquad y_{t+1} = y_t + g(x_t, y_t)\Delta + \sigma_y\Delta^{1/2}\xi_{y,t+1}, \tag{8}$$

where Δ denotes the time step. Here, $\xi_x, \xi_y \sim \mathcal{N}(0, \mathcal{I})$ are samples of the standard Gaussian white noises. When g and σ_y are unknown, we can directly estimate these terms and obtain a sharper error bound:

Theorem 2. *blueLet $\mathcal{F}$ and $\mathcal{G}$ be Lipschitz in x and y. Let x_{t+1} be the solutions of* (8) *and $\hat{x}_{t+1}$ be the solutions of,*

$$\begin{aligned} \hat{x}_{t+1} &= \hat{x}_t + f(\hat{x}_t, \hat{y}_t)\,\Delta + \sigma_x\Delta^{1/2}\xi_{x,t+1}, \\ \hat{y}_{t+1} &= \hat{y}_t + \Delta\mathbb{E}^\epsilon\Big[Y^\Delta_{t+1}|\hat{x}_t, \hat{y}_t\Big] + \hat{\sigma}_y\Delta^{1/2}\xi_{y,t+1}, \end{aligned} \tag{9}$$

under the same initial conditions $x_0 = \hat{x}_0, y_0 = \hat{y}_0$. Here, we have defined $Y^\Delta_{t+1} := \frac{Y_{t+1}-Y_t}{\Delta}$ and the noise variance,

$$\hat{\sigma}_y^2\Delta := \mathbb{E}\Big[\Big(Y_{t+1} - Y_t - \Delta\mathbb{E}^\epsilon\Big[Y^\Delta_{t+1}|X_t, Y_t\Big]\Big)^2\Big]$$

is estimated from the training data. Suppose that the learning variance error rate is,

$$\mathbb{E}[(\mathbb{E}[Y^\Delta_{t+1}|X_t, Y_t] - \mathbb{E}^\epsilon[Y^\Delta_{t+1}|X_t, Y_t])^2] \le C\epsilon^2,$$

for some $C > 0$. Let f and g be Lipschitz continuous in x and y, then,

$$\mathbb{E}\Big[\max_{t\in\{0,\dots,T\}} |\hat{x}_t - x_t|\Big] = O(\epsilon T^2\Delta^2).$$

Proof. See Appendix B. □

This result suggests that the solution of the proposed approximate dynamics in (9) strongly converges to that of (8) up to finite time. The convergence rate suggests that one can expect a path-wise prediction with an accuracy of order learning rate error, ϵ, up to order-one model unit time, $(T\Delta)^2 = \mathcal{O}(1)$. In other words, the length of accurate path-wise prediction is inversely proportional to the square root of the learning error rate, $T\Delta \approx \epsilon^{-1/2}$. Using a consistent learning algorithm, $\epsilon \to 0$, one can control the polynomial error growth. In the next section, we will use this error rate to estimate an accurate prediction time as a function of the number of training data, using a class of hypothesis space with well-known optimal (in the sense of bias-variance tradeoff) learning rate.

2.2 Parametric versus nonparametric closure models

The essence of parametric closure modeling is to simply specify $\mathbb{E}^\epsilon$ in (3) with a specific choice of function $\mathcal{P}(\hat{\boldsymbol{z}}_{t,m}; W)$ that depends on a finite-dimensional parameter W. The choice of ansatz $\mathcal{P}$ is usually based on physical intuition; see e.g., [45, 27, 6, 39, 41]. Once the model is specified, the hyper-parameter W can be obtained by regressing the pairs $\{\boldsymbol{z}_{t,m}, \theta_{t+1}\}$. Subsequently, the variance of Ξ is estimated as in (4). In the later section, we will consider the Long-Short-Term-Memory model for $\mathcal{P}$.

In [30], a non-parametric closure model is considered. Specifically, the conditional expectation in (3) is estimated using the kernel mean embedding of conditional distribution formula [66, 65]. In the implementation, they assume that $\mathbb{E}[\theta_{t+1}|\cdot]$ belongs to the reproducing kernel Hilbert space (RKHS) $\mathcal{H} \subset L^2(\mathcal{Z}, q)$ with a well-defined Mercer-type kernel, constructed based on orthonormal basis functions $\{\varphi_k\}$ of this L^2-space, weighted by an arbitrary positive $q \in L^1(\mathcal{Z})$ [30]. For non-compact domain, such construction was studied in [74]. The advantage of RKHS beyond inheriting the orthogonality from the L^2-space is that any sequence of functions that converges in $\mathcal{H}$-norm also converges uniformly and any function in $\mathcal{H}$ inherits the regularity of the kernel. In fact, one can show that for appropriate choice of kernel, the RKHS $\mathcal{H}$ is dense in $C_b(\mathcal{Z})$ for compact $\mathcal{Z}$ [67] or $C_0(\mathcal{Z})$ for non-compact $\mathcal{Z}$ [74]. Then, for any $\boldsymbol{z}_t \in \mathcal{Z}$, we can represent:

$$\mathbb{E}[\theta_{t+1}|\boldsymbol{z}_t] = \sum_{k=0}^{\infty} c_k \varphi_k(\boldsymbol{z}_t), \tag{10}$$

where the coefficients c_k are precomputed from the available training data $\{\theta_t, x_t\}$; see [30] for details. The key point is that this nonparametric formulation turns the problem of choosing the closure model into a problem of constructing basis functions of a Hilbert space (i.e., choosing an appropriate hypothesis space). Since $\mathcal{H}$ is dense in the space of continuous bounded functions, then any bounded continuous parametric function $\mathcal{P}(\boldsymbol{z}; W)$ can be consistently approximated by a truncation of the series expansion in (10) for appropriate choice of basis functions. In this sense, choosing a parametric-based model is somewhat equivalent to specifying an appropriate RKHS space.

Now, let us demonstrate the effectiveness of this approach in the following simple yet nontrivial example.

Example: Consider a Langevin dynamics

$$\begin{aligned} dx &= y\,dt, \\ dy &= (-\nabla V(x) - \gamma y)\,dt + \sigma_y\,dW_t, \end{aligned} \tag{11}$$

where $x \in \mathbb{R}$ is the displacement, $y \in \mathbb{R}$ is the velocity, $V(x) = -x^2/2 + x^4/4$ is the double-well potential, $\gamma = 1$ is the damping coefficient, dW is a standard Gaussian white noise, and $\sigma_y = 3\sqrt{2}/10$ is the driving strength. We observe the trajectories of the variables (x_t, y_t) at every time step $\Delta = 0.01$, obtained using the Euler-Maruyama discretization scheme. In our previous notation, $\theta(y) = y$ and we consider a closure

Table 1: Comparisons of mean exit time $\bar{\tau}_0$ and reaction rate ν_R between the full model and closure models.

	True	RKHS $N = 50,000$	RKHS $N = 500,000$
$\bar{\tau}_0$	99.2	69.1	102.7
ν_R	0.0079	0.0040	0.0075

model in (3) with $\mathbb{E}[\theta_{t+1}|\hat{\boldsymbol{z}}_{t,0}] = \mathbb{E}\left[Y_{t+1}|\hat{x}_t, \hat{y}_t\right]$. This example is nontrivial due to the transition state induced by the double-well potential $V(x)$. Note that only when the driving strength σ_y is in a reasonable region, the transition state phenomenon can be observed for this double-well potential system (see Fig. 1 for example).

In Fig. 1, we compare the result obtained using the RKHS approximation in (10) with the true trajectories and the statistics from the full model in (11). In this comparison, we apply the formula in (10) with a tensor product of 50×50 Hermite polynomials. We compare the prediction of the trajectory up to finite-time, marginal density of x, and auto-covariance function $\mathbb{E}[x_\tau x_0]$ of the true dynamics in (11) with those from the closure models, trained using $N = 5 \times 10^4$ and 5×10^5 data points. Notice that using large enough training data, we are not only accurately recovering the trajectory path-wise longer in time but also the density and auto-covariance function. Compare to the optimal learning rate $\epsilon = O(N^{-\frac{\beta}{2\beta+d}})$ of [68] for very smooth function with $\beta = \infty$, the empirical prediction length (as shown in Fig. 1) is on the order of the theoretical prediction length $T\Delta = \epsilon^{-1/2} \approx 14.95$ for $N = 5 \times 10^4$ and is slightly longer than the conservative estimate $T\Delta = \epsilon^{-1/2} \approx 26.59$ for $N = 5 \times 10^5$. In Table 1, we also see the agreement of several statistics that are commonly used to characterize metastable dynamics. When the training data set is large, we see a relatively accurate estimation of the mean exit time $\bar{\tau}_0$ of a particle to escape one of the wells [21] and the reaction rate ν_R, defined as the limit of N_T/T as $T \to \infty$ where N_T is the number of trajectories to escape a well in the time interval T [69].

2.3 Partially identifiable unresolved variables

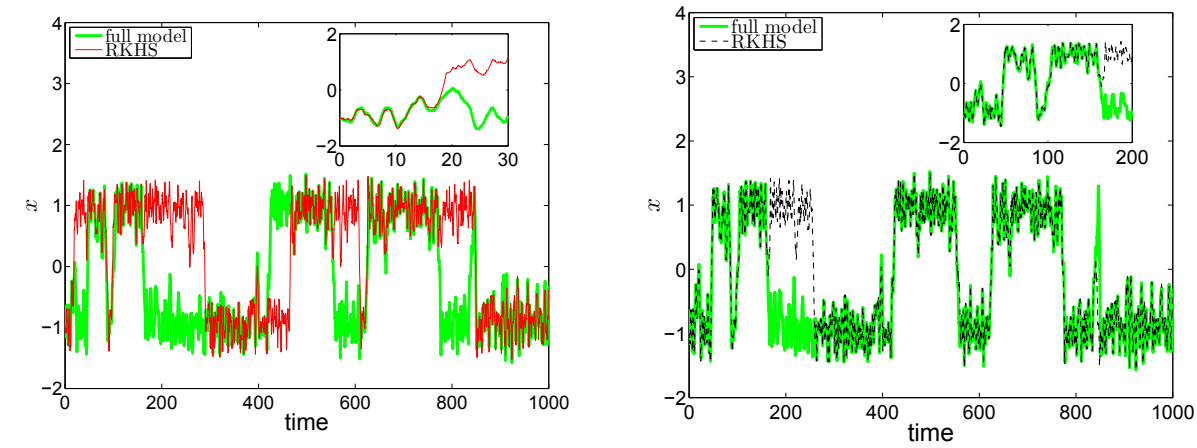
While the closure model in (6) is theoretically consistent and the example above shows a very promising approach, in real applications, the function $\Theta(x, y) \neq y$ since the unresolved variables, y, are usually not identifiable from the data $\{x_t\}$ and the map $\mathcal{F}(x_t, \theta_t)$. Even if the full data of y are available, they are very high-dimensional relative to θ. What is usually identified is θ in the sense of (2), which yields the same dimensionality as the resolved variable x.

In this case, let us rewrite the underlying dynamics in (1) as a function of (x, θ). To do this, we consider the Koopman operator $S : \mathcal{H} \to \mathcal{H}$ defined as, $S \circ f = f \circ \Phi$, for $f \in \mathcal{H}$, a space of $(\mathcal{X}\times\mathcal{W})$-valued functions $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{W}$, equipped with an inner product, $\langle f, g\rangle_{\mathcal{H}} := \int_\Omega \langle f(\omega), g(\omega)\rangle_{\mathcal{X}\times\mathcal{W}}\, d\mu(x)$, weighted by the invariant measure μ. The main interest is of observable function $\pi \in \mathcal{H}$ defined as $\pi := (X, \Theta)$, such that $\pi(x, y) = (X(x, y), \Theta(x, y)) = (x, \theta), \forall (x, y) \in \mathcal{X} \times \mathcal{Y}$, which can be interpreted as a map that changes coordinate. For identifiable unresolved component as defined in (2), then $\mathcal{X} = \mathcal{W} = \mathbb{R}^{d_x}$ for d_x-dimensional real-valued observable; if $\mathcal{Y} = \mathbb{R}^{d_y}$, where $d_y \gg d_x$, then the random variable π maps vectors in $\mathbb{R}^{d_x+d_y}$ into vectors in $\mathbb{R}^{2d_x}$. Let $P : \mathcal{H} \to \mathcal{H}$ be an orthogonal projection operator to a closed set of functions of (x_0, θ_0), namely $\mathcal{V} = \{f \in \mathcal{H} : f = g \circ \pi \,|\, g : \mathcal{X} \times \mathcal{W} \to \mathcal{X} \times \mathcal{W}\}$. To this end, we also define $Q := I - P$ be the projection map to the orthogonal space $\mathcal{V}^\perp \subset \mathcal{H}$. We now rewrite the dynamics of the observables (x, θ) by applying the Dyson formula [17, 38],

$$S^{t+1} = \sum_{k=0}^{t} S^{t-k} PS\,(QS)^k + (QS)^{t+1},$$

(a) Trajectories ($N = 50,000$)

(b) Trajectories ($N = 500,000$)



(c) Densities

(d) ACVs

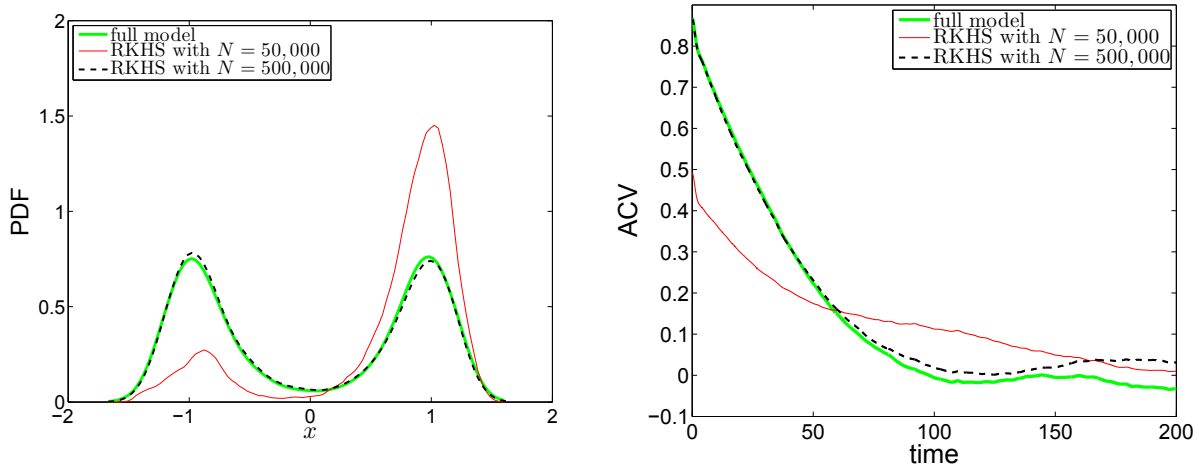


Figure 1: The top panels display comparison of the trajectories between the full model (green) and RKHS closure model using different size of training dataset. The bottom panels compare PDFs and auto-covariance functions (ACVs) among the full model (green), RKHS closure models.

on π, evaluated at initial condition (x_0, y_0) with distribution μ. The left-hand-side term, $S^{t+1}\pi(x_0, y_0) = \pi(\Phi^{t+1}(x_0, y_0)) = \pi(x_{t+1}, y_{t+1}) = (X(x_{t+1}, y_{t+1}), \Theta(x_{t+1}, y_{t+1})) = (x_{t+1}, \theta_{t+1})$. Evaluating the first term on the right-hand-side on $\pi(x_0, y_0)$, we obtain $S^{t-k}PS(QS)^k\pi(x_0, y_0) = PS(QS)^k\pi(x_{t-k}, y_{t-k})$. For $k = 0$, we obtain the Markovian term,

$$PS\pi(x_t, y_t) = P(\pi \circ \Phi(x_t, y_t)) = P(\pi \circ (\mathcal{F}(x_t, y_t), \mathcal{G}(x_t, y_t))). \tag{12}$$

In order to write (12) in terms of (x_t, θ_t), we use the definition of $\mathcal{V}$. Particularly, since $PS\pi \in \mathcal{V}$, there exists a function $\bar{\Phi}_0 : \mathcal{X} \times \mathcal{W} \to \mathcal{X} \times \mathcal{W}$ such that $PS\pi = \bar{\Phi}_0 \circ \pi$. Note that PS is a linear operator whereas Φ_0 is a function (possibly nonlinear) where they share the same range in $\mathcal{V}$. Let $\bar{\Phi}_0 = (\bar{\mathcal{F}}_0, \bar{\mathcal{G}}_0)$, where the maps $\bar{\mathcal{F}}_0 : \mathcal{X} \times \mathcal{W} \to \mathcal{X}$ and $\bar{\mathcal{G}}_0 : \mathcal{X} \times \mathcal{W} \to \mathcal{W}$ denote the $x-$ and $\theta-$components of the Markovian term in (12), respectively. Since π is identity in the x-direction, then the $x-$component of (12) is simply $P\mathcal{F}(x_t, y_t)$. But since $\mathcal{F}(x, y) := \mathcal{F}(x, \theta)$ as we explained prior to Eq. (2), it is clear that $P\mathcal{F}(x_t, y_t) = \mathcal{F}(x_t, \theta_t)$, which is nothing but $\bar{\mathcal{F}}_0(x_t, \theta_t)$.

For the memory terms, $k > 0$, since $PS(QS)^k\pi \in \mathcal{V}$, there exists $\bar{\Phi}_k : \mathcal{X} \times \mathcal{W} \to \mathcal{X} \times \mathcal{W}$ such that $PS(QS)^k\pi = \bar{\Phi}_k \circ \pi$, by the definition of $\mathcal{V}$. As before, we let $\bar{\Phi}_k = (\bar{\mathcal{F}}_k, \bar{\mathcal{G}}_k)$ such that the maps $\bar{\mathcal{F}}_k : \mathcal{X} \times \mathcal{W} \to \mathcal{X}$ and $\bar{\mathcal{G}}_k : \mathcal{X} \times \mathcal{W} \to \mathcal{W}$ denote the $x-$ and $\theta-$components of the memory functions, respectively. Since $Q\mathcal{F} = 0$, the $x-$component of the dynamics has no memory term and therefore $\bar{\mathcal{F}}_k = 0, k > 0$. Putting all these together, we have

$$\begin{aligned} x_{t+1} &= \mathcal{F}(x_t, \theta_t), \\ \theta_{t+1} &= \bar{\mathcal{G}}_0(x_t, \theta_t) + \sum_{k=1}^{t} \bar{\mathcal{G}}_k(x_{t-k}, \theta_{t-k}) + (QS)^{t+1}(x_0, y_0). \end{aligned} \tag{13}$$

Here the dynamics of θ inherits Markovian, memory and orthogonal terms, where the last term is usually regarded as noise with the randomness corresponding to the distribution of the initial condition, μ.

Conceptually, one can consider learning the entire dynamics of θ by considering the target function $\mathbb{E}[\Theta_{t+1}|\boldsymbol{z}_{t,t}]$, conditioned to the entire history of $\boldsymbol{z}_{t,t} = (\boldsymbol{x}_{0:t}, \boldsymbol{\theta}_{0:t})$, which is not practical. In applications, however, the length of the memory is often finite, $0 < m < t$, and it can be estimated as shown in [24, 58]. In fact, for nonlinear systems with linear dual closure (that is, the vector field is linear in terms of θ), one can determine the minimum length of m to guarantee the identifiability of θ_{t+1} from observed data of $\boldsymbol{z}_{t,m} := (\boldsymbol{x}_{t-m:t}, \boldsymbol{\theta}_{t-m:t}) \in \mathcal{Z}$ [57]. We should point out that while the linear dual closure condition is satisfied in the Langevin example above, it is not necessarily satisfied for all examples in Section 4, that is, while the dependence of $\mathcal{F}$ on θ is linear as in (2), the dependence of $\bar{\mathcal{G}}_0$ on θ in (13) may not be linear. Also, the second fluctuation-dissipation theorem [76] states that the time correlation of the orthogonal dynamics is proportional to the memory function. This suggests that if the memory term dissipates beyond m-lags, then the orthogonal dynamics should also depend on state variables of not longer than m-lags. With this in mind, let us rewrite (13) as,

$$\theta_{t+1} = \bar{\mathcal{G}}_0(x_t, \theta_t) + \sum_{k=1}^{m} \bar{\mathcal{G}}_k(x_{t-k}, \theta_{t-k}) + (QS)^{m+1}\pi(x_{t-m}, y_{t-m}) + R_{t+1},$$

where

$$R_{t+1} := \sum_{k=m+1}^{t} \bar{\mathcal{G}}_k(x_{t-k}, \theta_{t-k}) + (QS)^{m+1}\left((QS)^{t-m}\pi(x_0, y_0) - \pi(x_{t-m}, y_{t-m})\right), \tag{14}$$

denotes the modeling error due to finite memory assumption. Depending on the choice of learning algorithm and model (e.g., width or depth of the deep neural network), the estimator $\mathbb{E}^\epsilon[\Theta_{t+1}|\boldsymbol{z}_{t,m}]$ is usually

found by minimizing the variance of $R_t \sim \Sigma$, where Σ denotes the random variable of the error. From the learning perspective, this modeling error is usually minimized in the training phase with the hope that the corresponding generalization error will be not much larger than the training error. For the convenient of the analysis below, we first assume that the finite memory approximation holds such that $R_{t+1} = 0$.

For the convergence analysis, we require that the following assumption.

Assumption 2. *Let PS and QS be bounded linear operators defined with respect to functions $f \in C(\mathcal{X}\times\mathcal{Y})$.*

With these assumptions, let $(x,\theta),(x',\theta') \in \mathcal{X}\times\mathcal{W}$, where $(x,\theta) = \pi(x,y)$, $(x',\theta') = \pi(x',y')$, and $y, y' \in \mathcal{Y}$. For $\pi \in C(\mathcal{X}\times\mathcal{Y}) \cap \mathcal{H}$, notice that $|\bar{\Phi}_k(x,\theta) - \bar{\Phi}_k(x',\theta')| = |PS(QS)^k\pi(x,y) - PS(QS)^k\pi(x',y')| \le |PS(QS)^k||\pi(x,y) - \pi(x',y')| \le C|(x,\theta) - (x',\theta')|$, where the $|\cdot|$ denotes the appropriate uniform norms. Therefore, this technical assumption says that $\bar{\Phi}_k$ are Lipschitz continuous on x and θ, which is needed to bound the errors in term of x and θ. Notice that if P and π are both identity maps and $k = 0$, this assumption is equivalent to saying $\Phi = (\mathcal{F},\mathcal{G})$ is Lipschitz in x and y, which is assumed in both theorems in previous section. For the decomposition in (2), where $\mathcal{F}$ is linear in θ, this assumption means that $\mathcal{F}$ is also Lipschitz in the x.

Theorem 3. *Let x_{t+1} be the solution of*

$$\begin{aligned} x_{t+1} &= \mathcal{F}(x_t,\theta_t),\\ \theta_{t+1} &= \bar{\mathcal{G}}_0(x_t,\theta_t) + \sum_{k=1}^{m}\bar{\mathcal{G}}_k(x_{t-k},\theta_{t-k}) + (QS)^{m+1}\pi(x_{t-m},y_{t-m}) = \mathbb{E}[\Theta_{t+1}|\boldsymbol{z}_{t,m}], \end{aligned} \tag{15}$$

where $\pi \in C(\mathcal{X}\times\mathcal{Y}) \cap \mathcal{H}$ and the Assumption 2 is satisfied. Let $\hat{x}_{t+1}$ be the solution of

$$\hat{x}_{t+1} = \mathcal{F}(\hat{x}_t,\hat{\theta}_t), \qquad \hat{\theta}_{t+1} = \mathbb{E}^{\epsilon}[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}] + \xi_{t+1}, \tag{16}$$

under the same initial conditions $x_{-m:0} = \hat{x}_{-m:0}, \theta_{-m:0} = \hat{\theta}_{-m:0}$. In (16)*, the noise amplitude $\xi_t \sim \Xi$ is estimated according to* (4) *using the training data. Under the Assumptions 1, if $\mathbb{E}^{\epsilon}[\Theta_{t+1}|\cdot] \to \mathbb{E}[\Theta_{t+1}|\cdot]$ as $\epsilon \to 0$ with variance error of order ϵ^2, then*

$$\mathbb{E}\Big[\max_{t\in\{0,\ldots,T+1\}}|\hat{x}_t - x_t|\Big] = O(a^T\epsilon)$$

where $a > 1$ is a constant that is independent of T and ϵ.

Proof. See Appendix C. □

We should point out that if the underlying dynamic solves the full Mori-Zwanzig equation in (13) and the approximate dynamic in (16) commits modeling errors R_{t+1} with error variance of order ϵ_m^2, then the error bound above becomes $O(a^T(\epsilon + \epsilon_m))$. The only change in the proof in Appendix C is that the Eqn. 40 has an additional order-ϵ_m term due to the model error.

For dynamical systems driven by stochastic noises, one can rewrite the full dynamics as an autonomous dynamical system by augmenting (x_n, y_n) with the entire history of the noises. See [34] for the details or the Appendix of [38] for the key idea. Subsequently, one can apply the Dyson formula to the resulting autonomous dynamics, defined on appropriate state space that includes $\mathcal{X}\times\mathcal{Y}$ and the space of the history of the noises, and derive an analogous representation as in (13). We suspect that the result is not different from that in Theorem 3 and thus we will not pursue this derivation.

Again, this error rate is rather loose with unknown coefficients $a > 1$ since no other assumptions are included in $\mathcal{F},\mathcal{G}$. One might achieve an improved error rate by analyzing the eigenvalue problem corresponding to the autoregressive model of order-m, which bounds the dynamical equation for the errors between (13) and (3). Another plausible approach is to consider a class of dynamical systems with spatially short-range interaction as studied in [12], which will require further investigation.

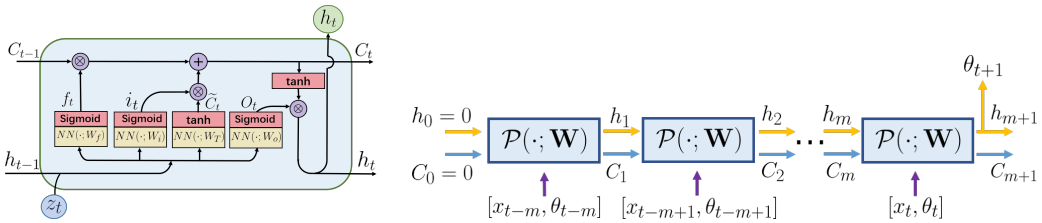


Figure 2: Left: the basic computational flow of a LSTM recurrence. + and $\bigotimes$ are element-wise addition and multiplication respectively. Right: a sequence of LSTM cells applied compositionally.

3 Deep Learning via Long-Short-Term-Memory

As a nonlinear type parametric regression method, deep learning outperforms kernel methods including the RKHS approach in terms of generalization error when the target functions are sufficiently smooth [18]. Though it is theoretically unclear whether there are any advantages of using deep learning over other non-parametric regression methods for general continuous functions in terms of overcoming the curse of dimension [5, 51, 53, 52], deep learning has practical advantages over the RKHS approaches. A significant challenge with the RKHS approximation in (10) is that there is no a priori guideline for choosing the appropriate hypothesis space. If the orthogonal basis is used, it is practically difficult to even construct these basis functions on very high-dimensional variables $\boldsymbol{z} \in \mathcal{Z}$. On the other hand, if arbitrary radial functions are used as a basis, the evaluation of the resulting model on a new point $\boldsymbol{z}$ requires evaluating the basis functions on $\|\boldsymbol{z} - \boldsymbol{z}_{t,m}\|$ for all training data $t = 1, \ldots, N$, predicting with (3) becomes too costly since we need to evaluate the conditional expectation in (10) on a new point in each iteration. In contrast, deep learning as a nonlinear parametric regression method is not hampered by these issues, since it is practically just a nonlinear interpolation technique using a composition of nonlinear activation functions and linear transforms. Of course, the main issue with nonlinear regression is whether one can obtain the minimum on such a non-convex optimization problem in the training phase. Recent advances in optimization theory show that simple gradient descent can identify a local minimizer with an arbitrarily small loss and a generalization error without the curse of dimensionality when the network size is sufficiently large for classification problems [9]. Though there is no existing optimization theory that guarantees good local minimizers in general settings, motivated by many positive numerical results shown in other closure modeling approaches [70, 50], we will consider realizing the closure model in (3) using recurrent neural networks.

As a special case of recurrent neural networks, Long-Short-Term-Memory (LSTM) is capable of learning multi-scale temporal effects and hence is adopted in our method. The computational flow of the LSTM consists of a sequence of computational cells, each of which is

$$\begin{aligned} f_t &= \sigma \circ NN(h_{t-1}, \boldsymbol{z}_t; W_f), \quad i_t = \sigma \circ NN(h_{t-1}, \boldsymbol{z}_t, W_t) \\ o_t &= \sigma \circ NN(h_{t-1}, \boldsymbol{z}_t; W_o), \quad \tilde{C}_t = \tanh(NN(h_{t-1}, \boldsymbol{z}_t; W_T)), \\ C_t &= f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t, \quad h_t = o_t \otimes \tanh(C_t), \end{aligned}$$

where σ denotes the sigmoid function, $\otimes$ is the pointwise product, and NN denotes a fully connected network which stacks layers of linear transformation and nonlinear activation function. See Fig. 2 (left) for an illustration of an LSTM cell. For simplicity, let us denote the above flow as $(h_t, C_t) = \mathcal{P}(\boldsymbol{z}_t, h_{t-1}, C_{t-1}; W)$ with parameters W, inputs $(\boldsymbol{z}_t, h_{t-1}, C_{t-1})$, and outputs (h_t, C_t). LSTM cells can be applied compositionally and we denote the LSTM sequence with $m+1$ cells as $(h_{m+1}, C_{m+1}) = \mathcal{P}_m(\{\boldsymbol{z}_t\}_{t=1}^{m+1}, h_0, C_0; W)$ (see Fig. 2 (right) for an illustration).

Now let us apply the LSTM to approximate the closure model in (3) with the given training data $\{\boldsymbol{z}_{t,m}, \theta_t\}$, where $\boldsymbol{z}_{t,m} := (\boldsymbol{x}_{t-m:t}, \boldsymbol{\theta}_{t-m:t}) \in \mathcal{Z}$. We train an $(m+1)$-cell LSTM $(h_{m+1}, C_{m+1}) = \mathcal{P}_m(\boldsymbol{z}_{t,m}, h_0, C_0; W)$ with

an input in the j-th cell as $(x_{t-m+j-1}, \theta_{t-m+j-1})$ such that h_{m+1} predicts θ_{t+1} well. The parameters h_0 and C_0 are set to be 0 for simplicity and W is identified via minimizing a mean squared error (MSE) function specified below. In what follows, we adopt the notation $h_{m+1} = \mathcal{P}_m(\boldsymbol{z}_{t,m}; W)$ for simplicity. Define the MSE loss as

$$\mathcal{L}(W) := \frac{1}{N-m-1} \sum_{s=m+1}^{N-1} (\mathcal{P}_m(\boldsymbol{z}_{s,m}; W) - \theta_{s+1})^2, \tag{17}$$

i.e., we aim to identify a predictor $\mathcal{P}_m(\boldsymbol{z}_{s,m}; W)$ such that it can predict the $(s+1)$-th sample in the time series given $(m+1)$ preceding samples. Minimizing (17) can be achieved efficiently via a mini-batch stochastic gradient descent (SGD) and backpropagation through time (BPTT) [55, 62, 72]. Though the global minimizer of the above highly non-convex optimization might not be available, empirically numerical results [59] and partial theoretical analysis show that gradient-based algorithms are able to provide a minimizer W^* with a reasonably good generalization capacity in the case of over-parametrized networks [3, 1, 11]. Once W^* has been identified, $\theta_{t+1} = \mathcal{P}_m(\boldsymbol{z}_{t,m}; W^*)$ is applied instead of the conditional expectation in (3).

The computational cost of the proposed framework mainly consists of two parts: training and evaluation. In the training part, suppose the SGD with a batch size K is applied to minimize the loss in (17), then the computational cost for evaluating an approximate loss and gradient via BPTT is essentially $O(Km)$ matrix-vector multiplications with a matrix of size $d_L \times d_L$, where d_L is the dimension of the hidden layer of NN. According to the approximation theory of DNNs [43, 63], d_L is required to be larger than d; based on the optimization analysis of DNNs [8, 2, 9], a larger d_L admits a simpler optimization problem in deep learning. In the numerical examples shown below, we empirically set $d_L = 500$ for problems with dimensions $d = 40, 80$, and set $d_L = 1000$ for a problem that involves $d = 480$. Hence, the total computational cost is $O(Kmd_L^2)$ for one iteration in the SGD, which has been parallelized via GPU computing in standard machine learning packages. Recent theoretical analysis shows that the convergence of SGD is linear under mild conditions [3]. Therefore, the upper bound of the iteration number in the SGD to guarantee a loss less than ϵ is $O(\log(\frac{1}{\epsilon}))$. In our numerical examples below, the total numbers of iterations required for accurate performance are 4×10^3 for the first problem and 4×10^4 for the last two problems. In the evaluation part, the cost estimation for predicting one data sample with m pairs of historical data is $O(md_L^2)$, which has been parallelized via GPU.

4 Numerical Examples

In this section, we numerically demonstrate the effectiveness of our proposed closure framework on severely truncated dynamical systems in three prototypical applications. First, the topographic mean flow interaction that mimics the blocked and unblocked patterns observed in the atmosphere [10, 25, 47, 20] is considered. In our test, we will use the stochastic version of the 57-mode model studied in [61]. Second, the nonlinear Schrödinger equation that finds many applications in optics and Bose-Einstein-Condensate (see the references in [31]) is studied. Finally, the Kuramoto-Shivashinsky equation with a spatiotemporal chaotic pattern formation with applications in trapped ion modes in plasma [37] and phase dynamics in reaction-diffusion systems [35]. We shall see that the closure models in these three examples progressively involve approximations of functions of dimensions 40 to 480.

4.1 Topographic mean flow interaction

We consider the topographic mean flow interaction that solves a barotropic quasi-geostrophic equation with a large-scale zonal mean flow $u(t)$ on a two-dimensional $2\pi \times 2\pi$ periodic domain, formulated as in [61]:

$$\begin{aligned} &\frac{du}{dt} + \oint \frac{\partial h}{\partial x}\psi = -\bar{d}u + \sigma\mu^{-1/2}\dot{W}_0, \\ &\frac{\partial \omega}{\partial t} + \nabla^{\perp}\psi \cdot \nabla q + u\frac{\partial q}{\partial x} + \beta\frac{\partial \psi}{\partial x} = -\mathcal{D}\psi + \Sigma\dot{W}. \end{aligned} \tag{18}$$

Here, $q = \omega + h$ is the small-scale potential vorticity which is advected by the velocity field $\mathbf{v} = \nabla^{\perp}\psi \equiv \left(-\partial_y\psi, \partial_x\psi\right)$; $\omega = \Delta\psi$ and ψ are the relative potential vorticity and the stream function, respectively; $h(\mathbf{x}) = h(x, y)$ is the topography. The parameter β is associated with the β-plane approximation to the Coriolis force. The integral in (18) is a two-dimensional integral over a periodic box of $[-\pi, \pi] \times [-\pi, \pi]$. On the right hand side of (18), the dissipation and forcing operators are applied on both the small and the large scales. On the small scale, the dissipation operator is in the form of $\mathcal{D} = -\bar{d}\Delta$ with $\bar{d} \geq 0$ and Δ the Laplace operator corresponding to the Ekman drag dissipation. On the large scale, operator $-\bar{d}u$ represents the momentum damping. The forcing terms are represented by random Gaussian white noises (e.g. unresolved baroclinic instability processes on small scales, random wind stress, etc), where $W(t)$ and $W_0(t)$ are standard Wiener processes; $\sigma\mu^{-1/2} > 0$ is a constant amplitude and Σ is spatially dependent.

Following [47, 20, 61], we construct a set of special solutions to (18), which inherit the nonlinear coupling of the small-scale vortical modes with the large-scale mean flow via topographic stress. Consider the truncated spectral expansion of the state variables for ψ and ω with high wavenumber truncation $1 \leq |\mathbf{k}| \leq K$ using standard Fourier basis $\exp(i\mathbf{k} \cdot \mathbf{x})$ with $\mathbf{k} = \left(k_x, k_y\right)$. We can rewrite (18) for the large-scale mean flow $u(t)$ in a truncated Fourier form, as:

$$\begin{aligned} \frac{du}{dt} &= i \sum_{1 \leq |\mathbf{k}| \leq K} \frac{k_x}{|\mathbf{k}|^2}\hat{h}^*_{\mathbf{k}}\omega_{\mathbf{k}} - d\left(u - u_{eq}\right) + \sigma\mu^{-1/2}\dot{W}_t, \\ \frac{d\omega_{\mathbf{k}}}{dt} &= \mathcal{P}_{K,\mathbf{k}}\left(\nabla^{\perp}\psi_N \cdot \nabla q_N\right) + ik_x\left(\frac{\beta}{|\mathbf{k}|^2} - u\right)\omega_{\mathbf{k}} - ik_x\hat{h}_{\mathbf{k}}u - d\left(\omega_{\mathbf{k}} - \omega_{eq,\mathbf{k}}\right) + \sigma_{\mathbf{k}}\dot{W}_{\mathbf{k},t}, \quad 1 \leq |\mathbf{k}| \leq K, \end{aligned} \tag{19}$$

Here, $\hat{h}_{\mathbf{k}}$ and $\omega_{\mathbf{k}}$ are the Fourier transform of the topography $h(\mathbf{x})$ and the relative potential vorticity ω, respectively; $u_{eq} = -\beta/\mu$ is the equilibrium mean of $u(t)$; $\omega_{eq,\mathbf{k}} = -|\mathbf{k}|^2 \hat{h}_{\mathbf{k}}/\left(\mu + |\mathbf{k}|^2\right)$ is the mean relative vorticity; $\sigma_{\mathbf{k}} = \sigma\left(1 + \mu|\mathbf{k}|^{-2}\right)^{-1/2}$ is the forcing strength for each mode $\mathbf{k}$. The parameter σ is chosen such that $\sigma^2_{eq} = \frac{\sigma^2}{2\bar{d}} = 1$. The parameters $\beta = 1$ and $\mu = 2$ are fixed in our simulation.

In our implementation, we consider the ground truth as the solution corresponding to the truncation $1 \leq |\mathbf{k}| \leq K$ with $K = 17$ such that there are 57 degrees of freedom for integers $\mathbf{k} = \left(k_x, k_y\right)$. In this topographic 57-mode model, we use the standard 4th order Runge-Kutta method for the time integration up to 5×10^7 time iterations with a time step $\delta t = 2.5\text{E}{-3}$, which is small enough to capture the small-scale dynamics. For the nonlinear advection term, $\mathcal{P}_{K,\mathbf{k}}(\nabla^{\perp}\psi_K \cdot \nabla q_K)$, the 2/3 rule is applied for de-aliasing [61]. The noise is added using the standard Euler-Maruyama scheme. Here, the initial condition, $\psi(\boldsymbol{x}, 0)$, is a sample of Gaussian distribution with random phases and amplitudes consistent with the ensemble mean and enstrophy as in [46]. The observed data are recorded at every 20 time steps, that is, we observe the data at every $\Delta = 0.05$ time unit. Taking half of this data set for training, $N = 1.25\times10^6$ samples. For the topography $h(\mathbf{x})$, we use a simple layered topography with variation only in the x-direction, $h(\mathbf{x}) = H(\cos(x) + \sin(x))$, where H denotes the topography amplitude.

We now present the closure model for the large-scale mean flow $u(t)$ in Eq. (19). The application of the

Euler-Maruyama scheme for the large-scale mean flow $\hat{u}(t)$ gives

$$\begin{aligned}\hat{u}_{t+1} &= \hat{u}_t + \Delta\,\hat{\theta}_t - \Delta\bar{d}\left(\hat{u}_t - u_{eq}\right) + \sqrt{\Delta}\sigma\mu^{-1/2}\eta_{t+1},\\ \hat{\theta}_{t+1} &= \mathbb{E}^{\epsilon}\Big[\Theta_{t+1}|\hat{u}_{t-m:t},\hat{\theta}_{t-m:t}\Big] + \xi_{t+1},\end{aligned} \tag{20}$$

where the time step $\Delta = 0.05$ and $\hat{\theta}_t$ is an estimator of the identifiable unresolved variable. In this case, $\theta = \theta(\omega_{\mathbf{k}}) = i\sum_{1\le|\mathbf{k}|\le K}\frac{k_x}{|\mathbf{k}|^2}\hat{h}^*_{\mathbf{k}}\omega_{\mathbf{k}}$, is a function of the unresolved variables alone. The noises η_t are i.i.d. standard Gaussian while the noises ξ_t are Gaussian with mean zero and variance Ξ determined from the training residual using Eq. (4). We will approximate the conditional expectation $\mathbb{E}^{\epsilon}$ in (20) with the LSTM model with $m = 19$, which involves an approximation of a forty dimensional function. We should point out that the results become inaccurate when m is too small. Taking the efficiency of computation into consideration, we empirically found that $m = 19$ is a convenient choice. In fact, in all LSTM examples in the remaining of this paper, we use this same number of m on the purpose of showing that our proposed framework is not sensitive to m. We will also include an experiment mimicking the existing approach in [44, 70, 50], where the conditional expectation in (20) is replaced with $\mathbb{E}^{\epsilon}\Big[\Theta_{t+1}|\hat{u}_{t-m_u:t}]$, a function that depends only on the memory of the resolved scale with memory length $m_u = 19$. In this case, the conditional expectation is a twenty dimensional function. In addition, we also report the RKHS approximation to (20) with $m = 2$, which involves only an approximation of a six-dimensional function $(\hat{u}_{t-2:t},\hat{\theta}_{t-2:t}) \mapsto \mathbb{E}^{\epsilon}[\Theta_{t+1}|\hat{u}_{t-2:t},\hat{\theta}_{t-2:t}]$. Here, we use a tensor product of Hermite polynomials to represent this six dimensional function. The curse of dimensionality makes the RKHS model with orthogonal polynomials prohibitive in higher dimensions.

To compare the pathwise trajectories for short-time forecasting in the verification phase, we need to drive the closure model in (20) with the appropriate realization of noises ξ_t and η_t, that respects the realization of the noises, $\dot{W}_{t+1}, \dot{W}_{\mathbf{k},t}$ of (19), which drive the verification trajectory. Since the closure model depends on the identifiable unresolved variable, θ, one can, in principle, account for the realization of the noise $\sigma_k\dot{W}_{\mathbf{k},t}$ corresponding to the verification trajectory by first mapping it to $\mathcal{W}$-space via the linear mapping $\theta(\cdot)$ (defined short after (20)) and use it as a realization of ξ_t. Based on our numerical inspection, we found that the variance of this noise (in $\mathcal{W}$-space) is on the order of 10^{-8}, whereas the variance of θ is on the order of 10^{-2}, so there is a large signal-to-noise ratio in the true θ-dynamics. With such a distinct signal-to-noise in the true θ-dynamics, one can expect that the trajectory of u can be recovered up to some accuracy even if this noise realization is neglected in the closure model so long as the conditional expectation model in (20) can accurately approximate the deterministic part of the unresolved dynamics. Furthermore, we found that the residuals in the training phase are small (the variances, Ξ, are on the order of 10^{-5}) in most of the dynamical regimes that we tested. Based on these observations, we will show numerical results obtained without additional noise ξ_t in (20), to verify the prediction skill of the conditional expectation model alone. We should note that in separate experiments (not reported here), we found that the differences between the results without and with noise, ξ_t, where the latter uses a completely random realization, are negligible.

As for the stochastic noise, $\dot{W}_t$, corresponds to the resolved dynamics in (19), we notice that the full model and the closure models are integrated with different time steps. The full model is integrated with a relatively small time step $\delta t = 2.5\text{E}{-3}$ in order to resolve all the small-scale vorticity modes. Nevertheless, the closure models are integrated with a relatively large time step $\Delta = 0.05$ for resolving only the large-scale mean flow $u(t)$. To compare the pathwise trajectories, we first generate a realization of the noise $\dot{W}_{t+1}$ from identifiable variables using finite difference method

$$\eta_{t+1} = \frac{u_{t+1} - \Big[u_t + \Delta\theta_t - \Delta d\left(u_t - u_{eq}\right)\Big]}{\sqrt{\Delta}\sigma\mu^{-1/2}},$$

where u_t and θ_t are the identifiable variables from the dataset for verification of the full model observed with time interval $\Delta = 0.05$. Using such realization of $\dot{W}_{t+1}$ for the noise η_{t+1} in (20), we can now compare

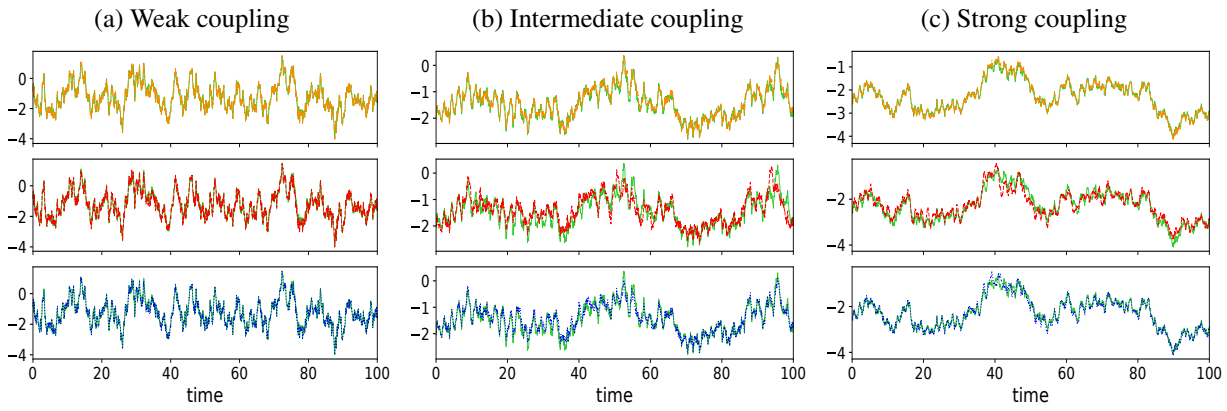


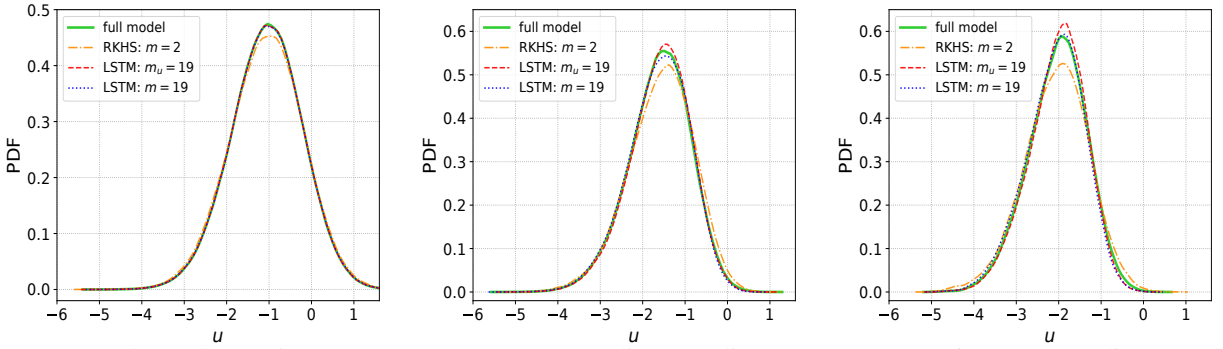
Figure 3: Comparison of trajectories between the full and closure models for the mean velocity u. The true trajectory (green solid) in all panels; RKHS $m = 2$ (yellow dash-dotted line) in the first row; LSTM $m_u = 19$ (red dashes) in the second row; LSTM $m = 19$ (blue dotted line) in the third row.

the path of the closure model with $\Delta = 0.05$ to the path of the true trajectory, starting at the same initial condition.

In Fig. 3, we present the short-time predictions of u in three regimes: weak coupling ($H = 3\sqrt{2}/4$, $\bar{d} = 0.5$), intermediate coupling ($H = 5\sqrt{2}/4, \bar{d} = 0.1$), and strong coupling ($H = 7\sqrt{2}/4$, $\bar{d} = 0.1$). These three regimes were considered in [61]. We would like to emphasize that we verify the closure model in (20) using initial conditions not in the training data set. In each regime, we compare the trajectories of the full model and the three closure models discussed above. In the weak coupling regime, one can see that the short-time predictions are all excellent among three closure models. For the other two regimes, the LSTM without the unresolved variables ($m_u = 19$) proposed in [44, 70, 50] produces the worst prediction, which justifies the importance of considering both the resolved and identifiable unresolved variables in the closure model proposed in this paper. Surprisingly, the RKHS prediction is quite accurate considering that it requires less computational effort relative to LSTM. As we discussed before, given the large signal-to-noise ratio in the θ-dynamics, the reasonably accurate trajectory recovery of the closure model in (20) without incorporating the noise realization $\sigma_k \dot{W}_{\mathbf{k},t}$ suggests that the conditional expectation model alone has captured the bulk of the deterministic part of the unresolved dynamics.

In Fig. 4, we show the comparison of the equilibrium density and Auto-Correlation Functions (ACFs) of the full model and three prediction methods. The auto-correlation function (ACF) for the large-scale mean flow u is calculated as $\langle U_t U_0 \rangle \,/\, \langle U_0 U_0 \rangle$, where $U_t = u_t - \langle u_t \rangle$ with $\langle \cdot \rangle$ being the temporal average over 1.25×10^6 data for verification, which is different from the $N = 1.25 \times 10^6$ training dataset. The probability density function (PDF) for u is obtained from the same verification dataset using the kernel density estimation (KDE) method. For the long-time statistics, both the LSTM methods provide a better approximation than the RKHS model. This is because not enough memory terms are used in the RKHS model. In the strong coupling regime, one can see that the LSTM method with $m = 19$ is the best approximation. This is because the variance of the training residual is small about 10^{-5} for the LSTM with $m = 19$ and the variance is relatively large about 10^{-1} for the LSTM with $m_u = 19$. This result confirms the robustness of our framework with a closure model that depends on, both, the memories of the resolved and identifiable unresolved variables.

(a) Weak coupling (b) Intermediate coupling (c) Strong coupling



(d) Weak coupling (e) Intermediate coupling (f) Strong coupling

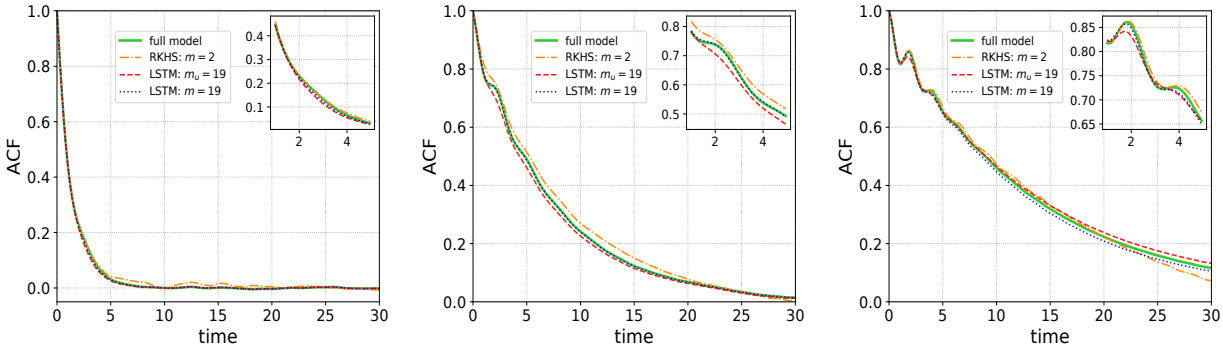


Figure 4: Comparison of densities and Auto-Correlation functions (ACF) between the full and closure models for the mean velocity u.

4.2 Nonlinear Schrödinger equation

We consider the cubic nonlinear Schrödinger (NLS) equation defined on a periodic boundary condition $[0, 2\pi]$, which dynamical equation can be written in terms of the Fourier modes as,

$$\frac{du_k}{dt} = -ik^2 u_k - i \sum_{k_1 \in \mathbb{Z}} \sum_{k_2 \in \mathbb{Z}} u_{k_1} u_{k_2} u^*_{k_1+k_2-k}. \tag{21}$$

Numerically, we generate the truth by integrating (21) on finite wavenumbers $|k| \leq K$ and Strang's splitting method in time [4]. Here, the number of modes $K = 32$ and the observation time interval $\Delta = 0.02$. The observation data length is 10^6, obtained from a single trajectory. Taking half of this data set for training, $N = 5 \times 10^5$ samples.

We simulate the initial conditions by sampling from the Gibbs distribution $\pi = \exp(-\frac{E}{k_B T})$, where E denotes the Hamiltonian of the ODE system resulting from the Fourier representation (21); k_B and T denote the Boltzmann constant and temperature, respectively. In this case, the Hamiltonian is given by $E = E_2 + E_4$,

$$E_2 = \sum_{k \in \mathbb{Z}} k^2 |u_k|^2, \qquad E_4 = \frac{1}{2} \sum_{k_1 \in \mathbb{Z}} \sum_{k_2 \in \mathbb{Z}} \sum_{k_3 \in \mathbb{Z}} u_{k_1} u_{k_2} u^*_{k_3} u^*_{k_1+k_2-k_3}.$$

We should point out that the qualitative solutions for higher temperature have larger amplitudes and frequencies. Since smaller time steps are required for accurate solutions with higher amplitude as well as the faster frequency, the problem is numerically stiff as the temperature increases. To keep the presentation short, we only show the numerical results for the zeroth mode u_0 in a high-temperature regime with $k_B T = 10$. Our numerical test on lower temperature regime (not shown) do not change the conclusion in this section. In fact, a parametric closure proposed in [26] has shown accurate recovery for extremely low temperature regime, $k_B T = 10^{-4}$, and less accurate as the temperature increases. The stiffness of high-temperature regime will also be manifested in the numerical scheme that is used in integrating the closure model as we will explained below.

In this example, we are interested in constructing a closure model for the dynamics of the zeroth mode u_0 of the NLS equation. Given the dynamical equation of the resolved variable, u_0, we can rewrite it in the form of (2) as,

$$\frac{du_0}{dt} = -i \sum_{k_1 \in \mathbb{Z}} \sum_{k_2 \in \mathbb{Z}} u_{k_1} u_{k_2} u^*_{k_1+k_2} := -i\Big(|u_0|^2 u_0 + \theta\Big), \tag{22}$$

where θ is basically the full vector field without the cubic term that involves only u_0 in the right-hand-side of (22). The closure model is obtained by concatenating a discretization of (22) with time step Δ with a map, $\mathbb{E}^\epsilon\Big[\Theta_{t+1}|\cdot\Big] : \mathbb{R}^{(m+1)\times 4} \to \mathbb{R}^2$, defined as,

$$\theta_{t+1} \quad = \mathbb{E}^\epsilon\Big[\Theta_{t+1} | u_{0,t-m:t}, \boldsymbol{\theta}_{t-m:t}\Big]. \tag{23}$$

Here θ_{t+1} denotes the unresolved identifiable component at discrete time $t + 1$. To train this model, we need a time series of $\{\theta_t\}$ in addition to $\{u_{0,t}\}$. Based on the form of the resolved dynamics in (22), given a training time series of $\{u_{0,t}\}$, we extract $\{\theta_t\}$ by a direct subtraction and a finite difference approximation to the derivative. However, we should point out that if we reverse-engineer this step, that is, solve (22) with the true initial condition of $u_0(0)$ using a lower-order scheme (such as Euler method) and directly use the data $\{\theta_t\}$ that we just obtained from direct subtraction, the solution for $u_0(t)$ will blow up in finite-time. This is a manifestation of the stiffness of this problem. To avoid this issue in the closure model, we apply the following time-splitting method in our numerical discretization of (22). That is, we use the Euler scheme

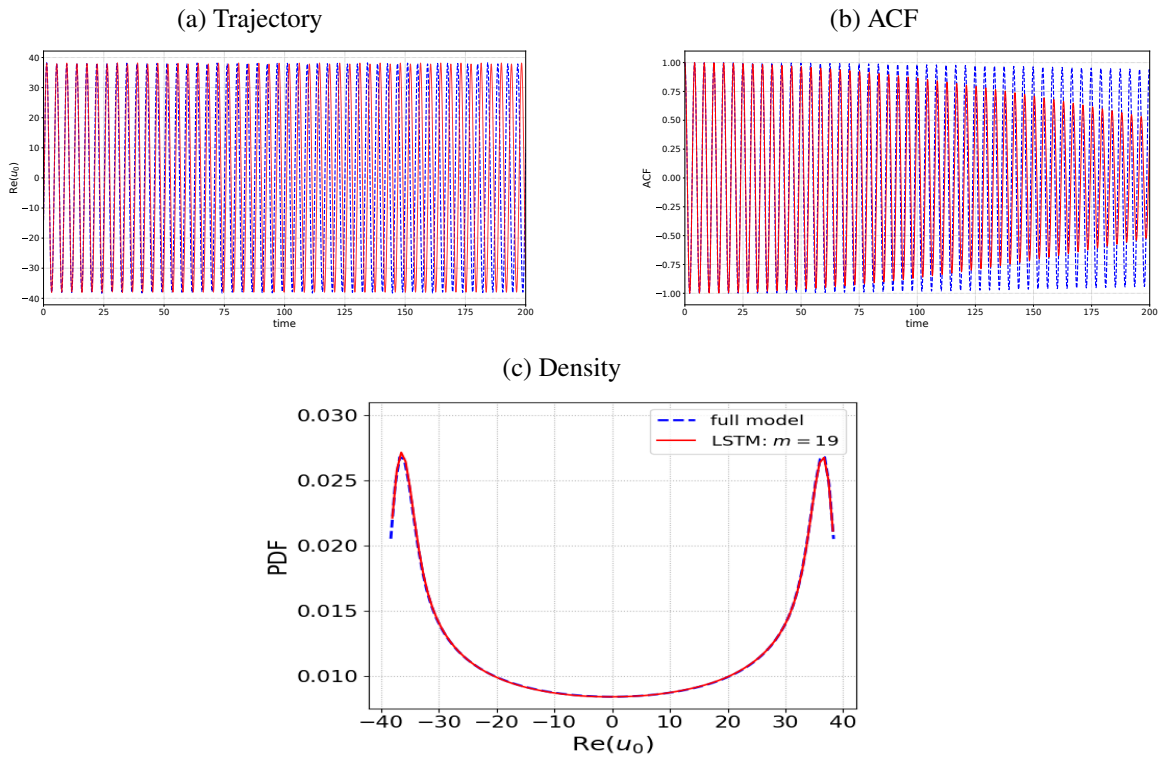


Figure 5: Comparison of (a) trajectories, (b) ACFs, and (c) densities of the full and closure LSTM models in the high temperature regime with $k_BT = 10$. The time step for the closure model is $\Delta = 0.02$ and the LSTM method uses $m = 19$ memory terms for $(\boldsymbol{\theta}, u_0)$ in Eq. (23). Full model (blue dashes); LSTM $m = 19$ (red solid).

to solve the linear ODE, $du_0/dt = -i\theta$, since we only have discrete estimates of θ, and we use the explicit solution $u_0(t) = u_0(t_0)\exp(-i|u_0(t_0)|^2 t)$ for the nonlinear ODE, $du_0/dt = -i|u_0|^2 u_0$.

In this numerical experiment, we fix the memory length to be $m = 19$ in the LSTM method, resulting in an approximation of 80-dimensional function $\mathbb{E}^{\epsilon}\left[\Theta_{t+1}|\cdot\right]$. No residual term is added in (23). For the short-time forecasting, we observe from Fig. 5(a) that the path-wise solution of Re(u_0) is well captured for a sufficiently long time; the discrepancies in the frequencies are noticeable as time increases. In Fig. 5(b) and (c), we also reported the ACF for the Re(u_0), calculated by a temporal average over 5×10^5 verification data, which is different from the $N = 5\times 10^5$ training dataset. The PDF for Re(u_0) is obtained from the same verification dataset using the kernel density estimation (KDE) method. Notice that both the ACF (below time 40 unit) and the density of the true u_0 are also well reproduced. Therefore, for the first mode u_0 of the NLS equation, the proposed closure model using the LSTM method can reasonably replicate the short-time forecasting skill and long-time statistics in the high-temperature regime.

We should point out that the resulting model is only valid in predicting the evolution of the system on the same energy level since the underlying Hamiltonian system is not ergodic. This implies that the verification will only be valid to predict the evolution of the system with initial conditions sampled from the same Gibbs distribution where the training data is generated from.

4.3 The Kuramoto-Sivashinsky equation

We consider the Kuramoto-Sivashinsky equation (KSE) on an $L-$periodic domain, the Fourier representation of which can be written as

$$\frac{d}{dt}v_k = \left(q_k^2 - q_k^4\right)v_k - \frac{iq_k}{2}\sum_{l=-\infty}^{\infty} v_l v_{k-l}, \tag{24}$$

where $q_k = 2\pi k/L$ with $k \in \mathbb{Z}$, and v_k denotes the kth Fourier mode.

In our numerical implementation, we let the full dynamics to be the Galerkin truncation of (24) for $|k| \leq K/2$, where $K = 96$. Notice that in the linearized equations of (24), each Fourier mode has an eigenvalue $q_k^2 - q_k^4$ so that high k modes with $|q_k| > 1$ are linearly stable whereas low k modes with $|q_k| \leq 1$ are not. We set the spatial length $L = 2\pi/\sqrt{0.085}$ so that the number of linearly unstable modes is $\left\lfloor 1/\sqrt{0.085} \right\rfloor = 3$. In this case, the energy is transferred from the linearly unstable low 3 modes to the damped high $K/2 - 3 = 45$ modes through the nonlinear terms so that the KSE is well-posed and the solutions remain globally bounded in time [23]. This regime is exactly the same as the one considered in [40, 38].

We predict the six leading modes of the KSE with the following partial dynamics,

$$\frac{d}{dt}\hat{v}_k = \left(q_k^2 - q_k^4\right)\hat{v}_k - \frac{iq_k}{2}\sum_{1\leq|l|,|k-l|\leq 6} \hat{v}_l\hat{v}_{k-l} + \hat{\theta}_k, \quad k = 1, \ldots, 6. \tag{25}$$

In this case, since the nonlinear terms in (25) only involve summation of terms that are restricted to $1 \leq |l|, |k-l| \leq 6$, the identifiable unresolved variables, θ_k, depends also on the resolved modes, in addition to the unresolved modes. The proposed closure model is to concatenate the numerical discretization of (25) with the discrete nonparametric closure model,

$$\hat{\boldsymbol{\theta}}_{t+1} = \mathbb{E}^{\epsilon}\left[\Theta_{t+1}|\hat{\boldsymbol{v}}_{t-m:t}, \hat{\boldsymbol{\theta}}_{t-m:t}\right], \tag{26}$$

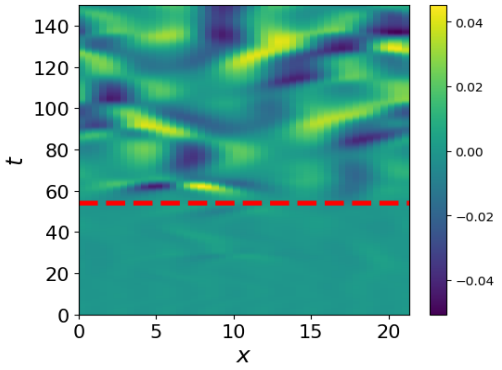
where $\hat{\boldsymbol{\theta}}_t = (\hat{\theta}_{1,t}, \ldots, \hat{\theta}_{6,t}) \in \mathbb{C}^6$ and $\hat{\boldsymbol{v}}_t = (\hat{v}_{1,t}, \ldots, \hat{v}_{6,t}) \in \mathbb{C}^6$. In our numerical experiment, we set $m = 19$ such that $\mathbb{E}^{\epsilon}$ in (26) is a function that maps a real-valued vector of size $(19 + 1) \times 12 \times 2 = 480$ to a 12-dimensional vector consisting of the real and imaginary values of $\hat{\boldsymbol{\theta}}_{t+1}$. To evolve the dynamics in (25)-(26), we discretize (25) with the midpoint rule and a time step Δ.

In our numerical experiment, the true time series for training are obtained by integrating the full dynamics, that is, (24) truncated on $1 \leq |k| \leq 48$ with a time step $\delta t = 0.005$. We observe only the first 6 modes at a time step $\Delta = 0.05$. The size of the training data set is $N = 2.5 \times 10^5$. The identifiable unresolved variable, $\boldsymbol{\theta}_t$, is estimated by fitting the time series $\boldsymbol{v}_t$ to the dynamics in (25). Subsequently, we use the pair $\{\boldsymbol{\theta}_t, \boldsymbol{v}_t\}$ to train the LSTM model for (26); for training, we add Gaussian noises of variance 1% relative to that of the original data to avoid overfitting that tends to occur when the hypothesis space is rather complex and the amount of data is finite.

Fig. 6(a) displays the difference of the short-time spatiotemporal manifestation between the full and the closure models. One can see that the spatio-temporal pattern of the proposed closure model is consistent with that of the full KS model up to roughly time $t = 54$. A close inspection shows an accurate path-wise prediction of the real component of the leading six Fourier modes up to time 54 (see Fig. 8). In Fig. 7, we report the root-mean-square-error (RMSE) and (anomaly correlation) ANCR as defined in [16] that characterize the lead-time prediction skill, averaged over 1000 initial conditions out-of-samples and the spatial domain. Notice that both metrics show a substantial improvement in the prediction skill relative to that of the bare truncated model which is a result of using only (25) with $\hat{\theta}_k = 0$.

For this regime $L \approx 21.55$, the leading Lyapunov exponent is roughly $\lambda_1 \approx 0.04$ [19], which suggests that the accurate prediction length is roughly $54 \times \lambda_1 = 2.16$ Lyapunov time units. In other words, the length of the prediction is on the same order as the Lyapunov time. While this empirical result suggests that the constant a in Theorem 3 is roughly e^{λ_1}, a theoretical justification for such a tighter bound will require more thorough investigation with possibly additional assumptions on the dynamics.

(a) Difference of the full and closure models solutions



(b) Energy spectrum

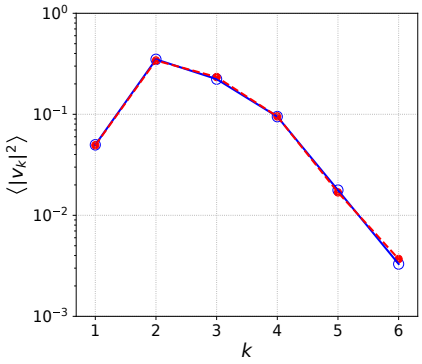
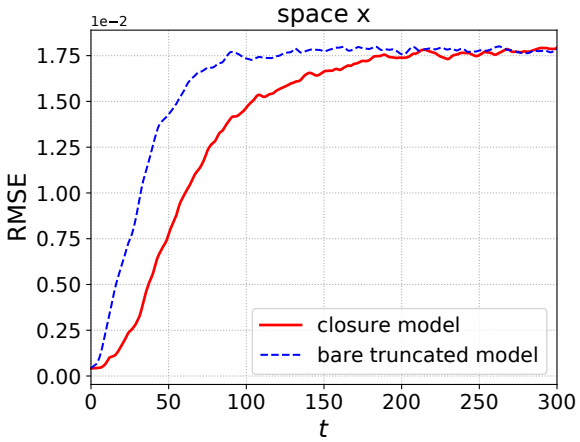


Figure 6: (a) Difference of spatiotemporal manifestation of KS solutions starting from the same initial conditions between the full model and the closure model using the LSTM method; (b) The energy spectra $\langle |v_k| \rangle^2$ for the KS solutions between the full (blue solid) and the closure (red dashes) LSTM models.

(a) RMSE



(b) ANCR

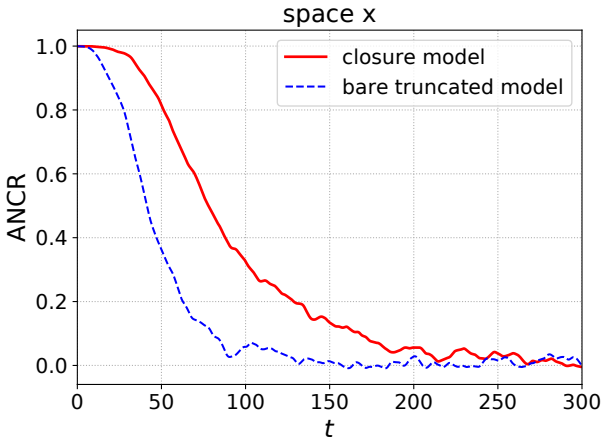


Figure 7: Prediction error: RMSE and ANCR as a function of time. These metric are estimated by a spatial and temporal average over 1000 initial conditions out-of-sampling.

In Fig. 6(b), we show the accurate recovery of the energy spectra. Fig. 8 also displays the results for the comparison of ACFs and PDFs for all the Fourier modes $v_1, \ldots, v_6$ and CCF's defined as the cross-correlation functions between $|v_k|^2$ and $|v_4|^2$. All of these long-time statistics are computed using the Monte-Carlo estimation over 2.5×10^5 data samples. We can see that ACFs, CCFs, and PDFs can be well reproduced by the LSTM for all modes. Therefore, the closure model using the LSTM can provide an accurate recovery for both the short-time forecasting and the long-time statistics of the KSE.

To summarize, we should also mention that while such an accurate recovery in path-wise and statistical prediction has also been achieved with the NARMAX parametric closure in [40, 38], careful choice of parametric ansatz is necessary with the NARMAX model. Here, an accurate recovery is obtained with a much simpler nonparametric model in (26).

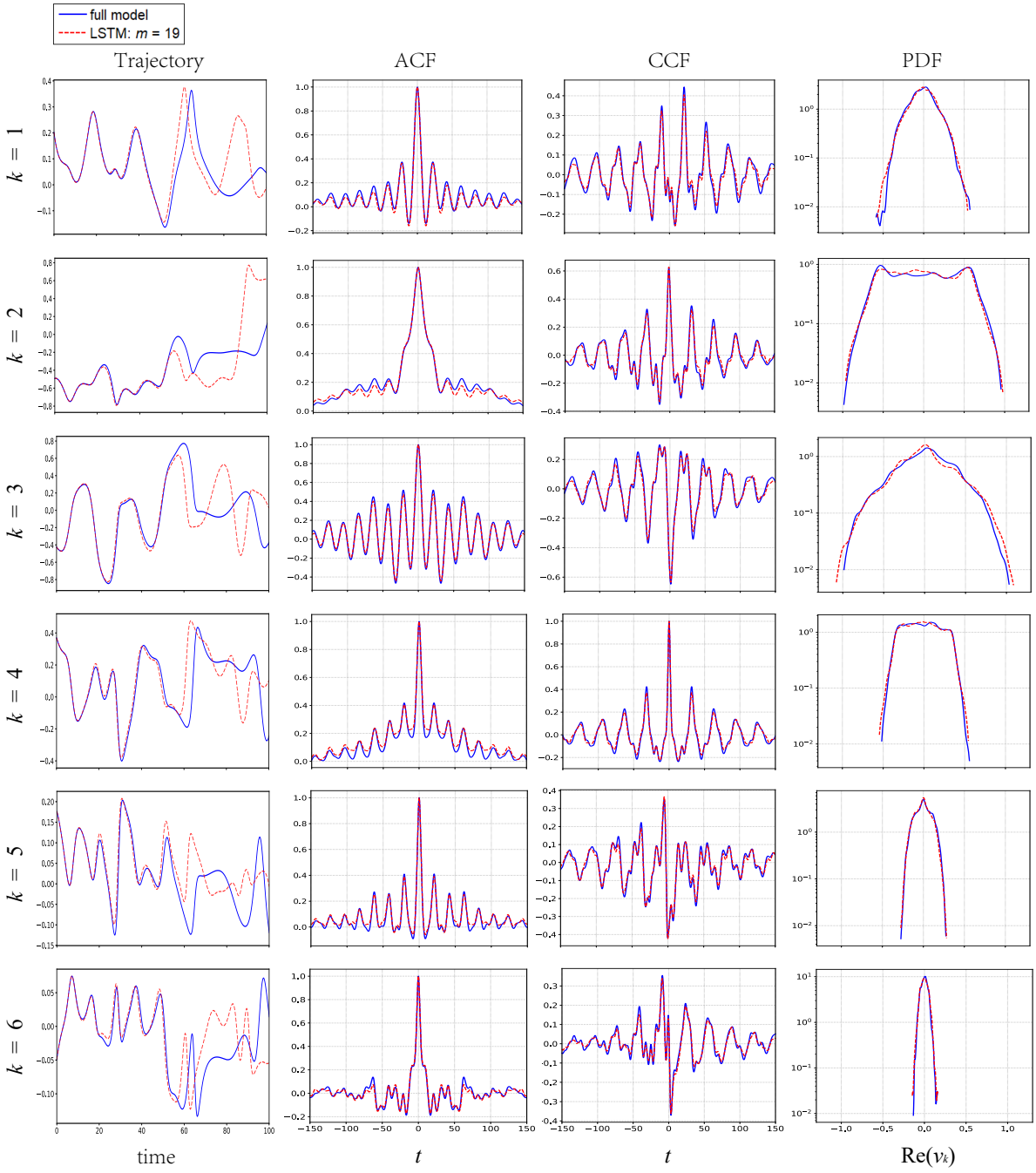


Figure 8: (Color online) Comparison of trajectories for $\mathrm{Re}(v_k)$, ACFs for $\mathrm{Re}(v_k)$, CCFs between $|v_k|^2$ and $|v_4|^2$, and PDFs for $\mathrm{Re}(v_k)$ between the full model and the closure model. Solid blue line corresponds to the full model and dashed red line corresponds to the closure model.

5 Summary

We have presented a general nonparametric framework for prediction with missing dynamics. The proposed framework reformulates the closure model as a supervised learning problem in which the task is to approximate a high-dimensional map that takes the history of resolved and identifiable unresolved variables to the missing components in the resolved dynamics. Mathematically, we validate the approach with an error bound which implies that the closure framework converges when a consistent learning algorithm is used. Numerically, we demonstrate the effectiveness of our framework in replicating severely truncated complex nonlinear problems arising in many applications. While the framework can be realized using any machine learning technique, we found that the LSTM as a special class of RNN is robust for this particular task.

From the positive numerical tests, several open questions deserve further investigation. For example, justifying the existence of the equilibrium distribution of the closure model; demonstrating the convergence to the underlying equilibrium distribution; characterizing the prediction error using Lyapunov exponents for chaotic dynamics; clarifying the condition under which we can achieve a stable closure model.

Acknowledgments. The research of J.H. was partially supported by the ONR Grant N00014-16-1-2888 and NSF Grant DMS-1854299. J.H. thanks Di Qi for sharing the codes for the 57-mode barotropic stress model. S. L. and H. Y. gratefully acknowledge the support of National Supercomputing Center Singapore (NSCC) and High-Performance Computing (HPC) of the National University of Singapore for providing computational resources, and the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. H. Y. was partially supported by the US National Science Foundation under award DMS-1945029.

A Proof of Theorem 1

Before we prove Theorem 1 in the main text, let us review the following bound which will be used below as well as in the proof of Theorem 3.

Lemma 1. *Let $\alpha, c > 0$ be real numbers and $m, T \geq 0$ be integers. Suppose that,*

$$E_{T+1} \leq \alpha \sum_{j=T-m}^{T} E_j + c,$$

If $E_j = 0$ for $j = -m, \ldots, 0$, then for all integer $T \geq 0$.

$$E_{T+1} \leq c(1+\alpha)^T.$$

Proof. We proceed by induction. one can verify that, $E_1 \leq c$, $E_2 \leq c(1+\alpha)$ and so on. In fact, we can verify for $j = 0, \ldots, m$ one by one that

$$E_j \leq c(1+\alpha)^{j-1}. \tag{27}$$

By induction, for $T \geq m$, we have

$$E_{T+1} \leq c\alpha \sum_{j=T-m}^{T} (1+\alpha)^{j-1} + c \leq c\alpha \sum_{j=1}^{T} (1+\alpha)^{j-1} + c = c(1+\alpha)^T. \tag{28}$$

□

Now we proceed with the proof of Theorem 1. As we mentioned before, since

$$\mathbb{E}[y_{t+1}|x_t, y_t] = \mathcal{G}(x_t, y_t), \tag{29}$$

we can rewrite the full dynamics as,

$$\begin{aligned} x_{t+1} &= \mathcal{F}(x_t, y_t), \\ y_{t+1} &= \mathbb{E}[Y_{t+1}|x_t, y_t]. \end{aligned}$$

We consider an approximate dynamics given as,

$$\begin{aligned} \hat{x}_{t+1} &= \mathcal{F}(\hat{x}_t, \hat{y}_t) \\ \hat{y}_{t+1} &= \mathbb{E}^{\epsilon}[y_{t+1}|\hat{x}_t, \hat{y}_t] + \xi_{t+1}, \end{aligned}$$

where $\xi_{t+1} \sim \Xi$ are Gaussian white noises with variance,

$$\mathbb{E}[\Xi^2] := \mathbb{E}\left[(Y_{t+1} - \mathbb{E}^{\epsilon}[Y_{t+1}|X_t, Y_t])^2\right] = \mathbb{E}\left[(\mathbb{E}[Y_{t+1}|X_t, Y_t] - \mathbb{E}^{\epsilon}[Y_{t+1}|X_t, X_t])^2\right] = O(\epsilon^2). \tag{30}$$

Define $E_{x,t} := |x_{t+1} - \hat{x}_{t+1}|$ and $E_{y,t} := |y_{t+1} - \hat{y}_{t+1}|$, using the consistency in (29) and the Lipschitz conditions of $\mathcal{F}$ and $\mathcal{G}$, we deduce

$$\begin{aligned} E_{y,t+1} &\leq \left|\mathbb{E}[Y_{t+1}|x_t, y_t] - \mathbb{E}^{\epsilon}[Y_{t+1}|\hat{x}_t, \hat{y}_t]\right| + |\xi_{t+1}| \\ &\leq |\mathbb{E}[Y_{t+1}|x_t, y_t] - \mathbb{E}[Y_{t+1}|\hat{x}_t, \hat{y}_t]| + \left|\mathbb{E}[Y_{t+1}|\hat{x}_t, \hat{y}_t] - \mathbb{E}^{\epsilon}[Y_{t+1}|\hat{x}_t, \hat{y}_t]\right| + |\xi_{t+1}| \\ &< |\mathcal{G}(x_t, y_t) - \mathcal{G}(\hat{x}_t, \hat{y}_t)| + \left|\mathbb{E}[Y_{t+1}|\hat{x}_t, \hat{y}_t] - \mathbb{E}^{\epsilon}[Y_{t+1}|\hat{x}_t, \hat{y}_t]\right| + |\xi_{t+1}| \\ &\leq L_1 E_{y,t} + L_2 E_{x,t} + \left|\mathbb{E}[Y_{t+1}|\hat{x}_t, \hat{y}_t] - \mathbb{E}^{\epsilon}[Y_{t+1}|\hat{x}_t, \hat{y}_t]\right| + |\xi_{t+1}| \end{aligned}$$

Define $E^*_{x,T+1} := \mathbb{E}[\max_{t=\{0,\ldots,T+1\}} E_{x,t}]$ and $E^*_{y,T+1} := \mathbb{E}[\max_{t=\{0,\ldots,T+1\}} E_{y,t}]$. Then, by the Burkholder-Davis-Gundy inequality [60],

$$E^*_{y,T+1} \leq L_1 E^*_{y,T} + L_2 E^*_{x,T} + C\epsilon, \tag{31}$$

in which we have used (30) to bound the last two terms. This bound can be explicitly written as,

$$\begin{aligned} E^*_{y,T+1} &\leq L_1^{T+1} E^*_{y,0} + \sum_{j=0}^{T} L_1^j (L_2 E^*_{x,T-j} + C\epsilon) \\ &\leq L_1^{T+1} E^*_{y,0} + (L_2 E^*_{x,T} + C\epsilon) \sum_{j=0}^{T} L_1^j \\ &= (L_2 E^*_{x,T} + C\epsilon) \frac{L_1^{T+1} - 1}{L_1 - 1}. \end{aligned} \tag{32}$$

where we have used the fact that $L_2 E^*_{x,t} + C\epsilon$ is non-decreasing to get the second inequality and $E^*_{y,0} = 0$ to obtain the last equality.

Using similar algebra, we have

$$E^*_{x,T+1} \leq L_3 E^*_{x,T} + L_4 E^*_{y,T} \tag{33}$$

Inserting (32) into (33), we obtain

$$\begin{aligned} E^*_{x,T+1} &\leq L_3 E^*_{x,T} + L_4 (L_2 E^*_{x,T-1} + C\epsilon) \frac{L_1^T - 1}{L_1 - 1} \\ &\leq \alpha (E^*_{x,T} + E^*_{x,T-1}) + C L_4 \epsilon \frac{L_1^T - 1}{L_1 - 1} \end{aligned} \tag{34}$$

where we have define $\alpha = \max\{L_3, L_2 L_4 \frac{L_1^T - 1}{L_1 - 1}\}$.

Given $E^*_{x,0} = 0$, we apply the bound in Lemma 1 for $m = 1$,

$$E^*_{x,T+1} \quad \leq \quad CL_4\epsilon \frac{L_1^T - 1}{L_1 - 1}(1+\alpha)^T = O(a^T\epsilon),$$

for some constant $a > 1$ and the proof is complete.

B Proof of Theorem 2

Let $Y^\Delta_{t+1} := \frac{Y_{t+1} - Y_t}{\Delta}$ such that,

$$\mathbb{E}\Big[Y^\Delta_{t+1} | x_t, y_t\Big] = g(x_t, y_t). \tag{35}$$

With this definition, we can rewrite the full dynamics as,

$$\begin{aligned} x_{t+1} &= x_t + f(x_t, y_t)\Delta + \Delta^{1/2}\sigma_x \xi_{x,t+1}, \\ y_{t+1} &= y_t + \mathbb{E}[Y^\Delta_{t+1} | x_t, y_t]\Delta + \Delta^{1/2}\sigma_y \xi_{y,t+1}. \end{aligned}$$

We consider an approximate dynamics given as,

$$\begin{aligned} \hat{x}_{t+1} &= \hat{x}_t + f(\hat{x}_t, \hat{y}_t)\Delta + \Delta^{1/2}\sigma_x \xi_{x,t+1}, \\ \hat{y}_{t+1} &= \hat{y}_t + \mathbb{E}^\epsilon[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t]\Delta + \Delta^{1/2}\hat{\sigma}_y \xi_{y,t+1}. \end{aligned}$$

First, notice that

$$\begin{aligned} \Delta\hat{\sigma}_y^2 &= \mathbb{E}\left[\Big(Y_{t+1} - Y_t - \Delta\mathbb{E}^\epsilon[Y^\Delta_{t+1} | X_t, Y_t]\Big)^2\right] \\ &\leq \mathbb{E}\left[\Big(Y_{t+1} - Y_t - \Delta\mathbb{E}[Y^\Delta_{t+1} | X_t, Y_t]\Big)^2\right] + \Delta^2\mathbb{E}\Big(\mathbb{E}[Y^\Delta_{t+1} | X_t, Y_t] - \mathbb{E}^\epsilon[Y^\Delta_{t+1} | X_t, Y_t]\Big)^2 \dots \\ &\quad + 2\Delta\mathbb{E}\Big[\Big(Y_{t+1} - Y_t - \Delta\mathbb{E}[Y^\Delta_{t+1} | X_t, Y_t]\Big)\Big(\mathbb{E}[Y^\Delta_{t+1} | X_t, Y_t] - \mathbb{E}^\epsilon[Y^\Delta_{t+1} | X_t, Y_t]\Big)\Big] \\ &= \Delta\sigma_y^2 + O(\Delta^2\epsilon^2), \end{aligned} \tag{36}$$

where the last term vanishes since the mean of $y_{t+1} - y_t - \Delta\mathbb{E}[Y^\Delta_{t+1} | x_t, y_t] = \Delta^{1/2}\sigma_y \xi_{x,t+1}$ is zero.

Define $E_{x,t+1} := |x_{t+1} - \hat{x}_{t+1}|$ and $E_{y,t+1} := |y_{t+1} - \hat{y}_{t+1}|$, using the consistency in (35) and the Lipschitz conditions of f and g, we deduce

$$\begin{aligned} E_{y,t+1} &\leq E_{y,t} + \Delta\left|\mathbb{E}[Y^\Delta_{t+1} | x_t, y_t] - \mathbb{E}^\epsilon[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t]\right| + \Delta^{1/2}\left|\sigma_y - \hat{\sigma}_y\right| |\xi_{y,t+1}| \\ &\leq E_{y,t} + \Delta\left|\mathbb{E}[Y^\Delta_{t+1} | x_t, y_t] - \mathbb{E}[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t]\right| + \Delta\left|\mathbb{E}[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t] - \mathbb{E}^\epsilon[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t]\right| + \Delta^{1/2}\left|\sigma_y - \hat{\sigma}_y\right| |\xi_{y,t+1}| \\ &< E_{y,t} + \Delta\,|g(x_t, y_t) - g(\hat{x}_t, \hat{y}_t)| + \Delta\left|\mathbb{E}[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t] - \mathbb{E}^\epsilon[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t]\right| + \Delta^{1/2}\left|\sigma_y - \hat{\sigma}_y\right| |\xi_{y,t+1}| \\ &\leq (1 + \Delta\ell)E_{y,t} + \Delta\ell E_{x,t} + \Delta\left|\mathbb{E}[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t] - \mathbb{E}^\epsilon[Y^\Delta_{t+1} | \hat{x}_t, \hat{y}_t]\right| + \Delta^{1/2}\left|\sigma_y - \hat{\sigma}_y\right| |\xi_{y,t+1}|. \end{aligned}$$

where $\ell = O(1)$ denotes the largest Lipschitz constant in all directions. Define $E^*_{x,T+1} := \mathbb{E}[\max_{t=\{0,\dots,T+1\}} E_{x,t}]$ and $E^*_{y,T+1} := \mathbb{E}[\max_{t=\{0,\dots,T+1\}} E_{y,t}]$. Then, by the Burkholder-Davis-Gundy inequality [60], we have

$$E^*_{y,T+1} \leq (1 + \Delta\ell)E^*_{y,T} + \Delta\ell E^*_{x,T} + C\Delta\epsilon, \tag{37}$$

where we have used (36). Concatenate this with

$$E^*_{x,T+1} \leq (1 + \Delta\ell)E^*_{x,T} + \Delta\ell E^*_{y,T},$$

we have

$$E^*_{T+1} \leq (I + A + A^2 + \ldots + A^T)b,$$

where

$$E^*_{T+1} = \begin{pmatrix} E^*_{x,T+1} \\ E^*_{y,T+1} \end{pmatrix}, \qquad A = \begin{pmatrix} 1+\Delta\ell & \Delta\ell \\ \Delta\ell & 1+\Delta\ell \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ C\Delta\epsilon \end{pmatrix}.$$

Using the fact that,

$$A = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1+2\ell\Delta \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

one can deduce that,

$$\begin{aligned} E^*_{x,T+1} &\leq C\Delta\epsilon\Big(-(T+1) + \frac{1-(1+2\ell\Delta)^{T+1}}{-2\ell\Delta}\Big) \\ &\leq C\Delta\epsilon\left(-(T+1) + \left[\frac{-1+\Big(1+2\ell\Delta\,(T+1) + \frac{T(T+1)}{2}\,(2\ell\Delta)^2 + O(\Delta^3T^3)\Big)}{2\ell\Delta}\right]\right) \\ &= C\Delta\epsilon\frac{T\,(T+1)}{2}2\ell\Delta + O(\epsilon\Delta^3T^3) \\ &= O(\epsilon\Delta^2T^2). \end{aligned}$$

where we use Taylor expansion over small $2\ell\Delta$ and the proof is complete.

C Proof of Theorem 3

In this case, we have

$$\mathbb{E}[\Theta_{t+1}|\boldsymbol{z}_{t,m}] = \bar{\mathcal{G}}_0(x_t,\theta_t) + \sum_{k=1}^{m}\bar{\mathcal{G}}_k(x_{t-k},\theta_{t-k}) + (QS)^{m+1}\pi(x_{t-m},y_{t-m}), \tag{38}$$

where $\boldsymbol{z}_{t,m} = (\boldsymbol{x}_{t-m:t}, \boldsymbol{\theta}_{t-m:t})$.

Define $E_{\theta,t+1} := |\theta_{t+1} - \hat{\theta}_{t+1}|$ and $E_{x,t+1} := |x_{t+1} - \hat{x}_{t+1}|$. By the Assumption 2, $\mathcal{F}$ and $\mathcal{G}_k$ are Lipschitz continuous on x and θ, and $(QS)^{m+1}$ is a bounded linear operator in uniform sense. Thus, we have

$$\begin{aligned} E_{\theta,t+1} &\leq \Big|\mathbb{E}[\Theta_{t+1}|\boldsymbol{z}_{t,m}] - \mathbb{E}^\epsilon[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}]\Big| + |\xi_{t+1}| \\ &\leq \Big|\mathbb{E}[\Theta_{t+1}|\boldsymbol{z}_{t,m}] - \mathbb{E}[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}]\Big| + \Big|\mathbb{E}[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}] - \mathbb{E}^\epsilon[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}]\Big| + |\xi_{t+1}| \\ &\leq \sum_{k=0}^{m}\Big|\mathcal{G}_k(x_{t-k},\theta_{t-k}) - \mathcal{G}_k(\hat{x}_{t-k},\hat{\theta}_{t-k})\Big| + \Big|(QS)^{m+1}\pi(x_{t-m},y_{t-m}) - (QS)^{m+1}\pi(\hat{x}_{t-m},\hat{y}_{t-m})\Big| \\ &\quad + \Big|\mathbb{E}[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}] - \mathbb{E}^\epsilon[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}]\Big| + |\xi_{t+1}| \\ &\leq \sum_{s=t-m}^{t}K_{s-(t-m)}E_{\theta,s} + \sum_{s=t-m}^{t}L_{s-(t-m)}E_{x,s} + \Big|\mathbb{E}[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}] - \mathbb{E}^\epsilon[\Theta_{t+1}|\hat{\boldsymbol{z}}_{t,m}]\Big| + |\xi_{t+1}|, \end{aligned} \tag{39}$$

where K_s, L_s are Lipschitz constants.

Define $E^*_{\theta,T+1} := \mathbb{E}[\max_{t=\{0,\ldots,T+1\}} E_{\theta,t}]$ and $E^*_{x,T+1} := \mathbb{E}[\max_{t=\{0,\ldots,T+1\}} E_{x,t}]$. Then we have,

$$E^*_{\theta,T+1} \leq \sum_{s=T-m}^{T} K_{s-(T-m)} E^*_{\theta,s} + \sum_{s=T-m}^{T} L_{s-(T-m)} E^*_{x,s} + C\epsilon, \tag{40}$$

where the expectation of the last term in (39) is bounded using the Burkholder-Davis-Gundy inequality [60]. We should point out that since the expectation in

$$\mathbb{E}\Big[(\mathbb{E}[\theta_{t+1}|Z_{t,m}] - \mathbb{E}^{\epsilon}[\Theta_{t+1}|Z_{t,m}])^2\Big] = O(\epsilon^2),$$

is defined with respect to the pushforward measure $\nu := Z_{t,m*}\mu$, that is, $\nu(B) = \mu(Z^{-1}_{t,m}(B))$, for all $B \in \mathcal{B}(\mathcal{Z})$ in the σ-algebra, associated to the random variable $Z_{t,m} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Since $\nu(\mathcal{Z}) = \int_{\mathcal{Z}} d\nu(z) < \infty$, it is clear that expectation of the third term in (39),

$$\mathbb{E}\Big[\,\big|\mathbb{E}[\theta_{t+1}|Z_{t,m}] - \mathbb{E}^{\epsilon}[\Theta_{t+1}|Z_{t,m}]\big|\,\Big] \leq \mathbb{E}\Big[\,\big|\mathbb{E}[\theta_{t+1}|Z_{t,m}] - \mathbb{E}^{\epsilon}[\Theta_{t+1}|Z_{t,m}]\big|^2\,\Big]^{1/2} \nu(\mathcal{Z})^{1/2} = C\epsilon. \tag{41}$$

is also bounded by order-ϵ.

Let $0 < K := \max\{K_0, \ldots, K_m\}$, applying the bound in Lemma 1, we can obtain from (40)

$$E^*_{\theta,T+1} \quad \leq \quad \Big(\sum_{s=T-m}^{T} L_{s-(T-m)} E^*_{x,s} + C\epsilon\Big)(1+K)^T. \tag{42}$$

Using similar algebra, we have

$$E^*_{x,T+1} \leq L_{m+1} E^*_{x,T} + K_{m+1} E^*_{\theta,T}, \tag{43}$$

for some constants $K_{m+1}, L_{m+1} > 0$. Inserting (42) into (43), let $0 < L := \max_{j=0,\ldots,m}\{L_{m+1}, K_{m+1}L_j(1+K)^{T-1}\}$, applying the bound (28), we obtain

$$\begin{aligned}
E^*_{x,T+1} &\leq L_{m+1} E^*_{x,T} + K_{m+1}\Big(\sum_{s=T-m-1}^{T-1} L_{s-(T-m-1)} E^*_{x,s} + C\epsilon\Big)(1+K)^{T-1} && (44)\\
&\leq L \sum_{s=T-m-1}^{T} E^*_{x,s} + K_{m+1} C\epsilon (1+K)^{T-1} && (45)\\
&\leq K_{m+1} C\epsilon (1+K)^{T-1}(1+L)^T && (46)\\
&= O(a^T\epsilon),
\end{aligned}$$

for some $a > 1$ and the proof is completed.

References

[1] Z. Allen-Zhu and Y. Li. Can SGD learn recurrent neural networks with provable generalization? *CoRR*, abs/1902.01028, 2019.

[2] Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *CoRR*, abs/1811.04918, 2018.

[3] Z. Allen-Zhu, Y. Li, and Z. Song. On the convergence rate of training recurrent neural networks. *CoRR*, abs/1810.12065, 2018.

[4] W. Bao, S. Jin, and P. A. Markowich. Numerical study of time-splitting spectral discretizations of nonlinear schrödinger equations in the semiclassical regimes. *SIAM Journal on Scientific Computing*, 25(1):27–64, 2003.

[5] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

[6] T. Berry and J. Harlim. Linear Theory for Filtering Nonlinear Multiscale Systems with Model Error. *Proc. Roy. Soc. A 20140168*, 2014.

[7] T. Berry and J. Harlim. Semiparametric modeling: Correcting low-dimensional model error in parametric models. *J. Comput. Phys.*, 308:305–321, 2016.

[8] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. *CoRR*, abs/1710.10174, 2017.

[9] Y. Cao and Q. Gu. A generalization theory of gradient descent for learning over-parameterized deep relu networks. *CoRR*, abs/1902.01384, 2019.

[10] G. F. Carnevale and J. S. Frederiksen. Nonlinear stability and statistical mechanics of flow over topography. *Journal of Fluid Mechanics*, 175:157–181, 1987.

[11] M. Chen, X. Li, and T. Zhao. On generalization bounds of a family of recurrent neural networks, 2019.

[12] N. Chen, A. J. Majda, and X. T. Tong. Spatial localization for nonlinear dynamical stochastic models for excitable media. *arXiv preprint arXiv:1901.07318*, 2019.

[13] A. Chorin, O. Hald, and R. Kupferman. Optimal prediction with memory. *Physica D: Nonlinear Phenomena*, 166(3):239–257, 2002.

[14] A. Chorin and P. Stinis. Problem reduction, renormalization, and memory. *Communications in Applied Mathematics and Computational Science*, 1(1):1–27, 2007.

[15] C. K. Chui and G. Chen. *Linear systems and optimal control*, volume 18. Springer Science & Business Media, 2012.

[16] D. Crommelin and E. Vanden-Eijnden. Subgrid-scale parameterization with conditional markov chains. *Journal of the Atmospheric Sciences*, 65(8):2661–2675, 2008.

[17] E. Darve, J. Solomon, and A. Kia. Computing generalized langevin equations and generalized fokker–planck equations. *Proceedings of the National Academy of Sciences*, 106(27):10884–10889, 2009.

[18] W. E, C. Ma, and L. Wu. A priori estimates of the generalization error for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

[19] R. A. EDSON, J. E. BUNDER, T. W. MATTNER, and A. J. ROBERTS. Lyapunov exponents of the kuramoto?sivashinsky pde. *The ANZIAM Journal*, 61(3):270?285, 2019.

[20] C. Franzke, I. Horenko, A. J. Majda, and R. Klein. Systematic metastable atmospheric regime identification in an agcm. *Journal of the Atmospheric Sciences*, 66(7):1997–2012, 2009.

[21] T. Gao, J. Duan, X. Li, and R. Song. Mean exit time and escape probability for dynamical systems driven by lvy noises. *SIAM Journal on Scientific Computing*, 36(3):A887–A906, 2014.

[22] D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity*, 17(6):R55, 2004.

[23] J. Goodman. Stability of the kuramoto-sivashinsky and related systems. *Communications on Pure and Applied Mathematics*, 47(3):293–306, 1994.

[24] A. Gouasmi, E. J. Parish, and K. Duraisamy. A priori estimation of memory effects in reduced-order models of nonlinear systems using the mori–zwanzig formalism. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170385, 2017.

[25] M. J. Grote, A. J. Majda, and C. G. Ragazzo. Dynamic mean flow and small-scale interaction through topographic stress. *Journal of Nonlinear Science*, 9(1):89–130, 1999.

[26] J. Harlim and X. Li. Parametric reduced models for the nonlinear Schrödinger equation. *Phys. Rev. E.*, 91:053306, 2015.

[27] J. Harlim, A. Mahdi, and A. Majda. An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. *J. Comput. Phys.*, 257, Part A:782–812, 2014.

[28] M. W. Hirsch, S. Smale, and R. L. Devaney. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2012.

[29] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997.

[30] S. Jiang and J. Harlim. Modeling of missing dynamical systems: Deriving parametric models using a nonparametric framework. *arXiv:1905.08082*, 2019.

[31] P. Kevrekidis and D. Frantzeskakis. Solitons in coupled nonlinear schrödinger models: a survey of recent developments. *Reviews in Physics*, 1:140–153, 2016.

[32] B. Khouider, J. A. Biello, and A. J. Majda. A stochastic multicloud model for tropical convection. *Comm. Math. Sci.*, 8:187–216, 2010.

[33] D. Kondrashov, M. D. Chekroun, and M. Ghil. Data-driven non-markovian closure models. *Physica D: Nonlinear Phenomena*, 297:33–55, 2015.

[34] H. Kunita. *Stochastic flows and stochastic differential equations*, volume 24. Cambridge university press, 1997.

[35] Y. Kuramoto and T. Tsuzuki. Persistent Propagation of Concentration Waves in Dissipative Media Far from Thermal Equilibrium. *Progress of Theoretical Physics*, 55(2):356–369, 02 1976.

[36] F. Kwasniok. Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1962):1061–1086, 2012.

[37] R. E. LaQuey, S. Mahajan, P. Rutherford, and W. Tang. Nonlinear saturation of the trapped-ion mode. *Physical Review Letters*, 34(7):391, 1975.

[38] K. K. Lin and F. Lu. Data-driven model reduction, wiener projections, and the mori-zwanzig formalism. *arXiv preprint arXiv:1908.07725*, 2019.

[39] F. Lu, K. Lin, and A. Chorin. Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems. *Communications in Applied Mathematics and Computational Science*, 11(2):187–216, 2016.

[40] F. Lu, K. K. Lin, and A. J. Chorin. Data-based stochastic model reduction for the kuramoto–sivashinsky equation. *Physica D: Nonlinear Phenomena*, 340:46–57, 2017.

[41] F. Lu, X. Tu, and A. J. Chorin. Accounting for model error from unresolved scales in ensemble kalman filters by stochastic parameterization. *Monthly Weather Review*, 145(9):3709–3723, 2017.

[42] J. Lu, Z. Shen, H. Yang, and S. Zhang. Deep Network Approximation for Smooth Functions. *arXiv e-prints*, page arXiv:2001.03040, Jan. 2020.

[43] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. *CoRR*, abs/1709.02540, 2017.

[44] C. Ma and J. Wang. Model reduction with memory and the machine learning of dynamical systems. *arXiv preprint arXiv:1808.04258*, 2018.

[45] A. Majda and J. Harlim. Physics constrained nonlinear regression models for time series. *Nonlinearity*, 26:201–217, 2013.

[46] A. Majda, I. Timofeyev, and E. Vanden-Eijnden. Systematic strategies for stochastic mode reduction in climate. *Journal of the Atmospheric Sciences*, 60:1705–1722, 2003.

[47] A. Majda and X. Wang. *Nonlinear dynamics and statistical theories for basic geophysical flows*. Cambridge University Press, UK, 2006.

[48] A. J. Majda, I. Timofeyev, and E. V. Eijnden. Models for stochastic climate prediction. *Proceedings of the National Academy of Sciences*, 96(26):14687–14691, 1999.

[49] A. J. Majda, I. Timofeyev, and E. Vanden Eijnden. A mathematical framework for stochastic climate models. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 54(8):891–974, 2001.

[50] R. Maulik, A. Mohan, B. Lusch, S. Madireddy, and P. Balaprakash. Time-series learning of latent-space dynamics for reduced-order model closure. *arXiv preprint arXiv:1906.07815*, 2019.

[51] H. Montanelli and Q. Du. New error bounds for deep networks using sparse grids. 2017.

[52] H. Montanelli and H. Yang. Error bounds for deep relu networks using the kolmogorovarnold superposition theorem. *Neural Networks*, 129:1 – 6, 2020.

[53] H. Montanelli, H. Yang, and Q. Du. Deep relu networks overcome the curse of dimensionality for bandlimited functions. *arXiv:1903.00735 [math.NA]*, 2019.

[54] H. Mori. Transport, collective motion, and Brownian motion. *Prog. Theor. Phys.*, 33:423 – 450, 1965.

[55] M. C. Mozer. Backpropagation. chapter A Focused Backpropagation Algorithm for Temporal Pattern Recognition, pages 137–169. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1995.

[56] R. Nakada and M. Imaizumi. Adaptive approximation and estimation of deep neural network to intrinsic dimensionality. *arXiv:1907.02177 [stat.ML]*, 2019.

[57] S. Pan and K. Duraisamy. Data-driven discovery of closure models. *SIAM Journal on Applied Dynamical Systems*, 17(4):2381–2413, 2018.

[58] E. J. Parish and K. Duraisamy. A dynamic subgrid scale model for large eddy simulations based on the mori–zwanzig formalism. *Journal of Computational Physics*, 349:154–175, 2017.

[59] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1310–III–1318. JMLR.org, 2013.

[60] G. Pavliotis and A. Stuart. *Multiscale Methods: Averaging and Homogenization*, volume 53 of *Texts in Applied Mathematics*. Springer, 2000.

[61] D. Qi and A. J. Majda. Low-dimensional reduced-order models for statistical response and uncertainty quantification: Barotropic turbulence with topography. *Physica D: Nonlinear Phenomena*, 343:7–27, 2017.

[62] A. J. Robinson and F. Fallside. The utility driven dynamic error propagation network. Technical Report CUED/F-INFENG/TR.1, Engineering Department, Cambridge University, Cambridge, UK, 1987.

[63] Z. Shen, H. Yang, and S. Zhang. Deep network approximation characterized by number of neurons. *arXiv:1906.05497 [math.NA]*, 2019.

[64] Z. Shen, H. Yang, and S. Zhang. Deep network approximation with discrepancy being reciprocal of width to power of depth. *ArXiv*, abs/2006.12231, 2020.

[65] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Process. Mag.*, 30(4):98–111, 2013.

[66] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.

[67] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

[68] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[69] E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annual review of physical chemistry*, 61:391–420, 2010.

[70] P. R. Vlachas, W. Byeon, Z. Y. Wan, T. P. Sapsis, and P. Koumoutsakos. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213):20170844, 2018.

[71] E. Weinan, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. Heterogeneous multiscale methods: a review. *Commun. Comput. Phys*, 2(3):367–450, 2007.

[72] P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4):339 – 356, 1988.

[73] D. S. Wilks. Effects of stochastic parametrizations in the lorenz'96 system. *Quarterly Journal of the Royal Meteorological Society*, 131(606):389–407, 2005.

[74] H. Zhang, J. Harlim, and X. Li. Computing linear response statistics using orthogonal polynomial based estimators: An RKHS formulation. *arXiv:1912.11110*, 2019.

[75] R. Zwanzig. Statistical mechanics of irreversiblity. *Lectures in Theoretical Physics*, 3:106–141, 1961.

[76] R. Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, 2001.