MULTISCALE MODEL. SIMUL.
Vol. 19, No. 2, pp. 1011–1040

KERNEL ANALOG FORECASTING: MULTISCALE TEST PROBLEMS*

DMITRY BUROV†, DIMITRIOS GIANNAKIS‡, KRITHIKA MANOHAR†, AND ANDREW STUART†

Abstract. Data-driven prediction is becoming increasingly widespread as the volume of data available grows and as algorithmic development matches this growth. The nature of the predictions made and the manner in which they should be interpreted depend crucially on the extent to which the variables chosen for prediction are Markovian or approximately Markovian. Multiscale systems provide a framework in which this issue can be analyzed. In this work kernel analog forecasting methods are studied from the perspective of data generated by multiscale dynamical systems. The problems chosen exhibit a variety of different Markovian closures, using both averaging and homogenization; furthermore, settings where scale separation is not present and the predicted variables are non-Markovian are also considered. The studies provide guidance for the interpretation of data-driven prediction methods when used in practice.

Key words. data-driven prediction, multiscale systems, kernel methods, analog forecasting, averaging, homogenization

AMS subject classifications. 37M10, 34E13, 58J65

DOI. 10.1137/20M1338289

1. Introduction.

Data-driven prediction holds great promise in many areas of science and engineering. Growth in the volume of data available in numerous application areas has been matched by advances in computational methodologies which are designed to utilize these data for prediction. However, fundamental questions arise in this field relating to the choice of variables on which to base prediction and whether the system is Markovian in the chosen variables. While delay embedding can be used to enhance the choice of variables in which Markovian structure is present, prediction is often undertaken using variables in which there is no Markovian closure or in which this closure is only approximate. The objective of the paper is to use multiscale systems to provide a framework in which the fundamental issue of the role of Markovianity in data-driven prediction can be studied. We work within the setting of kernel analog forecasting (KAF), a methodology that has seen success in a number of application domains and that is backed by a mature theory. Subsection 1.1 provides an overview of relevant literature in data-driven prediction for dynamical systems and the multiscale setting in which we work. We outline our con-

*Received by the editors May 18, 2020; accepted for publication (in revised form) February 2, 2021; published electronically June 16, 2021.
https://doi.org/10.1137/20M1338289
Funding: The first and fourth authors were supported by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program, by Earthrise Alliance, Mountain Philanthropies, the Paul G. Allen Family Foundation, and the National Science Foundation (NSF) (award AGS1835860). The second author is supported by NSF (awards 1842538 and DMS-1854383) and ONR (awards N00014-16-1-2649 and N00014-19-1-242). The third author is supported by the NSF Mathematical Sciences Postdoctoral Research Fellowship (award 1803663). The fourth author is also supported by NSF (award DMS-1818977) and by the Office of Naval Research (award N00014-17-1-2079).

†Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 USA (dburov@caltech.edu, kmanohar@caltech.edu, astuart@caltech.edu).

‡Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA (dimitris@cims.nyu.edu).

tributions to the understanding of data-driven prediction within multiscale systems in subsection 1.2.

1.1. Background and literature review. In 1969, Lorenz originally introduced the idea of *analog forecasting* for prediction of dynamical systems using historical data [30]. Given initial data, the method locates its closest analog among the historical points and reports the historical value of the corresponding observable, shifted by the desired lead time. By construction, analog forecasting avoids model error, but the resulting forecast is not continuous with respect to initial data. It is therefore non-physical, and this fact, when combined with the paucity of data available at the time the method was proposed, limited the value of the methodology in practice. An exponential growth in data volume has precipitated the development of improved methodologies which build on Lorenz's original idea, leading to algorithms which are backed by large data theories and which depend smoothly on initial condition. This has been achieved through the use of kernel-based methods which result in data-driven prediction based on weighting all the historical data according to their similarity to the initial data; this leads to algorithms which enforce continuity of the forecast with respect to initial data [55] and, building on the theory of reproducing kernel Hilbert spaces (RKHSs) [3], to algorithms which can be theoretically justified in the large data limit [17, 1].

KAF draws on several fundamental ideas rooted in kernel methods for machine learning. First, the choice of kernel function is guided by the need for dimension reduction of big data and the specific learning task at hand. For clustering, similarity kernels [43] are used to construct graphs over the data and clusters determined by its graph Laplacian eigenvectors [4]. This graph Laplacian construction is generalized in [10] to characterize diffusion operators on the manifold on which the data lie via their eigenfunctions, known as diffusion maps, and further generalized in [7] to a class of variable-bandwidth kernels that control for variations in the density of the sampling distribution. Under different choices of kernel and normalization, the resulting eigenmaps can describe slow coordinates in dynamical systems [36] and can also be used as a basis to approximate evolution operators in SDEs [6]. Algorithmic development has been aided by advances in the theory for pointwise [45] and spectral [50, 47] convergence of these coordinates in the large data limit. Second, Markov operators constructed from kernels map into RKHSs in which kernel evaluation corresponds to function evaluation and allow evaluating these eigencoordinates on out-of-sample data [11] in a procedure known as Nyström extension. This has found use in semi-supervised classification using support vector machines [44] to extend labels to new data, spline interpolation [51], and forecasting [55].

In addition to exploiting ideas from kernel methods for machine learning, KAF may exploit additional structure from time-ordered data arising from dynamical systems. For example, delay embedding is frequently used to identify Markovian structure. In diffusion forecasting [6], diffusion maps are time-shifted to approximate the action of a shift operator on observables in SDEs. Alternatively, time-shifted diffusion maps can be used to approximate the reduction coordinates of this shift operator directly [17], or temporal structure can be directly embedded into specialized cone kernels for analog forecasting [55]. The latter takes the Nyström extension perspective of KAF, while in fact KAF evaluates a conditional expectation of this shift operator, conditioned on the observations [1]. The aforementioned shift operator, known as the *Koopman* operator, acts by composing observables with the dynamical flow map and is a linear operator on these function spaces. Hence, the data-driven approximation of the Koopman operator is an exciting area of research.

Bernhard Koopman introduced the linear operator that carries his name in the 1930s as part of his study of ergodic and Hamiltonian dynamics [27]. In data-driven identification of coherent structures, spectral decompositions of the Koopman operator [34] and the related transfer operator [15] play a central role, driven by the fact that in such a basis, forecasting of nonlinear dynamics amounts to scalar multiplication by eigenvalues. Algorithms used in practice compute finite-dimensional regression onto precomputed libraries consisting of functions of the time-lagged snapshots [42, 41, 26]. However, convergence guarantees are limited, requiring stringent assumptions on the libraries and spectrum [2]. In particular, mixed spectra resulting from chaotic/mixing systems pose a challenge for numerical methods. Recent data-driven methods which leverage infinite-dimensional feature spaces provided by kernels [13], as well as kernel constructions in spectral space [28], are able to tackle the continuous part of the spectrum of the Koopman operator. For forecasting purposes, pointwise evaluation of the Koopman operator acting on observables is the natural setting, rather than spectral approximation, and is the perspective we take.

Multiscale analysis provides a setting in which to understand the role of rapidly varying (in space or time) system components on the slowly varying variables used for predictive models [53]. In this paper we will work in the framework of averaging and homogenization for PDEs and SDEs, as developed in [5]. Chapters 9–11 of the book [38] contain a pedagogical exposition of the subject that is adapted to the chaotic ODE setting that is the focus of this paper. However, the rigorous extension of the theory of averaging and homogenization to ODEs, rather than SDEs, is nontrivial and less well developed. Early work in this direction was contained in [37]. However, it was not until the fundamental work of Melbourne and coworkers that a theoretical approach with verifiable conditions was developed [32, 33, 24, 23].

We will use the example developed in [33], which exploits the (proven) chaotic properties of the Lorenz 63 model [29, 48], to provide an example of a chaotic ODE which homogenizes to give an SDE. And we will use the multiscale Lorenz 96 model [31] to provide an example of an ODE to which the averaging principle may be applied to effect dimension reduction, as pioneered and exploited in [16]; we note, however, that the mixing properties required to *prove* the averaging principle for the Lorenz 96 model have not been established, even though numerical evidence strongly suggests that it applies in certain parameter regimes. The work of Jiang and Harlim [22] studies data-informed model-driven prediction in partially observed ODEs, using ideas from kernel-based approximation. In the paper [21] the idea is generalized to discrete time dynamical systems, and neural networks and LSTM modeling is used in place of kernel methods. In both the papers [22, 21] multiscale systems are used to test their methods in certain regimes.

Data-driven analog forecasting, kernel methods, and Koopman methodologies have each individually found widespread use in real-world forecasting and coherent pattern extraction applications. Analog forecasting, albeit without kernels, has been used to predict weather patterns [8, 14, 49] yet is known to have limitations predicting chaotic behavior. Khodkar, Hassenzadeh, and Antoulas recently developed a Koopman-based framework using delay embedded observables to predict chaotic dynamics [25]. Nonlinear Laplacian spectral analysis, which applied kernel and delay embeddings akin to Koopman observables, successfully recovers coherent oscillatory phenomena such as the seasonal/diurnal cycles and the El Niño Southern oscillation [46] through kernel eigenfunctions [12]. Implemented with such kernels, KAF would naturally capture the fundamental oscillatory components of the predictand variable and thus interpolate between an initial-value ("weather") forecast at short

times and a climatology forecast at asymptotic times when the mixing component of the predictand has decayed; see, e.g., [52]. Koopman operator approximation has also been widely adopted to study high-dimensional complex, even turbulent fluid flows [18]; see [35] for a study of these applications.

In this paper we use KAF as developed in the papers [1, 17, 55]. Technical details on the implementation, including pseudocode and further relevant references, are collected in Appendix A.

1.2. Our contribution. We use multiscale methodology to introduce four classes of ODE test problems which exhibit Markovian dynamics after elimination of fast variables; stochastic, chaotic, quasiperiodic, and periodic behavior may be obtained in the slow variable, depending on the setting considered. Using these test problems, our four main contributions to the understanding of KAF are as follows:

1. We apply KAF techniques to data generated by each of these four classes of multiscale test problems and use the behavior of the averaged or homogenized slow system to interpret the resulting predictions. In particular, KAF methods converge, in the large data limit, to a conditional expectation defined via the Koopman operator of the multiscale systems; we use this as the basis for our interpretation. Moreover, we demonstrate the use of KAF to estimate the variance of the predictor itself. In each of the four cases the 2σ-interval captures the real trajectory, even when the KAF predictor does not track the trajectory itself. This can be viewed as a separate application of the KAF methodology which will be useful in cases when forecasting of high probability bounding sets suffices even when the trajectory itself is hard to predict. In all cases we also study problems in which the scale separation is not present, but KAF prediction of mean and variance is attempted on the basis of data from only a subset of the variables.
2. The KAF method is based on data-driven approximation of the eigenvalue problem arising from a kernel integral operator. In the setting in which the multiscale ODE homogenizes to produce an SDE corresponding to a bistable gradient system with additive noise, a limiting analytical expression is available for the eigenfunctions; we demonstrate that these limiting eigenfunctions are well approximated by the data-driven method. This comparison gives insight into the empirical methods used to tune free parameters within KAF.
3. In the setting in which the multiscale ODE averages to produce an ODE of lower dimension, we use alternative data-driven ODE closures as a benchmark against which to compare the purely data-driven KAF methods. This gives insight into the relative merits of purely data-driven prediction and prediction which combines model-based knowledge with data.
4. We use the insights from these carefully constructed numerical experiments to make recommendations about deployment and parameter tuning of KAF methods to real data.

The paper is organized as follows. In section 2, we outline the data-driven construction of the prediction function using KAF methodology. We explain the sense in which the construction converges to a conditional expectation defined via the Koopman operator associated to a measure-preserving dynamical system assumed to underlie the data. We also describe two kinds of canonical multiscale systems which give rise to homogenization and averaging effects and which we use to provide interpretation of this conditional expectation. Section 3 introduces a test problem in the form of a double-well gradient flow driven by chaotic Lorenz 63 dynamics which

homogenizes to give an SDE in the scale-separated regime; numerical results applied to prediction of the slow variable exhibit contributions 1 and 2. In section 4, we introduce the multiscale Lorenz 96 system, which averages to give an ODE in the slow variables; three different parametric regimes give rise to periodic, quasiperiodic, and chaotic responses in the slow variable. The behavior of KAF-based prediction in these three regimes is studied to illustrate contribution 1, and a slow-variable closure model, built using Gaussian process (GP) regression (GPR), is compared with the KAF to illustrate contribution 3. In section 5, we provide an overview the insights obtained by studying KAF methods through the lens of multiscale systems; and we then make concrete recommendations about interpreting the output of KAF techniques when applied to naturally occurring data, contribution 4.

2. Methodology. In subsection 2.1 we provide an overview of the two key ideas which interact to underpin the studies in this paper: KAF and multiscale methods, tailoring the exposition to the use of the latter as a tool to understand the former. We then give more details on KAF. The two primary components of the KAF methodology are (i) viewing forecasting as evaluation of a conditional expectation of the Koopman operator applied to the desired observable and (ii) approximation of this conditional expectation in a data-driven fashion. Subsections 2.2 and 2.3 describe (i) and (ii) respectively, while subsection A.2 is devoted to a key practical component of the data-driven approximation, namely, construction of the kernel, and subsection A.3 to the data-driven choice of integer ℓ, the number of (approximate) eigenfunctions used in the data-driven forecast.

2.1. Overview of methodology. The problem setting for prediction is as follows. We assume that we are given N time-ordered data samples

$$\{x_n\}_{n=0}^{N-1} \subset \mathcal{X},$$

where $x : \mathbb{R} \to \mathcal{X}$ is a continuous time process, $x_n = x(n\Delta t)$, and Δt is the sampling rate. We assume that the continuous time process x in $\mathcal{X}$ is derived from Markovian dynamics for a coupled pair (x, y) evolving in the larger state space $\mathcal{X} \times \mathcal{Y}$. Assume that the desired prediction *lead time* τ is an integer multiple of the sampling interval, that is, $\tau = q\Delta t$. Included with the data are values of the associated *prediction observable* advanced by τ time units

$$\{f_{n+q}\}_{n=0}^{N-1} \subset \mathbb{R},$$

defined by the Markovian dynamics via an unknown map $F\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$; thus, $f_n = F(x_n, y_n)$. The goal of KAF is to predict $F(x(\tau), y(\tau))$ given only partial information, $x(0) = x$, and the N data samples x_n. We view the data-driven predictor as a map $Z_\tau\colon \mathcal{X} \to \mathbb{R}$ which takes initial condition x as input.

Given initial data x and lead time τ, the KAF predictor averages over the τ-shifted observable values provided in the training data and weighted by a kernel $p : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ constructed from the data; the resulting algorithm has the following form:

$$\begin{aligned} Z_\tau(x) &= \frac{1}{N} \sum_{n=0}^{N-1} p(x, x_n) f_{n+q}, \\ p(x, x_n) &= \sum_{j=0}^{\ell(\tau)-1} \frac{\psi_j(x)\phi_j(x_n)}{\lambda_j^{1/2}}. \end{aligned} \tag{2.1}$$

The weighting kernel $p(x, x_n)$ determines how much weight to attach to a time series initialized at point x_n, according to its proximity to x, the desired initial point. The features ϕ_j are computed from an eigenvalue problem associated with a data-driven approximation of a kernel integral operator, constructed from x_n; in the large data limit this provides an orthonormal basis for the entire space. The function ψ_j is an out-of-sample Nyström extension of ϕ_j, orthonormalized with respect to an underlying RKHS structure. The method may be seen as a smoothed version of Lorenz's original proposal for data-driven prediction—analog forecasting [30]. Analog forecasting, by contrast, predicts the trajectory in the training data obtained by finding the training data point nearest to the given initial condition in some metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$Z_\tau(x) = f_{n^\star+q}, \qquad n^\star = \underset{n=0,\dots,N-1}{\arg\min}\, d(x, x_n). \tag{2.2}$$

It can be seen that Lorenz's method will result in predictions discontinuous with respect to initial data, especially for systems that exhibit sensitive dependence on initial conditions. In particular, KAF addresses the issue of continuity of the prediction with respect to the initial condition, and it does so in a framework which is provably statistically consistent in the large data limit [1, 17, 55]. Further details of the methodology are given in the next two subsections and the attendant information in Appendix A.

An important challenge addressed by this methodology is that, since the y component of the system is not observed, the sequences $\{x_n\}$ and $\{f_{n+q}\}$ are non-Markovian. As a consequence, the standard idea of constructing a Markov chain from the data is not natural. The KAF method evaluates a conditional expectation of the forecast conditioned, using the observed data $\{x_n\}$, explicitly incorporating information loss resulting from unobserved y; it is hence a natural approach to the problem at hand. Multiscale systems provide a natural setting for the study of KAF methods and in particular the issue of prediction of non-Markovian or approximately Markovian systems. In this paper we will consider the variable x as the slow component of a Markovian system for pair (x, y) in which y evolves as a fast variable. We consider averaging and homogenization settings in which the dynamics for x is approximately Markovian, and the conditional expectation arising in the KAF method may be understood explicitly. This will enable us to obtain a deeper understanding of how KAF works and help users of the methodology interpret it. We now outline the averaging and homogenization settings that we will use. Details of the theory underlying them may be found in [38].

We will study multiscale systems which exhibit averaging, in the form

$$\text{(A)} \qquad \begin{cases} \dot{x} = v_0(x) + By, \\ \dot{y} = \frac{1}{\varepsilon} g(x, y), \end{cases}$$

where $B \colon \mathcal{Y} \to \mathcal{X}$ is linear. The average of By under the invariant measure of the y dynamics, with x frozen, provides a closed approximate ODE dynamics for x when ε is small. If we denote the x-parameterized invariant measure for the y dynamics with x frozen by $\nu^x(dy)$, then for $\varepsilon \ll 1$ we obtain $x \approx X$, where

$$\text{(A0)} \qquad \begin{aligned} \dot{X} &= v_0(X) + v(X), \\ v(\zeta) &= \int_{\mathcal{Y}} By\, \nu^\zeta(dy). \end{aligned}$$

To guarantee uniqueness of solutions in (A0), it suffices that the conditional measure has enough continuity, as a function of x, so that $v(\zeta)$ is Lipschitz.

When the variable By averages to zero, a different scaling is required to elicit the effect of the fast variable on the slow one. To this end we also consider multiscale systems which exhibit homogenization in the form

$$\text{(H)} \qquad \begin{cases} \dot{x} = v_0(x) + \frac{1}{\varepsilon} By, \\ \dot{y} = \frac{1}{\varepsilon^2} g(y). \end{cases}$$

Here we assume that

$$\int_{\mathcal{Y}} By \, \nu(dy) = 0,$$

where ν is the invariant measure of the y dynamics. The approximate dynamics for x, when ε is small, is then governed by an SDE in this setting; the work of Melbourne and coworkers provides the sharpest results in this context [32, 33, 24, 23]. If this is the case, then, invoking the homogenization principle, $x \approx X$, where X is governed by an SDE of the form

$$\text{(H0)} \qquad \dot{X} = v_0(X) + \sqrt{2\sigma}\dot{W},$$

where W denotes the Wiener process and σ is a uniquely determined positive constant that can be computed numerically from the mixing properties of the y process.

2.2. Koopman formulation of prediction. We let $\Omega = \mathcal{X} \times \mathcal{Y}$ and assume that $\Phi^t \colon \Omega \to \Omega$ is an ergodic dynamical system with invariant probability measure μ; we assume $t \in \mathbb{R}^+$, but the extension to discrete time is straightforward. Define the continuous observation map $\Pi \colon \Omega \to \mathcal{X}$ and the prediction observable $F \colon \Omega \to \mathbb{R}$; we assume that F is square-integrable with respect to the invariant measure:

$$F \in L^2_\mu(\Omega; \mathbb{R}) \coloneqq \left\{ F \colon \Omega \to \mathbb{R} \;\middle|\; \int_\Omega |F(\omega)|^2 \mu(d\omega) < \infty \right\}.$$

We define the Koopman operator $U^t \colon L^2_\mu(\Omega; \mathbb{R}) \to L^2_\mu(\Omega; \mathbb{R})$ by $U^t g = g \circ \Phi^t$. We seek, in a sense to be made precise, the function $Z_\tau \colon \mathcal{X} \to \mathbb{R}$ such that $Z_\tau \circ \Pi$ is the best approximation to a perfect prediction of $F \circ \Phi^\tau$. We formalize this by introducing the Hilbert subspace $V \subseteq L^2_\mu(\Omega; \mathbb{R})$ given by

$$V \coloneqq \left\{ g \in L^2_\mu(\Omega; \mathbb{R}) \;\middle|\; g = g' \circ \Pi,\ g' \colon \mathcal{X} \to \mathbb{R} \right\}.$$

This Hilbert space captures the notion of making predictions based only on information in $\mathcal{X}$. Note that the perfect forecast would satisfy $Z_\tau \circ \Pi = U^\tau F$ but that such a forecast will not be possible in general because $\mathcal{X}$ is a proper subset of Ω and information needed for perfect prediction of F will be missing. Among all elements of V, the minimal prediction error in $L^2_\mu(\Omega; \mathbb{R})$ is attained by the conditional expectation

$$\mathbb{E}(U^\tau F \mid \Pi) = \operatorname*{arg\,min}_{g' \in V} \|g' - U^\tau F\|_{L^2_\mu} = \operatorname*{proj}_V U^\tau F. \tag{2.3}$$

This formulation of prediction encapsulates the inherent loss of information incurred through observing only a set of functionals of an ergodic dynamical system and the effect of this loss of information on prediction. In subsequent sections of this paper we will assume that $F = F' \circ \Pi$ for some $F' : \mathcal{X} \to \mathbb{R}$ because it is often natural to try to predict only functionals of the slow variables. Note, however, that the methodology is not restricted to such F, and in this subsection, the next subsection, and Appendix A, we describe the more general setting for completeness.

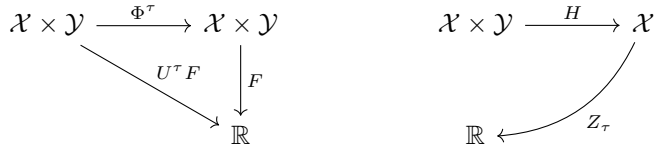

FIG. 2.1. *Comparison of exact Koopman picture and data-driven approximation. The approximate mapping $Z_\tau(\cdot)$ is constructed from the data streams $\{H\Phi^t(x_0, y_0)\}_{0\leqslant t\leqslant T}$ and $\{F(\Phi^t(x_0, y_0))\}_{0\leqslant t\leqslant T}$.*

2.3. Data-driven approximation. The formulation in the preceding subsection encapsulates the inherent loss of predictive power incurred through observing only a set of functionals of an ergodic dynamical system. This is formalized by seeking the best approximation of the Koopman evolution from within a Hilbert subspace capturing the notion of depending only on specified functionals on the state space of the dynamical system. We now demonstrate how data may be used to further approximate this best approximation and to do so in a manner which is refineable as more data are acquired. The approach is summarized in Figure 2.1.

For observation time $\Delta t > 0$ we define

$$\omega_n = \Phi^{t_n}(\omega_0), \quad t_n = n\Delta t,$$
$$x_n = \Pi(\omega_n), \quad f_n = F(\omega_n).$$

We assume that we are given time-ordered pairs

$$\{(x_0, f_q), (x_1, f_{1+q}), \ldots, (x_{N-1}, f_{N-1+q})\}, \tag{2.4}$$

and the objective is to construct, from these data, a function $Z_\tau \colon \mathcal{X} \to \mathbb{R}$ which predicts F at lead time τ so that $Z_\tau \circ \Pi \approx g^\star$, where $g^\star$ solves the minimization problem in (2.3). Furthermore, we wish to carry this out in a manner which ensures that, in an appropriate topology, $Z_\tau \circ \Pi \to g^\star$ as $N \to \infty$.

To this end we introduce a hypothesis space $\mathcal{H}_{\ell,N}$, of dimension ℓ and depending on the N-dependent data set (2.4), and seek to solve the minimization problem

$$Z_\tau = \operatorname*{arg\,min}_{g\in\mathcal{H}_{\ell,N}} \|g \circ \Pi - U^\tau F\|_{L^2_\mu}. \tag{2.5}$$

The choice of the hypothesis space is constrained by the need to be able to solve the minimization problem (2.5) explicitly, using only the data (2.4), and by the requirement that $Z_\tau \circ \Pi$ recovers $g^\star$ in the large data limit $N \to \infty$. Moreover, in order to be practically useful forecast functions, elements of $\mathcal{H}_{\ell,N}$ should allow pointwise evaluation at any $x \in \mathcal{X}$, which is not defined in arbitrary subspaces of $L^2_\mu(\Omega; \mathbb{R})$.

With these considerations in mind, we introduce a kernel function $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and RKHS $\mathcal{K}$ with the properties

$$f(x) = \langle k_x, f\rangle_{\mathcal{K}}, \quad k_x = k(x, \cdot), \quad \langle k_x, k_{x'}\rangle_{\mathcal{K}} = k(x, x').$$

We then define $\mathcal{H}_{\ell,N}$ as an ℓ-dimensional subspace of $\mathcal{K}$, to be described below. We also note that the kernel k is constructed from a data stream of length N, but we suppress the explicit dependence of k on N in the notation. In Appendix A we

discuss our data driven construction of k and choice of ℓ. For now we proceed on the assumption that we have a kernel and hence a RKHS as well as a method for choosing ℓ.

Let $\mu_N = \frac{1}{N}\sum_{n=0}^{N-1}\delta_{\omega_n}$ be the sampling measure underlying the training data (2.4), and define

$$L^2_{\mu_N}(\Omega;\mathbb{R}) \coloneqq \left\{ F\colon \Omega \to \mathbb{R} \;\middle|\; \int_\Omega |F(\omega)|^2 \mu_N(d\omega) = \frac{1}{N}\sum |F(\omega_n)|^2 < \infty \right\}.$$

Associated with μ_N is an integral operator $G\colon L^2_{\mu_N}(\Omega;\mathbb{R}) \to L^2_{\mu_N}(\Omega;\mathbb{R})$, which we identify with a symmetric, positive-semidefinite, $N \times N$ kernel matrix $G \in \mathbb{R}^{N\times N}$ with entries

$$G_{mn} = k(x_m, x_n), \quad x_n = \Pi(\omega_n), \quad 0 \le m, n \le N-1. \tag{2.6}$$

The eigenvectors of this matrix lead to an orthogonal basis $\{\phi_j\}_{j=0}^{N-1}$ of $\mathbb{R}^N$ such that

$$G\phi_j = \lambda_j \phi_j, \quad \lambda_0 \ge \lambda_1 \ge \cdots \ge \lambda_{N-1}, \quad \|\phi_j\|_2 = \sqrt{N}.$$

We may also identify each element $\phi_j \in \mathbb{R}^N$ with element $\phi_j \circ \Pi \in L^2_{\mu_N}(\Omega;\mathbb{R})$ via the definition $\phi_j(x_n)$ as the nth entry of the vector $\phi_j \in \mathbb{R}^N$, so that $\{\phi_j \circ \Pi\}_{j=0}^{N-1}$ is an orthonormal basis of $L^2_{\mu_N}(\Omega;\mathbb{R})$. Using the same symbols for elements of $\mathbb{R}^N$ and $L^2_{\mu_N}(\Omega;\mathbb{R})$, as well as for linear transformations on those spaces, is a useful economy of notation. Then the following functions $\psi_j\colon \mathcal{X} \to \mathbb{R}$ form an orthonormal set in $\mathcal{K}$:

$$\psi_j = \frac{1}{N\lambda_j^{1/2}} \sum_{n=0}^{N-1} k(\cdot, x_n)\phi_j(x_n), \quad \lambda_j > 0. \tag{2.7}$$

This is a form of Nyström extension [11].

As hypothesis space we take

$$\mathcal{H}_{\ell,N} = \operatorname{span}\{\psi_0, \ldots, \psi_{\ell-1}\} \subseteq \mathcal{K}, \tag{2.8}$$

noting that the basis functions themselves depend on the data set and hence on N. We may now solve the optimization problem (2.5), and an explicit computation yields, for $\tau = q\Delta t$,

$$Z_\tau(x) = \sum_{j=0}^{\ell-1} \frac{c_j(\tau)}{\lambda_j^{1/2}}\psi_j(x), \quad c_j(\tau) = \langle \phi_j \circ \Pi, U^\tau F\rangle_{L^2(\mu_N)} = \frac{1}{N}\sum_{n=0}^{N-1}\phi_j(x_n) f_{n+q}. \tag{2.9}$$

Note that this construction of the predictor Z_τ is entirely data-driven: The basis functions ψ_j and the eigenvalues λ_j are found from the eigenvalues and eigenvectors of the data-defined kernel matrix, and the coefficients c_j are computed as sums over the data set. Furthermore, Theorem 14 in [1] proves that $Z_\tau \circ X$ converges to $g^\star$, the solution of the minimization problem (2.3), as $N \to \infty$, followed by $\ell \to \infty$, in an L^2 sense with respect to the invariant measure μ on Ω.

More generally, any function of the observable can be predicted in this data-driven manner, which provides a convenient framework for uncertainty quantification. The conditional variance between forecast and ground truth can also be computed in the hypothesis space as in (2.9) using the coefficients

$$\hat{c}_j(\tau) = \langle \phi_j \circ \Pi, (U^\tau F - Z_\tau)^2\rangle_{L^2(\mu_N)} = \frac{1}{N}\sum_{n=0}^{N-1}\phi_j(x_n)(f_{n+q} - Z_\tau(x_n))^2. \tag{2.10}$$

For detail on the data-driven kernel construction, the data-driven choice of ℓ, and the conditional variance estimator, see subsections A.2, A.3, and A.4, respectively.

3. Homogenization: Lorenz 63–driven system. This section is devoted to the setting in which a chaotic ODE of form (H) is approximated by an SDE of form (H0). The goal is to make predictions of the x variable, using data concerning only the x variable from (H); the role of (H0) is simply to help us interpret those predictions. This setting presents unique challenges for forecasting, as one cannot expect the outcome of any method to predict a sample path of a stochastic process without knowledge of the driving noise. This fact has direct bearing on prediction in (H) using x-data alone since (H0) demonstrates that the time series of the $x-$data is approximately that of an SDE; without knowledge of the noise, which is governed by the unobserved y variable, prediction of the trajectory of x is not possible. In subsection 3.2 we examine instead the long-term statistics predicted by KAF from data generated by (H)—the conditional expectation and variance of the stochastic process—and compare them with estimates computed from (H0) using Monte Carlo simulation of the SDE. This illustrates our main contribution 1 from the list in subsection 1.2. Then, in subsection 3.3, exploiting the fact that the limiting process is one-dimensional, we find explicit expressions for the kernel eigenfunctions in the limit problem (H0) and compare these with the eigenfunctions obtained from data-driven techniques applied to (H), our main contribution 2 from subsection 1.2. Subsection 3.4 is concerned with non-Markovian prediction, in which there is no scale separation between observed and unobserved variables. We start, however, in subsection 3.1, introducing the concrete model around which our experiments are organized.

3.1. The model. The first test problem arises from driving a double-well gradient flow with a chaotic signal obtained from the Lorenz 63 model [19]:

$$\begin{aligned} \dot{x} &= x - x^3 + \frac{0.059}{\varepsilon} y_2, \\ \dot{y}_1 &= \frac{10}{\varepsilon^2}(y_2 - y_1), \\ \dot{y}_2 &= \frac{1}{\varepsilon^2}(28y_1 - y_2 - y_1 y_3), \\ \dot{y}_3 &= \frac{1}{\varepsilon^2}\left(y_1 y_2 - \frac{8}{3} y_3\right). \end{aligned} \tag{3.1}$$

This is of form (H). In [33] it is proved that as $\varepsilon \to 0$, this system converges weakly in $C([0,T];\mathbb{R})$, when projected onto the x variable alone, to the solution of the SDE

$$\begin{aligned} \dot{X} &= -\Xi'(X) + \sqrt{2\sigma}\dot{W}, \\ \Xi(x) &= \frac{1}{4}(1 - X^2)^2. \end{aligned} \tag{3.2}$$

Thus, this white-noise–driven gradient system is of form (H0). The value of the constant σ is identified in [19]. For the current work the key point to appreciate is that for small ε, the variable x in (3.1) exhibits (approximately) Markovian behavior, but this behaviour is stochastic. The SDE is ergodic and has invariant probability density function

$$\rho_\infty(x) \propto \exp\left(-\frac{1}{\sigma}\Xi(x)\right) \tag{3.3}$$

with respect to Lebesgue measure.

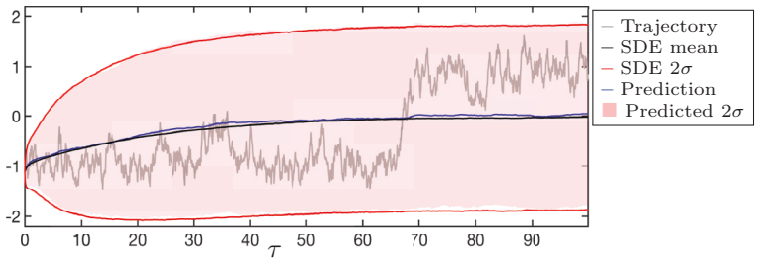

FIG. 3.1. Long-term forecast convergence. *Gray is the trajectory of the SDE started from* $x = -1.10$*, blue is the KAF predictor* $Z_\tau(-1.10)$ *with pink shades giving two-standard-deviation confidence bands computed from the conditional variance, black is the Monte Carlo approximation of the conditional mean using the SDE, and red is the Monte Carlo approximation of the conditional variances using the SDE. KAF computations of mean and variance agree with the true conditional expectation and mean computed from* 10000 *Monte Carlo realizations of the SDE.*

3.2. Conditional expectation and variance. We aim to predict the x variable from historical data of a long trajectory of x alone. Thus, the observation and observable maps are $\Pi(\omega) = x, F(\omega) = x$. We will also estimate second moments, enabling us to compute conditional variance, for which $F(\omega) = x^2$. Observation data are generated by using an implicit time-stepping scheme with time step 0.01 in the slow variable and built-in MATLAB solvers to integrate the fast variables with $\varepsilon^2 = 0.001$. Note that the coefficient of y_2 driving the x dynamics in (3.1), 0.059, differs from that of Givon et al [19], and approximates the invariant measure of the SDE (3.2) under a different value of σ than the value predicted in [19]. Our predicted value of σ is 0.1091. Source data for the slow variable x are gathered for $N = 40000$ points sampled at the macroscopic time interval $\Delta t = 0.05$. Then $\hat{N} = 7500$ out-of-sample points from a new trajectory $\{\hat{x}_n\}$ are gathered at the same resolution, providing the ground truth for assessing forecast error. The natural error metric is the root mean square error (RMSE), the L^2 norm of $Z_\tau - U^\tau F$. To account for differences in scale, we normalize the RMSE by the standard deviation of the trajectory:

$$\mathsf{RMSE}(\tau) = \left(\frac{1}{\hat{N}-q}\sum_{n=0}^{\hat{N}-q-1}|Z_\tau(\hat{x}_n) - \hat{x}_{n+q}|^2\right)^{1/2}\left(\frac{1}{\hat{N}-q}\sum_{n=0}^{\hat{N}-q-1}|\hat{x}_{n+q} - \bar{x}|^2\right)^{-1/2}.$$

Figure 3.1 depicts the behavior of the KAF forecast $Z_\tau(x)$ as a function of lead time τ for fixed x. This forecast exhibits two interesting properties which can be understood through the small-ε homogenization limit. The first relates to the fact that the trajectory itself is not well predicted; the second explains what is well predicted.

First, the predictor tracks the conditional mean initialized at $x = -1.10$ and not the trajectory itself. This is predicted by the theory since what is predicted is the long-term conditional expectation $\mathbb{E}[U^\tau F|\Pi]$. Indeed, this latter quantity necessarily converges for large τ to a constant, under mixing assumptions on the (x, y) system, while individual trajectories in x exhibit stochastic dynamics, approximately that of X. This explains the growth in RMSE seen in Figure 3.2. Second, exploiting the fact that we expect the system (3.1) to behave like (3.2), when projected onto the x coordinate, we can provide an objective evaluation of the KAF forecasts by running Monte Carlo simulations of the SDE for X; to do this we compute the sample mean and variance over 10000 sample paths initialized at the same initial point $x = -1.10$. Figure 3.1 compares the KAF forecast mean and variance with that predicted by

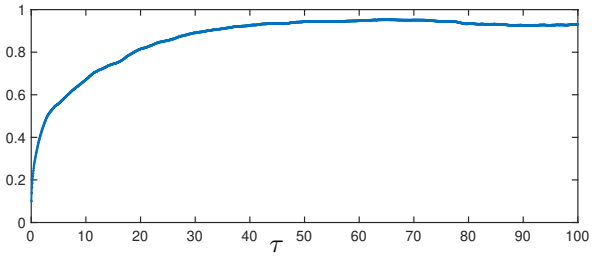

FIG. 3.2. Long-term RMSE *for the forecast in Figure* 3.1 *saturates at* $\tau = 50$*, as the forecast converges to the long-term mean at* $X = 0$.

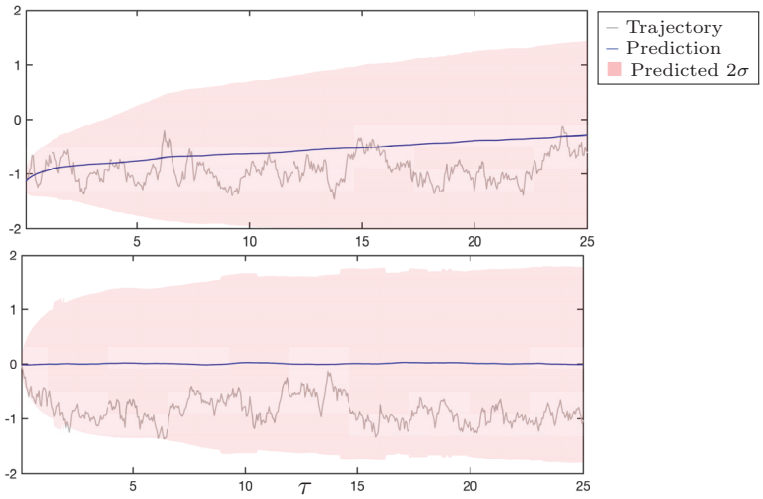

FIG. 3.3. Comparison of uncertainty for different initial data. *Blue is the predictor (conditional mean), gray is the trajectory, and pink gives two-standard-deviation bounds computed from the conditional variance. The forecast uncertainty when initialized at* $x = 0$ *(right) displays more rapid growth in uncertainty over short lead times than that when* $x = -1.10$ *(left). This sensitivity to initial conditions is a desirable feature of KAF in light of the fact that both plots share the same training set.*

Monte Carlo mean and variance for the SDE, and they are seen to agree very well over the entire window of computation.

Finally, in Figure 3.3 we use the possibility of varying the initial condition in the KAF to demonstrate that the variability encapsulated in the conditional variance is able to pick up different sensitivities, depending on initial condition. The panel on the left shows a trajectory initialized at $x = -1.10$, and the panel on the left shows a trajectory initialized at $x = 0$. As can be expected from the limiting SDE, the uncertainty when initialized at $x = 0$ is greater, and this is manifest in the conditional variance.

Examination of the KAF technique in this homogenization setting thus clearly reveals the inability of the method to predict trajectories but shows that it can accurately approximate statistics of trajectories, averaged over the unobserved component of the system. Furthermore, analysis of the SDE provides a means of characterizing the geometry of the underlying hypothesis space, as seen in the next subsection.

3.3. Insights into the hypothesis space. By studying the large data and small kernel bandwidth limit of the matrix G defined in (2.6), we get insights into the structure of the hypothesis space (2.8). The theory in [17], building on the pa-

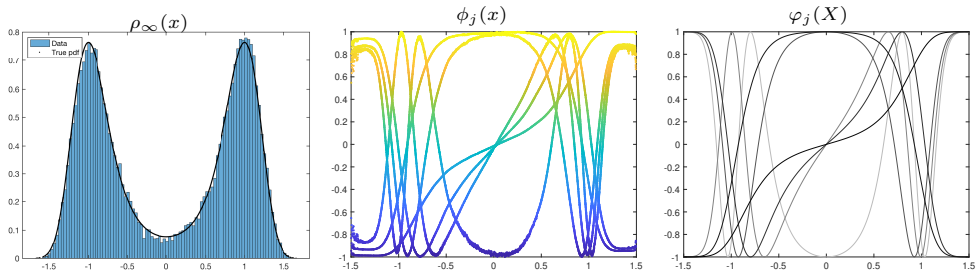

FIG. 3.4. **Comparison of invariant densities and eigenfunctions.** *Left: invariant density (3.3) and histogram of x from (3.1). Middle and right: six empirically computed eigenfunctions using x from (3.1) and using theory associated with (3.2), respectively.*

pose [10, 7], demonstrates that for the choice of kernel described in Appendix A, the vectors of G are approximated in the large data and small bandwidth limit by eigenfunctions of the Laplace–Beltrami operator Δ_h on $\mathbb{R}$, with $h = \rho^{-2}$. We now demonstrate that this is indeed the case.

The action of the Laplace–Beltrami operator on $f \in C^\infty(M)$ is given by

$$\Delta_h f = -\mathrm{div}_\mu \mathrm{grad}_h f = -\mathrm{div}_\mu \left(\frac{1}{\rho} \frac{df}{dX} \right) = -\frac{1}{\rho} \frac{d}{dX} \left(\frac{1}{\rho} \frac{df}{dX} \right),$$

where div_μ and grad_h are the divergence and gradient operators associated with μ and h, respectively. Using the above, we solve the eigenvalue problem $-\Delta_h \varphi = \lambda^2 \varphi$ directly. We make the substitution $dY = \rho dX$, mapping $X \in \mathbb{R}$ into $Y \in [0,1]$; we note that Y has interpretation as the cumulative distribution function coordinate of ρ. In terms of Y we have

$$\Delta_h f = \frac{1}{\rho} \frac{d}{dX} \left(\frac{1}{\rho} \frac{df}{dX} \right) = \frac{d^2 f}{dY^2}.$$

Noting that the natural boundary conditions for the Laplace–Beltrami operator are of no-flux type, it follows that, when viewed as functions of Y, the eigenfunctions of Δ_h satisfy a boundary value problem of the form

$$-\varphi''(Y) = \lambda^2 \varphi(Y)$$
$$\varphi'(0) = \varphi'(1) = 0.$$

The solutions are the well-known harmonics $\cos(k\pi Y)$, and corresponding eigenvalues $\lambda_k = k\pi, k \in \mathbb{N}$. Changing back to variable x we obtain

$$\varphi_k(X) = \cos\left(k \int_{-\infty}^{X} \rho(a) da \right). \tag{3.4}$$

We now verify that, for large data sets and small bandwidth, the eigenfunctions of G are indeed close to those associated with the Laplace–Beltrami operator Δ_h. This is demonstrated in the middle and right panels of Figure 3.4. The middle panel shows the first six eigenfunctions of G, computed from data derived from the x variable in

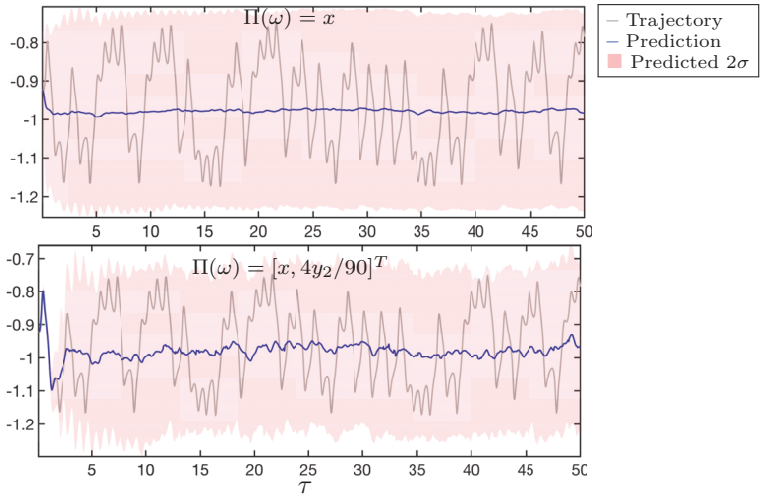

FIG. 3.5. Prediction in non-Markovian regime, $\varepsilon = 1$, *requires observations augmented by the forcing term to achieve short-term predictability.*

are indeed close to those associated with the Laplace–Beltrami operator Δ_h. This is demonstrated in the middle and right panels of Figure 3.4. The middle panel shows the first six eigenfunctions of G, computed from data derived from the x variable in (3.1); the right panel shows the eigenfunctions of Δ_h for the diffusion process (3.2) in variable X, which governs the limiting behavior of x in the scale-separated case. The agreement is excellent, demonstrating that the heuristics underlying parameter choices within the kernel-based methodology (see Appendix A) work well in this setting.

3.4. Non-Markovian regime. In the preceding subsections we studied predictors for x, based only on time-series data in the x coordinate, for the equation (3.1). We studied the scale-separated regime where $\varepsilon \ll 1$ and x is approximately Markovian—it is approximately governed by an SDE. Here we study the behavior of identical predictors when $\varepsilon = 1$; the system (3.1) then no longer exhibits homogenization, and x is no longer Markovian in view of the lack of scale separation. As a result, the prediction of x, shown in the top of Figure 3.5, is poor even at very short times, and the two-standard-deviation confidence bands reflect this rapid initial error growth and then remain large throughout the time window. Z_τ converges rapidly to the conditional mean. To render the problem Markovian, we may include more data and in particular the forcing term. To this end we take $\Pi(\omega) = [x, 4y_2/90]^T$; see the bottom of Figure 3.5. The prediction of x with these augmented observations yields very accurate predictions over a lead time of $\tau = 3$, considerably larger than in the preceding case, where $\Pi(\omega) = x$. After $\tau = 3$, however, pathwise predictability again fails due to the sensitive dependence of solutions with respect to the forcing function. Once again Z_τ converges to the conditional mean. However, the uncertainty quantification provided by the two-standard-deviation error bars consistently captures the true trajectory, even in this non-Markovian regime.

This non–scale-separated pair of examples illustrates that prediction of non-Markovian variables is inherently harder than Markovian variables but that sensitive dependence of trajectories with respect to perturbations also limits predictability, even in the Markovian setting. This second point will be prominent, too, in the next section.

4. Averaging: Lorenz 96 multiscale system. This section is devoted to the setting in which a chaotic ODE of form (A) is approximated by an ODE of form (A0). The goal is to make predictions of the x variable using data concerning the x variable alone from (A). The role of (A0) is primarily to help us interpret those predictions; however, it also serves to motivate a different prediction methodology which mixes model-based and data-driven methodologies against which we will compare KAF.

The averaging setting presents interesting opportunities to understand forecasting. In particular, by tuning a parameter in (A), which is also present in (A0), we are able to create settings in which the variable to be predicted is, in turn, approximately periodic, quasiperiodic, and chaotic. These different settings give rise to different types of forecasts, and we demonstrate this. In subsection 4.2 we examine the long-term statistics predicted by KAF from data generated by (A)—the conditional expectation and variance of one component of the slow variable—and compare them with estimates computed from (A0) using Monte Carlo simulation. This illustrates our main contribution 1 from the list in subsection 1.2. Then, in subsection 4.3, we compare the purely data-driven method of prediction with a hybrid data-model predictor. This hybrid is computed by using GPR to compute an approximate closure $v_{GP}(x) \approx v(x)$ in the notation of (A0); for more details on such approximate closures in the context of the specific model that we study in the following four subsections, see [16] and the references therein. The work in subsection 4.3 illustrates our main contribution 3 from subsection 1.2. Subsection 4.4 is concerned with non-Markovian prediction, in which there is no scale separation between observed and unobserved variables. We start, however, in subsection 4.1, introducing the concrete model around which our experiments are organized.

4.1. The model. In this section we focus our attention on another chaotic dynamical system, colloquially known as Lorenz 96 multiscale [31], which we will simply abbreviate to L-96. Following the notation established in [16], the L-96 equations model K slow variables $\{x_k\}_{k=1}^K$ coupled to JK fast variables $\{y_{j,k}\}_{j,k=1,1}^{J,K}$ with evolution given as follows:

$$\begin{aligned} \dot{x}_k &= -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F_x + \frac{h_x}{J}\sum_{j=1}^{J} y_{j,k}, \\ \dot{y}_{j,k} &= \frac{1}{\varepsilon}\Big(-y_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - y_{j,k} + h_y x_k\Big), \\ x_{k+K} &= x_k, \quad y_{j,k+K} = y_{j,k}, \quad y_{j+J,k} = y_{j,k+1}. \end{aligned} \tag{4.1}$$

This is of the form (A). On the assumption that the y variables, with x frozen, are ergodic, the averaging principle shows the existence of a function $C\colon \mathbb{R}^K \to \mathbb{R}^K$ such that, for small ε, the x variables are approximated by $X = (X_1, \ldots, X_k)$ solving

$$\dot{X}_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F_x + h_x C_k(X), \quad k \in \{1, \ldots, K\}, \tag{4.2}$$

with the periodic boundary conditions $X_{k+K} = X_k$ and $C_k\colon \mathbb{R}^K \to \mathbb{R}$ denoting the kth component of vector-valued function C. This system is of form (A0). Since system (4.1) is index-shift-invariant, it is clear that the closure C_k, if it exists, satisfies $C_{k+1}(X) = C_k(\pi X)$, where π shifts the vector indices by adding one unit, invoking periodicity at the end points. Furthermore, when J is large, empirical evidence [54, 16] suggests that there is a function $c\colon \mathbb{R} \to \mathbb{R}$ such that the approximation $C_k(X) = c(X_k)$ is a good one. For the current work the key point to appreciate is

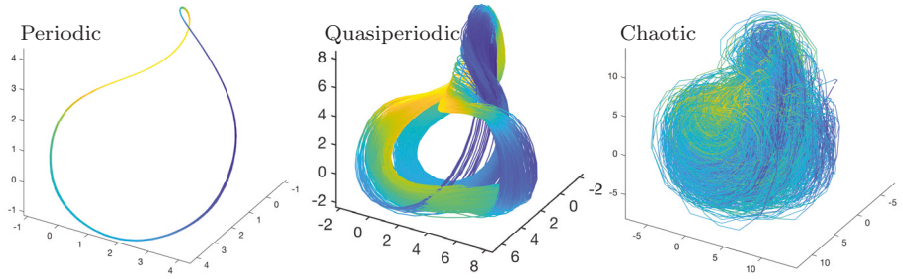

FIG. 4.1. Lorenz 96 regimes of increasing complexity (left to right). *Phase portraits show* (x_1, x_2, x_3) *coordinates shaded by* x_4. *The parameter* F_x *takes values* 5.0, 6.9, *and* 10.0, *respectively, from left to right, and all other parameters are as in* (4.3).

that for small ε, the variables x in (4.1) exhibit (approximately) Markovian behavior, and this behavior is deterministic and governed by X. However, by tuning F_x, different responses arise in the deterministic variable. In the following we fix parameters $\varepsilon^{-1}, K, J, h_x, h_y$ throughout all our experiments as follows:

$$\varepsilon^{-1} = 128,\ K = 9,\ J = 8,\ h_x = -0.8,\ h_y = 1.0. \tag{4.3}$$

We then choose F_x to distinguish three cases as follows:

periodic	quasiperiodic	chaotic
$F_x = 5.0,$	$F_x = 6.9,$	$F_x = 10.0.$

Figure 4.1 demonstrates the three responses within system (4.1) resulting from these parameter choices.

4.2. Conditional expectation and variance. We aim to predict the x_1 variable from historical data of a long trajectory of x alone. Thus, the observation and observable maps are $\Pi(\omega) = x$, $F(\omega) = x_1$. We will also use $F(\omega) = x_1^2$ when estimating conditional variance. By tuning the scalar parameter F_x (not to be confused with function F) as outlined in the preceding subsection, we can obtain periodic, quasiperiodic, and chaotic responses in the averaged variable X. It is intuitive that the ability of the KAF to track the true trajectory of the slow variables decreases with increasing complexity; in other words, predictions in the periodic case should be the most accurate, while those in the chaotic case present a significant challenge. In the experiments that follow, the size and sampling interval of the source (training) data remain fixed at $(40000, 0.05)$, and the out-of-sample (test) data set is fixed at $\hat{N} = 7000$.

The space of observables $\mathcal{X}$ in the current example is the space of all slow variables. Since under the small-ε limit an ODE closure of the slow dynamics is obtained, the variable x behaves (approximately) like a deterministic Markov process, and the expectation in (2.3) disappears; the predictor is expected to track the actual trajectory $x_1(t)$. To see this another way, note that by simply knowing the initial values of the x variables (recall that $\mathcal{X}$ is precisely all x variables) and the closure $C(X)$ in (4.2), we are able to predict x_1 (or indeed any x_k) exactly, given the initial conditions for all x variables.

However, this picture is greatly affected by the sensitivity of the system to initial conditions and sampling errors due to high dimensionality of the attractor. We now

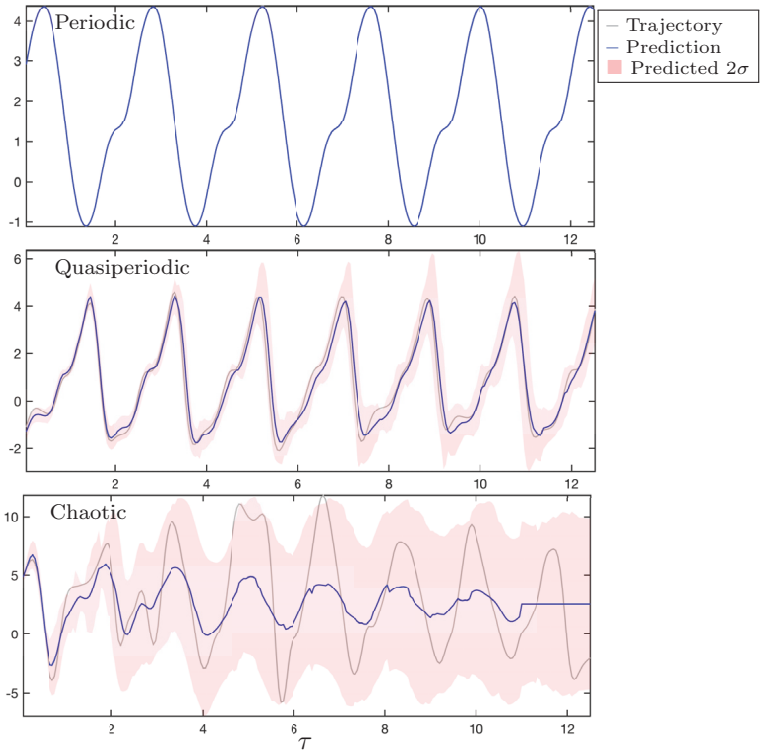

FIG. 4.2. Predictability: Periodic, quasiperiodic, and chaotic regimes. *Prediction $Z_\tau(x)$ of observable $F(\omega) = x_1$ across three different regimes. In each figure, gray is the true trajectory, blue is the predictor using KAF, and pink gives two-standard-deviation confidence bands, computed using the conditional variance. The parameter F_x takes values* $5.0, 6.9$, *and* 10.0, *respectively, from top to bottom, and all other parameters are as in* (4.3). *In the first, periodic response regime, the trajectory is predicted almost perfectly, and this accuracy is reflected in the narrow confidence bands. In the second, quasiperiodic response regime, the trajectory is predicted very well but with growing error reflected accurately in the slowly growing confidence bands. In the third, chaotic response regime, the predictive capability is lost due to sensitivity to initial conditions, and this is reflected in the rapidly growing confidence bands and in the convergence of the predictor to a constant for large τ.*

describe how these predictions work in practice in the three regimes shown in Figure 4.1. We display our results in Figure 4.2, where x_1 and standard deviation bands are predicted and compared with the true signal starting from the same point. The long-term predictability in each regime is constrained by the complexity of the underlying Markovian, deterministic, slow dynamics. In the periodic regime, since chaos is absent in the slow variables, a perfect predictor is obtained via the partially observed dynamics; one interpretation of why this occurs is because the eigenfunctions of the Koopman operator lie in a finite span of the diffusion coordinate observables [2]. Observe that Z_τ remains in phase, and the forecast variance is negligible for long lead times up to the length of the entire out-of-sample trajectory ($\tau = 350$). The quasiperiodic trajectory is tracked imperfectly, but with significant accuracy over the same range of times; errors are visible mainly around the extrema of x_1 as suggested by the phase portrait; and the conditional variance reflects the significant accuracy present. Prediction in the fully chaotic regime only tracks the trajectory, however, until a lead time of approximately 1 time unit, exhibiting behavior at long lead times which is somewhat similar to that seen in the previous, homogenization section in

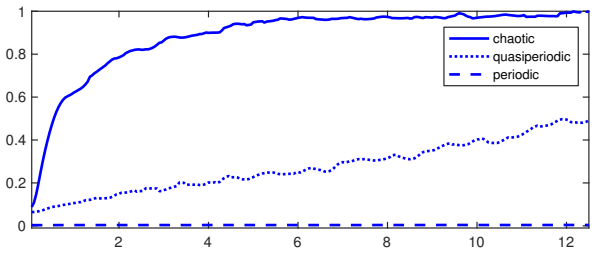

FIG. 4.3. RMSE. *This figure depicts the RMSE of the predictor $Z_\tau(x)$ for* (4.1)*, for different F_x, as a function of τ. The parameter F_x takes values* 5.0, 7.1, *and* 10.0*, respectively, from smaller to larger error, corresponding to periodic, quasiperiodic, and chaotic response; all other parameters are as in* (4.3).

which the predicted variable behaved as if drawn from a Markov stochastic process. In particular, the long-term predictor in the chaotic regime converges to a constant by construction, assuming mixing, and this is consistent with the inherent unpredictability of chaotic dynamics. It is notable that the size of the conditional variance (and the resulting confidence bands) is a useful guideline as to the pathwise accuracy of the data-driven predictor. The observations about the predictability of the system by KAF methods are also manifest in Figure 4.3, which shows the RMSE in each of the periodic, quasiperiodic, and chaotic regimes.

We mention that in the quasiperiodic case, the presence of multiple attractors (or multiple lobes of the same attractor) and resulting intermittent switching between these attractors leads to a loss of predictability that is significant on timescales much longer than those shown here. For the figure shown here we have ensured that training points and out-of-sample points are gathered from the same (part of the) attractor to maintain accuracy. We train using two different trajectories to gather ample training data.

Recall that at each lead time τ along the horizontal axis, there is a potentially different number of eigenfunctions $\ell(\tau)$ used in the data-driven method. See Appendix A.3 for details on the choice of ℓ. In the chaotic regime the optimal $\ell(\tau)$ tends to 1 for large times (see subsubsection A.3.2 for an explanation), while ℓ fluctuates around 50 in the quasiperiodic regime; we obtain $\ell \approx 9$ for all τ in the periodic regime.

4.3. Comparison of data-driven and model-data–driven prediction. The previous subsection concerned purely data-driven prediction of variable x from (A), using only data in the form of a time series for x. In this subsection we provide a comparison with a different forecasting technique based on a combination of model and data-driven prediction, using data in the form of a time series for (x, By). Knowledge of By enables the use of GPR [40] (or kriging) to approximate $v(\cdot)$ by $v_{GP}(\cdot)$ in (A0). Our approach is motivated by the paper [16], which looked at finding such closures for the L-96 model in form (4.1). When applied to (4.1), the methodology leads to an approximate closure for the slow variable X, which takes the form

$$\dot{X}_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F_x + h_x c_{GP}(X_k), \quad k \in \{1, \dots, K\} \tag{4.4}$$

subject to periodic boundary conditions $X_{k+K} = X_k$. This should be compared with (4.2), which arises from application of the averaging principle; note that, in addition, we have invoked the hypothesis that $C_k(X)$ can be well approximated by function of $c(X_k)$, as discussed directly after (4.2), and we will determine an approximation c_{GP} for c by GPR.

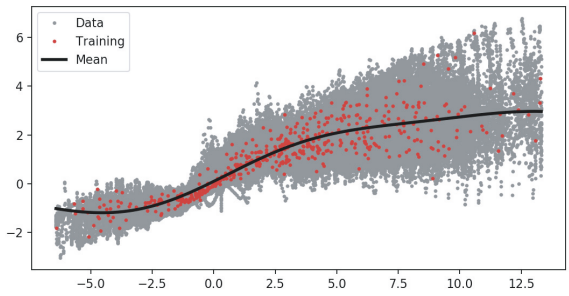

FIG. 4.4. Mean of Gaussian process regression as a closure. *Function c_{GP} and data used to determine it from data generated by* (4.1) *with parameters as in* (4.3) *and* $F_x = 10.0$.

Explicit details of the procedure we use to build a GP closure are described in Appendix B; here we observe that for training we use tuples $\{x_k(t_n), (By)_k(t_n)\}_{n=1}^N$, over all $k = 1, \ldots, 9$. See Figure 4.4 to see the data used (red random subsamples, without replacement, of the total grey data set) and an approximate GP closure c_{GP} determined from those data.

Once we have the closed model appearing in (4.4), we may use it to predict the variable x appearing in (4.1), and we may compare that prediction with the one made by KAF. Figure 4.5 shows the result of doing so. It shows that the KAF approach is superior in the periodic and quasiperiodic settings but that for predictions of the trajectory itself, the model-data–based predictor (4.4) is superior to KAF in the chaotic case. Note that the model-data–based predictor has access to more data than does the KAF and requires model knowledge; the KAF is entirely data-driven.

We now dig a little deeper into the comparison. We do this in a systematic way in the periodic, quasiperiodic, and chaotic regimes. In each of these three cases we show four RMSE error curves, labeled as follows: (a) the standard KAF based on x data alone; (b) an enhanced KAF using (x, By) data, the same data used to train the ODE (4.4); (c) a prediction using the ODE (4.4); and (d) a KAF prediction trained on X data alone, generated by the ODE (4.4). Figure 4.6 shows that KAF (a) is the ideal predictor in the periodic regime and is near ideal in the quasiperiodic regime; on the other hand, the ODE (4.4) predictor (c) is ideal for short-term predictability in the chaotic case. Augmenting observations with By within KAF, as in (b), gives errors similar to those arising from (a) when observing x alone; thus, knowledge of By provides little extra information. In the chaotic case, the RMSEs of KAF trained on x, (a), and on X, (d) are very close, confirming that the ODE (4.4) for X captures the invariant measure of the approximately Markovian variables x as intended.

We emphasize the difference between averaging and homogenization here: In the averaging case, observing the fast variables adds nothing to our prediction because there is a closed system determined only by the slow variables (see Figure 4.6, graphs (a) and (b) in all three plots). By contrast, in the homogenization case, observing the fast variables improves short-term predictions because it provides further information about the driving stochastic process entering the homogenized limit (see Figure 3.5).

4.4. Non-Markovian regime. In the preceding subsections we studied predictors for x, based only on time-series data in the x coordinate, for (4.1). We studied the scale-separated regime where $\varepsilon \ll 1$ and x is approximately Markovian and deterministic—it is approximately governed by an ODE. Here we study the behavior

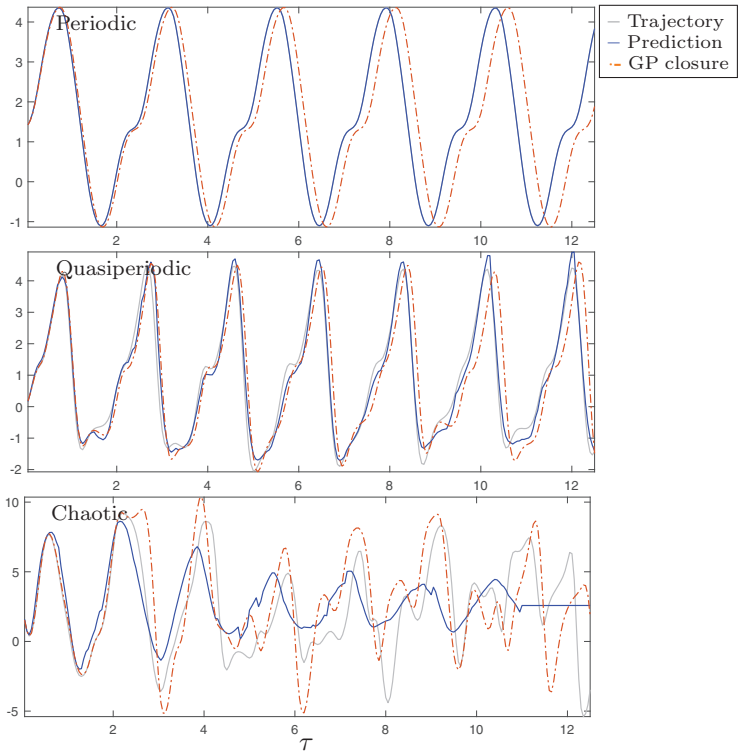

FIG. 4.5. Comparison of data-driven and model-data–driven prediction. *The true trajectory is shown in gray, the KAF data-driven prediction is in blue, and the model-data–driven predictions based on* (4.4) *are in dotted red; the periodic, quasiperiodic, and chaotic regimes are considered in turn.*

of identical predictors when $\varepsilon = 1$; the system (4.1) then no longer exhibits averaging, and x is no longer Markovian because there is no scale separation between x and y. This experiment is conducted with $F_x = 10$. Because of the lack of Markovian behavior, we expect rapid loss of predictability in time when $\Pi(\omega) = x, F(\omega) = x_1$. The resulting conditional mean and variance, shown in Figure 4.7, confirms this intuition. Indeed, the conditional mean is out of phase with the truth at lead time $\tau = 1$, and this is also reflected in the large growth of the conditional variance. Furthermore, the conditional mean tapers to a constant at $\tau = 6$, twice as quickly as it does in the $\varepsilon \ll 1$ setting, in which this tapering occurs at $\tau \approx 11$ (Figures 4.2 and 4.5).

4.5. Comparison with Lorenz's method. We illustrate the advantage of KAF over Lorenz's original method of analog forecasting (2.2), which can produce predictions that are discontinuous with respect to initial condition. In particular, this occurs when data are partially observed from a larger state space and different states map to identical partial observations. To study this, we observe a single coordinate x_1 of the periodic regime ($F_x = 5, \varepsilon^{-1} = 128$) so that the observed data are highly non-Markovian. We select initial conditions that are $\mathcal{O}(10^{-3})$ apart but are separated in time by integer multiples of the period. Figure 4.8 plots the resulting predictions from Lorenz's method and KAF. Although Lorenz's method is accurate for one initial condition (right), it gives a diverging prediction for a nearly identical point (left). By contrast, KAF is continuous with respect to initial condition and displays theo-

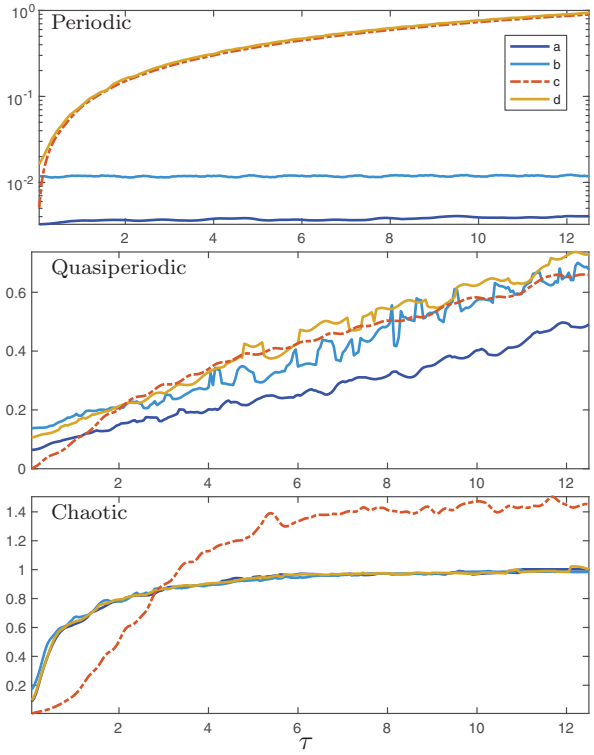

FIG. 4.6. RMSE comparison for the four cases (a)–(d) described in the text in the periodic, quasiperiodic, and chaotic regimes. *In the periodic regime, KAF* (a) *is an ideal predictor with negligible growth in error (note the logarithmic scale). In the quasiperiodic response regime, the growth in* RMSE *with KAF* (a), (b) *is significantly slower than that of the GP-based ODE prediction* (c). *In the chaotic response regime, the GP-based ODE prediction* (c) *is more accurate in the near term, yet KAF error stabilizes as the prediction converges to the conditional mean.*

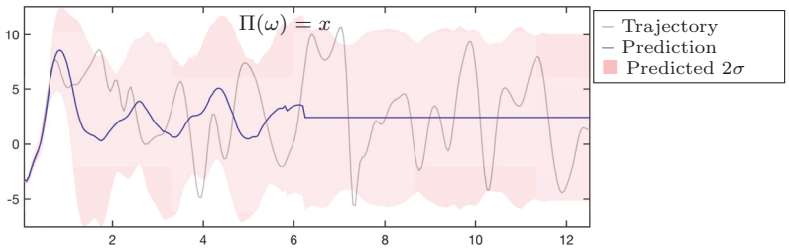

FIG. 4.7. Prediction in the non-Markovian regime, $F_x = 10$, $\varepsilon = 1$, *results in much shorter accurate trajectory predictability, followed by rapid convergence of the conditional mean to a constant. Note, however, that the uncertainty prediction bands contain the true trajectory for all time.*

retically optimal predictive skill (in an RMSE sense) for even highly non-Markovian observation data. This experiment illuminates a key feature of KAF: that it gives consistent predictions that are continuous with respect to initial conditions. Note also that KAF uncertainty predictions of a periodic observable are also periodic and vanish at every half period when predictions intersect the ground truth.

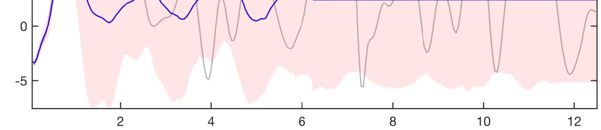

FIG. 4.7. **Prediction in non-Markovian regime,** $F_x = 10,\ \varepsilon = 1$**,** *results in much shorter accurate trajectory predictability, followed by rapid convergence of the conditional mean to a constant. Note, however, that the uncertainty prediction bands contain the true trajectory for all time.*

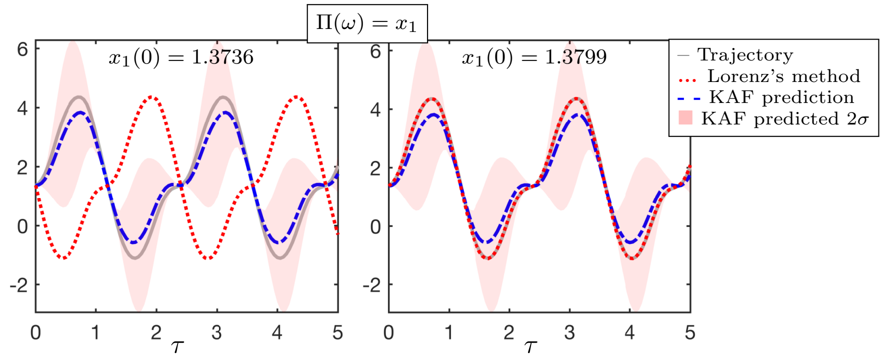

FIG. 4.8. **Lorenz' method (2.2) sensitivity to initial conditions,** *in the partially-observed periodic* $F_x = 5$, $\varepsilon^{-1} = 128$ *regime, results in diverging predicted trajectories for nearby initial conditions, separated only by 0.006. By contrast, KAF, which is continuous with respect to initial condition, shows moderate predictive skill and makes nearly identical predictions for nearby initial conditions.*

4.4. Non-Markovian regime. In the preceding subsections we studied predictors for x, based only on time-series data in the x coordinate, for the equation (4.1). We studied the scale-separated regime $\varepsilon \ll 1$ and dynamics approximately Markovian and deterministic or stochastic. In this section we study the behavior of identical predictors when $\varepsilon = 1$; the system (4.1) then no longer exhibits averaging and x is not Markovian because there is no scale-separation between x and y. This experiment is conducted with $F_x = 10$. Because of the lack of Markovian behaviour we expect rapid loss of predictability in time. The resulting conditional mean and variance are shown in Figure 4.7, confirming this intuition. Indeed the conditional mean is not plotted for the true trajectory beyond $\tau = 1$, and this is also reflected in the large KAF variance. Furthermore, the conditional mean KAF reaches a constant at $\tau = 6$, twice as quickly as it does in the $\varepsilon \ll 1$ setting in which the dynamics are chaotic (Figure 4.4).

4.5. Comparison with Lorenz' method. We illustrate the advantage of KAF over Lorenz' original method of analog forecasting (2.2), which can produce predic-

5. Conclusions.

1. We have studied KAF for data-driven prediction, for the equation (4.1).
   - We treated multiscale systems in which a subset of the variables (*slow variables*) evolve in an approximately Markovian fashion.
   - We study KAF performance for a range of separation between the slow variables, described by approximately stochastic, chaotic, quasiperiodic and periodic behaviour in time, when $\Pi(\omega) = x, F(\omega) = x_1$.
   - In the stochastic case we use the homogenized equations for the slow variables to obtain explicit formula for the eigenfunctions of the operator underlying KAF and use this to validate the performance of the KAF method.
   - In the chaotic, quasiperiodic and periodic cases (Figures 4.2–4.5) we use the averaged equations for the slow variables to obtain a GPR-based approximate closure model and use hybrid data-model predictions from this closure in order to evaluate the KAF method.
2. What we illustrate about use of the KAF:
   - When the variable being predicted is (approximately) a component of a stochastic Markov process, prediction of individual trajectories is not possible, while the mean and variance, averaged over possible realizations of the stochastic behaviour, can be accurately predicted by KAF.
   - When the variable being predicted is (approximately) a component of a deterministic but chaotic Markov process, prediction of individual trajectories is also not possible, except over short time horizons.
   - In both the chaotic and stochastic cases, while prediction of individual trajectories is not to be expected, simply bounding the future trajectory may be useful in applications—to this end, we show that in all cases two standard deviation bands around the predicted conditional mean always reliably capture the truth.
   - In both the stochastic and chaotic settings a signature of the lack of predictability is the convergence of the predictor to a constant, for large τ, accompanied by the data-driven choice of parameter ℓ converging to one.

- When the variable being predicted is (approximately) a component of a deterministic quasiperiodic or periodic process, the prediction of individual trajectories over long time horizons is possible; in this case parameter ℓ stays away from one for significantly large τ.
- In all cases the predicted standard deviations around the predictor provide a reliable indicator of the time scale on which the predictor is accurate trajectorywise and on longer timescales provide a reliable indicator of the scale of the errors incurred, trajectorywise.
- The choice of kernel and the interpretation of eigenfunctions as harmonics over the observed submanifold are well adapted to capturing periodic or quasiperiodic structure, while the design of eigenfunctions to include the constant eigenfunction permits convergence to the conditional mean for chaotic, mixing dynamics.
- The choice of ℓ indicates the number of harmonics contained in the predicted variable and is often approximately constant in τ in periodic and quasiperiodic settings, while it converges to one in the chaotic or stochastic mixing cases.
- When the variable being observed does not evolve in an approximately Markovian fashion, the KAF conditional mean cannot track x for even short times, as observed for $\varepsilon = 1$ for both the Lorenz 63 and Lorenz 96 examples.

3. The work also suggests a number of directions for future study in the area of KAF:
   - Delay embedding can be used to deal with non-Markovian behavior, and it would be of interest to automate the choice of delay embedding dimension within the KAF framework to get closer to Markovianity.
   - Combining KAF with data assimilation holds the possibility of greater predictability—for work in this direction, see [20], which studies the EnKF with a data-driven model update.
   - It would be of interest to have algorithms to identify slow subspaces, using data living in larger spaces, and hence to conduct KAF using approximately Markovian variables.
   - It would be of interest to extend KAF predictor Z_τ so that it acts on the joint space of initial conditions and key parameters, enabling prediction at as yet unseen parameters.

Appendix A. Computation. This appendix contains details of the implementation of KAF. Subsection A.1 describes the algorithms for computing the kernel, diffusion eigenbasis, RKHS basis functions, and, finally, the prediction. Our specific choice of kernel, which endows the RKHS structure, is explained and motivated in subsection A.2. We outline procedures for choosing the truncation parameter ℓ and approximating the conditional variance of the forecast in subsections A.3 and A.4.

A.1. Linear algebra.

A.1.1. Algorithm 1 (diffusion eigenbasis). The starting point of this algorithm is the variable-bandwidth diffusion kernel function $\kappa_N : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\kappa_N(x, x') = \exp\left(\frac{-|x - x'|^2}{\epsilon r_N(x) r_N(x')}\right),$$

described further in Appendix A.2. An automated procedure for estimating the data-dependent bandwidth function r_N and width ϵ are given in [6].

- Inputs
  - Training data $x_0, x_1, \ldots x_{N-1} \in \mathcal{X}$
  - Desired maximum number of eigenvectors $L \leq N$
- Outputs
  - Diffusion eigenvectors $\phi_0, \ldots, \phi_{L-1} \in \mathbb{R}^N$, stacked in matrix $\Phi \in \mathbb{R}^{N \times L}$
  - Right singular vectors $\gamma_0, \ldots, \gamma_{L-1} \in \mathbb{R}^N$
  - Diffusion eigenvalues $\lambda_0, \ldots, \lambda_{L-1} \in \mathbb{R}$ in diagonal matrix $\Lambda \in \mathbb{R}^{L \times L}$
- Steps
  1. Form the matrix $\hat{K} \in \mathbb{R}^{\hat{N} \times N}$ with entries $K_{ij} = \kappa_N(x_i, x_j)$.
  2. Comput the $v_N, w_N \in \mathbb{R}^N$ using $v_N = K\vec{1}$, and $w_N = KV^{-1}\vec{1}$, where $V = \mathrm{diag}(v_N)$.
  3. Form the normalized kernel matrix $S = V^{-1}KW^{-1/2}$, where $W = \mathrm{diag}(w_N)$.
  4. Compute the L largest singular values $\sigma_0, \ldots, \sigma_{L-1}$, the corresponding left singular vectors $\phi_0, \ldots, \phi_{L-1}$, and right singular vectors $\gamma_0, \ldots, \gamma_{L-1}$ of S. Normalize the singular vectors such that $\|\phi_j\|_2 = \|\gamma_j\|_2 = \sqrt{N}$, where $\|\cdot\|_2$ is the standard 2-norm on $\mathbb{R}^N$. Stack eigenvectors columnwise into $N \times L$ matrix $\Phi := [\phi_0, \ldots, \phi_{L-1}]$, and form the $L \times L$ diagonal matrix of eigenvalues $\Lambda := \mathrm{diag}(\lambda_j)$, with $\lambda_j := \sigma_j^2$.

Remark. Recall that the key idea of KAF is the eigendecomposition of the Markov operator G, which can be represented by an $N \times N$ matrix eigenvalue problem,

$$G\phi = \lambda\phi.$$

The matrix G is never explicitly formed; instead, we exploit the fact that $G = SS^T$, and hence ϕ are the left singular vectors of S. Thus, we bypass the eigendecomposition step with a reduced singular value decomposition (SVD) in step (4). Note that the SVD approach is natural when working with kernels with an explicit factorization $G = SS^T$, including the bistochastic kernels described in Appendix A.2. For more general kernels, one typically performs direct eigendecomposition of G. KAF can also be implemented with nonsymmetric kernels satisfying a detailed balance condition making them conjugate to positive-definite kernels; see section 4.2 in [1] for more details.

A.1.2. Algorithm 2 (RKHS basis functions). RKHS basis functions are computed using the Nyström extension (2.7) reproduced below

$$\psi_j = \frac{1}{N\lambda_j^{1/2}} \sum_{n=0}^{N-1} k(\cdot, x_n)\phi_j(x_n), \quad \lambda_j > 0.$$

- Inputs
  - Out-of-sample data $\hat{x}_0, \hat{x}_1, \ldots \hat{x}_{\hat{N}-1} \in \mathcal{X}$
  - Right singular vectors $\gamma_0, \ldots, \gamma_{L-1}$ from Algorithm 1
  - W from Algorithm 1
- Outputs
  - RKHS basis $\Psi = [\psi_0, \ldots, \psi_{L-1}]$
- Steps
  1. Form matrix $\hat{K} \in \mathbb{R}^{\hat{N} \times N}$ with entries $\hat{K}_{ij} = \kappa_N(\hat{x}_i, x_j)$. Note that this requires another kernel density estimation step to evaluate the sampling density on out-of-sample data.

2. Compute $\hat{v}_{\hat{N}} \in \mathbb{R}^{\hat{N}}$ using $\hat{v}_{\hat{N}} = \hat{K}\vec{1}$.
3. Form the $\hat{N} \times N$ kernel matrix $\bar{K} = \hat{V}^{-1}\hat{K}W^{-1/2}$, where $\hat{V} = \mathrm{diag}(\hat{v}_{\hat{N}})$.
4. Output $\psi_j = \bar{K}\gamma_j$, stacked columnwise in the $\hat{N} \times L$ matrix $\Psi := [\psi_0, \ldots, \psi_{L-1}]$.

A.1.3. Algorithm 3 (predictor). The final predictor is constructed according to (2.1), reproduced here for convenience:

$$Z_\tau(x) = \frac{1}{N}\sum_{n=0}^{N-1}\left(\sum_{j=0}^{\ell(\tau)-1}\frac{\psi_j(x)\phi_j(x_n)}{\lambda_j^{1/2}}\right) f_{n+q}.$$

- Inputs
  - Lead time $\tau = q\Delta t$
  - Truncation parameter $\ell \le L$
  - Vector of sampled observable $f_\tau = [f_q, \ldots, f_{N-1+q}]^T$
  - Diffusion eigenvectors $\Phi_\ell \in \mathbb{R}^{N\times\ell}$ (first ℓ columns of Φ)
  - Diffusion eigenvalues $\Lambda_\ell = \mathrm{diag}(\lambda_0, \ldots, \lambda_{\ell-1})$
  - RKHS basis $\Psi_\ell \in \mathbb{R}^{N\times\ell}$ (first ℓ columns of Ψ)
- Output
  - Prediction Z_τ at $\hat{x}_0, \hat{x}_1, \ldots \hat{x}_{\hat{N}-1}$
- Steps
  1. Form ℓ-dimensional coefficient vector $c := \Phi_\ell^T f_\tau / N$.
  2. Compute $z := \Psi_\ell \Lambda_\ell^{-1/2} c$. Report prediction for i^{th} initial point $Z_\tau(\hat{x}_i) := z_i$.

A.2. Choice of kernel. Details concerning the kernel choice may be found in section 5 of [1]. Here we briefly summarize the key ideas. Our starting point is the Gaussian kernel

$$\kappa(x, x'; \epsilon) = \exp\bigl(-|x - x'|^2/\epsilon\bigr),$$

where we assume that $\mathcal{X}$ is a subset of $\mathbb{R}^d$, $x, x' \in \mathcal{X}$ and $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$. Note that this kernel is data independent. From it we can generalize to a data-dependent kernel defined by [7]

$$\begin{aligned}
\kappa_N(x, x') &= \exp\bigl(-|x - x'|^2/(r_N(x)r_N(x')\epsilon)\bigr),\\
r_N(x) &= q_N(x)^{\frac{1}{m}},\\
q_N(x) &= \frac{1}{(\pi\delta)^{m/2}}\int_\Omega \kappa(x, \Pi(\omega); \delta)\mu_N(d\omega).
\end{aligned}$$

Here, ϵ, δ and m are lengthscale and dimension parameters, respectively, estimated from the data. The role of including r_n and hence a variable bandwidth kernel is to compensate for variations in sampling density across the space. A key conceptual idea underlying the construction of q_N is that it approximates the Lebesgue density of the measure μ in the large data limit. Thus, the kernel κ_N weights the distance of x and x' in a manner which reflects the sampling density of the data.

From κ_N, the kernel k_N is constructed as follows, invoking a second principle which is to design a bistochastic Markov kernel [9]. Doing so ensures that the top eigenvalue of G is 1 with corresponding eigenvector a constant; then the hypothesis space also contains constants. This is useful for capturing the mean of the predicted quantity $U^\tau F$ and, in particular, plays a central role in the large τ asymptotics for

mixing systems. To achieve the Markov property, we proceed as follows. First, we define

$$v_N(x) = \int_\Omega \kappa_N\big(x, \Pi(\omega)\big)\mu_N(d\omega),$$
$$w_N(x) = \int_\Omega \frac{\kappa_N\big(x, \Pi(\omega)\big)}{v_N\big(\Pi(\omega)\big)}\mu_N(d\omega).$$

Since the above empirically determined quantities only take values at the N sampled points, they are isomorphic to vectors in $\mathbb{R}^N$ identified by the same name within Algorithm A.1.1. Finally, define

$$k_N(x, x') = \int_\Omega \frac{\kappa_N\big(x, \Pi(\omega)\big)\kappa_N\big(\Pi(\omega), x'\big)}{v_N(x)w_N\big(\Pi(\omega)\big)v_N(x')}\mu_N(d\omega).$$

The operators constructed from the unnormalized and normalized kernels, κ_N and k_N, are likewise represented by $N \times N$ matrices K and SS^T, respectively, in Algorithm A.1.1. It may be verified that the Markov property is satisfied, and so too are the positivity conditions required for the aforementioned large data convergence result.

A.3. Choice of ℓ. Recall that the predictor (2.9) actually corresponds to a *family* of predictors parameterized by the desired lead time τ and by the truncation parameter ℓ. Thus, we write $Z_{\tau,\ell}$. Here we describe how to choose ℓ for a fixed lead time τ. In practice, the choice of ℓ is determined from the minimizer of the empirical RMSE based on (2.5), computed from a validation data set with $\tilde{N}$ samples $\hat{x}_0, \ldots, \hat{x}_{\tilde{N}-1}$:

$$\ell = \operatorname*{arg\,min}_{\ell'=1,\ldots,L} \mathsf{RMSE}\,(Z_{\tau,\ell'}).$$

A.3.1. Algorithm 4 (tuning ℓ).

- Inputs
  - Forecast lead time $\tau = q\Delta t$
  - Validation out-of-sample data $\hat{x}_0, \ldots, \hat{x}_{\tilde{N}-1} \in \mathcal{X}$
  - Ground truth vector of observables $\hat{f}_\tau = [\hat{f}_q, \ldots, \hat{f}_{\tilde{N}-1+q}]^T$
  - Diffusion eigenvectors Φ from Algorithm 1
  - Diffusion eigenvalues Λ from Algorithm 1
- Outputs
  - Truncation parameter ℓ
- Steps
  1. Compute RKHS basis functions Ψ_L using Algorithm 2. Set $R := \infty$.
  2. For $\ell' = 1$ to L,
     (a) compute predictor $Z_{\tau,\ell'}(\hat{x}) := z$ using Algorithm 3;
     (b) compute $\mathsf{RMSE}_{\ell'}$ for $Z_{\tau,\ell'}(\hat{x})$ as $\mathsf{RMSE}_{\ell'} := \|z - \hat{f}_\tau\|_2$;
     (c) if $\mathsf{RMSE}_{\ell'} \le R$, set $\ell := \ell'$ and $R := \mathsf{RMSE}_{\ell'}$.
  3. Return ℓ.

This tuning procedure must be carried out for each desired lead time τ.

A.3.2. Long-time behavior of ℓ. It should be noted that in the presence of mixing or chaotic dynamics, for long lead times τ, the projected subspaces become one-dimensional; i.e., $\ell = 1$, and the predictor converges to a constant (this occurs only if the subspace includes the constant eigenfunction from a Markov kernel operator).

The weak convergence of the conditional expectation of mixing dynamics to the mean of the observable F is described in [1],

$$\lim_{\tau\to\infty}\langle g, \mathbb{E}[U^\tau F|\Pi] - \mathbb{E}[F]\mathbf{1}_\Omega\rangle_{L^2_\mu} = 0, \tag{A.1}$$

a property that is a direct consequence of a measure-theoretic definition of mixing

$$\lim_{\tau\to\infty}\langle U^{\tau *} g, h\rangle_{L^2_\mu} = \mathbb{E}[g]\,\mathbb{E}[h], \tag{A.2}$$

where the expectations are taken over the invariant measure μ.

A.4. Formula for variance. The uncertainty associated with the prediction at each lead time can be estimated via the conditional variance

$$\mathrm{var}[U^\tau F|\Pi] = \mathbb{E}[(U^\tau F - \mathbb{E}[U^\tau F|\Pi])^2|\Pi] \approx \mathbb{E}[(U^\tau F - Z_\tau)^2|\Pi].$$

The variance is yet another observable in $L^2_\mu(\Omega;\mathbb{R})$ that can be evaluated using the same basis functions as the predictor, and, once the predictor is computed, it only remains to compute the expansion coefficients

$$\hat{c}_j(\tau) = \langle \phi_j \circ \Pi, (U^\tau F - Z_\tau)^2\rangle_{L^2(\mu_N)} = \frac{1}{N}\sum_{n=0}^{N-1}\phi_j(x_n)(f_{n+q} - Z_\tau(x_n))^2. \tag{A.3}$$

Note that this computation requires a different optimal truncation $\ell(\tau)$ for the variance and hence another validation set for parameter tuning:

1. (Tuning ℓ) Run Algorithm 4 on a separate validation data set for which the predictor is already computed, using the new observable $\hat{g}_\tau$ instead of $\hat{f}_\tau$:
$$\hat{g}_\tau = [(\hat{f}_q - Z_\tau(\hat{x}_0))^2, \ldots, (\hat{f}_{\tilde{N}-1+q} - Z_\tau(\hat{x}_{\tilde{N}-1}))^2]^T.$$
2. Run Algorithm 3 on the initial prediction data using the observable g_τ instead of f_τ,
$$g_\tau = [(f_q - Z_\tau(x_0))^2, \ldots, (f_{N-1+q} - Z_\tau(x_{N-1}))^2]^T,$$
denoting the output predicted variance as V_τ.
3. Finally, the uncertainty bands of the predictor at each lead time are given by two standard deviations, or twice the square root of the variance, $Z_\tau \pm 2\sqrt{|V_\tau|}$.

Appendix B. Details of the GP closure. In this section we describe details of the construction of the GP underlying the model-data–driven approach and leading to the approximate closed equation (4.4). We use L-96 explicitly, but the methodology easily generalizes to other multiscale systems.

Construction of a GP closure is performed using the following steps:

(a) Choose random initial conditions for L-96.
(b) Numerically integrate for time T_{conv} to determine an initial condition on the global attractor.
(c) Numerically integrate from this initial condition with a fixed time step Δt for time T_{learn} to collect pairs of data:
$$\left\{\left(x_k(t_n), (By)_k(t_n)\right)\right\}_{n=1}^{N}$$
for $k = 1, \ldots 9$ and where $t_n = n\Delta t$, $N = \lfloor \frac{T_{\mathrm{learn}}}{\Delta t} \rfloor$.
(d) Train a GP using collected pairs of data, including optimization over hyperparameters, such as lengthscale, and set c_{GP} to be the mean of this GP.

Note that in step (c) we exploit the statistical invariance w.r.t. circular index shifting; this enables us to collect K pairs in one time step. For numerical implementation of the GPR, we used the `scikit-learn` package [39]; the kernel of the GP was chosen as the sum of a standard radial basis function and white noise kernels, setting the noise level in the latter to 0.5, and the length scale parameter was optimized. Out of roughly 30000 points obtained in step (c), we subsampled 500 uniformly at random, without replacement, to train the GP. The result of such a procedure is shown in Figure 4.4.

Acknowledgment. DG is grateful to the Department of Computing and Mathematical Sciences at the California Institute of Technology for hospitality and for providing a stimulating environment during a sabbatical where part of this work was completed.

REFERENCES

[1] R. Alexander and D. Giannakis, *Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques*, Phys. D, 409 (2020), 132520, https://doi.org/10.1016/j.physd.2020.132520.

[2] H. Arbabi and I. Mezic, *Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the koopman operator*, SIAM J. Appl. Dyn. Syst., 16 (2017), pp. 2096–2126, https://doi.org/10.1137/17M1125236.

[3] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[4] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396, https://doi.org/10.1162/089976603321780317.

[5] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Vol. 374, American Mathematical Society, Providence, RI, 2011.

[6] T. Berry, D. Giannakis, and J. Harlim, *Nonparametric forecasting of low-dimensional dynamical systems*, Phys. Rev. E, 91 (2015), 032915, https://doi.org/10.1103/PhysRevE.91.032915.

[7] T. Berry and J. Harlim, *Variable bandwidth diffusion kernels*, Appl. Comput. Harmon. Anal., 40 (2016), pp. 68–96, https://doi.org/10.1016/j.acha.2015.01.001.

[8] A. Chattopadhyay, E. Nabizadeh, and P. Hassanzadeh, *Analog forecasting of extreme-causing weather patterns using deep learning*, J. Adv. Model. Earth Syst., 12 (2020), e2019MS001958, https://doi.org/10.1029/2019MS001958.

[9] R. Coifman and M. Hirn, *Bi-stochastic kernels via asymmetric affinity functions*, Appl. Comput. Harmon. Anal., 35 (2013), pp. 177–180, https://doi.org/10.1016/j.acha.2013.01.001.

[10] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30, https://doi.org/10.1016/j.acha.2006.04.006.

[11] R. R. Coifman and S. Lafon, *Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 31–52, https://doi.org/10.1016/j.acha.2005.07.005.

[12] S. Das and D. Giannakis, *Delay-coordinate maps and the spectra of Koopman operators*, J. Stat. Phys., 175 (2019), pp. 1107–1145, https://doi.org/10.1007/s10955-019-02272-w.

[13] S. Das, D. Giannakis, and J. Slawinska, *Reproducing kernel Hilbert space compactification of unitary evolution groups*, Appl. Comput. Harmon. Anal., 54 (2021), pp. 75–136, https://doi.org/10.1016/j.acha.2021.02.004.

[14] L. Delle Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, *Probabilistic weather prediction with an analog ensemble*, Monthly Weather Rev., 141 (2013), pp. 3498–3516.

[15] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM J. Numer. Anal., 36 (1999), pp. 491–515, https://doi.org/10.1137/S0036142996313002.

[16] I. Fatkullin and E. Vanden-Eijnden, *A computational strategy for multiscale systems with applications to Lorenz 96 model*, J. Comput. Phys., 200 (2004), pp. 605–638, https://doi.org/10.1016/j.jcp.2004.04.013.

[17] D. Giannakis, *Data-driven spectral decomposition and forecasting of ergodic dynamical systems*, Appl. Comput. Harmon. Anal., 62 (2019), pp. 338–396, https://doi.org/10.1016/j.acha.2017.09.001.

[18] D. Giannakis, A. Kolchinskaya, D. Krasnov, and J. Schumacher, *Koopman analysis of the long-term evolution in a turbulent convection cell*, J. Fluid Mech., 847 (2018), pp. 735–767, https://doi.org/10.1017/jfm.2018.297.

[19] D. Givon, R. Kupferman, and A. Stuart, *Extracting macroscopic dynamics: Model problems and algorithms*, Nonlinearity, 17 (2004), R55, https://doi.org/10.1088/0951-7715/17/6/R01.

[20] F. Hamilton, T. Berry, and T. Sauer, *Ensemble Kalman filtering without a model*, Phys. Rev. X, 6 (2016), 011021, https://doi.org/10.1103/PhysRevX.6.011021.

[21] J. Harlim, S. W. Jiang, S. Liang, and H. Yang, *Machine learning for prediction with missing dynamics*, J. Comput. Phys., 428 (2021), 109922, https://doi.org/10.1016/j.jcp.2020.109922.

[22] S. W. Jiang and J. Harlim, *Modeling of Missing Dynamical Systems: Deriving Parametric Models Using a Nonparametric Framework*, Res. Math. Sci., 7 (2020), pp. 1–25, https://doi.org/10.1007/s40687-020-00217-4.

[23] D. Kelly and I. Melbourne, *Deterministic homogenization for fast–slow systems with chaotic noise*, J. Funct. Anal., 272 (2017), pp. 4063–4102, https://doi.org/10.1016/j.jfa.2017.01.015.

[24] D. Kelly and I. Melbourne, *Smooth approximation of stochastic differential equations*, Ann. Probab., 44 (2016), pp. 479–520, https://doi.org/10.1214/14-AOP979.

[25] M. A. Khodkar, P. Hassanzadeh, and A. Antoulas, *A Koopman-Based Framework for Forecasting the Spatiotemporal Evolution of Chaotic Dynamics with Nonlinearities Modeled as Exogenous Forcings*, preprint, arXiv:1909.00076, 2019.

[26] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, *Data-driven model reduction and transfer operator approximation*, J. Nonlinear Sci., 28 (2018), pp. 985–1010, https://doi.org/10.1007/s00332-017-9437-7.

[27] B. O. Koopman, *Hamiltonian systems and transformation in Hilbert space*, Proc. Natl. Acad. Sci. USA, 17 (1931), p. 315.

[28] M. Korda, M. Putinar, and I. Mezić, *Data-driven spectral analysis of the Koopman operator*, Appl. Comput. Harmon. Anal., 48 (2020), pp. 599–629, https://doi.org/10.1016/j.acha.2018.08.002, in press.

[29] E. N. Lorenz, *Deterministic nonperiodic flow*, J. Atmos. Sci., 20 (1963), pp. 130–141, https://doi.org/10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2.

[30] E. N. Lorenz, *Atmospheric predictability as revealed by naturally occurring analogues*, J. Atmos. Sci., 26 (1969), pp. 636–646, https://doi.org/10.1175/1520-0469(1969)26⟨636:APARBN⟩2.0.CO;2.

[31] E. N. Lorenz, *Predictability: A problem partly solved*, in Proceedings of the Seminar on Predictability, Vol. 1, ECMWF, 1996.

[32] I. Melbourne and M. Nicol, *Almost sure invariance principle for nonuniformly hyperbolic systems*, Commun. Math. Phys., 260 (2005), pp. 131–146, https://doi.org/10.1007/s00220-005-1407-5.

[33] I. Melbourne and A. Stuart, *A note on diffusion limits of chaotic skew-product flows*, Nonlinearity, 24 (2011), pp. 1361–1367, https://doi.org/10.1088/0951-7715/24/4/018.

[34] I. Mezić, *Spectral properties of dynamical systems, model reduction and decompositions*, Nonlinear Dyn., 41 (2005), pp. 309–325, https://doi.org/10.1007/s11071-005-2824-x.

[35] I. Mezić, *Analysis of fluid flows via spectral properties of the koopman operator*, Annu. Rev. Fluid Mech., 45 (2013), pp. 357–378, https://doi.org/10.1146/annurev-fluid-011212-140652.

[36] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, *Diffusion maps, spectral clustering and reaction coordinates of dynamical systems*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 113–127, https://doi.org/10.1016/j.acha.2005.07.004.

[37] G. C. Papanicolaou and W. Kohler, *Asymptotic theory of mixing stochastic ordinary differential equations*, Commun. Pure Appl. Math., 27 (1974), pp. 641–668, https://doi.org/10.1002/cpa.3160270503.

[38] G. Pavliotis and A. Stuart, *Multiscale Methods: Averaging and Homogenization*, Springer-Verlag, Berlin, 2008, https://doi.org/10.1007/978-0-387-73829-1.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, J. Mach. Learn. Res., 12 (2011), pp. 2825–2830.

[40] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, MIT Press, Cambridge, MA, 2005.

[41] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, *Spectral analysis of nonlinear flows*, J. Fluid Mech., 641 (2009), pp. 115–127, https://doi.org/10.1017/S0022112009992059.
[42] P. J. Schmid, *Dynamic mode decomposition of numerical and experimental data*, J. Fluid Mech., 656 (2010), pp. 5–28, https://doi.org/10.1017/S0022112010001217.
[43] B. Schölkopf, A. Smola, and K. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural Comput., 10 (1998), pp. 1299–1319, https://doi.org/10.1162/089976698300017467.
[44] B. Schölkopf, A. J. Smola, and F. Bach, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2002.
[45] A. Singer, *From graph to manifold laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134, https://doi.org/10.1016/j.acha.2006.03.004.
[46] J. Slawinska and D. Giannakis, *Indo-Pacific variability on seasonal to multidecadal time scales. Part* I: *Intrinsic SST modes in models and observations*, J. Climate, 30 (2017), pp. 5265–5294, https://doi.org/10.1175/JCLI-D-16-0176.1.
[47] N. G. Trillos, M. Gerlach, M. Hein, and D. Slepčev, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs towards the Laplace–Beltrami operator*, Found. Comput. Math., 20 (2020), pp. 827–887, https://doi.org/10.1007/s10208-019-09436-w.
[48] W. Tucker, *The Lorenz attractor exists*, C. R. Acad. Sci. Ser. I Math., 328 (1999), pp. 1197–1202, https://doi.org/10.1016/S0764-4442(99)80439-X.
[49] H. Van den Dool, *A new look at weather forecasting through analogues*, Monthly Weather Rev., 117 (1989), pp. 2230–2247, https://doi.org/10.1175/1520-0493(1989)117⟨2230:ANLAWF⟩2.0.CO;2.
[50] U. Von Luxburg, M. Belkin, and O. Bousquet, *Consistency of spectral clustering*, Ann. Statist., (2008), pp. 555–586, http://doi.org/10.1214/009053607000000640.
[51] G. Wahba, *Spline Models for Observational Data*, Vol. 59, SIAM, Philadelphia, 1990.
[52] X. Wang, J. Slawinska, and D. Giannakis, *Extended-range statistical ENSO prediction through operator-theoretic techniques for nonlinear dynamics*, Sci. Rep., 10 (2020), 2636, https://doi.org/10.1038/s41598-020-59128-7.
[53] E. Weinan, *Principles of Multiscale Modeling*, Cambridge University Press, Cambridge, 2011.
[54] D. S. Wilks, *Effects of stochastic parametrizations in the Lorenz* '96 *system*, Quart. J. Roy. Meteorol. Soc., 131 (2005), pp. 389–407, https://doi.org/10.1256/qj.04.03.
[55] Z. Zhao and D. Giannakis, *Analog forecasting with dynamics-adapted kernels*, Nonlinearity, 29 (2016), 2888, https://doi.org/10.1088/0951-7715/29/9/2888.