Modeling of Missing Dynamical Systems: Deriving Parametric Models using a Nonparametric Framework

Shixiao W. Jiang [a] and John Harlim * [a,b,c]

[a]Department of Mathematics, the Pennsylvania State University, 109 McAllister Building, University Park, PA 16802-6400, USA
[b]Department of Meteorology and Atmospheric Science, the Pennsylvania State University, 503 Walker Building, University Park, PA 16802-5013, USA
[c]Institute for Computational and Data Sciences, the Pennsylvania State University, 224B Computer Building, University Park, PA 16802, USA

June 22, 2020

Abstract

In this paper, we consider modeling missing dynamics with a nonparametric non-Markovian model, constructed using the theory of kernel embedding of conditional distributions on appropriate Reproducing Kernel Hilbert Spaces (RKHS), equipped with orthonormal basis functions. Depending on the choice of the basis functions, the resulting closure model from this nonparametric modeling formulation is in the form of parametric model. This suggests that the success of various parametric modeling approaches that were proposed in various domains of applications can be understood through the RKHS representations. When the missing dynamical terms evolve faster than the relevant observable of interest, the proposed approach is consistent with the effective dynamics derived from the classical averaging theory. In the linear Gaussian case without the time-scale gap, we will show that the proposed non-Markovian model with a very long memory yields an accurate estimation of the nontrivial autocovariance function for the relevant variable of the full dynamics. Supporting numerical results on instructive nonlinear dynamics show that the proposed approach is able to replicate high-dimensional missing dynamical terms on problems with and without the separation of temporal scales.

Keywords. Missing dynamical systems, closure model, nonparametric non-Markovian model, kernel embedding

Mathematics Subject Classification. 37M10,37M25,41A10,41A45,60J22

1 Introduction

One of the long-standing issues in modeling dynamical systems is model errors arising from incomplete understanding of the physics. The progress in tackling this problem goes under different names depending on the scientific fields. In applied mathematics and engineering sciences, some of these approaches are known as the *reduced-order modeling*, which ultimate goal is to derive an effective model with low computational complexity from the first principle, assuming that the full dynamics is known. They include the Mori-Zwanzig formalism [44, 53, 54] and its approximations [4, 5, 12, 17, 25]; the averaging/homogenization when there is an apparent scale separation between the relevant and irrelevant variables [10, 42, 43, 50]. In domain sciences, various methods for *subgrid-scale parameterization* were proposed to handle the same problem that arises in applications such as material science, molecular dynamics, climate dynamics, just to name a few.

*Corresponding author: jharlim@psu.edu

They include the Markov chain type modeling [7, 23]; stochastic parameterization [1, 7, 18, 29, 32, 34, 37, 51]; superparameterization in cloud modeling [13, 24, 36] and in combustion problems [20, 21]; Direct-Interaction Approximation (DIA) for parameterizing sub-grid scale processes in isotropic turbulence [26] and its extensions [9], for modeling non-Markovian memory in inhomogeneous turbulence over topography. We should point out that this list is incomplete and these approaches share some commonality despite being developed independently and having different implementation details. Namely, the key unifying theme in these aforementioned methods is the parametric modeling assumption with specific choices of class of functions/distributions and typically having finite number of parameters.

In this paper, we consider a nonparametric modeling framework to compensate for the missing dynamical components. One of the main goals in this paper is to show that parametric modeling approaches can be understood and systematically derived from a nonparametric framework as opposed to the empirical choices of parametric models. In our setup, suppose that the underlying full dynamics is an ergodic system of Itô diffusion with relevant components $x \in \mathcal{X}$ and irrelevant components $y \in \mathcal{Y}$. The objective is to predict the evolution of $x \in \mathcal{X}$ and its statistics, given only the $x-$component of the full dynamics,

$$\begin{aligned} dx &= a(x,y)\,dt + b(x,y)\,dW_t, \\ dy &= c(x,y)\,dt + d(x,y)\,dV_t, \end{aligned} \tag{1}$$

and a historical data set $\{x_i := x(t_i), y_i = y(t_i)\}_{i=1,\dots,N}$. In (1), a and b denote, respectively, the $x-$component of the drift and diffusion terms that are known, while c and d denote, respectively, the $y-$component of the drift and diffusion terms that are not known. Also, W_t and V_t denote the standard (uncorrelated) Wiener processes.

While the core of the problem is similar to that in the reduced-order modeling framework, the fact that we have no knowledge of the full dynamics prohibits us to derive an effective equation from the first principle as in the standard averaging theory or Mori-Zwanzig formalism. Motivated by the practical applications where the underlying dynamics are not fully understood, instead, we will use the available historical data to reconstruct the missing dynamical components. We should point out that the restriction of knowing historical measurement of the irrelevant component, $y_i \in \mathcal{Y}$, can be relaxed in some cases. When $\{x_i\}_{i=1\dots,N}$ is the only available measurement, one can use, for example, likelihood maximum estimate [27, 33] in the deterministic case or an adaptive Bayesian filtering [2] (when b is constant and the training data is noisy) to extract the "identifiable" components of y_i. By identifiable components, we refer to variables that depend on y that appear in a and b, as we shall see in our numerical examples. Abusing the notation, we will denote y as the identifiable components. We will clarify this notion in our numerical examples.

Given the pair of historical time series $\{x_i, y_i\}_{i=1,\dots,N}$ with time lag $\tau = t_{i+1} - t_i$, let us define $\boldsymbol{z}_t := (\boldsymbol{x}_{t-m:t}, \boldsymbol{y}_{t-n:t-1}) \in \mathcal{Z}$ with $\boldsymbol{x}_{t-m:t} := (x_{t-m}, x_{t-m+1}, \dots, x_t)$ and $\boldsymbol{y}_{t-n:t-1} := (y_{t-n}, y_{t-m+1}, \dots, y_{t-1})$ for some integers $m, n \in \{-1, 0, \dots\}$. When $m = -1$, $\boldsymbol{z}_t$ has only $\boldsymbol{y}$ components (similarly for $n = 0$, $\boldsymbol{z}_t$ has only $\boldsymbol{x}$ components). We should point out that we have reserved the index i for the training data and used a different index t for an arbitrary prediction time with the same lag τ. Given these time series, our modeling approach is to approximate the conditional expectations,

$$\hat{a}(x, \boldsymbol{z}_t) := \mathbb{E}[a(x,Y)|\boldsymbol{z}_t] \qquad \hat{B}(x, \boldsymbol{z}_t) := \mathbb{E}[b(x,Y)b(x,Y)^\top|\boldsymbol{z}_t], \tag{2}$$

where the expectations are defined with respect to the equilibrium conditional density $p(y|\boldsymbol{z}_t)$ of the random variable $Y|\boldsymbol{z}_t$, a short hand for $Y|\boldsymbol{z} = \boldsymbol{z}_t$. Here, $Y|\boldsymbol{z}$ is nothing but the stationary random variable $Y_t|\boldsymbol{z}_t$. Throughout this paper, we will not use the notation $Y_t|\boldsymbol{z}_t$ to avoid a potential confusion with the non-stationary time-dependent distributions.

Given these conditional statistics, the closure model is given by

$$\hat{x}_{t+1} = \hat{x}_t + \int_{t\tau}^{(t+1)\tau} \hat{a}(\hat{x}(s), \hat{\boldsymbol{z}}_t)\,ds + \int_{t\tau}^{(t+1)\tau} \hat{B}(\hat{x}(s), \hat{\boldsymbol{z}}_t)^{1/2} dW_s, \tag{3}$$

where $\hat{\boldsymbol{z}}_t := (\hat{\boldsymbol{x}}_{t-m:t}, \hat{\boldsymbol{y}}_{t-n:t-1})$. To proceed the forecast at the next time step-$(t+1)$, one needs to update $\hat{\boldsymbol{z}}_{t+1}$. This variable is obtained by concatenating the components from previous time steps, $(\hat{\boldsymbol{x}}_{t+1-m:t+1}, \hat{\boldsymbol{y}}_{t+1-n:t-1})$ and $\hat{y}_t = \mathbb{E}[Y|\hat{\boldsymbol{z}}_t]$ that is estimated at time t.

Notice that if the $x-$component is slow and the missing $y-$component is fast with a scale gap denoted by a small parameter ϵ, the closure model in (3) is identical to the effective dynamics deduced by the averaging

theory [22, 28, 46] when the conditional expectation in (2) is defined with respect to the invariant density of the fast dynamics $\rho_\infty(y;\hat{x}_t)$ for a fixed $\hat{x}_t$, if such density exists. In this specific situation (fast-slow system), by setting $\hat{\boldsymbol{z}}_t = \hat{x}_t$, that is $m = 0, n = 0$, our framework effectively closes the dynamics by averaging over $p(y|\hat{x}_t) = p_{eq}(\hat{x}_t, y)/\int p_{eq}(\hat{x}_t, y)dy$, where p_{eq} denotes the invariant density of the full dynamics. We will show that averaging over $p(y|\hat{x}_t)$ is consistent with averaging over $\rho_\infty(y;\hat{x}_t)$ up to order ϵ. In general case where there is no separation of scales, the choice of m, n will be problem dependent. In this case, the predictive skill of certain statistics will depend on the specific choices of $\boldsymbol{z}_t$. For example, in the linear Gaussian case without a time-scale gap, we will show the existence of a conditional density $p(y|\boldsymbol{z})$ which allows for Eq. (3) to accurately estimate one-point and two-point statistics of the x-components of the full dynamics.

The main idea in this paper is to consider a nonparametric representation for $p(y|\boldsymbol{z})$ using the theory of kernel embedding of conditional distributions, which was introduced in the machine learning community [48, 49]. The kernel embedding of conditional distributions [48, 49] suggests that one can represent probability distribution as an element of a reproducing kernel Hilbert space (RKHS). In this paper, we will show that if $\mathcal{H}$ is an RKHS induced by an orthonormal basis $\{\phi_k : \mathcal{Y} \to \mathbb{R}\}$ of an appropriate L^2-space, then any $p(\cdot|\boldsymbol{z}) \in \mathcal{H}$ for any fixed $\boldsymbol{z} \in \mathcal{Z}$ can be represented as $p(\cdot|\boldsymbol{z}) = \sum_{k=1}^{\infty} c_k(\boldsymbol{z})\phi_k(\cdot)$, where the coefficients in c_k will be pre-computed using the historical data set and the kernel embedding of conditional distributions formula. Here, the convergence of the series representation is in uniform sense. In this paper, we will consider parametric orthonormal basis functions such as the Hermite polynomials for low-dimensional $\mathcal{Z}$ as well as the proper orthogonal decomposition (POD) modes for high-dimensional $\mathcal{Z}$. In the latter case, we shall see that the resulting closure model in (3) is a parametric model that is well-known, namely the linear non-autonomous autoregressive model. In general, the form of parametric closure models depends on the ansatz of ϕ_k as a function of $\boldsymbol{z}$. We should point out that one can also leave it entirely nonparametric with the data-driven basis functions constructed by the diffusion maps algorithm as in [3, 19]. While this is theoretically sound, the construction of data-driven basis functions requires an elaborate computational effort and is limited to problems with intrinsically low-dimensional $\mathcal{Y}$. In addition to constructing the basis, the main computational cost arises when evaluating the estimated basis functions on new points $\hat{z}_t$ for future-time prediction. Given these constraints, we will not explore the data-driven nonparametric basis in this paper.

The remaining of the paper is organized as follows. In Section 2, we briefly review the theory of kernel embedding of conditional distributions for estimating $p(y|\boldsymbol{z})$ using an orthonormal basis representation and discuss the proposed closure models in detail. In Section 3, we provide an intuition for choosing the density $p(y|\boldsymbol{z})$ by discussing missing dynamics in a linear Gaussian dynamics with and without temporal scale gaps. In Section 4, we numerically demonstrate the proposed approach on two nonlinear high-dimensional test problems, where m, n are small in the first example and large in the second example. In Section 5, we conclude the paper with a brief summary and discussion. We supplement the paper with two Appendices: Appendix A supplements Section 2 with a more detailed derivation of the kernel embedding of conditional distributions; Appendix B shows the consistency of the proposed approach in estimating autocovariance functions in the linear Gaussian case without the time-scale gap.

2 A nonparametric formulation of modeling missing dynamics

In this section, we first give a brief review on the kernel embedding of conditional distributions introduced in [48, 49], formulated using an orthonormal basis of appropriate square-integrable function spaces as in [3, 19, 52]. Subsequently, we present the proposed nonparametric modeling approach for missing dynamics.

2.1 Kernel embedding of conditional distributions

Let $\mathcal{Y}$ be a compact set and define $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ to be a kernel, which means it is symmetric positive definite and let it be bounded. By Moore-Aronszajn theorem, there exists a unique Hilbert space $\mathcal{H} = \overline{\text{span}\{K(y,\cdot), \forall y \in \mathcal{Y}\}}$. Let $q : \mathcal{Y} \to \mathbb{R}$ be a positive weight function and $\{\psi_k q\}_{k\geq 1}$ be a set of eigenfunctions corresponding to eigenvalues $\{\lambda_k\}$ of the following integral operator,

$$\mathcal{K}f(y) = \int_{\mathcal{Y}} K(y,y')f(y')q^{-1}(y')dy', \qquad f \in L^2(\mathcal{Y}, q^{-1}). \tag{4}$$

We should note that $\{\psi_k q\}$ forms an orthonormal basis of $L^2(\mathcal{Y}, q^{-1})$ and by Mercer's theorem, the kernel K has the following representation,

$$K(y, y') = \sum_{k=1}^{\infty} \lambda_k \psi_k(y) q(y) \psi_k(y') q(y'). \tag{5}$$

We should point out that if $\mathcal{Y}$ is not a compact domain such as $\mathbb{R}^n$, with an exponentially decaying q, one can construct a bounded Mercer-type kernel as in (5) with an appropriate choice of decreasing sequence $\{\lambda_k\}$ (see Lemma 3.2 in [52]) and it is a reproducing kernel corresponding to the RKHS $\mathcal{H}$ (see Proposition 3.4 in [52]). This result provides a justification for the use of Hermite polynomials $\{\psi_k\}$ with Gaussian weight q in one of our numerical examples.

We should point out that the RKHS $\mathcal{H}$ induced by the Mercer-type kernel in (5) is a subspace of $L^2(\mathcal{Y}, q^{-1})$ with the reproducing property corresponding to the inner product defined as $\langle f, g\rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \frac{f_k g_k}{\lambda_k}$, where $f_k = \langle f, \psi_k q\rangle_{L^2(\mathcal{Y}, q^{-1})}$ and $g_k = \langle g, \psi_k q\rangle_{L^2(\mathcal{Y}, q^{-1})}$. Then for any $f \in \mathcal{H}$ and $y \in \mathcal{Y}$, we can represent,

$$f(y) = \langle f, K(y, \cdot)\rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \frac{f_k \lambda_k \psi_k(y) q(y)}{\lambda_k} = \sum_{k=1}^{\infty} f_k \psi_k(y) q(y),$$

with the orthonormal basis of $L^2(\mathcal{Y}, q^{-1})$ where the convergence of the series holds uniformly (or in $C_0(\mathbb{R}^n)$ for non-compact $\mathcal{Y} = \mathbb{R}^n$).

Let Y and Z be random variables on $\mathcal{Y}$ and $\mathcal{Z}$, respectively, with distribution $P(Y, Z)$. Assuming that the conditional density $p(\cdot|\boldsymbol{z}) \in \mathcal{H}$ for any fixed $\boldsymbol{z} \in \mathcal{Z}$, we have the representation for $p(\cdot|\boldsymbol{z})$ in the RKHS $\mathcal{H}$ of real-valued functions on $\mathcal{Y}$:

$$p(y|\boldsymbol{z}) = \sum_{k=1}^{\infty} \langle p(\cdot|\boldsymbol{z}), \psi_k q\rangle_{L^2(\mathcal{Y}, q^{-1})} \psi_k(y) q(y) = \sum_{k=1}^{\infty} \langle p(\cdot|\boldsymbol{z}), \psi_k\rangle_{L^2(\mathcal{Y})} \psi_k(y) q(y), \tag{6}$$

where the convergence of the series is in the uniform sense and the coefficients are to be determined. The theory of kernel embedding of conditional distributions [48, 49], implemented also in [3, 19], suggests that the coefficients can be expressed as,

$$\langle p(\cdot|\boldsymbol{z}), \psi_k\rangle_{L^2(\mathcal{Y})} = \mathbb{E}_{Y|\boldsymbol{z}}[\psi_k(Y)] = \sum_{l=1}^{\infty} \left[\boldsymbol{C}_{YZ} \boldsymbol{C}_{ZZ}^{-1}\right]_{kl} \varphi_l(\boldsymbol{z}), \tag{7}$$

where $\{\varphi_l\}_{l \geq 1}$ forms an orthonormal basis of $L^2(\mathcal{Z}, \hat{q})$ and

$$[\boldsymbol{C}_{YZ}]_{ks} = \mathbb{E}_{YZ}\left[\psi_k(Y) \otimes \varphi_s(Z)\right], \qquad [\boldsymbol{C}_{ZZ}]_{sl} = \mathbb{E}_{ZZ}\left[\varphi_s(Z) \otimes \varphi_l(Z)\right].$$

See the detailed derivation of (7) in Appendix A. Substituting (7) to (6), we obtain,

$$p(y|\boldsymbol{z}) = \sum_{k,l=1}^{\infty} \psi_k(y) q(y) \left[\boldsymbol{C}_{YZ} \boldsymbol{C}_{ZZ}^{-1}\right]_{kl} \varphi_l(\boldsymbol{z}). \tag{8}$$

Notice that this representation can be understood as a linear regression in infinite-dimensional spaces with respect to the basis functions $\psi_k q$ and φ_l. Connecting to the notation in the introduction, $c_k(\boldsymbol{z}) = \sum_{l=1}^{\infty} \left[\boldsymbol{C}_{YZ} \boldsymbol{C}_{ZZ}^{-1}\right]_{kl} \varphi_l(\boldsymbol{z})$ and $\phi_k = \psi_k q$. The representation in (8) is nonparametric in the sense that we do not assume any particular distribution for the density.

Given pairs of data $\{y_i, \boldsymbol{z}_i\}_{i=1,\ldots,N}$, where $\boldsymbol{z}_i := (\boldsymbol{x}_{i-m:i}, \boldsymbol{y}_{i-n:i-1}) \in \mathcal{Z}$, distributed according to $P(Y, Z)$, we can estimate these coefficients via Monte-Carlo averages:

$$[\boldsymbol{C}_{YZ}]_{ks} \approx \frac{1}{N} \sum_{i=1}^{N} \psi_k(y_i) \varphi_s(\boldsymbol{z}_i), \qquad [\boldsymbol{C}_{ZZ}]_{sl} \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_s(\boldsymbol{z}_i) \varphi_l(\boldsymbol{z}_i). \tag{9}$$

We should point out that if the weight $\hat{q}$ in $L^2(\mathcal{Z}, \hat{q})$ is the sampling density of the data in $\mathcal{Z}$, since $\{\varphi_s\}$ is orthonormal under the corresponding inner product, then $\boldsymbol{C}_{ZZ}$ is an identity matrix. While a representation

on this Hilbert space is desirable, finding the corresponding orthonormal basis for high-dimensional $\mathcal{Z}$ is computationally challenging. In addition to constructing the basis, the main computational cost arises when evaluating the estimated basis functions on new points $\hat{z}_t$ for future-time prediction as shown in the next section. To avoid these expensive computations, we will adopt simpler basis functions, namely the Hermite polynomial basis for low-dimensional $\mathcal{Z}$ and the proper orthogonal decomposition (POD) basis for high-dimensional $\mathcal{Z}$.

2.2 Modeling the missing dynamics

Given the pre-computed conditional density in (8), the closure modeling approach proposed in (3) requires estimating the following statistical quantities,

$$\begin{aligned} \hat{a}(\hat{x}, \hat{\boldsymbol{z}}_t) &:= \mathbb{E}[a|\hat{\boldsymbol{z}}_t] := \int_{\mathcal{Y}} a(\hat{x}, y) p(y|Z = \hat{\boldsymbol{z}}_t)\, dy, \\ \hat{B}(\hat{x}, \hat{\boldsymbol{z}}_t) &:= \mathbb{E}[bb^\top|\hat{\boldsymbol{z}}_t] := \int_{\mathcal{Y}} b(\hat{x}, y) b(\hat{x}, y)^\top p(y|Z = \hat{\boldsymbol{z}}_t)\, dy. \end{aligned} \tag{10}$$

In the discussion below, we will just focus on the expectation of a (the calculation of the expectation of $bb^\top$ will be similar). In our formulation, we set the weight q in the Hilbert space $L^2(\mathcal{Y}, q^{-1})$ to be the sampling density of the data in $\mathcal{Y}$. In particular, substituting (8) into (10), we obtain,

$$\mathbb{E}[a|\hat{\boldsymbol{z}}_t] = \sum_{k,l=1}^{\infty} \int_{\mathcal{Y}} a(\hat{x}, y) \psi_k(y)\, q(y)\, dy \left[\boldsymbol{C}_{YZ}\boldsymbol{C}_{ZZ}^{-1}\right]_{kl} \varphi_l(\hat{\boldsymbol{z}}_t) = \sum_{l=1}^{\infty} A_l(\hat{x}) \varphi_l(\hat{\boldsymbol{z}}_t), \tag{11}$$

where

$$\begin{aligned} A_l(\hat{x}) &:= \sum_{k=1}^{\infty} \int_{\mathcal{Y}} a(\hat{x}, y) \psi_k(y)\, q(y)\, dy \left[\boldsymbol{C}_{YZ}\boldsymbol{C}_{ZZ}^{-1}\right]_{kl} \\ &\approx \frac{1}{N} \sum_{i=1}^{N} a(\hat{x}, y_i) \sum_{k=1}^{\infty} \psi_k(y_i) \left[\boldsymbol{C}_{YZ}\boldsymbol{C}_{ZZ}^{-1}\right]_{kl} \\ &= \frac{1}{N} \sum_{i=1}^{N} a(\hat{x}, y_i) \sum_{k=1}^{\infty} \psi_k(y_i) \sum_{s=1}^{\infty} \left[\boldsymbol{C}_{YZ}\right]_{ks} \left[\boldsymbol{C}_{ZZ}^{-1}\right]_{sl} \\ &\approx \frac{1}{N^2} \sum_{i,j=1}^{N} a(\hat{x}, y_i) \sum_{k=1}^{\infty} \psi_k(y_i) \sum_{s=1}^{\infty} \psi_k(y_j) \varphi_s(\boldsymbol{z}_j) \left[\boldsymbol{C}_{ZZ}^{-1}\right]_{sl} \\ &\approx \frac{1}{N} \sum_{i=1}^{N} a(\hat{x}, y_i) \sum_{s=1}^{\infty} \varphi_s(\boldsymbol{z}_i) \left[\boldsymbol{C}_{ZZ}^{-1}\right]_{sl} \end{aligned} \tag{12}$$

can be pre-computed. In this derivation, the second line is due to the Monte-Carlo average using data $y_i \sim q$, the fourth line above used (9), and the last line is due to the truncation in the summation of the index$-k$ up to order N, and the fact that,

$$\frac{1}{N} \sum_{k=1}^{N} \psi_k(y_i) \psi_k(y_j) = \delta_{ij}, \tag{13}$$

whenever $\{\psi_k\}$ is orthonormal in $L^2(\mathcal{Y}, q)$, where the weight q is exactly the sampling density of $\{y_i\}$. To see (13), define an $N \times N$ matrix with components $A_{ij} = \psi_j(y_i)$, then the orthonormality condition means that $A^\top A = I$, where I denotes an $N \times N$ identity matrix. Thus, (13) is the (i, j)th component of $AA^\top = I$. Since the resulting coefficients in (12) are independent to $\psi_k(y)$, in practice, we only need to choose the basis $\varphi_l(\boldsymbol{z})$.

Notice that the resulting representation in (11) arising from the proposed nonparametric formulation in (8) is a parametric model when the summation term is truncated, where the parametric ansatz is determined

by how φ_l depends on $\boldsymbol{z}$. For example, when $\mathcal{Z}$ is low-dimensional, we will consider Hermite polynomial basis functions for $\{\varphi_l(\boldsymbol{z})\}_{l=1,\ldots,L}$ in a numerical example in Section 4.1. In this case, the resulting parametric model is a polynomial of degree$-L$ and the coefficients in $A_l(\hat{x})$ are directly estimated via the kernel embedding formula.

We should point out that when we use the Hermite polynomial basis, we set the weight $\hat{q}$ to be Gaussian with mean and covariance determined empirically from the training data $\{\boldsymbol{z}_i\}_{i=1,\ldots,N}$. In our numerics, we also employ a regularization $(\boldsymbol{C}_{ZZ}+\lambda\boldsymbol{I})^{-1}$ replacing $\boldsymbol{C}_{ZZ}^{-1}$ in (12), with a small parameter λ to compensate for the conditional density that is not in $\mathcal{H}$ (as suggested in [48, 49]). Basically, this regularization is the penalty of not building the appropriate RKHSs that respect the sampling distribution and geometry of the data.

For high-dimensional $\mathcal{Z}$, we will consider using the proper orthogonal decomposition (POD) as a basis for $\varphi_l(\boldsymbol{z})$. Conceptually, this choice of basis corresponds to using an empirical covariance as the kernel in (4) (see e.g., Chapter 5 of [15] for more detailed discussion). Computationally, define a matrix $\boldsymbol{Z}\in\mathbb{R}^{N\times n_z}$, where the ith row consists of the training data, $\boldsymbol{z}_i-\bar{\boldsymbol{z}}\in\mathbb{R}^{n_z}$, centered about its empirical mean, $\bar{\boldsymbol{z}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_i$, such that its row sum is zero. In this case, the function value $\varphi_j(\boldsymbol{z}_i)$ will be determined by the (i,j)th component of the orthonormal matrix $\boldsymbol{U}$ defined as,

$$\boldsymbol{U}=\boldsymbol{Z}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}, \tag{14}$$

where $\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$ is the singular value decomposition (SVD). These basis functions are called the Proper Orthogonal Decomposition (POD) modes or a discrete version of the Karhunen-Loève basis expansion (see e.g., Chapter 5 of [15]).

From the orthonormality of $\boldsymbol{U}$, we have $\boldsymbol{C}_{ZZ}=\boldsymbol{I}/N$ such that Eq. (11) can be further simplified to,

$$\mathbb{E}\left(a|\hat{\boldsymbol{z}}_t\right)=\sum_{i=1}^{N}\sum_{s=1}^{L}a(\hat{x},y_i)\varphi_s\left(\boldsymbol{z}_i\right)\varphi_s\left(\hat{\boldsymbol{z}}_t\right), \tag{15}$$

where we used L basis functions. Suppose that $a(\hat{x},y)=y$, then Eq. (15) can be equivalently rewritten in a matrix form as,

$$\mathbb{E}\left(Y|\hat{\boldsymbol{z}}_t\right)=\boldsymbol{Y}^{\top}\boldsymbol{U}\boldsymbol{U}_{\text{new}}^{\top}, \tag{16}$$

where the matrix $\boldsymbol{Y}=[y_1,\ldots,y_N]^{\top}$ is $N\times n_y$ with $\{y_i\}_{i=1}^{N}$ denoting the training data with dimension n_y. Here, the matrix $\boldsymbol{U}_{\text{new}}:=(\hat{\boldsymbol{z}}_t-\bar{\boldsymbol{z}})\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\in\mathbb{R}^{1\times L}$ is the Nyström extension for SVD [45], whose components approximate the basis function values at a new point $\hat{\boldsymbol{z}}_t$, that is, $\boldsymbol{U}_{\text{new}}\approx[\varphi_1\left(\hat{\boldsymbol{z}}_t\right),\ldots,\varphi_L\left(\hat{\boldsymbol{z}}_t\right)]$. Substituting Eq. (14) into the conditional expectation (16), we obtain

$$\mathbb{E}\left(Y|\hat{\boldsymbol{z}}_t\right)=\boldsymbol{Y}^{\top}\boldsymbol{Z}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1\top}\boldsymbol{V}^{\top}(\hat{\boldsymbol{z}}_t-\bar{\boldsymbol{z}})^{\top}=\left(\boldsymbol{Y}^{\top}\boldsymbol{Z}\right)\left(\boldsymbol{Z}^{\top}\boldsymbol{Z}\right)^{-1}(\hat{\boldsymbol{z}}_t-\bar{\boldsymbol{z}})^{\top}. \tag{17}$$

The formula in (17) is exactly a linear regression between observations $\{y_i\}_{i=1}^{N}$ and $\{\boldsymbol{z}_i\}_{i=1}^{N}$. This means that the nonparametric RKHS representation reduces to the parametric linear regression when POD bases are used to represent functions defined on the $\mathcal{Z}$ space. In the case where $\boldsymbol{z}_i:=(\boldsymbol{x}_{i-m:i},\boldsymbol{y}_{i-n:i-1})\in\mathcal{Z}$, $n_z=m+n+1$, the resulting closure model in (17) is nothing but a linear autoregressive model for variable x with a linear non-autonomous variable y.

While the POD representation is convenient for high dimensional problems, we should point out these basis functions may not be adequate for systems with nonlinear and/or non-Gaussian nature. In fact, we will show in Section 4.2 that the POD basis representation is not sufficient to recover the missing terms in a nonlinear system even when the invariant density is close to Gaussian. In this case, we will find that an additional Gaussian white noise term can be used to compensate for the residual space (orthogonal to POD).

3 A linear Gaussian example

In this section, we provide an intuitive argument for the choice of conditional density function $p(y_t|\boldsymbol{z}_t)$ in compensating the missing dynamical terms as proposed in (3). Specifically, we will build our intuition for

choosing variables $\boldsymbol{z}_t$ by studying the missing dynamics of an analytically tractable linear Gaussian problem with and without temporal scale gaps. That is, we consider a linear multi-scale dynamical model,

$$dx = (a_{11}x + a_{12}y)\,dt + \sigma_x dW_x, \tag{18}$$

$$dy = \frac{1}{\epsilon}(a_{21}x + a_{22}y)\,dt + \frac{\sigma_y}{\sqrt{\epsilon}}dW_y, \tag{19}$$

for a slow variable $x \in \mathbb{R}$ and a fast variable $y \in \mathbb{R}$ [11]. Here, W_x and W_y are independent Wiener processes. The parameters $\sigma_x, \sigma_y \neq 0$ and the eigenvalues of the matrix

$$A_\epsilon = \begin{pmatrix} a_{11} & a_{12} \\ \frac{1}{\epsilon}a_{21} & \frac{1}{\epsilon}a_{22} \end{pmatrix}$$

are strictly negative, to assure the existence of a unique invariant joint density $p_{eq}(x, y)$. The parameter $\epsilon > 0$ characterizes the time-scale separation between variables x and y. Moreover, we assume the coefficient

$$\widetilde{a} = a_{11} - a_{12}a_{22}^{-1}a_{21} < 0, \quad a_{22} < 0, \tag{20}$$

to assure that the leading-order slow dynamics supports an invariant measure $\hat{\rho}_{eq}(x)$.

When there is a time-scale gap, in the limit of $\epsilon \to 0$, the leading-order dynamics,

$$d\hat{x}_t = a\hat{x}_t dt + \sigma_x dW_x, \tag{21}$$

with $a = \widetilde{a}$ as defined in (20), can be obtained by averaging the slow component of the vector field, $(a_{11}x + a_{12}y)$, with respect to the invariant density $\rho_\infty(y; \hat{x}_t)$ of the fast dynamics in (19) for a fixed $\hat{x}_t := \hat{x}(t\tau)$. For this simple example, it is clear that $\rho_\infty(y; x) = \mathcal{N}(-a_{21}a_{22}^{-1}x, -.5\sigma_y^2 a_{22}^{-1})$. The effective equation in (21) is deduced using the averaging theory [22, 28, 46], which approximates the density of the full dynamics as,

$$p(x, y, t) = \hat{\rho}(x, t)\rho_\infty(y; x) + \mathcal{O}(\epsilon), \qquad t \geq 0, \tag{22}$$

where $\hat{\rho}(x, t)$ denotes the evolution density corresponding to the leading-order dynamics. First, we should point out that when the fast dynamics for y in (19) is not available, we have no information about the invariant density $\rho_\infty(y; x)$ and we also cannot generate samples of this density. Thus, $\tilde{a}$ is not computable since a_{21} and a_{22} are unknown.

Our proposed model in (3) for the closure is motivated by the following observation. Here, we first provide the theoretical validity of our closure model. Taking $t \to \infty$ in (22), the invariant density of the full dynamics can be approximated by that of the leading-order dynamics up to order-ϵ, that is, $p_{eq}(x, y) = \hat{\rho}_{eq}(x)\rho_\infty(y; x) + \mathcal{O}(\epsilon)$. Therefore,

$$p(y|x) := \frac{p_{eq}(x, y)}{\int_{\mathcal{Y}} p_{eq}(x, y)\,dy} = \frac{\hat{\rho}_{eq}(x)\rho_\infty(y; x) + \mathcal{O}(\epsilon)}{\hat{\rho}_{eq}(x) + \mathcal{O}(\epsilon)} = \rho_\infty(y; x) + \mathcal{O}(\epsilon). \tag{23}$$

This equation basically suggests that one can approximate $\rho_\infty(y; x)$ with the following conditional density $p(y|x)$. For this linear Gaussian example, one can solve the Lyapunov equation of the full system in (18)-(19) for the equilibrium covariance matrix $S = (s_{ij})_{i,j=1,2}$ and deduce that $p(y|x) = \mathcal{N}(s_{21}s_{11}^{-1}x, s_{22} - s_{21}s_{11}^{-1}s_{12})$. Expanding the mean and variance statistics in terms of ϵ, we obtain

$$\mathbb{E}[Y|x] := \int_{\mathcal{Y}} yp(y|x)\,dy = -a_{22}^{-1}a_{21}x + \mathcal{O}(\epsilon),$$

$$\mathbb{E}[Y^2|x] := \int_{\mathcal{Y}} y^2 p(y|x)\,dy = -\frac{\sigma_y^2}{2a_{22}} + \mathcal{O}(\epsilon),$$

which means that the order-ϵ expansion error in (23) is in the sense of the mean and variance.

Averaging the slow Eq. (18) with respect to this conditional density, $p(y|\hat{x}_t)$, we obtain a closure model of the form (21) with

$$a\hat{x}_t = \overline{a}\hat{x}_t = \int_{\mathcal{Y}} (a_{11}\hat{x}_t + a_{12}y)\,p(y|\hat{x}_t)\,dy = \left(a_{11} + a_{12}s_{21}s_{11}^{-1}\right)\hat{x}_t = \tilde{a}\hat{x}_t + \mathcal{O}(\epsilon), \tag{24}$$

which means that the proposed closure obtained by averaging over $p(y_t|x_t)$ is consistent (up to an order-ϵ error) with the reduced model obtained from the classical averaging theory. However, in general, such an analytical expression in (24) will not be available since we have no access to s_{21} and s_{11}. Numerically, we will approximate the conditional density, $p(y|x)$, by applying the kernel embedding of the conditional distributions discussed in the previous section on the training data set $\{x_i, y_i\}_{i=1}^N$. In this case, it is clear that $\boldsymbol{z}_t = x_t$ is the natural choice. In the remainder of this section, we will refer to this closure model as the "RKHS $p(y_t|x_t)$".

Now we turn to the discussion of our closure model for large ϵ. When there is no time-scale gap, i.e., $\epsilon = \mathcal{O}(1)$ is large, the approximation via the averaging theory is not valid, and thus, averaging over $p(y_t|x_t)$ will not work. In this case, let us consider $\boldsymbol{z}_t = \boldsymbol{x}_{t-m:t}$ such that our closure model is an average over a non-Markovian conditional density function $p(y_t|\boldsymbol{x}_{t-m:t})$. That is,

$$d\hat{x}_t = (a_{11}\hat{x}_t + a_{12}\mathbb{E}\left[Y|\hat{\boldsymbol{x}}_{t-m:t}\right])\, dt + \sigma_x dW_x(t), \tag{25}$$

where the conditional average is evaluated at a new data point $\hat{\boldsymbol{x}}_{t-m:t} := \Big(\hat{x}((t-m)\tau), \hat{x}((t-m+1)\tau), \ldots, \hat{x}(t\tau)\Big)$ for the time lag interval $\tau > 0$, resulting from the integration of (25) at previous time steps. Since the random variables Y of y_t and X of $\boldsymbol{x}_{t-m:t}$ are both Gaussian with mean zero and covariance,

$$\mathrm{Cov}\left(\begin{bmatrix} Y \\ X \end{bmatrix}, \begin{bmatrix} Y \\ X \end{bmatrix}\right) := \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{26}$$

we can deduce that

$$\mathbb{E}\left[Y|\hat{\boldsymbol{x}}_{t-m:t}\right] = \Sigma_{12}\Sigma_{22}^{-1}\hat{\boldsymbol{x}}_{t-m:t}. \tag{27}$$

When the covariance components Σ_{12} and Σ_{22} are empirically estimated from the training data, notice that (27) is identical to the conditional expectation with respect to the kernel embedding of the conditional distributions formulated using the POD basis in (17). More importantly, one can analytically show that the autocovariance function (ACV) of the proposed non-Markovian model in (25) with $m \to \infty$ agrees with the ACV of the $x-$component of the full model (see Appendix B for the detailed proof of this statement). The consistency of the ACV prediction as well as the closure in (27) with the RKHS formulation in (17) justifies the choice of $\boldsymbol{z}_t = \boldsymbol{x}_{t-m:t}$ when ϵ is large. In the numerics below, we will verify the robustness of the non-Markovian closure model resulted from this choice of $\boldsymbol{z}_t$ in terms of the short-time prediction skill and the long-time statistics of ACVs for any $\epsilon > 0$.

In Figure 1, we compare our proposed closure model in (27) , which we will refer to as "RKHS $p(y_t|\boldsymbol{x}_{t-m:t})$", with the standard averaging model in (21) with $a = \widetilde{a}$ and the RKHS $p(y_t|x_t)$ as well. In this numerical simulation, we build the closure models using the simulated data at discrete time step $\tau = 0.01$. When ϵ is small, one can observe the pathwise convergence of the solutions of the closure models to those of the full model (18)-(19) [Fig. 1(a)]. For small $\epsilon = 0.01$, the ACVs of all the closure models are in good agreement with the ACV of the full model [Fig. 1(b)]. These results agree with the invariant manifold theory for small ϵ [47]. However, when ϵ is large, the short-time predictions and the long-time ACVs become quite different among the three closure models [Figs. 1(c) and (d)]. In term of short-time predictions, the closure model (25) with $m = 500$ memory terms provides a slightly better RMSE than the other two closure models [Fig. 1(c)]. In term of long-time statistics, only the closure model (25) with long memory terms produces an accurate approximation of the ACV, whereas the other two closure models do not [Fig. 1(d)]. This consistency of ACVs can be verified explicitly as we mentioned before (see Appendix B).

The analysis over this simple example shows that the proposed modeling framework using the kernel embedding of the conditional density formulation provides accurate short-time predictions and consistent long-term statistical recoveries in the limit of the memory length $m \to \infty$. This consistency is robust whether the underlying full system has or does not have any temporal scale gap. Using this result as a guideline, a natural extension for compensating missing components in nonlinear systems is to consider $\boldsymbol{z}_t := (\boldsymbol{x}_{t-m:t}, \boldsymbol{y}_{t-n:t-1})$, that allows for the missing dynamical components to also depend on the history of y in addition to that of x. In practice, the key parameters which will be determined case-by-case are the memory length, m and n, as we shall see in the nonlinear examples in the next section.

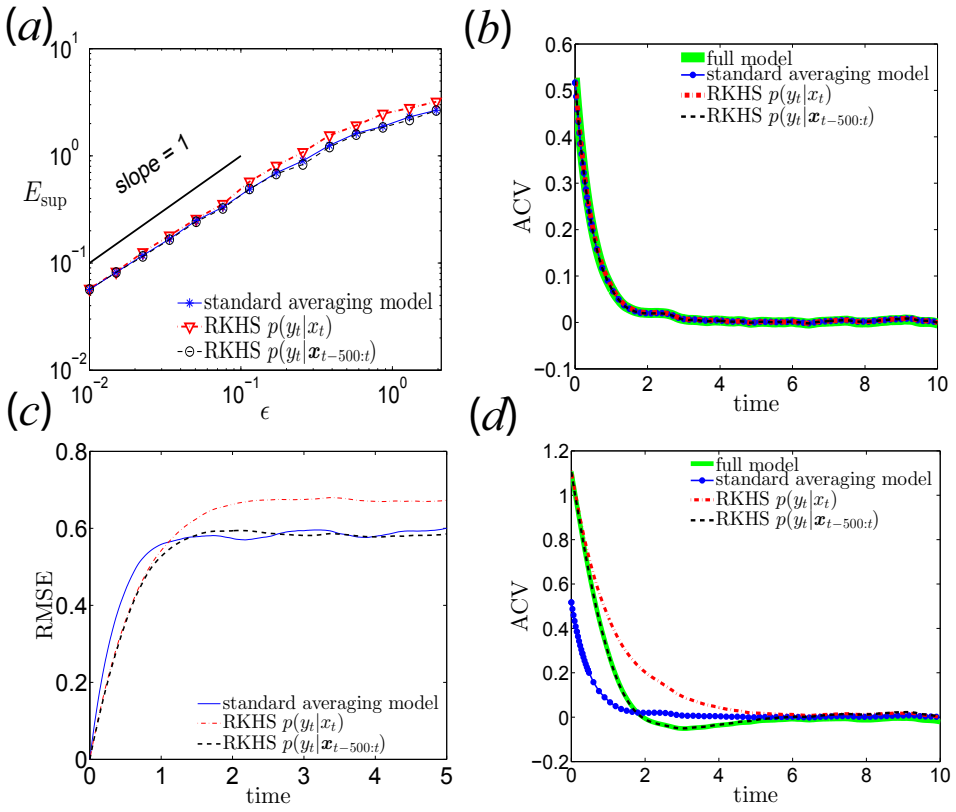

Figure 1: (Color online) (a) Supremum errors $E_{\sup}$ as functions of parameters ϵ, where $E_{\sup} \equiv \mathbb{E}\left(\sup_{0\le t\le 40}|e(t)|^2\right)$ with $e(t) = x(t) - \hat{x}(t)$. Here, $x(t)$ are solutions of the full model (18)-(19) and $\hat{x}(t)$ are solutions of the closure models. Trajectories are averaged over 100 realizations. The parameters are $a_{11} = a_{21} = a_{22} = -1$, $a_{12} = 1$, and $\sigma_x = \sigma_y = \sqrt{2}$. When ϵ is small, the solutions of all the closure models are pathwise convergent nearly on the order of ϵ. (c) Comparison of RMSEs averaged over 1000 realizations for large $\epsilon = 1.30$ regime. Comparison of ACVs for (b) the small $\epsilon = 0.01$ regime and (d) the large $\epsilon = 1.30$ regime.

4 Nonlinear examples

In this section, we study the short-time prediction and long-time statistical properties of two nonlinear examples: the Lorenz-96 (L96) model [31] possessing a short memory effect and the truncated Burgers-Hopf (TBH) model [39, 35, 38, 41] possessing a long memory effect.

4.1 Two-layer Lorenz-96 model

Consider the two-layer Lorenz-96 (L96) model [31],

$$\begin{aligned} \dot{X}^k &= X^{k-1}\left(X^{k+1} - X^{k-2}\right) - X^k + F + B^k, \\ \dot{Y}^{j,k} &= \frac{1}{\varepsilon}\left[Y^{j+1,k}\left(Y^{j-1,k} - Y^{j+2,k}\right) - Y^{j,k} + h_y X^k\right], \end{aligned} \tag{28}$$

for $k = 1, \ldots, K$, and $j = 1, \ldots, J$, where each relevant variable X^k is coupled to J irrelevant variables $Y^{j,k}$, and

$$B^k = \frac{h_x}{J}\sum_{j=1}^{J} Y^{j,k}. \tag{29}$$

The indices of the variables X^k and $Y^{j,k}$ are cyclic, $X^k = X^{k+K}, Y^{j,k} = Y^{j,k+K}, Y^{j+J,k} = Y^{j,k+1}$. The parameters are taken to be $K = 18$, $J = 20$, $F = 10$, $h_x = -1$, and $h_y = 1$ [7]. The parameter ε characterizes the time scale separation between the relevant component X^k and the irrelevant component $Y^{j,k}$. In this example, we will show the results for a small $\varepsilon = 1/128$ and a large $\varepsilon = 0.5$ (the large $\varepsilon = 0.5$ regime was studied in [7, 8, 29, 34]). We integrate the full L96 model using a 4th-order Runge-Kutta method for 10^3 time units with a time step $\delta t = 0.001$. We observe the trajectories of the variables (X^k, B^k) every 10 time steps, so that the observation time step is $\tau = 0.01$ and the dataset contains $N = 10^5$ observation points.

In the following numerical simulations, we compare our proposed closure RKHS models with the deterministic parametric formulation suggested by Wilk's method [51]. In particular, the Wilk's deterministic parameterization scheme is a closure model obtained by fitting the data $\{(X_i^k, B_i^k)\}_{i=1}^N$ with the following polynomial,

$$B^k = b_0 + b_1 X^k + b_2\left(X^k\right)^2 + b_3\left(X^k\right)^3 + b_4\left(X^k\right)^4 + b_5\left(X^k\right)^5. \tag{30}$$

We should point out that if we are restricted to only observing $\{X_i^k\}$, then $\{B_i^k\}$ are the identifiable components that can be extracted, for example, using a likelihood maximum estimate [27, 33, 51] or an adaptive Bayesian filtering [2], as we pointed out in the introduction. The key point is that we cannot extract the detailed components $Y^{j,k}$ if the fast dynamical components in (28) are unknown and, in fact, we are not interested in constructing a closure model by averaging over conditional density that depends directly on $Y^{j,k}$ since this can be very expensive. Instead, we will consider a closure model based on averaging over the conditional density $p\left(B_t^k | X_t^k\right)$ for small ϵ, where $B_t^k := B^k(t\tau)$ and $X_t^k := X^k(t\tau)$. For the large ε regime, we will consider $p\left(B_t^k | X_t^k, B_{t-1}^k\right)$. While conditioning to other variables (e.g., spatial neighbors of X^k or B^k or longer temporal history) can be considered, we do not find any meaningful improvement over the results that are presented below. These densities will be constructed using the kernel embedding formulation discussed in Section 2 for each k; connecting to the notation in the previous section, $y_t := B_t^k$ and $\boldsymbol{z}_t$ is either X_t^k or (X_t^k, B_{t-1}^k). To clarify, the full problem in (30) is $K + KJ = 18 + 18 \times 20 = 378$ dimension, and the closure model for the missing components, $\{Y^{j,k}\}_{k=1,\ldots,K, j=1,\ldots,J}$, is defined through a set of either one-dimensional or two-dimensional conditional densities; i.e., for each k, the density takes either X_t^k or (X_t^k, B_{t-1}^k) as inputs. Since these densities are low-dimensional functions with respect to the conditional variables (either X_t^k or (X_t^k, B_{t-1}^k)), we will represent the kernel embedding formula in (8) using the Hermite polynomials, expansion truncated at order $L = 50$ for each memory term.

To validate the proposed approach, we compare the short-time predictions and long-time statistics of the X^k-components between the full model and the closure models. Particularly, we compare several standard long-time statistical quantities as in [7, 33]:

- The probability density function (PDF) for X^k.
- The autocorrelation function (ACF) for X^k, $\left\langle X_t^k X_0^k \right\rangle / \left\langle X_0^k X_0^k \right\rangle$, where $\langle\cdot\rangle$ denotes the temporal average over $N = 10^5$ data points.

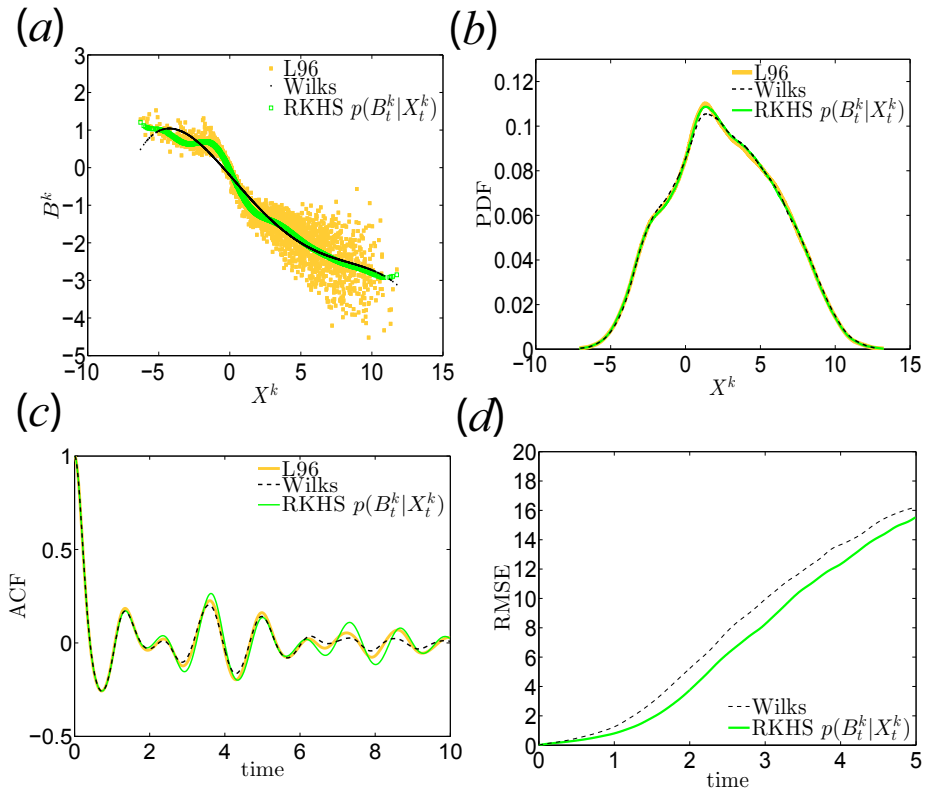

Figure 2: (Color online) Long-time statistics and short-time predictions for the small $\epsilon = 1/128$ regime of the L96 model. (a) The yellow dots are the scatter plot of B^k vs. X^k for the full L96 model. The black dots are the fifth-order polynomial fit used for the deterministic parametrization of B_k using Wilks's method [51]. The green squares are the closure model using the conditional density $p\left(B_t^k|X_t^k\right)$. Comparison of (b) PDFs and (c) ACFs among the full L96 model and the closure models. (d) Comparison of RMSEs from ensemble averages. The number of ensembles is 1000 where each ensemble corresponds to an initial state.

- The cross-correlation function (CCF) between X^k and X^{k+1}, $\langle X_t^k X_0^{k+1} \rangle / \langle X_0^k X_0^k \rangle$.
- The mean wave amplitude $\langle |u^m| \rangle$, for $m = 0, \ldots, K/2$, where u^m is the Fourier transform of X^k.
- The wave variance $\left\langle |u^m - \langle u^m \rangle|^2 \right\rangle$.

For the PDFs, ACFs, and CCFs, we plot the average over all $k = 1, \ldots, K$. For small $\varepsilon = 1/128$, we only show the results for the PDFs and ACFs. To assess the short-time prediction skill, we calculate the root-mean-square error (RMSE) and the anomaly correlation (ANCR), where the RMSE measures the difference between the true trajectory and the forecast trajectory whereas the ANCR measures the correlation between them [7]. The definitions of RMSE and ANCR are the same as those in [7]. We take the average using the data from 1000 different ensembles, each starting from a different initial state over five time units.

We first report the small $\varepsilon = 1/128$ regime of the L96 model. Figure 2(a) displays the scatter plot of B_t^k vs. X_t^k for the full L96 model, the polynomial fit (30) for the deterministic parametrization of B_k (Wilks's method), and the expectation $\mathbb{E}\left[B_t^k | X_t^k\right]$ using the RKHS representation (method referred to as the RKHS $p\left(B_t^k | X_t^k\right)$). For long-time statistics, one can see from Figs. 2(b) and 2(c) that the PDFs and ACFs for X^k can be well reproduced by both closure models. For short-time predictions, one can see from Fig. 2(d) that the RKHS $p\left(B_t^k | X_t^k\right)$ provides a better approximation of the trajectory compare to the Wilks's deterministic parametrization scheme. These results can be expected due to the validity of the classical averaging theory on dynamical systems with time-scale separation (small ε regime) [47].

We now report the L96 model for the large $\varepsilon = 0.5$ regime in which there is no significant time-scale separation between the relevant, X^k, and irrelevant variables, B^k. By comparing Fig. 2(a) and 3(a), one can see that the patterns of the scatter plots differ substantially between the small and large ε regimes. Specifically, the scatter plot for the large ε regime is much broader in B^k direction compare to that for the small ε regime. This indicates that when ε is small, the irrelevant (fast) variable significantly relies on the relevant (slow) variable. When ε becomes large, such dependence of irrelevant variable B^k on the relevant variable X^k reduces.

For large $\varepsilon = 0.5$, one can observe from Fig. 3(a) that the RKHS representation of $\mathbb{E}\left[B_t^k | X_t^k, B_{t-1}^k\right]$ can nearly reproduce the scatter plot of the full model, whereas the Wilks's deterministic parametrization scheme and the RKHS representation $\mathbb{E}\left[B_t^k | X_t^k\right]$ cannot. The PDFs for X^k of the full model can be reproduced by all the closure models [Fig. 3(b)]. For the other long-time statistics, ACFs, CCFs, mean wave amplitudes, and wave variances can be well reproduced only by the closure model using the conditional density $p\left(B_t^k | X_t^k, B_{t-1}^k\right)$ [Figs. 3(c)(d)(e)(f)]. Notice also the significant improvement in terms of short-time predictions using the RKHS $p\left(B_t^k | X_t^k, B_{t-1}^k\right)$ (smaller RMSE and higher ANCR) over the Wilks's method and the RKHS $p\left(B_t^k | X_t^k\right)$ as shown in Fig. 4.

To determine the reliability of the ensemble forecasts, we also calculate the rank histograms from an ensemble of integrations [14]. A rank histogram is obtained by repeatedly tallying the rank of the true observation relative to the sorted N_{ens}-member ensemble [14]. We use the same method as in [7]. For every initial state $X_{t_0}^k$, we do N_{ens} integrations of the closure models over the lead T time units starting from the $X_{t_0}^k$ plus a small random perturbation. The random perturbations are Gaussian distribution with mean zeros and standard deviation 0.15. We sort the $N_{\text{ens}} + 1$ values for X_t^k for each grid point k and time t from the ensemble members and the full L96 model. Figure 5 displays the rank histograms for all the closure models with $N_{\text{ens}} = 9$ at lead time $T = 2$. An ideal rank histogram is flat. One can see that the rank histogram by the RKHS $p\left(B_t^k | X_t^k, B_{t-1}^k\right)$ is close to be flat, whereas rank histograms by Wilks's deterministic parametrization scheme and the RKHS $p\left(B_t^k | X_t^k\right)$ exhibit U-shape distributions. Therefore, the closure model with $p\left(B_t^k | X_t^k, B_{t-1}^k\right)$ performs better than the other two closure models.

4.2 The truncated Burgers-Hopf (TBH) model

Consider the truncated Burgers-Hopf (TBH) model [39, 35, 38, 41], which is described by a system of quadratic nonlinear equations for the complex Fourier modes, u^k, with $u^{-k} = (u^k)^*$ for $1 \le |k| \le \Lambda$,

$$\frac{du^k}{dt} = -\frac{\mathrm{i}k}{2} \sum_{\substack{k+p+q=0 \\ 1 \le |p|,|q| \le \Lambda}} (u^p)^* (u^q)^*. \tag{31}$$

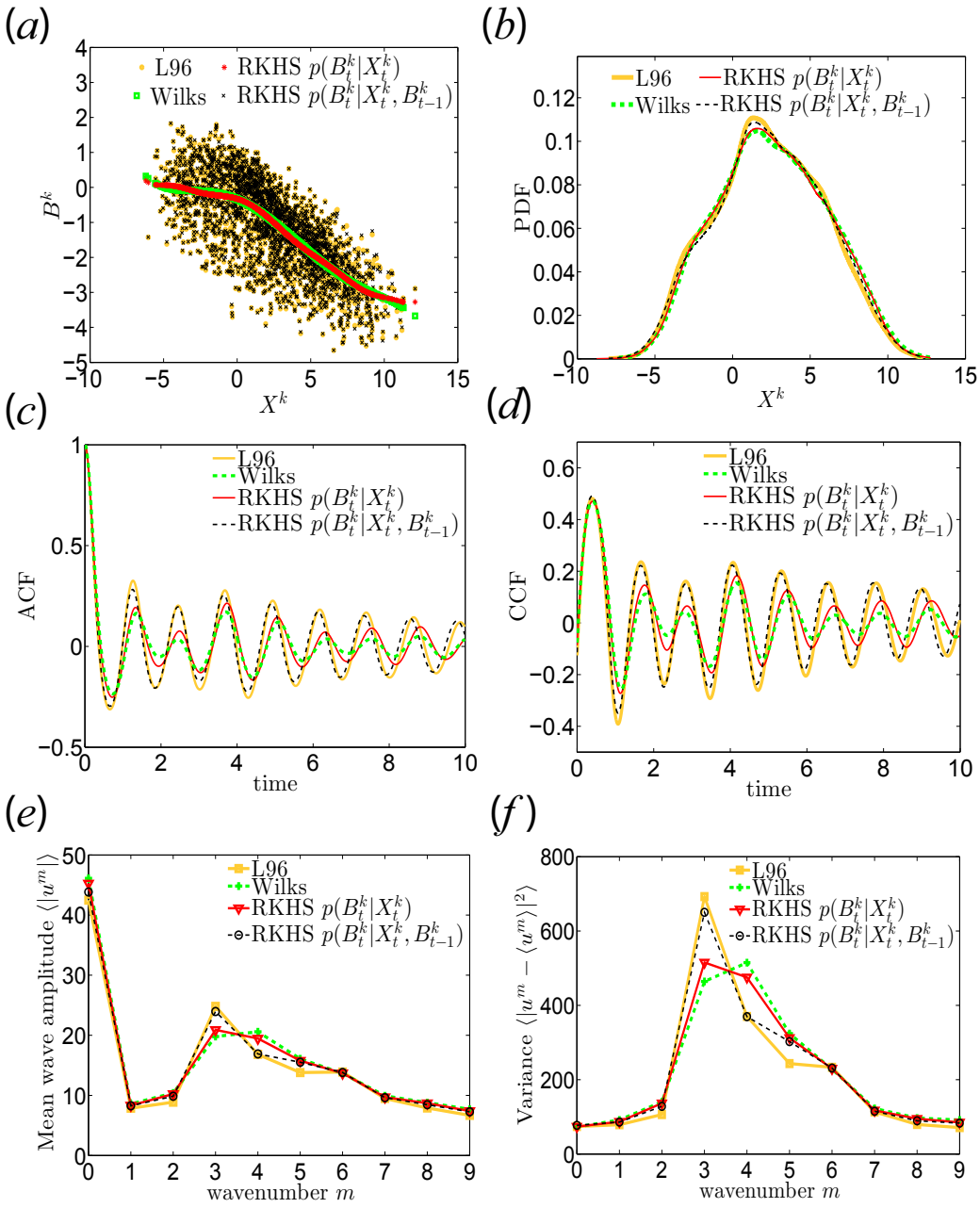

Figure 3: (Color online) Long-time statistics and short-time predictions for the large $\epsilon = 0.5$ regime of the L96 model. (a) The yellow dots are the scatter plot of B^k vs. X^k for the full L96 model. The green squares are the fifth-order polynomial fit using Wilks's method [51]. The red asterisks and black crosses correspond to the closure models using the conditional densitys $p\left(B_t^k|X_t^k\right)$ and $p\left(B_t^k|X_t^k, B_{t-1}^k\right)$, respectively. Comparison of (b) PDFs, (c) ACFs, (d) CCFs, (e) mean wave amplitudes, and (f) wave variances among the full model and closure models.

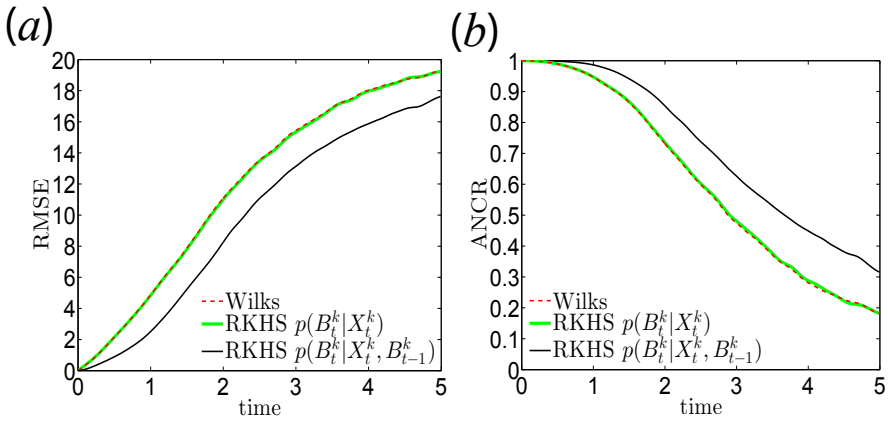

Figure 4: (Color online) Comparison of (a) RMSEs and (b) ANCRs for the large $\epsilon = 0.5$ regime. The number of ensembles is 1000 where each ensemble corresponds to an initial state.

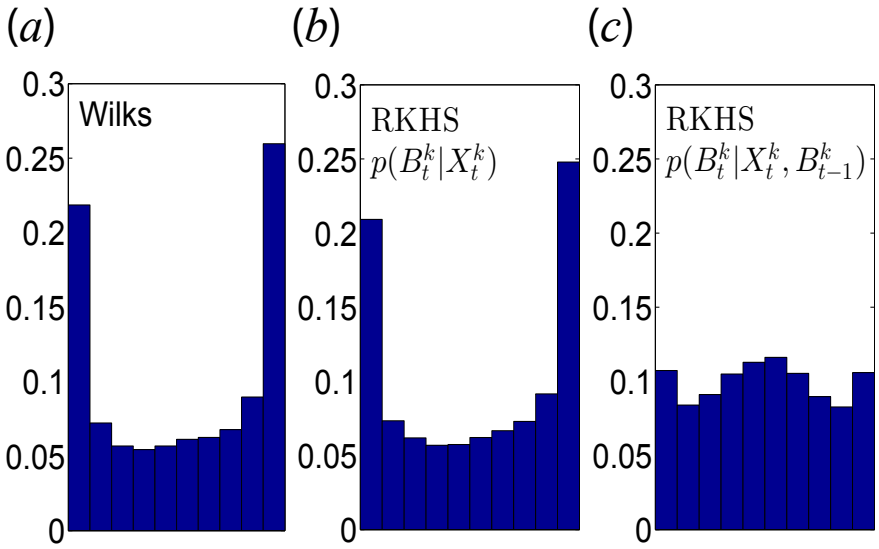

Figure 5: (Color online) Rank histograms for closure models with ensemble members $N_{\text{ens}} = 9$ at lead time $T = 2$. Ideally, the rank histogram is nearly flat. The rank histogram of the closure model using $p\left(B_t^k|X_t^k, B_{t-1}^k\right)$ is close to be flat, whereas rank histograms of Wilks's method and the closure model using $p\left(B_t^k|X_t^k\right)$ exhibit U-shape distributions.

This model is a Galerkin truncation of the inviscid Burgers equation on Fourier modes and we should point out that the dynamics of the truncated system is totally different from the inviscid Burgers equation. Particularly, the TBH exhibits intrinsic stochastic dynamics with ergodic behavior in a large deterministic system [39, 35, 38, 41]. We are interested in estimating the TBH model's first Fourier mode given only the dynamical component of this mode,

$$\frac{du^1}{dt} = -\mathrm{i}(u^1)^* u^2 + F, \tag{32}$$

where u_2 denotes the second Fourier mode and F denotes the forcing component obtained by subtracting $-\mathrm{i}(u^1)^* u^2$ from the right hand side of Eq. (31), that is,

$$F = -\frac{\mathrm{i}}{2} \sum_{\substack{1+p+q=0 \\ 2\le |p|,|q|\le \Lambda}} (u^p)^*(u^q)^*. \tag{33}$$

While u^2 and F may be identifiable from observing u^1 alone, in our experiment below, we assume that we are given the data set of $\{u_i^1, u_i^2, F_i\}_{i=1}^N$. We should point out that this model has an equipartition energy, that is, all of the Fourier modes in TBH have the same variances, and the first Fourier mode (which is of our interest) possesses the longest autocorrelation time and the largest statistical memory [40], which makes this example a tough test problem.

To compensate for the missing dynamics in (32), we substitute the irrelevant variables u^2 and F with their conditional expectations. In this case, the closure model involves $p(y|\boldsymbol{x}_{t-m:t-1}, \boldsymbol{y}_{t-n:t-1})$, where the irrelevant variable y is one of $\{u^{2,\mathrm{Re}}, u^{2,\mathrm{Im}}, F^{\mathrm{Re}}, F^{\mathrm{Im}}\}$, and the relevant variable $\boldsymbol{x}$ is one of $\{u^{1,\mathrm{Re}}, u^{1,\mathrm{Im}}\}$ such that $\boldsymbol{y}$ and $\boldsymbol{x}$ are both real or both imaginary parts. In particular, we employ the RKHS formulation to construct four conditional densities: $p(u_t^{2,\mathrm{Re}}|\boldsymbol{u}_{t-m:t-1}^{1,\mathrm{Re}}, \boldsymbol{u}_{t-n:t-1}^{2,\mathrm{Re}})$, $p(u_t^{2,\mathrm{Im}}|\boldsymbol{u}_{t-m:t-1}^{1,\mathrm{Im}}, \boldsymbol{u}_{t-n:t-1}^{2,\mathrm{Im}})$, $p(F_t^{\mathrm{Re}}|\boldsymbol{u}_{t-m:t-1}^{1,\mathrm{Re}}, \boldsymbol{F}_{t-n:t-1}^{\mathrm{Re}})$, and $p(F_t^{\mathrm{Im}}|\boldsymbol{u}_{t-m:t-1}^{1,\mathrm{Im}}, \boldsymbol{F}_{t-n:t-1}^{\mathrm{Im}})$. For the forcing F, an additional Gaussian noise term is added to compensate for the residual space. Since the conditional states are high-dimensional (when m, n are large), the conditional expectations over these densities are represented using the POD bases as in (17).

To conduct this numerical experiment, the training dataset is generated from the full TBH model (31), where F is calculated by Eq. (33). We integrate the full TBH model for 10^4 time units with time step $\Delta t = 10^{-3}$. We store the data at every 0.01 time unit and thereafter the dataset contains 10^6 points for all u^k. We compare the results generated by the full TBH model (31) and the closure models, resulted by averaging the partial dynamics in (32) over the pre-trained conditional densities. In this example, we consider the full TBH model (31) in a high-energy regime with $\beta = 10$ and $\Lambda = 50$ as in [41]. Here, Λ denotes number of modes in Eq. (31) and $\beta = \Lambda/\overline{E}$ with $\overline{E}$ being the mean energy per mode. Here, the full dynamics in (31) has 50-dimensional complex variables and the four conditional densities (in previous paragraph) are proposed as the closure model for the dynamics of the missing components, $\{u^2, \ldots, u^\Lambda\}$. Each of the four conditional densities above is a real-valued function that takes $m+n$ dimensional variables. In our numerical experiments, we will take $m + n \ge 20$.

We compare three closure models of (32) with different memory terms m and n and different temporal steps $\tau = 0.01$ or 0.1. Figure 6 displays long-time statistics and short-time predictions for these closure models. One can see from Figs. 6(a)(b)(c)(d) that the long-time statistics can be well reproduced by the proposed closure models of (32) when the irrelevant variables have long memory terms, that is, n is large enough. In terms of short-time predictions, all three closure models exhibit comparable results for RMSEs and ANCRs where the errors saturate at about the time when the autocovariance function diminishes [Figs. 6(e) and 6(f)]. The fact that the two choices of m, n, τ ($m = n = 20, \tau = 0.1$ and $m = n = 200, \tau = 0.01$), corresponding to the two models with the same memory length $n\tau = 2$ time units, produce comparable results (see the red dash-dotted and black dashed curves in Fig. 6) suggests that the temporal step τ does not affect the inference. Thus it is more economical to use the model with smaller n (and possibly coarser time lag τ) that gives the same accuracy. Finally, we should also point out that if the memory length $n\tau$ (in unit time) is small, the estimates become less accurate. Therefore, for this difficult test problem involving observations of the first Fourier mode of the TBH model, the proposed closure model can replicate the long-time statistics accurately and produce reasonable short-time prediction skills when there are long enough memory terms.

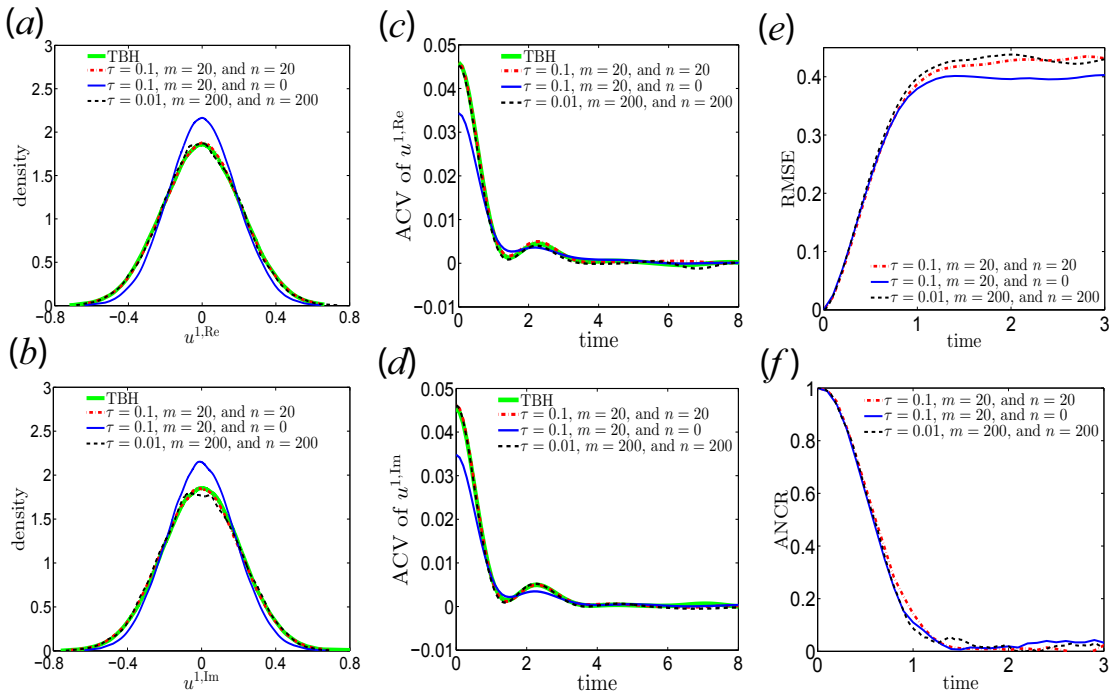

Figure 6: (Color online) Long-time statistics and short-time predictions for the TBH model. Comparison of PDFs of (a) $u^{1,\mathrm{Re}}$ and (b) $u^{1,\mathrm{Im}}$. Comparison of ACVs of (c) $u^{1,\mathrm{Re}}$ and (d) $u^{1,\mathrm{Im}}$. Comparison of (e) RMSEs and (f) ANCRs. The closure models use the conditional density $p(y|\boldsymbol{x}_{t-m:t-1}, \boldsymbol{y}_{t-n:t-1})$ with different observation time step τ and number of memory terms m and n [see text].

5 Summary and discussion

In this paper, we considered a data-driven nonparametric model for capturing the missing dynamics in the context of systems of ergodic SDE's and ODE's. The non-Markovian closure model is formulated as an averaging over an equilibrium conditional density function, $p(y_t|\boldsymbol{x}_{t-m:t}, \boldsymbol{y}_{t-n:t-1})$, that is approximated using the kernel embedding of conditional distribution formulation. In particular, we considered a representation of the conditional density on RKHS induced by an orthonormal basis of appropriate weighted Hilbert space. A thorough investigation of the modeling framework on a linear Gaussian problem shows the consistency with the classical averaging theory for fast-slow systems and justifies our use of long non-Markovian memory terms to obtain accurate two-point statistical predictions in the case of no temporal scale separation. Numerical simulations on nonlinear problems demonstrate the robustness of the framework in producing accurate short-term predictions as well as to recover two-point statistics even when the missing terms are high-dimensional and have no separation of scales.

Modeling of missing dynamics with parametric models (or closure) has a long history as we noted in the introduction. Practically, such modeling paradigm requires modeler to choose the parametric model (ansatz) and fit the proposed model to the data to estimate the parameters in the ansatz. The choice of the parametric model is typically problem specific. When the underlying full system is known, one can deduce the model from the first principle. For example, one can apply the Mori-Zwanzig formalism to deduce such parametric model (see e.g., [4, 5, 17, 25]) and then use various mathematical tools to estimate the memory integral terms as well as the parameters in the reduced model, which remains challenging if the resulting closure model is nonlinear or contains high-dimensional parameters. For example, when a rational approximation is used as a model for the memory kernel [30], while the parameters can be identified from derivatives of the kernel, it requires the availability of highly accurate time series (in the sense of accurate several order of derivatives) which is rare in practice.

In this paper, the proposed nonparametric formulation discovered some of the well-known parametric models, including the non-autonomous autoregressive linear models. An important feature of the proposed nonparametric framework in this paper is that it translates the problem of choosing parametric model into choosing the memory length m, n and constructing orthonormal basis of a weighted Hilbert space of functions that take values on $\boldsymbol{z} \in \mathcal{Z}$. For the memory length, our experience suggests that we can use the decaying time scale of processes x and y as a guideline. While the natural candidate of model is a representation on a Hilbert space spanned by the orthonormal basis of functions that respect the geometry and sampling density of the data as in [19], constructing such a basis is computationally challenging especially if $\mathcal{Z}$ is high-dimensional. In addition to the difficulty in the basis construction, the main computational cost arises as we evaluate the estimated basis functions on new points for future-time prediction. For very low-dimensional $\mathcal{Z}$, our numerical results suggest that we can avoid all of this practical issue with classical polynomial basis functions. In this case, the form of parametric model is polynomial functions. For very high-dimensional $\mathcal{Z}$, we showed the effectiveness of using the POD basis for representing linear problems. In nonlinear problems, we found that in some case, additional noise terms can be used to compensate for the orthogonal components that are not represented by the POD bases. In this case, the resulting parametric model is a linear non-autonomous autoregressive model. The second important feature is that the proposed nonparametric framework provides a linear technique for estimating the parameters in the resulting parametric models regardless of whether they are linear or nonlinear. This important feature is inherited from the kernel embedding formulation that allows one to "gain" linearity by representing nonlinear functions of a finite dimensional space $\mathcal{Z}$ with a basis of functions of infinite dimensional linear space. To summarize, the proposed framework and the results in this study suggest that one can understand the parametric modeling paradigm from a unified framework using appropriate Reproducing Kernel Hilbert Spaces. Such realization lies on the interpretation of the Mercer's type kernels in (5). While the so-called kernel "trick" uses Mercer kernels to avoid the evaluation of inner product in feature space, our view point is to use the Mercer kernels to construct the parametric model of interest by an appropriate choice of finite number of basis functions. Thus, this framework turns the problem of finding the right closure model into a problem of constructing a complete basis of the Hilbert space induced by the data, which remains challenging in general.

Finally, we should also point out that the modeling framework introduced here can also be realized with any supervised learning algorithm other than the kernel embedding discussed here. In a separate report, [16], we found that the closure modeling framework introduced here is effective for high-dimensional

nonlinear problems when it is realized with the Long-Short-Term-Memory (a special class of Recurrent Neural Network).

Acknowledgments

It is a great pleasure to dedicate this paper to Andrew Majda on the occasion of his 70th birthday. The research of J.H. was partially supported by the ONR Grant N00014-16-1-2888, NSF Grants DMS-1619661 and DMS-1854299. S.W.J. was supported as a postdoctoral fellow under the ONR Grant N00014-16-1-2888.

Conflict of interest

The authors declare that they have no conflict of interest.

A Kernel mean embedding of conditional distributions

The purpose of this review is to verify Eq. (7). While the derivation here follows closely the description in [49, 48], we taylor the discussion here for Mercer-type kernels induced by orthonormal basis of L^2-spaces. Some of the basic theory of RKHS can be found in many texts, such as [6].

First, let us repeat the discussion in Section 2.1 on $\mathcal{Z}$. Let $\mathcal{Z}$ be a compact set and define $\hat{K} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ to be a kernel, which means it is symmetric positive definite and let it be bounded. By Moore-Aronszajn theorem, there exists a unique Hilbert space $\mathcal{H}_Z = \overline{\text{span}\{\hat{K}(\boldsymbol{z}, \cdot), \forall \boldsymbol{z} \in \mathcal{Z}\}}$. Let $\hat{q} : \mathcal{Z} \to \mathbb{R}$ be a positive weight function and $\{\varphi_k\}_{k\geq 1}$ be a set of eigenfunctions corresponding to eigenvalues $\{\xi_k\}$ of the following integral operator $\hat{\mathcal{K}} : L^2(\mathcal{Z}, \hat{q}) \to L^2(\mathcal{Z}, \hat{q})$, defined as,

$$\hat{\mathcal{K}} f(\boldsymbol{z}) := \int_{\mathcal{Z}} \hat{K}(\boldsymbol{z}, \boldsymbol{z}') f(\boldsymbol{z}') \hat{q}(\boldsymbol{z}') d\boldsymbol{z}'. \tag{34}$$

By Mercer's theorem, the kernel $\hat{K}$ has the following representation,

$$\hat{K}(\boldsymbol{z}, \boldsymbol{z}') = \sum_{k=1}^{\infty} \xi_k \varphi_k(\boldsymbol{z}) \varphi_k(\boldsymbol{z}'). \tag{35}$$

We should point out that if $\mathcal{Z}$ is not a compact domain such as $\mathbb{R}^n$, with an exponentially decaying $\hat{q}$, one can construct a bounded Mercer-type kernel as in (35) with an appropriate choice of decreasing sequence $\{\xi_k\}$ (see Lemma 3.2 in [52]) and it is a reproducing kernel corresponding to the RKHS $\mathcal{H}_Z$ (see Proposition 3.4 in [52]).

In this case, the RKHS $\mathcal{H}_Z$ induced by the Mercer-type kernel in (35) is a subspace of $L^2(\mathcal{Z}, \hat{q})$ with the reproducing property corresponding to an inner product defined as $\langle f, g \rangle_{\mathcal{H}_Z} = \sum_{k=1}^{\infty} \frac{f_k g_k}{\xi_k}$, for all $f, g \in \mathcal{H}_Z$ where $f_k = \langle f, \varphi_k \rangle_{L^2(\mathcal{Z}, \hat{q})}$ and $g_k = \langle g, \varphi_k \rangle_{L^2(\mathcal{Z}, \hat{q})}$. Then for any $f \in \mathcal{H}_Z$ and $\boldsymbol{z} \in \mathcal{Z}$, we can represent

$$f(\boldsymbol{z}) = \langle f, \hat{K}(\boldsymbol{z}, \cdot) \rangle_{\mathcal{H}_Z} = \sum_{k=1}^{\infty} \frac{f_k \xi_k \varphi_k(\boldsymbol{z})}{\xi_k} = \sum_{k=1}^{\infty} f_k \varphi_k(\boldsymbol{z}), \tag{36}$$

with basis of $L^2(\mathcal{Z}, \hat{q})$, where the convergence of the series holds uniformly (or in $C_0(\mathbb{R}^n)$ for non-compact $\mathcal{Z} = \mathbb{R}^n$).

We called the Hilbert space of functions, $\mathcal{H}_Z$, as an RKHS induced by the orthonormal basis of $L^2(\mathcal{Z}, \hat{q})$. While we have discussed $\mathcal{H}$ as an RKHS induced by the orthonormal basis of $L^2(\mathcal{Y}, q^{-1})$ in Section 2.1, we can also repeat the argument above and construct $\mathcal{H}_Y$ as an RKHS induced by the orthonormal basis of $L^2(\mathcal{Y}, q)$. In this case, recall that while $\{\psi_k q\}$ are orthogonal eigenbasis of the integral operator in (4), the orthogonal basis $\psi_k \in L^2(\mathcal{Y}, q)$ are eigenfunctions of an adjoint integral operator of (4). That is, one can verify that

$$\langle \psi_k q, \mathcal{K}^* \psi_k \rangle_{L^2(\mathcal{Y})} = \langle \mathcal{K}(\psi_k q), \psi_k \rangle_{L^2(\mathcal{Y})} = \lambda_k \langle \psi_k q, \psi_k \rangle_{L^2(\mathcal{Y})}, \tag{37}$$

where for $f \in L^2(\mathcal{Y}, q)$,

$$\mathcal{K}^* f(x) := \int_{\mathcal{Y}} K^*(x,y) f(y) q(y)\, dy,$$

and $K^*(x,y) = q(x)^{-1} K(x,y) q^{-1}(y)$ is also a symmetric positive definite kernel. By Mercer's theorem, one can write

$$K^*(y,y') = \sum_{k=1}^{\infty} \lambda_k \psi_k(y) \psi_k(y'). \tag{38}$$

Let Y and Z be random variables on $\mathcal{Y}$ and $\mathcal{Z}$ with distribution $P(Y,Z)$, we define the cross-covariance operators, $\mathcal{C}_{YZ} : \mathcal{H}_Z \to \mathcal{H}_Y$ and $\mathcal{C}_{ZZ} : \mathcal{H}_Z \to \mathcal{H}_Z$ as,

$$\begin{aligned} \mathcal{C}_{YZ} &:= \mathbb{E}_{YZ}[K^*(Y,\cdot) \otimes \hat{K}(Z,\cdot)], \\ \mathcal{C}_{ZZ} &:= \mathbb{E}_{Z}[\hat{K}(Z,\cdot) \otimes \hat{K}(Z,\cdot)]. \end{aligned} \tag{39}$$

One can immediately see that for any $f \in \mathcal{H}_Y$ and $g \in \mathcal{H}_Z$,

$$\begin{aligned} \mathbb{E}_{YZ}[f(Y) \otimes g(Z)] &= \int_{\mathcal{Y}\times\mathcal{Z}} f(y) g(\boldsymbol{z}) dP(y,\boldsymbol{z}) = \int_{\mathcal{Y}\times\mathcal{Z}} \langle f, K^*(y,\cdot)\rangle_{\mathcal{H}_Y} \langle g, \hat{K}(\boldsymbol{z},\cdot)\rangle_{\mathcal{H}_Z} dP(y,\boldsymbol{z}) \\ &= \int_{\mathcal{Y}\times\mathcal{Z}} \langle f \otimes g, K^*(y,\cdot) \otimes \hat{K}(\boldsymbol{z},\cdot)\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_Z} dP(y,\boldsymbol{z}) = \langle f \otimes g, \mathcal{C}_{YZ}\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_Z}. \end{aligned} \tag{40}$$

Let us define feature maps $\Psi : \mathcal{Y} \to \mathcal{F}_Y \subset \ell_2$ and $\Phi : \mathcal{Z} \to \mathcal{F}_Z \subset \ell_2$, respectively,

$$\begin{aligned} \Psi(y) &= (\sqrt{\lambda_1}\psi_1(y), \sqrt{\lambda_2}\psi_2(y), \ldots), \\ \Phi(\boldsymbol{z}) &= (\sqrt{\xi_1}\varphi_1(\boldsymbol{z}), \sqrt{\xi_2}\varphi_2(\boldsymbol{z}), \ldots). \end{aligned} \tag{41}$$

Then we can write

$$\begin{aligned} \hat{K}(\boldsymbol{z},\boldsymbol{z}') &= \langle \Phi(\boldsymbol{z}), \Phi(\boldsymbol{z}')\rangle_{\ell_2} = \langle \hat{K}(\boldsymbol{z},\cdot), \hat{K}(\boldsymbol{z}',\cdot)\rangle_{\mathcal{H}_Z}, \\ K^*(y,y') &= \langle \Psi(y), \Psi(y')\rangle_{\ell_2} = \langle K^*(y,\cdot), K^*(y',\cdot)\rangle_{\mathcal{H}_Y}, \end{aligned}$$

where the inner products in $\mathcal{H}_Z$ and $\mathcal{H}_Y$ can be identified by ℓ_2 inner products in the corresponding feature spaces. Also, for any function $f \in \mathcal{H}_Z$ and $\boldsymbol{z} \in \mathcal{Z}$, we can rewrite the expansion in (36) as,

$$f(\boldsymbol{z}) = \langle f, \hat{K}(\boldsymbol{z},\cdot)\rangle_{\mathcal{H}_Z} = \sum_{k=1}^{\infty} \langle f, \varphi_k\rangle_{L^2(\mathcal{Z},\hat{q})} \varphi_k(\boldsymbol{z}) = \sum_{k=1}^{\infty} \frac{\langle f, \varphi_k\rangle_{L^2(\mathcal{Z},\hat{q})}}{\sqrt{\xi_k}} \Phi_k(\boldsymbol{z}) = \sum_{k=1}^{\infty} \langle f, \Phi_k\rangle_{\mathcal{H}_Z} \Phi_k(\boldsymbol{z}), \tag{42}$$

where we have defined the functions $\Phi_k = \sqrt{\xi_k}\varphi_k \in \mathcal{H}_Z$. For convenience of the discussion below, we also define the functions $\Psi_k := \sqrt{\lambda_k}\psi_k \in \mathcal{H}_Y$.

Using the identity in (40), we can represent the cross-operators in (39) on the basis coordinates $\Psi_k \in \mathcal{H}_Y$ and $\Phi_\ell \in \mathcal{H}_Z$ as follows:

$$\begin{aligned} [C_{YZ}]_{k\ell} &:= \mathbb{E}_{YZ}[\Psi_k(Y) \otimes \Phi_\ell(Z)] = \langle \Psi_k \otimes \Phi_\ell, \mathcal{C}_{YZ}\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_Z}, \\ [C_{ZZ}]_{k\ell} &:= \mathbb{E}_{ZZ}[\Phi_k(Z) \otimes \Phi_\ell(Z)] = \langle \Phi_k \otimes \Phi_\ell, \mathcal{C}_{ZZ}\rangle_{\mathcal{H}_Z \otimes \mathcal{H}_Z} = \langle \Phi_k, \mathcal{C}_{ZZ}\Phi_\ell\rangle_{\mathcal{H}_Z}. \end{aligned} \tag{43}$$

Thus, the components of the following matrix multiplication are given as,

$$\begin{aligned} \left[C_{YZ}C_{ZZ}^{-1}\right]_{k\ell} &= \sum_j \left[C_{YZ}\right]_{kj} \left[C_{ZZ}^{-1}\right]_{j\ell} \\ &= \sum_j \langle \Psi_k \otimes \Phi_j, \mathcal{C}_{YZ}\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_Z} \langle \Phi_j, \mathcal{C}_{ZZ}^{-1}\Phi_\ell\rangle_{\mathcal{H}_Z} \\ &= \left\langle \mathcal{C}_{YZ}, \Psi_k \otimes \left(\sum_j \langle \Phi_j, \mathcal{C}_{ZZ}^{-1}\Phi_\ell\rangle_{\mathcal{H}_Z} \Phi_j\right)\right\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_Z} \\ &= \left\langle \mathcal{C}_{YZ}, \Psi_k \otimes \mathcal{C}_{ZZ}^{-1}\Phi_\ell\right\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_Z} \\ &= \left\langle \mathcal{C}_{YZ}\mathcal{C}_{ZZ}^{-1}, \Psi_k \otimes \Phi_\ell\right\rangle_{\mathcal{H}_Y \otimes \mathcal{H}_Z} \\ &= \left\langle \mathcal{C}_{YZ}\mathcal{C}_{ZZ}^{-1}\Phi_\ell, \Psi_k\right\rangle_{\mathcal{H}_Y}. \end{aligned} \tag{44}$$

To clarify this derivation, the second equality used the definition in (43), the fourth line used the fact that $\mathcal{C}_{ZZ}^{-1}\Psi_\ell \in \mathcal{H}_Z$ can be expanded as in (42), and the rest of the lines used the standard tensor identity.

The theory of kernel mean embedding of conditional distributions (see [49, 48]) suggests that,

$$\mathbb{E}_{Y|\boldsymbol{z}}[\Psi_k(Y)] = \langle \Psi_k, \mathcal{C}_{YZ}\mathcal{C}_{ZZ}^{-1}\hat{K}(\boldsymbol{z},\cdot)\rangle_{\mathcal{H}_Y}. \tag{45}$$

Since $\hat{K}(\boldsymbol{z},\cdot) \in \mathcal{H}_Z$, we can employ the expansion in (42) and deduce,

$$\begin{aligned}
\mathbb{E}_{Y|\boldsymbol{z}}[\Psi_k(Y)] &= \langle \Psi_k, \mathcal{C}_{YZ}\mathcal{C}_{ZZ}^{-1}\sum_{j=1}^{\infty}\frac{\langle \hat{K}(\boldsymbol{z},\cdot),\varphi_j\rangle_{L^2(\mathcal{Z},\hat{q})}}{\sqrt{\xi_j}}\Phi_j\rangle_{\mathcal{H}_Y} \\
&= \sum_{j=1}^{\infty}\frac{\langle \hat{K}(\boldsymbol{z},\cdot),\varphi_j\rangle_{L^2(\mathcal{Z},\hat{q})}}{\sqrt{\xi_j}}\langle \Psi_k, \mathcal{C}_{YZ}\mathcal{C}_{ZZ}^{-1}\Phi_j\rangle_{\mathcal{H}_Y} \\
&= \sum_{j=1}^{\infty}\frac{1}{\sqrt{\xi_j}}\left[C_{YZ}C_{ZZ}^{-1}\right]_{kj}\int_{\mathcal{Z}}\hat{K}(\boldsymbol{z},\boldsymbol{z}')\varphi_j(\boldsymbol{z}')\hat{q}(\boldsymbol{z}')\,d\boldsymbol{z}' \\
&= \sum_{j=1}^{\infty}\left[C_{YZ}C_{ZZ}^{-1}\right]_{kj}\Phi_j(\boldsymbol{z}),
\end{aligned} \tag{46}$$

where we have used (44) to deduce the third equality above and used the fact that φ_j and ξ_j are eigenfunction and eigenvalue of the integral operator in (34). Define,

$$[\boldsymbol{C}_{YZ}]_{ks} = \mathbb{E}_{YZ}\left[\psi_k(Y)\otimes\varphi_s(Z)\right], \qquad [\boldsymbol{C}_{ZZ}]_{sl} = \mathbb{E}_{ZZ}\left[\varphi_s(Z)\otimes\varphi_l(Z)\right],$$

then from (43) and the definitions of the corresponding feature maps in (41),

$$[C_{YZ}]_{ks} = \sqrt{\lambda_k\xi_s}\,[\boldsymbol{C}_{YZ}]_{ks}, \qquad [C_{ZZ}]_{sl} = \sqrt{\xi_s\xi_l}\,[\boldsymbol{C}_{ZZ}]_{sl}, \qquad \left[C_{YZ}C_{ZZ}^{-1}\right]_{k\ell} = \frac{\sqrt{\lambda_k}}{\sqrt{\xi_l}}\left[\boldsymbol{C}_{YZ}\boldsymbol{C}_{ZZ}^{-1}\right]_{k\ell}.$$

Substituting the third equation above to (46) and using the definitions of the feature maps in (41), we obtain

$$\mathbb{E}_{Y|\boldsymbol{z}}[\psi_k(Y)] = \frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{\infty}\left[C_{YZ}C_{ZZ}^{-1}\right]_{kj}\Phi_j(\boldsymbol{z}) = \sum_{j=1}^{\infty}\left[\boldsymbol{C}_{YZ}\boldsymbol{C}_{ZZ}^{-1}\right]_{kj}\varphi_j(\boldsymbol{z}),$$

which is exactly the claim in (7).

B ACV of the multi-scale linear Gaussian model

The full model (18)-(19) can be rewritten as

$$\begin{aligned}
\dot{x} &= (a_{11}x + a_{12}y) + \sigma_x\xi_x, \\
\dot{y} &= \frac{1}{\epsilon}(a_{21}x + a_{22}y) + \frac{\sigma_y}{\sqrt{\epsilon}}\xi_y,
\end{aligned} \tag{47}$$

where ξ_x and ξ_y are independent standard Gaussian noises. Similarly, the closure model (25) can be rewritten as

$$\dot{x}_t = \left(a_{11}x_t + a_{12}\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}\right) + \sigma_x\xi_x, \tag{48}$$

where $\boldsymbol{x} := \boldsymbol{x}_{t-m:t} = [x_{t-m}, x_{t-m+1}, \ldots, x_t]^\top$ and Σ_{12} and Σ_{22} are defined in Eq. (26). To simplify the notation, we drop the time indices $t-m:t$. We also drop the "hat"-notation in x_t and $\boldsymbol{x}_t$ since we will use it to denote the Fourier coefficient in this section. In this Appendix, we prove that the autocovariance (ACV) function of the closure model (48) is approximately equal to that of the full model (47) for any value of ϵ.

The Fourier transform and inverse Fourier transform is defined as

$$\widehat{f}(\omega) = \int f(t)\,e^{-i\omega t}dt, \quad f(t) = \frac{1}{2\pi}\int \widehat{f}(\omega)\,e^{i\omega t}d\omega.$$

The Fourier transforms of variables x and y of the full model (47) can be obtained as

$$\widehat{x} = \frac{\left(i\omega - \frac{1}{\epsilon}a_{22}\right)\sigma_x\widehat{\xi}_x + a_{12}\frac{\sigma_y}{\sqrt{\epsilon}}\widehat{\xi}_y}{\left(i\omega - a_{11}\right)\left(i\omega - \frac{1}{\epsilon}a_{22}\right) - a_{12}\frac{1}{\epsilon}a_{21}}, \tag{49}$$

$$\widehat{y} = \frac{\left(i\omega - a_{11}\right)\frac{\sigma_y}{\sqrt{\epsilon}}\widehat{\xi}_y + \frac{1}{\epsilon}a_{21}\sigma_x\widehat{\xi}_x}{\left(i\omega - a_{11}\right)\left(i\omega - \frac{1}{\epsilon}a_{22}\right) - a_{12}\frac{1}{\epsilon}a_{21}}. \tag{50}$$

Then, for the full model (47), the resulting spectrum of x is

$$\left|\widehat{x}\left(\omega\right)\right|^2 = \frac{\left(\omega^2 + c_0^2\right)\sigma_x^2\left|\widehat{\xi}_x\right|^2 + d_0^2\sigma_y^2\left|\widehat{\xi}_y\right|^2}{\left(-\omega^2 + \omega_0^2\right)^2 + \gamma_0^2\omega^2},$$

where

$$\begin{aligned} c_0 &= \frac{a_{22}}{\epsilon}, \quad d_0 = \frac{a_{12}}{\sqrt{\epsilon}}, \quad \omega_0 = \sqrt{\frac{1}{\epsilon}\left(a_{11}a_{22} - a_{12}a_{21}\right)}, \\ \gamma_0 &= a_{11} + \frac{1}{\epsilon}a_{22}, \quad \left|\widehat{\xi}_x\right|^2 = 1, \quad \left|\widehat{\xi}_y\right|^2 = 1. \end{aligned}$$

Now we compute the Fourier transform of the closure model (48),

$$i\omega\widehat{X} = a_{11}\widehat{X} + a_{12}\Sigma_{12}\Sigma_{22}^{-1}\begin{bmatrix} 1 \\ e^{-i\omega\tau} \\ \vdots \\ e^{-i\omega m\tau} \end{bmatrix}\widehat{X} + \sigma_x\widehat{\xi}_x, \tag{51}$$

where $\widehat{X}$ is the Fourier transform of x_t in Eq. (48). We need to simplify the quantity $\Sigma_{12}\Sigma_{22}^{-1}\begin{bmatrix} 1 & e^{-i\omega\tau} & \cdots & e^{-i\omega m\tau} \end{bmatrix}^\top$ in Eq. (51). Let $S = \Sigma_{12}\Sigma_{22}^{-1}$ be the $1 \times (m+1)$ vector with components denoted by $S\left[n\right]$ for $n = 0, \ldots, m$. Then, we can write

$$\Sigma_{12}\Sigma_{22}^{-1}\begin{bmatrix} 1 \\ e^{-i\omega\tau} \\ \vdots \\ e^{-i\omega m\tau} \end{bmatrix} = S\begin{bmatrix} 1 \\ e^{-i\omega\tau} \\ \vdots \\ e^{-i\omega m\tau} \end{bmatrix} = \sum_{n=0}^{m} S\left[n\right]e^{-i\omega n\tau} := \widehat{S}_m\left(\omega\right), \tag{52}$$

which is nothing but the discrete Fourier transform of S. Notice that, for any $n = 0, \ldots, m$,

$$\sum_{k=0}^{m} S\left[k\right]\gamma_{xx,m}\left[n-k\right] = \sum_{k=0}^{m} S\left[k\right]\Sigma_{22}\left[k,n\right] = \Sigma_{12}\left[n\right] = \gamma_{xy,m}\left[n\right] \tag{53}$$

where the first equality is due to the fact that the process is stationary such that $\Sigma_{22}[k,n] = \gamma_{xx,m}[n-k]$, the second equality is due to $S\Sigma_{22} = \Sigma_{12}$, and the last equality is by the definition of the covariance function. By the discrete convolution theorem, we have

$$\widehat{S}_m\left(\omega\right)\widehat{\gamma}_{xx,m}\left(\omega\right) = \widehat{\gamma}_{xy,m}\left(\omega\right), \tag{54}$$

where $\widehat{\gamma}_{xx,m}$ and $\widehat{\gamma}_{xy,m}$ are the discrete Fourier transforms of $\gamma_{xx,m}$ and $\gamma_{xy,m}$, respectively. Substituting $\widehat{S}_m\left(\omega\right)$ in Eq. (54) into Eq. (52), we obtain

$$\Sigma_{12}\Sigma_{22}^{-1}\begin{bmatrix} 1 \\ e^{-i\omega\delta t} \\ \vdots \\ e^{-i\omega m\delta t} \end{bmatrix} = \frac{\widehat{\gamma}_{xy,m}\left(\omega\right)}{\widehat{\gamma}_{xx,m}\left(\omega\right)} \longrightarrow \frac{\widehat{\gamma}_{xy}\left(\omega\right)}{\widehat{\gamma}_{xy}\left(\omega\right)}, \qquad \text{as } m \to \infty, \tag{55}$$

where $\widehat{\gamma}_{xx}$ and $\widehat{\gamma}_{xy}$ denote the Fourier transform of the covariance functions γ_{xx} and γ_{xy}.

Substituting the limiting case of Eq. (55) into Eq. (51), we can simplify the Fourier transform of the closure model as follows,

$$i\omega\widehat{X} = a_{11}\widehat{X} + a_{12}\frac{\widehat{\gamma}_{xy}(\omega)}{\widehat{\gamma}_{xx}(\omega)}\widehat{X} + \sigma_x\widehat{\xi}_x. \tag{56}$$

Moreover, based on the Wiener-Khinchin theorem and the cross-correlation theorem, we can further simplify Eq. (56) as

$$i\omega\widehat{X} = a_{11}\widehat{X} + a_{12}\frac{\widehat{y}}{\widehat{x}}\widehat{X} + \sigma_x\widehat{\xi}_x. \tag{57}$$

Substituting Eqs. (49) and (50) into above Eq. (57), we obtain the Fourier transform of the relevant variable, $\widehat{X}$, of the closure model,

$$\widehat{X} = \frac{\left(i\omega - \frac{1}{\epsilon}a_{22}\right)\sigma_x\widehat{\xi}_x + a_{12}\frac{\sigma_y}{\sqrt{\epsilon}}\widehat{\xi}_y}{\left(i\omega - a_{11}\right)\left(i\omega - \frac{1}{\epsilon}a_{22}\right) - a_{12}\frac{1}{\epsilon}a_{21}}, \tag{58}$$

which is the same as the $\widehat{x}$ of the full model in Eq. (49). Therefore, the ACV of the closure model (48) is consistent with that of the full model (47) in the limit of $m \to \infty$. In the numerics, the error comes from the truncation of finite number of memory terms in Eq. (52).

References

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

[2] T. Berry and J. Harlim. Semiparametric modeling: Correcting low-dimensional model error in parametric models. Journal of Computational Physics, 308:305–321, 2016.

[3] T. Berry and J. Harlim. Correcting biased observation model error in data assimilation. Mon. Wea. Rev., 145(7):2833–2853, 2017.

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

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

[6] A. Christmann and I. Steinwart. Support vector machines. Springer, 2008.

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

[8] I. Fatkullin and E. Vanden-Eijnden. A computational strategy for multiscale systems with applications to lorenz 96 model. Journal of Computational Physics, 200(2):605–638, 2004.

[9] J. Frederiksen and T. O'Kane. Entropy, closures and subgrid modeling. Entropy, 10:635–683, 2008.

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

[11] G. A. Gottwald and J. Harlim. The role of additive and multiplicative noise in filtering complex dynamical systems. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 469(2155):20130096, 2013.

[12] 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.

[13] W. Grabowski. An improved framework for superparameterization. J. Atmos. Sci., 61:1940–1952, 2004.

[14] T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review, 129(3):550–560, 2001.

[15] J. Harlim. Data-Driven Computational Methods: Parameter and Operator Estimations. Cambridge University Press, UK, 2018.

[16] J. Harlim, S. Jiang, S. Liang, and H. Yang. Machine learning for prediction with missing dynamics. arXiv:1910.05861, 2019.

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

[18] J. Harlim, A. Mahdi, and A. Majda. An ensemble kalman filter for statistical estimation of physics constrained nonlinear regression models. Journal of Computational Physics, 257, Part A:782 – 812, 2014.

[19] S. W. Jiang and J. Harlim. Parameter estimation with data-driven nonparametric likelihood functions. Entropy, 21(6):559, 2019.

[20] A. Kerstein. A linear- eddy model of turbulent scalar transport and mixing. Combustion Science and Technology, 60(4-6):391–421, 1988.

[21] A. Kerstein. One-dimensional turbulence: model formulation and application to homogeneous turbulence, shear flows, and buoyant stratified flows. Journal of Fluid Mechanics, 392:277–334, 1999.

[22] R. Khasminskii. On averaging principle for Itô stochastic differential equations. Kybernetika, Chekhoslovakia (in Russian), 4(3):260–279, 1968.

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

[24] B. Khouider, A. St-Cyr, A. Majda, and J. Tribbia. The MJO and convectively coupled waves in a coarse-resolution GCM with a simple multicloud parameterization. Journal of the Atmospheric Sciences, 68:240–264, 2011.

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

[26] R. H. Kraichnan. The structure of isotropic turbulence at very high Reynolds numbers. Journal of Fluid Mechanics, 5:497–543, 1959.

[27] S. Kravtsov, D. Kondrashov, and M. Ghil. Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. Journal of Climate, 18(21):4404–4424, 2005.

[28] T. Kurtz. Semigroups of conditional shifts and approximations of Markov processes. Annals of Probability, 3:618–642, 1975.

[29] 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.

[30] H. Lei, N. A. Baker, and X. Li. Data-driven parameterization of the generalized langevin equation. Proceedings of the National Academy of Sciences, 113(50):14183–14188, 2016.

[31] E. Lorenz. Predictability: a problem partly solved. In Seminar on Predictability, 4-8 September 1995, volume 1, pages 1–18, Shinfield Park, Reading, 1995. ECMWF, ECMWF.

[32] 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.

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

[34] 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.

[35] A. Majda, R. V. Abramov, and M. J. Grote. Information theory and stochastics for multiscale nonlinear systems, volume 25. American Mathematical Soc., 2005.

[36] A. Majda and I. Grooms. New perspectives on superparameterization for geophysical turbulence. Journal of Computational Physics, 271:60–77, 2014.

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

[38] A. Majda, I. Timofeyev, and E. Vanden-Eijnden. Stochastic models for selected slow variables in large deterministic systems. Nonlinearity, 19(4):769, 2006.

[39] A. Majda and I. Tomofeyev. Statistical mechanics for truncations of the burgers-hopf equation: a model for intrinsic stochastic behavior with scaling. Milan Journal of Mathematics, 70(1):39–96, 2002.

[40] A. J. Majda and J. Harlim. Physics constrained nonlinear regression models for time series. Nonlinearity, 26(1):201, 2012.

[41] A. J. Majda and I. Timofeyev. Remarkable statistical behavior for truncated burgers–hopf dynamics. Proceedings of the National Academy of Sciences, 97(23):12413–12417, 2000.

[42] 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.

[43] 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.

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

[45] A. Nemtsov, A. Averbuch, and A. Schclar. Matrix compression using the nyström method. Intelligent Data Analysis, 20(5):997–1019, 2016.

[46] G. C. Papanicolaou et al. Some probabilistic problems and methods in singular perturbations. Rocky Mountain Journal of Mathematics, 6(4):653–674, 1976.

[47] G. Pavliotis and A. Stuart. Multiscale methods: averaging and homogenization. Springer Science & Business Media, 2008.

[48] 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.

[49] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In Proc. 26th Annual Int. Conf. on Machine Learning, pages 961–968. ACM, 2009.

[50] 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.

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

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

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

[54] R. Zwanzig. Nonlinear generalized Langevin equations. J. Stat. Phys., 9:215 – 220, 1973.