

# Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques

Romeo Alexander, Dimitrios Giannakis\*

Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA

## ARTICLE INFO

### Article history:

Received 22 June 2019

Received in revised form 14 March 2020

Accepted 8 April 2020

Available online 14 April 2020

Communicated by G. Froyland

### Keywords:

Statistical forecasting

Kernel methods

Conditional expectation

Koopman operators

## ABSTRACT

Kernel analog forecasting (KAF), alternatively known as kernel principal component regression, is a kernel method used for nonparametric statistical forecasting of dynamically generated time series data. This paper synthesizes descriptions of kernel methods and Koopman operator theory in order to provide a single consistent account of KAF. The framework presented here illuminates the property of the KAF method that, under measure-preserving and ergodic dynamics, it consistently approximates the conditional expectation of observables that are acted upon by the Koopman operator of the dynamical system and are conditioned on the observed data at forecast initialization. More precisely, KAF yields optimal predictions, in the sense of minimal root mean square error with respect to the invariant measure, in the asymptotic limit of large data. The presented framework facilitates, moreover, the analysis of generalization error and quantification of uncertainty. Extensions of KAF to the construction of conditional variance and conditional probability functions, as well as to non-symmetric kernels, are also shown. Illustrations of various aspects of KAF are provided with applications to simple examples, namely a periodic flow on the circle and the chaotic Lorenz 63 system.

© 2020 Elsevier B.V. All rights reserved.

## 1. Introduction

Forecasting dynamically generated time series is a challenging problem that often requires statistical methods, especially when the underlying equations are either unknown or computationally intractable. Data-driven forecasting methods have been sought after at least since Lorenz attempted to use naturally occurring historical analogs for climate predictions in the 1960s [1]. That early attempt was limited in success, but larger datasets and improved computing resources have made more recent analog-based nonparametric methods more viable [2–7]. Various types of ensemble analog forecasting are employed in short-term meteorological forecasts [8,9], and versions of analog forecasting that utilize kernels have been shown to have predictive value for certain weather and climate phenomena [10–15].

While naturally occurring analogs may be a point of emphasis for nonparametric methods in physical science applications, abstract statistical structures are the focus when situated in a more general machine learning context. Common nonparametric machine learning techniques include multilayer perceptrons [16], Bayesian neural networks [17], classification and regression trees (CART), and a variety of kernel methods [18]. Although each of these methods can provide value in unique ways to specific

problems, kernel methods are particularly well suited to problems where there may be a natural, a priori, notion of similarity between data points. Since analog methods rely on the possibility that the relevance of any historical analog to present day conditions can be quantitatively determined, formal understanding of such methods can improve when they are cast within the larger framework of kernel methods.

Kernel methods constitute a class of algorithms that perform classical calculations in a rich functional feature space in order to extract and predict nonlinear patterns. This central idea, commonly referred to as “the kernel trick”, was first proposed in 1964 [19], was popularized with the invention of nonlinear support vector machines (SVMs) in 1992 [20], and has since spread to a variety of machine learning applications [21,22]. Kernel methods for regression, such as support vector regression (SVR) [23], kernel ridge regression (KRR) [24], and kernel principal component regression (KPCR) [25], may be applied to appropriately lagged signals to produce time series forecasts, such as with SVR forecasting [26], KRR forecasting [27], and KPCR forecasting [25]. Such kernel forecasting methods have been frequently used in finance and econometrics [28], and have recently found use in climate science [7,11,12,14], where they were termed kernel analog forecasting (KAF).

Statistical learning theory [29] is the standard theoretical framework for deriving and analyzing kernel methods, among other machine learning algorithms. The learning guarantees and

\* Corresponding author.

E-mail address: [dimitris@cims.nyu.edu](mailto:dimitris@cims.nyu.edu) (D. Giannakis).

**Table 1**Pseudocode outlining the construction of the empirical KAF target function  $f_{\tau,\ell,n}$  from Definition 9.

---

• Inputs
- Covariate training data $x_1, \dots, x_n \in \mathcal{X}$ at sampling interval $\Delta t$
- Response training data $y_1, \dots, y_n \in \mathcal{Y}$ at sampling interval $\Delta t$
- Symmetric positive-definite kernel function $k_n : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$
- Forecast timesteps $q \in \mathbb{N}_0$
- Number of principal components (eigenfunctions) $\ell \in \mathbb{N}$
• Outputs
- Target function $f_{\tau,\ell,n} : \mathcal{X} \rightarrow \mathbb{R}$ for lead time $\tau = q \Delta t$
• Steps
1. Compute the leading $\ell$ eigenvectors $\phi_j \in \mathbb{R}^n$ of the $n \times n$ kernel matrix $\mathbf{K} = [k_n(x_i, x_j)]$ , arranged in order of decreasing corresponding eigenvalue $\lambda_{j,n}$ . Normalize the eigenvectors such that $\phi_i \cdot \phi_j / n = \delta_{ij}$ .
2. Form the $q$ -step shifted response vector $\mathbf{y}_\tau = (0, \dots, 0, y_1, \dots, y_{n-q}) \in \mathbb{R}^n$ , and compute the expansion coefficients $\alpha_j(\tau) = \phi_j \cdot \mathbf{y}_\tau / n$ for $j \in \{1, \dots, \ell\}$ .
3. Form the orthonormal RKHS functions $\psi_{j,n}(x) = \mathbf{k}(x) \cdot \phi_j / (n \lambda_{j,n}^{1/2})$ , where $\mathbf{k}(x) = (k_n(x, x_1), \dots, k_n(x, x_n))$ is the kernel vector.
4. Form the target function $f_{\tau,\ell,n}(x) = \sum_{j=1}^{\ell} \alpha_{j,n}(\tau) \psi_{j,n}(x)$ .

---

estimates of rates of convergence are well known when the underlying data are independently and identically distributed [30]. For time series, where the i.i.d. assumption is generally not valid, an extension of the standard i.i.d. statistical learning framework to that of stochastic processes has yielded softer guarantees that depend on mild conditions on the stationarity of the system [31,32]. Although trajectories of a dynamical system can be viewed as a special case of a stochastic process, it is also worthwhile to employ the measure and operator-theoretic perspectives of modern dynamical systems theory [33,34], where the induced action of the dynamical system on an intrinsically linear space of observables is given a more prominent role. This operator-theoretic perspective, although widespread in the study of dynamical systems [35], has yet to be fully exploited in conjunction with kernel forecasting methods.

The main contribution of this paper is a rigorous reformulation of KAF techniques within the framework of operator-theoretic ergodic theory and statistical learning theory. This view relies on the equivalence of forecasts with conditional expectation or, alternatively, geometric projection, both of which draw on the rich theory of functional analysis. One benefit from such a perspective is that it turns the problem of error analysis into the well studied problem of convergence in Hilbert spaces. Another benefit to this approach is that it demystifies the kernel functions somewhat by revealing their special role in bridging the gap between  $L^2$  Hilbert spaces and the continuous function space in which forecasts are ultimately expressed. A third benefit is the modularity and extensibility that comes from casting kernel forecasting algorithms as a composition of operators applied to a careful choice of observable. In particular, by expressing forecasts as a composition of a regressor operator and the Koopman operator [36], the latter being a construct representing the action of evolving forward in time, features of the statistics and the dynamics are more easily separated and studied independently. For example, approximations of Koopman and the related transfer operators have been the subject of recent research [37–41,35,42–52], and may be combined with approximations of the regressor operator to yield new formulations. Moreover, with appropriate choices of the response observable, forecasts can be obtained not just for the conditional mean of an observed quantity, but also that quantity's conditional variance and higher-order moments, which are important for uncertainty quantification. Conditional probability may also be approximated and predicted with a kernel analog approach. In this analysis, reproducing kernel Hilbert spaces (RKHSs) [53,54] play a central role as ambient hypothesis spaces of functions, with enough structure to enable an explicit representation of the forecasting function (also known as target function) in a fully empirical manner.

This paper is organized as follows. Section 2 introduces the forecasting problem under study, and describes the KAF framework. Section 3 studies the generalization error of constructed

forecasts, paying particular attention on how to quantify the discrepancy between empirical and ideal forecasts. Our main result on the convergence of KAF to the conditional expectation is stated as Theorem 14 in that section. Section 4 introduces a few extensions, including KRR, non-symmetric kernels, conditional variance, and conditional probability. In Section 5, we provide general guidelines for choosing the kernel. Section 6 shows the result of applying KAF to two examples, namely a periodic flow on the circle and the chaotic Lorenz 63 (L63) system [55]. Section 7 provides our principal conclusory remarks, and examines the applicability of KAF to various real-world problems. Technical results are collected in an Appendix.

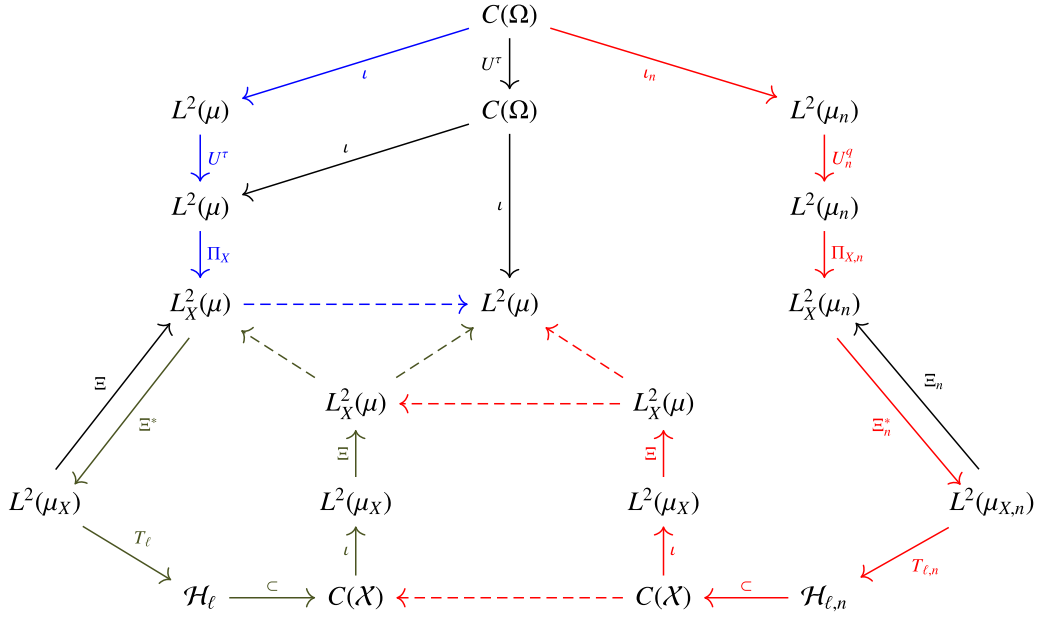
## 2. Kernel analog forecasting (KAF) techniques

In this section, we describe the mathematical framework underlying the KAF approach introduced in [7]. We start from a general formulation of forecasting as error minimization (Section 2.1), and gradually build onto that various dynamical systems and functional analytic tools, leading (in Section 2.4) to the construction of the RKHS-based KAF target function. It should be noted that our exposition differs substantially from [7], which focuses heavily on RKHS interpolation theory from the outset. In particular, an advantage of the perspective put forward here is that the RKHS formalism emerges as a natural consequence of seeking target functions in an explicitly constructible ambient hypothesis space with a Hilbert space structure, as opposed to the more “axiomatic” use of RKHSs in [7]. This perspective will also facilitate the error analysis in Section 3.

Fig. 1 depicts the relationships between the function spaces and operators involved in the construction of the KAF target function in the form of a commutative diagram. The basic steps of the construction are also summarized in pseudocode in Table 1. Fig. 2 shows an application of KAF to the L63 system under full and partial observations. This L63 application provides a guiding example of a number of challenges encountered in statistical forecasting, including partial state observations, mixing (i.e., chaotic) dynamics, and invariant measures supported on non-smooth attractors.

### 2.1. Mathematical background

**Measure-theoretic framework.** In the measure-theoretic setup that we wish to pursue here, the primary object is a probability space  $(\Omega, \mathcal{F}, \mu)$ , where  $\Omega$  is the state space,  $\mathcal{F}$  is a  $\sigma$ -algebra of distinguished subsets of  $\Omega$ , and  $\mu : \mathcal{F} \rightarrow \mathbb{R}$  is a probability measure. We also have a measurable covariate space  $(\mathcal{X}, \Sigma_{\mathcal{X}})$ , a measurable response space  $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ , and, for each time  $t \geq 0$ , data-producing measurable functions  $X_t : \Omega \rightarrow \mathcal{X}$  and  $Y_t : \Omega \rightarrow \mathcal{Y}$ . By data-producing, we mean that the covariate and response data,  $x_t$  and  $y_t$ , are regarded as the output of  $X_t$  and  $Y_t$ ,



**Fig. 1.** Commutative diagram illustrating the relationships between the function spaces and operators employed in KAF. We start from the space of continuous functions on  $\Omega$ , wherein the response variable  $Y$  lies. The path proceeding downwards from  $C(\Omega)$  along the black arrows represents the evolution of continuous functions under the Koopman operator  $U^\tau$  for lead time  $\tau$ , followed by inclusion in the  $L^2(\mu)$  space associated with the invariant measure. This leads to the observable  $U^\tau Y \in L^2(\mu)$ , which we seek to approximate with minimal error with respect to  $L^2(\mu)$  norm. The paths demarcated by blue, green, and red arrows represent three levels of approximation, each with its own errors (indicated by dashed arrows). The path colored in blue leads to the conditional expectation  $\mathbb{E}[U^\tau Y | X] = \Pi_X U^\tau Y$ , given by the orthogonal projection of  $U^\tau Y$  into the Hilbert space  $L^2_X(\mu) \subseteq L^2(\mu)$  consisting of pullbacks of functions on covariate space  $\mathcal{X}$ . The conditional expectation is associated with the regression function  $Z_\tau = \mathcal{E}^* \mathbb{E}[U^\tau Y | X]$  from Definition 1, and can exhibit an irreducible form of error relative to  $U^\tau Y$  if  $X$  is not injective (indicated by a dashed blue arrow). The green-colored path describes the approximation of the regression function by a continuous function  $f_{\tau,\ell}$  on covariate space (the ideal target function in Definition 2), lying in a finite-dimensional hypothesis space  $\mathcal{H}_\ell$ , which is a subspace of an RKHS  $\mathcal{H}$  on  $\mathcal{X}$ . The operator  $T_\ell$  leading to the target function is a spectrally truncated Nyström extension operator, where  $\ell$  is the number of kernel eigenfunctions employed. This imparts an additional error, which vanishes, however, as  $\ell \rightarrow \infty$  if the reproducing kernel of  $\mathcal{H}$  is strictly positive-definite. The path colored in red represents a data-driven approximation  $f_{\tau,\ell,n}$  of  $f_{\tau,\ell}$  (the empirical target function in Definition 9), obtained by replacing the invariant measure  $\mu$  by the sampling measure  $\mu_n$  associated with a training dataset consisting of  $n$  samples, and the Koopman operator  $U^\tau$  by the shift operator  $U_n^q$  on the  $n$ -dimensional Hilbert space  $L^2(\mu_n)$ . Here,  $\tau = q \Delta t$ , where  $\Delta t$  and  $q$  are the sampling interval and number of timesteps, respectively. Note that there is no path connecting the Koopman operator on  $C(\Omega)$  with the shift operator on  $L^2(\mu_n)$  as in the  $L^2(\mu)$  case since the dynamical flow is singular with respect to the sampling measure. The estimator  $f_{\tau,\ell,n}$  exhibits a sampling error relative to  $f_{\tau,\ell}$  (indicated by dashed red arrows) which vanishes almost surely as  $n \rightarrow \infty$  by ergodicity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

respectively, so that  $x_t = X_t(\omega)$  and  $y_t = Y_t(\omega)$  for some  $\omega \in \Omega$ . The space  $\mathcal{Y}$  is assumed to be a Hilbert space over the complex numbers, whose inner product,  $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$ , is taken to be conjugate-linear in its first argument. Note that we do not require that the space  $\mathcal{X}$  be linear.

The task of forecasting is to produce a measurable function  $f_\tau : \mathcal{X} \rightarrow \mathcal{Y}$  for any given lead time  $\tau \geq 0$ , referred to as the target function, such that  $f_\tau \circ X_t$  approximates  $Y_{t+\tau}$ . A heuristic for selecting such an approximation is the variational approach, wherein  $f_\tau$  is viewed as a minimizer of some global measure of error. The mean square error is a common such functional, given, as we will see below, its connection to Hilbert space theory. In particular, we regard  $Y_t$  as an element of the space  $\mathbb{L}^2(\mu)$  of functions  $\Omega \rightarrow \mathcal{Y}$  that are square-integrable with respect to  $\mu$ . The target function  $f_\tau$ , meanwhile, is sought in the space  $\mathbb{L}^2(\mu_{X_t})$  of functions  $f : \mathcal{X} \rightarrow \mathcal{Y}$  that are square-integrable with respect to  $\mu_{X_t}$ , where  $\mu_{X_t}$  is the pushforward of  $\mu$  along  $X_t$  (i.e.,  $\mu_{X_t}(S) = \mu(X_t^{-1}(S))$  for all  $S \in \Sigma_{\mathcal{X}}$ ). This implies that  $f \circ X_t$  is a square-integrable function in  $\mathbb{L}^2(\mu)$ .

In what follows,  $L^2(\mu)$  will denote the Hilbert space of equivalence classes of functions in  $\mathbb{L}^2(\mu)$  taking  $\mu$ -a.e. equal values, equipped with the standard inner product  $\langle g_1, g_2 \rangle_{L^2(\mu)} = \int_{\Omega} \langle g_1(\omega), g_2(\omega) \rangle_{\mathcal{Y}} d\mu(\omega)$ . We define the  $L^2(\mu_{X_t})$  Hilbert spaces associated with  $\mathbb{L}^2(\mu_{X_t})$  analogously. As is customary, we will oftentimes identify functions in  $\mathbb{L}^2$  with their corresponding  $L^2$  equivalence classes, but for the purpose of constructing concrete

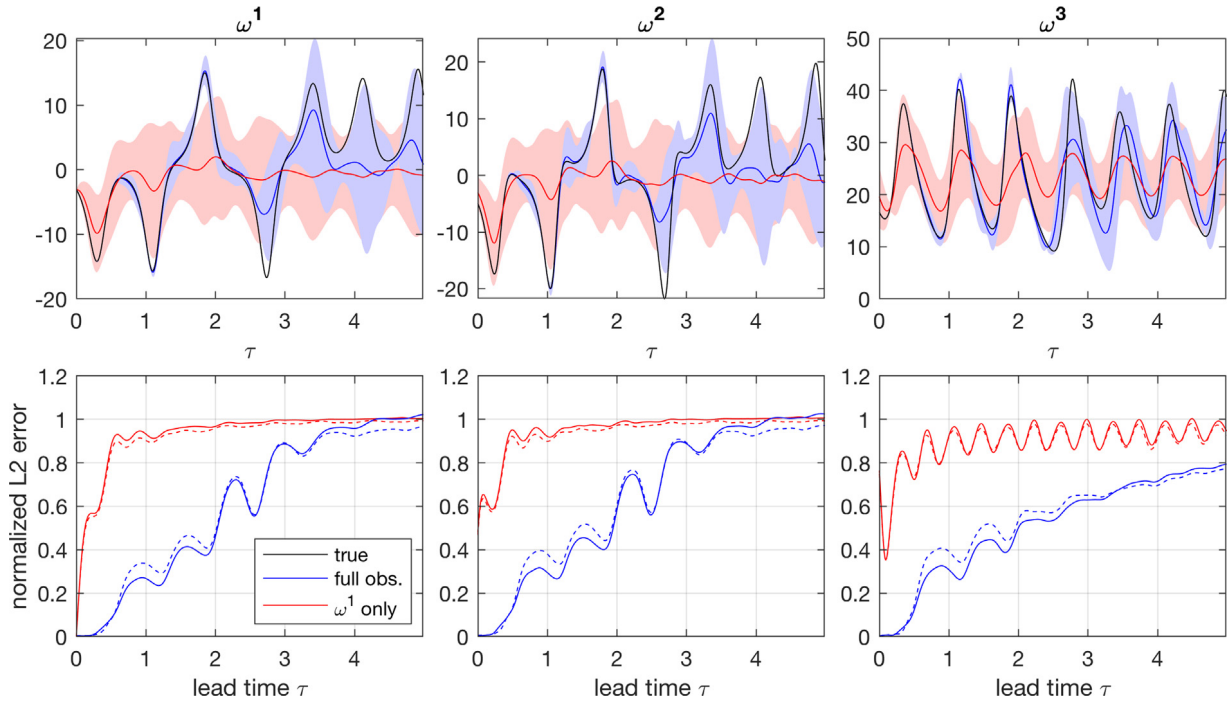
target functions we will keep elements of these spaces distinct. The mean square error of the target function  $f_\tau$ , given a lead time  $\tau \geq 0$ , may then be defined as

$$\|f_\tau \circ X_t - Y_{t+\tau}\|_{L^2(\mu)}^2 = \int_{\Omega} \|f_\tau \circ X_t(\omega) - Y_{t+\tau}(\omega)\|_{\mathcal{Y}}^2 d\mu(\omega).$$

**Dynamical system framework.** A dynamical system on the space  $\Omega$  is represented by a semigroup of measurable maps,  $\{\Phi^t : \Omega \rightarrow \Omega\}_{t \geq 0}$ , which evolve an initial state  $\omega_0$  to a new state  $\omega_t$ . The function  $X_t$  may then be represented by  $X \circ \Phi^t$ , where  $X : \Omega \rightarrow \mathcal{X}$ . The response function  $Y_{t+\tau}$  can be similarly broken up, but with the added step of using the flow map semigroup properties to split up  $\Phi^{t+\tau}$  into  $\Phi^\tau \circ \Phi^t$ , resulting in the expression  $Y_{t+\tau} = (Y \circ \Phi^\tau) \circ \Phi^t$ , where  $Y : \Omega \rightarrow \mathcal{Y}$ . It is frequently useful to express the composition  $Y \circ \Phi^\tau$  as the act of applying an operator  $U^\tau$ , known as the Koopman operator [36], on measurable  $\mathcal{Y}$ -valued functions on  $\Omega$ , so that  $U^\tau Y = Y \circ \Phi^\tau$ . Note that, unlike  $\Phi^\tau$ ,  $U^\tau$  is an intrinsically linear operator.

Henceforth, we will assume that the dynamical system is measure-preserving; that is, the pushforward measure of  $\mu$  along  $\Phi^t$ , denoted by  $\mu_t$ , is constant with respect to time, so that we may write  $\mu_t = \mu$  for all times  $t \geq 0$ . With such an assumption, the Koopman operator on measurable functions lifts to an isometry on  $L^2(\mu)$ , which we will denote using the same symbol  $U^\tau$ . Moreover, the mean square error is independent of the initialization time  $t$ , and is expressed as

$$\|f_\tau \circ X - U^\tau Y\|_{L^2(\mu)}^2. \quad (1)$$



**Fig. 2.** Results of KAF applied to prediction of the components of the state vector  $\omega = (\omega^1, \omega^2, \omega^3) \in \mathbb{R}^3$  of the L63 system, under full (blue lines) and partial (red lines) observations. In the fully observed case, the covariate  $X$  is the identity map on  $\Omega = \mathbb{R}^3$ . In the partially observed case,  $X(\omega) = \omega^1$  is the projection to the first coordinate. Both cases utilize training datasets of  $n = 64,000$  samples  $(x_j, y_j)$ , with  $x_j = X(\omega_j)$ ,  $y_j = Y(\omega_j)$ , obtained from the same underlying L63 trajectory  $\omega_j = \Phi^{(j-1)\Delta t}(\omega_1) \in \Omega$  at a sampling interval of  $\Delta t = 0.01$  natural time units. Top panels: True evolution  $U^\tau Y(\omega)$  of the response  $Y(\omega) = \omega^j$  (black lines) and forecast trajectories  $f_{\tau, \ell, n}(x)$  (solid blue and red lines) as a function of lead time  $\tau$ , obtained via the empirical target functions in Definition 9 with  $\ell = 3000$  (full observations) or  $\ell = 1000$  (partial observations). The forecasts are initialized from a fixed initial condition  $x = X(\omega)$  in the verification dataset. Shaded regions show error bounds obtained by adding  $\pm \varepsilon_{\tau, \ell, n}(x)$  to the forecast trajectories, where  $\varepsilon_{\tau, \ell, n}(x)$  is a KAF-derived estimate of the conditional standard deviation, given by (28) using the same training parameters as the  $Y$  forecasts. Bottom panels: Root mean square error (RMSE) as a function of  $\tau$  determined from a verification dataset of  $\tilde{n} = 64,000$  samples (solid lines). The RMSE is normalized by the empirical standard deviation of  $\omega^j$  on the training dataset. Dashed lines show the normalized RMS value of the estimated error  $\varepsilon_{\tau, \ell, n}$ . The agreement between actual and estimated errors indicates that  $\varepsilon_{\tau, \ell, n}$  provides useful uncertainty quantification. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Conditional expectation.** The random variable  $X$  induces a sub- $\sigma$  algebra  $\mathcal{G} \subseteq \mathcal{F}$ , defined by  $\mathcal{G} = X^{-1}(\Sigma_X)$ . This means that every function  $g : \Omega \rightarrow \mathcal{Y}$  that is measurable with respect to  $\mathcal{G}$  is such that  $g = f \circ X$  for some  $f : \mathcal{X} \rightarrow \mathcal{Y}$ . Thus,  $\mathcal{G}$ -measurable functions can be thought of as being “coarser” than  $\mathcal{F}$ -measurable functions, in the sense that they necessarily take constant values on subsets of  $\Omega$  where  $X$  is constant. We will denote the Hilbert subspace of  $L^2(\mu)$  consisting of  $\mathcal{G}$ -measurable equivalence classes of functions by  $L_X^2(\mu)$ . The composition map by  $X$ , i.e.,  $f \mapsto f \circ X$ , then describes an isometric embedding  $\mathcal{E} : L^2(\mu_X) \rightarrow L^2(\mu)$ , with range  $L_X^2(\mu)$ . It is a consequence of the Radon–Nikodym theorem [56, Chapter 5] that for  $U^\tau Y \in L^2(\mu)$ , there exists a unique  $\mathcal{G}$ -measurable element  $Z_\tau \circ X \in L^2(\mu)$ , such that for all  $g \in L_X^2(\mu)$ ,

$$\langle g, U^\tau Y \rangle_{L^2(\mu)} = \langle g, Z_\tau \circ X \rangle_{L^2(\mu)}. \quad (2)$$

It follows from this property that  $Z_\tau \circ X$  is the unique element in  $L^2(\mu)$ , or, equivalently, that  $Z_\tau$  the unique element in  $L^2(\mu_X)$ , that minimizes the mean square error in (1). The composition  $Z_\tau \circ X$  is known as the *conditional expectation* of  $U^\tau Y$  given the random variable  $X$ , and will be denoted by  $\mathbb{E}[U^\tau Y | X]$ . It follows from the Hilbert space projection theorem that  $\mathbb{E}[U^\tau Y | X]$  has the geometrical interpretation of being the orthogonal projection of  $U^\tau Y$  onto  $L_X^2(\mu)$ . That is,

$$\mathbb{E}[U^\tau Y | X] = \Pi_X U^\tau Y, \quad (3)$$

where  $\Pi_X : L^2(\mu) \rightarrow L^2(\mu)$  is the orthogonal projection mapping into  $L_X^2(\mu)$ . Because the conditional expectation lies in  $L_X^2(\mu)$ , there exists a unique observable  $Z_\tau \in L^2(\mu_X)$  on covariate space such

that

$$\mathbb{E}[U^\tau Y | X] = \mathcal{E} Z_\tau = Z_\tau \circ X.$$

This leads to the notion of the regression function, defined below through the adjoint map  $\mathcal{E}^* : L^2(\mu) \rightarrow L^2(\mu_X)$  with  $\ker \mathcal{E}^* = (L_X^2(\mu))^\perp$ .

**Definition 1 (Regression function).** The regression function at lead time  $\tau$  associated with the response  $Y$  and covariate  $X$  is the  $L^2(\mu_X)$  observable

$$Z_\tau = \mathcal{E}^* U^\tau Y = \mathcal{E}^* \mathbb{E}[U^\tau Y | X].$$

By virtue of its error-minimizing properties, it is natural to seek forecasting algorithms producing target functions that consistently approximate  $Z_\tau$ . In the ensuing sections, we will show that under suitable ergodicity assumptions, KAF naturally produces such consistent estimators of the regression function from time-ordered samples of  $X$  and  $Y$  along a dynamical trajectory, without requiring prior knowledge of the underlying equations of motion.

## 2.2. Hypothesis spaces

**Learning framework.** Constructing the target function requires distinguishing between the spaces  $L^2(\mu_X)$  and  $L^2(\mu)$ , which we do by way of the linear map  $\iota : L^2(\mu_X) \rightarrow L^2(\mu)$  that associates each concrete function  $f$  to its equivalence class  $\iota f$ . The mean square error is then represented with the functional



$\mathcal{E}_\tau : L^2(\mu_X) \rightarrow \mathbb{R}$ , known as the generalization error in machine learning contexts [30], defined by

$$\mathcal{E}_\tau(f) := \|lf \circ X - U^\tau Y\|_{L^2(\mu)}^2. \quad (4)$$

The Hilbert space structure of  $L^2(\mu)$ , as well as the error-minimizing property of the conditional expectation  $Z_\tau \circ X$ , allows the generalization error to be decomposed as

$$\mathcal{E}_\tau(f) = \mathcal{A}_\tau(f) + \sigma_\tau,$$

where  $\mathcal{A}_\tau(f)$  is the excess generalization error,

$$\mathcal{A}_\tau(f) = \|lf - Z_\tau\|_{L^2(\mu_X)}^2, \quad (5)$$

and  $\sigma_\tau$  is the error intrinsic to the system and choice of covariate and response functions,

$$\sigma_\tau = \|Z_\tau \circ X - U^\tau Y\|_{L^2(\mu)}^2. \quad (6)$$

Since  $\sigma_\tau$  does not depend on  $f$ , minimizing  $\mathcal{E}_\tau$  is equivalent to minimizing  $\mathcal{A}_\tau$ .

**Hypothesis space.** Constraints on the search for a minimizer of  $\mathcal{A}_\tau$  are characterized in terms of a hypothesis space  $\mathcal{H} \subseteq L^2(\mu_X)$  of functions. When the image  $H := \iota\mathcal{H}$  is a closed and convex subset of the Hilbert space  $L^2(\mu_X)$ , then there exists a unique  $g \in H$  such that  $\inf_{h \in H} \|h - Z_\tau\|_{L^2(\mu_X)} = \|g - Z_\tau\|_{L^2(\mu_X)}$ . Consequently, there exists  $f \in \mathcal{H}$  for which  $lf = g$ , and thus  $\inf_{f \in \mathcal{H}} \|lf - Z_\tau\|_{L^2(\mu_X)} = \|g - Z_\tau\|_{L^2(\mu_X)}$ . A sufficient condition for uniqueness of  $f$  is that  $\iota : \mathcal{H} \rightarrow L^2(\mu_X)$  is an injection.

**The pseudoinverse.** Assuming that  $H$  is closed and convex in  $L^2(\mu_X)$ , so that there exists a well-defined orthogonal projection map  $\Pi_H : L^2(\mu_X) \rightarrow L^2(\mu_X)$  mapping into  $H$ , the excess generalization error may be decomposed as

$$\mathcal{A}_\tau(f) = \|lf - \Pi_H Z_\tau\|_{L^2(\mu_X)}^2 + \|(I - \Pi_H)Z_\tau\|_{L^2(\mu_X)}^2.$$

The minimizer of  $\mathcal{A}_\tau$  over the hypothesis space  $\mathcal{H}$ , therefore, is found by minimizing the norm of  $lf - \Pi_H Z_\tau$ . When  $\iota$  is injective on  $\mathcal{H}$ , then the restriction  $\iota|_{\mathcal{H}}$  of  $\iota$  onto  $\mathcal{H}$  is invertible as a map  $\iota|_{\mathcal{H}} : \mathcal{H} \rightarrow H$ . In such a case, the unique minimizer of  $\mathcal{A}_\tau$  in  $\mathcal{H}$  is expressible as

$$f_{\tau, \mathcal{H}} = (\iota|_{\mathcal{H}})^{-1} \Pi_H Z_\tau, \quad (7)$$

and satisfies

$$\mathcal{A}_\tau(f_{\tau, \mathcal{H}}) = \|(I - \Pi_H)Z_\tau\|_{L^2(\mu_X)}^2. \quad (8)$$

We refer to this minimizer as the ideal target function since it does not depend on any training data.

**Definition 2 (Ideal target function).** The ideal target function  $f_{\tau, \mathcal{H}}$  at lead time  $\tau$  associated with the response  $Y$  and hypothesis space  $\mathcal{H}$  is the minimizer of the excess generalization error functional  $\mathcal{A}_\tau$  over  $\mathcal{H}$ , given by (7).

We shall refer to the map  $T : L^2(\mu_X) \rightarrow \mathcal{H}$ , with  $T = (\iota|_{\mathcal{H}})^{-1} \Pi_H$ , as the pseudoinverse of  $\iota$  on  $\mathcal{H}$ , in analogy with the Moore–Penrose pseudoinverse of bounded, closed-range linear maps between Hilbert spaces [57]. In particular, note that  $Tlf = f$  for every  $f \in \mathcal{H}$  and  $Tg = 0$  for every  $g \in H^\perp$ , which shows that  $T$  reduces to the Moore–Penrose pseudoinverse  $\iota^+$  of  $\iota$  if  $\mathcal{H}$  is a Hilbert space. In that case, the excess generalization error of the target function  $f_{\mathcal{H}}$  in (7) is due to the component  $(I - \Pi_H)Z_\tau$  of  $Z_\tau$  in the orthogonal complement of  $H$  in  $L^2(\mu_X)$ . See Appendix A.1 for additional details on the Moore–Penrose pseudoinverse.

**Ambient Hilbert space.** Explicit representations of  $T$  depend on the choice of  $\mathcal{H}$ , and among the many such possible choices, in KAF we focus on the case where  $\mathcal{H}$  is a finite-dimensional subspace of an ambient Hilbert space  $\mathcal{K}$  that  $\iota$  compactly embeds into  $L^2(\mu_X)$ . As  $\iota$  is a compact operator between Hilbert spaces, its adjoint  $\iota^* : L^2(\mu_X) \rightarrow \mathcal{K}$  is well-defined and compact. Consequently, the self-adjoint operator  $G := \iota^* \iota : L^2(\mu_X) \rightarrow L^2(\mu_X)$  is also compact. The spectral theorem for compact, self-adjoint operators thus guarantees the existence of an orthonormal basis  $\{\phi_i\}_{i=1}^\infty$  of  $L^2(\mu)$  consisting of eigenfunctions of  $G$ , with non-negative corresponding eigenvalues  $\lambda_i$  since  $\iota^*$  is positive.

**Remark 3.** As we will see in Section 2.3, under natural assumptions,  $\mathcal{K}$  has the structure of an RKHS. In that case, the adjoint operator  $\iota^*$  becomes an integral operator associated with the reproducing kernel of  $\mathcal{K}$ . Moreover, under appropriate continuity assumptions, the orthonormal functions  $\psi_i$  correspond to Mercer feature vectors, used, e.g., for unsupervised learning in kernel principal component analysis (KPCA) [58]. This perspective of expressing integral operators arising in learning problems as adjoints of restriction maps was also adopted by Rosasco et al. [59] in a study on spectral approximation of integral operators.

By convention, we order the eigenvalues  $\lambda_i$  in decreasing order, so that the sequence  $\lambda_1, \lambda_2, \dots$  only accumulates at zero by compactness of  $G$ . Defining

$$\psi_i = \iota^* \phi_i / \lambda_i^{1/2} \quad (9)$$

for each  $\lambda_i > 0$ , and choosing  $\ell \in \mathbb{N}$  such that  $\lambda_\ell > 0$ , we then select as a hypothesis space the  $\ell$ -dimensional subspace  $\mathcal{H}_\ell \subseteq \mathcal{K}$ , where

$$\mathcal{H}_\ell = \text{span}\{\psi_1, \dots, \psi_\ell\}. \quad (10)$$

It follows from orthonormality of the  $\phi_i$  and their definition in (9) that the  $\psi_i$  form an orthonormal set in  $\mathcal{K}$ , i.e.,  $\langle \psi_i, \psi_j \rangle_{\mathcal{K}} = \delta_{ij}$ . Here,  $\langle \cdot, \cdot \rangle_{\mathcal{K}}$  is the inner product of  $\mathcal{K}$ , taken conjugate-linear in its first argument. Moreover, the  $\psi_i$  are orthonormal eigenfunctions of the operator  $\tilde{G} := \iota^* \iota$  on  $\mathcal{K}$ , corresponding to the same eigenvalues  $\lambda_i$ ,  $\tilde{G}\psi_i = \lambda_i \psi_i$ . In fact, the square roots of the nonzero eigenvalues  $\lambda_i$  are the singular values of the compact operator  $\iota^*$ , and the corresponding  $\psi_i$  and  $\phi_i$  are left and right singular vectors, respectively, i.e.,

$$\iota^* h = \sum_{i: \lambda_i > 0} \psi_i \lambda_i^{1/2} \langle \phi_i, h \rangle_{L^2(\mu_X)}, \quad \forall h \in L^2(\mu_X).$$

With these definitions, it follows that  $\Pi_{\mathcal{H}_\ell}$ , where  $\mathcal{H}_\ell := \iota\mathcal{H}_\ell$ , is the  $L^2(\mu_X)$ -orthogonal projection with range  $\text{span}\{\phi_1, \dots, \phi_\ell\}$ . As for the inverse  $(\iota|_{\mathcal{H}_\ell})^{-1}$ , it acts as

$$(\iota|_{\mathcal{H}_\ell})^{-1} \phi_i = (\iota|_{\mathcal{H}_\ell})^{-1} \iota^* \phi_i / \lambda_i = \psi_i / \lambda_i^{1/2}$$

on each eigenfunction  $\phi_i$  corresponding to a nonzero eigenvalue  $\lambda_i$ . Consequently, by expanding  $Z_\tau \in L^2(\mu_X)$  as  $Z_\tau = \sum_{i=1}^\infty \alpha_i(\tau) \phi_i$ , where

$$\alpha_i(\tau) = \langle \phi_i \circ X, U^\tau Y \rangle_{L^2(\mu)}, \quad (11)$$

the target function  $f_{\tau, \ell}$  from Definition 2 associated with  $\mathcal{H}_\ell$  is given by

$$f_{\tau, \ell} = T_\ell Z_\tau = \sum_{i=1}^\ell \frac{\alpha_i(\tau)}{\lambda_i^{1/2}} \psi_i \quad (12)$$

where  $T_\ell : L^2(\mu_X) \rightarrow \mathcal{H}_\ell$  is the pseudoinverse operator from (7).

Considering now the image  $K = \iota\mathcal{K}$  of the ambient Hilbert space under  $L^2(\mu_X)$  inclusion, one can verify that it can be characterized as the subspace  $K = \left\{ \sum_{i: \lambda_i > 0} c_i \phi_i \in L^2(\mu_X) : \sum_{i: \lambda_i > 0} |c_i|^2 / \lambda_i < \infty \right\}$ . The following is then a direct consequence of the definition of the Moore–Penrose pseudoinverse in Definition 18.

**Lemma 4.** The operator  $\tilde{T} : D(\tilde{T}) \rightarrow \mathcal{K}$ , with dense domain  $D(\tilde{T}) = K \oplus \ker \iota \subseteq L^2(\mu_X)$ , defined as  $\tilde{T}f = \sum_{i:\lambda_i > 0} \langle \phi_i, f \rangle_{L^2(\mu_X)} \psi_i / \lambda_i^{1/2}$ , is a closed-range operator whose pseudoinverse  $\tilde{T}^+ : \mathcal{K} \rightarrow L^2(\mu_X)$  is equal to  $\iota$ . Moreover,  $\tilde{T}$  is equal to the pseudoinverse of  $\iota$ , and by (A.1) we have,

$$\tilde{T}f = \iota^* G^+ f, \quad \forall f \in D(G^+).$$

Lemma 4 shows that  $\tilde{T}$  maps each  $L^2(\mu_X)$  equivalence class in its domain to an everywhere-defined function in  $\mathcal{K}$ , and whenever  $f$  lies in  $K$ ,  $\tilde{T}f = f$ . That is,  $\tilde{T}$  is an extension operator, mapping  $f \in K$  to a representative in  $\mathcal{K}$ . Note that  $\tilde{T}$  is necessarily an unbounded operator if  $K$  is infinite-dimensional, and, moreover, if  $G$  is strictly positive-definite (so all  $\lambda_i$  are strictly positive), then  $K$  is a proper, dense subspace of  $L^2(\mu_X)$ . In fact,  $\tilde{T}$  is closely related to the Nyström extension operator employed in applications such as function interpolation and kriging [e.g., 60,59]. Noticing from (12) that  $T_\ell = \tilde{T}|_{H_\ell}$ , we may therefore interpret the target function  $f_{\tau,\ell}$  as a spectrally truncated Nyström extension of  $Z_\tau$ , which is well defined even if  $Z_\tau$  does not lie in  $K$ . In fact, it follows from (7) that  $\iota T_\ell$  is equal to the  $L^2(\mu_X)$  orthogonal projection  $\Pi_{H_\ell}$ . Moreover, we have:

**Lemma 5.** As  $\ell \rightarrow \infty$ ,  $\iota T_\ell$  converges strongly to the orthogonal projection  $\Pi_{\bar{K}} : L^2(\mu_X) \rightarrow L^2(\mu_X)$  onto the  $L^2(\mu_X)$ -closure of  $K$ ; that is,

$$\lim_{\ell \rightarrow \infty} \iota T_\ell f = \Pi_{\bar{K}} f, \quad \forall f \in L^2(\mu_X).$$

Lemma 5 indicates that even if the target functions  $f_{\tau,\ell} = T_\ell Z_\tau$  do not have a limit in the ambient space  $\mathcal{K}$ , they have a  $L^2(\mu_X)$  limit. In particular, if it can be arranged that  $K$  is a dense subspace of  $L^2(\mu_X)$ ,  $\Pi_{\bar{K}} = \text{Id}$ , and the  $f_{\tau,\ell}$  converge in  $L^2(\mu_X)$  norm to the regression function  $Z_\tau$ . Ensuring that  $\mathcal{K}$  is an empirically constructible space with dense image  $K = \iota \mathcal{K}$  in  $L^2(\mu_X)$  is a key consideration in KAF, which will occupy us in the ensuing sections.

### 2.3. Reproducing kernel Hilbert spaces

For the remainder of the paper, we will restrict attention to the case that the response variable  $Y$  is complex-valued, i.e.,  $\mathcal{Y} = \mathbb{C}$ . In this setting, the ambient Hilbert space  $\mathcal{K}$  naturally acquires the structure of an RKHS [54,53], as we describe below.

**Definition 6 (RKHS).** For each point  $x \in \mathcal{X}$ , let  $L_x : \mathcal{K} \rightarrow \mathbb{C}$  be the evaluation functional on the ambient Hilbert space, defined by  $L_x f = f(x)$ . The space  $\mathcal{K}$  is said to be an RKHS if  $L_x$  is bounded, and therefore continuous, at every  $x \in \mathcal{X}$ .

It is a known fact that no unbounded linear functional on a Banach space can be constructed without the axiom choice. Therefore, all explicitly constructible Hilbert spaces of complex-valued functions are necessarily RKHSs. Consequently, all explicitly representable target functions  $f_{\tau,\ell}$  from (7) necessarily lie in an RKHS. Note that by boundedness of  $L_x$  at every  $x \in \mathcal{X}$ , convergence of two functions in RKHS norm implies pointwise convergence on  $\mathcal{X}$ .

**Basic properties of RKHSs.** It follows from the Riesz representation theorem that for every  $x \in \mathcal{X}$  there exists a function  $k_x \in \mathcal{K}$  such that

$$f(x) = L_x f = \langle k_x, f \rangle_{\mathcal{K}}, \quad \forall f \in \mathcal{K}.$$

The above is known as the reproducing property of  $\mathcal{K}$ . The *reproducing kernel*  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$  of  $\mathcal{K}$  is then defined as the bivariate function

$$k(x_1, x_2) = \langle k_{x_1}, k_{x_2} \rangle_{\mathcal{K}}.$$

It follows from the defining properties of inner products that  $k$  is (i) conjugate-symmetric, i.e.,  $k(x_1, x_2) = k(x_2, x_1)^*$  for all  $x_1, x_2 \in \mathcal{X}$ ; and (ii) positive-definite, i.e., for all  $x_1, \dots, x_m \in \mathcal{X}$  and  $a_1, \dots, a_m \in \mathbb{C}$ ,

$$\sum_{i=1}^m \sum_{j=1}^m a_i^* a_j k(x_i, x_j) \geq 0. \quad (13)$$

Conversely, the Moore–Aronszajn theorem [61] states that for any conjugate-symmetric, positive-definite kernel function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ , there exists a unique RKHS on  $\mathcal{X}$  for which  $k$  is the reproducing kernel. Thus, there is a one-to-one correspondence between kernels and RKHSs.

Let  $\rho : \Sigma_{\mathcal{X}} \rightarrow [0, \infty]$  be any measure such that there exists a compact embedding  $\iota_\rho$  of the RKHS  $\mathcal{K}$  into  $L^2(\rho)$ . The practical utility of RKHSs manifests in the adjoint  $\iota_\rho^* : L^2(\mathcal{X}) \rightarrow \mathcal{K}$  being representable in terms of the kernel as

$$\iota_\rho^* f(x) = \langle k_x, \iota_\rho^* f \rangle_{\mathcal{K}} = \langle \iota_\rho k_x, f \rangle_{L^2(\rho)} = \int_{\mathcal{X}} k^*(x, \cdot) g d\rho,$$

where  $f$  is any element of  $L^2(\rho)$ . Thus, the adjoint of the embedding of the RKHS  $\mathcal{K}$  into  $L^2(\rho)$  is a compact integral operator on the latter space. Similarly,  $G_\rho := \iota_\rho \iota_\rho^*$  is a positive-semidefinite, self-adjoint, compact integral operator on  $L^2(\rho)$ .

**Definition 7 (Particular classes of kernels).** We will say that a positive-definite kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$  is:

- Strictly positive-definite if the inequality in (13) is strict whenever the  $x_i$  are all distinct and at least one of the  $a_i$  is nonzero.
- $L^2(\rho)$ -strictly-positive if  $G_\rho$  is a strictly-positive operator. In that case,  $K_\rho := \iota_\rho \mathcal{K}$  is a dense subspace of  $L^2(\rho)$ .
- $L^2(\rho)$ -Markov if  $G_\rho$  is a Markov operator, i.e., (i)  $\int_{\mathcal{X}} G_\rho f d\rho = \int_{\mathcal{X}} f d\rho$  for all  $f \in L^2(\rho)$ ; and (ii)  $G_\rho f = f$  if  $f$  is constant. As a result, the leading largest eigenvalue of  $G_\rho$  is equal to 1, and the corresponding eigenspace contains constant functions.
- $L^2(\rho)$ -Markov ergodic if it is  $L^2(\rho)$ -Markov and the eigenvalue 1 of  $G_\rho$  is simple.

In the case  $\rho = \mu_X$ , we will abbreviate  $\iota_{\mu_X} = \iota$  and  $G_{\mu_X} = G$  as in Section 2.2. The evaluation of the target function from (7) at a point  $x \in \mathcal{X}$  is then expressible as

$$f_{\tau,\ell}(x) = \sum_{i=1}^{\ell} \frac{\alpha_i(\tau)}{\lambda_i} \langle \iota k_x, \phi_i \rangle_{L^2(\mu_X)} = \sum_{i=1}^{\ell} \frac{\alpha_i(\tau)}{\lambda_i^{1/2}} \psi_i(x). \quad (14)$$

**Topological framework and Mercer kernels.** Henceforth, we will assume that  $\mathcal{X}$  has the structure of a metric space, equipped with its Borel  $\sigma$ -algebra  $\Sigma_{\mathcal{X}}$ , and  $\mu_X$  is a Borel probability measure with compact support  $\mathcal{X}_\mu \subseteq \mathcal{X}$ . Given any subset  $S$  of  $\mathcal{X}$ , we will use the notation  $\mathcal{K}(S)$  to represent the RKHS on  $S$  with reproducing kernel  $k|_{S \times S}$ . Note that  $\mathcal{K}(S)$  embeds naturally and isometrically into  $\mathcal{K}$ , so we may view it as a subspace of the latter space. We also let  $C(S)$  be the space of complex-valued continuous functions on  $S$ , and  $C_b(S)$  the Banach space of bounded functions in  $C(S)$ , equipped with the uniform norm. Note that  $C(S) = C_b(S)$  if  $S$  is compact.

In this setting, continuous kernel functions on  $\mathcal{X}$ , also known as Mercer kernels, have the property that their associated RKHS is a subset of  $C(\mathcal{X})$  [53]. Moreover, for any compact set  $S \subseteq \mathcal{X}$ , the embedding  $\mathcal{K}(S) \hookrightarrow C(S)$  is bounded. If, in addition,  $S$  is the support  $\mathcal{X}_\rho$  of a finite Borel measure  $\rho$  on  $\mathcal{X}$ ,  $C(\mathcal{X}_\rho)$  embeds into  $L^2(\rho)$  via a bounded linear map, and thus  $\iota_\rho : \mathcal{K}(\mathcal{X}_\rho) \rightarrow L^2(\rho)$  is a bounded, injective operator. It also follows by continuity of  $k$  and compactness of  $\mathcal{X}_\rho$  that  $G_\rho = \iota_\rho \iota_\rho^*$  is a trace-class (and therefore compact) operator, with trace norm equal to

$\text{tr } G_\rho = \int_{\mathcal{X}} k(x, x) d\rho(x)$  [62]. In particular, the compactness of  $G_\rho$  is equivalent to  $\iota_\rho$  being compact.

Mercer's theorem [54, Section 11.4] also states that for any  $x, x' \in X_\rho$  the kernel  $k(x, x')$  can be expressed through the series expansion,

$$k(x, x') = \sum_{i: \lambda_i > 0} \psi_i^*(x) \psi_i(x'), \quad (15)$$

where the  $\psi_i$  are orthonormal functions in  $\mathcal{K}$  associated with eigenvalue  $\lambda_i$  of  $G_\rho$ , defined analogously to (9), and convergence of the sum over  $i$  is uniform on  $X_\rho \times X_\rho$ . This result then implies that the restrictions of the  $\psi_i$  on  $X_\rho$  form an orthonormal basis of  $\mathcal{K}(X_\rho)$  (as opposed to merely an orthonormal set). It can also be shown that every strictly positive-definite Mercer kernel is  $L^2(\rho)$ -strictly positive for any compactly supported, finite Borel measure  $\rho$ . See [63] for a detailed study on the relationships between the RKHSs associated with different kernel classes (including those in Definition 7) and spaces of functions and measures.

By virtue of the above properties, Mercer kernels provide a convenient practical means of generating hypothesis spaces that are compactly embeddable into  $L^2(\mu_X)$ , as required for the hypothesis spaces in Section 2.2. Note that the target function in (14) associated with a Mercer kernel is an RKHS (and thus continuous) function defined on the whole of  $\mathcal{X}$ , but its behavior outside of the support  $X_\mu$  makes no contribution to the excess generalization error from (5) determined through the  $L^2(\mu_X)$  norm.

**Remark 8.** The Mercer expansion in (15) allows one to evaluate inner products between distinguished elements of the RKHS, namely the kernel sections  $k_x$  simply by evaluation of a known kernel function,  $\langle k_x, k_{x'} \rangle_{\mathcal{K}} = k(x, x')$  (i.e., the left-hand side of (15)), without having to compute a potentially infinite set of basis vectors for the space (i.e., the eigenfunctions  $\psi_i$  on the right-hand side). This well known “kernel trick” is employed in a variety of learning techniques, including kernel KRR and SVMs [22]. In contrast, KAF/KPCR methods incur a potentially significant computational cost associated with computing (in the training phase) and evaluating (in the prediction phase) a set of orthonormal basis functions  $\{\psi_1, \dots, \psi_\ell\}$  with  $\ell \ll n$ , with the benefit of controlling the regularity of the target functions through the spectral truncation parameter  $\ell$ . As we will see in Section 3.2, this is an effective means of controlling the sample error, allowing the method to operate stably in environments with small training datasets.

#### 2.4. Data-driven target function

We are now ready to construct the empirical target function employed in KAF. In this construction we consider a standard supervised learning scenario, where we have access to a training dataset consisting of pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , where  $x_j = X(\omega_j)$  and  $y_j = Y(\omega_j)$  are the values of the covariate and response variables, respectively, on an (unknown) collection of points  $\omega_1, \dots, \omega_n$  on the state space  $\Omega$ , sampled along a single dynamical trajectory

$$\omega_j = \Phi^{t_j}(\omega_1), \quad t_j = (j-1)\Delta t, \quad (16)$$

at a fixed sampling interval  $\Delta t > 0$ . Alternatively, the  $(x_j, y_j)$  may be generated by an ensemble of (shorter) trajectories on  $\Omega$ , so long as the timespan of each of these trajectories is not smaller than the desired lead time  $\tau$ .

**Sampling measures.** Associated with every dataset from (16) is an empirical probability measure  $\mu_n : \mathcal{F} \rightarrow [0, 1]$ , defined as  $\mu_n = \sum_{j=1}^n \delta_{\omega_j}/n$ , where  $\delta_{\omega_j}$  is the Dirac  $\delta$ -measure supported on  $\{\omega_j\} \subset \Omega$ . Similarly, the empirical probability measure  $\mu_{X,n} : \Sigma_{\mathcal{X}} \rightarrow [0, 1]$  is defined as  $\mu_{X,n} = \sum_{j=1}^n \delta_{x_j}/n$ . Intuitively, we view  $\mu_n$  and  $\mu_{X,n}$  as empirical approximations to  $\mu$  and  $\mu_X$ , respectively; a connection which will be made precise in Section 3.

Next, as empirical analogs of  $L^2(\mu)$  and  $L^2(\mu_X)$ , we employ the Hilbert spaces  $L^2(\mu_n)$  and  $L^2(\mu_{X,n})$ , consisting of equivalence classes of complex-valued, measurable functions on  $\Omega$  and  $X$  having common values at the sampled points  $\omega_j$  and  $x_j$ , respectively. As Hilbert spaces,  $L^2(\mu_n)$  and  $L^2(\mu_{X,n})$  have dimension at most  $n$  (with equality if all  $\omega_j$  and  $x_j$  are distinct, respectively), and can be homomorphically embedded into  $\mathbb{C}^n$ , equipped with the normalized dot product  $\mathbf{f} \cdot \mathbf{g}/n$ . That is, for every measurable function  $f : \Omega \rightarrow \mathbb{C}$ , the corresponding  $L^2(\mu_n)$  equivalence class can be represented by a column vector  $\mathbf{f} \in L^2(\mu_n)$  with  $\mathbf{f} = [f(\omega_1), \dots, f(\omega_n)]^T$ , storing in its components the values of  $f$  on  $\omega_j$ . Elements of  $L^2(\mu_{X,n})$  are represented by  $\mathbb{C}^n$  vectors in a similar manner, while operators on  $L^2(\mu_n)$  and  $L^2(\mu_{X,n})$  are represented by  $n \times n$  complex matrices.

As in the case of the  $L^2(\mu)$  and  $L^2(\mu_X)$  spaces, there is an isometric embedding  $\mathcal{E}_n : L^2(\mu_{X,n}) \rightarrow L^2(\mu_n)$ , given by composition by the covariate map,  $\mathcal{E}_n \mathbf{f} = \mathbf{f} \circ X$ , whose image we denote by  $L_X^2(\mu_n) = \mathcal{E}_n L^2(\mu_{X,n})$ . We also let  $\Pi_{X,n} : L^2(\mu_n) \rightarrow L^2(\mu_n)$  be the orthogonal projection mapping into  $L_X^2(\mu_n)$ . Note that in typical applications involving distinct training data,  $\Pi_{X,n}$  is the identity map and  $\mathcal{E}_n$  is unitary, even if  $X$  is non-injective on sets of positive  $\mu$ -measure (in which case,  $\Pi_X$  is not the identity). This disparity between  $\Pi_X$  and  $\Pi_{X,n}$  highlights the risk of overfitting commonly faced by data-driven techniques, which KAF addresses by employing hypothesis spaces of significantly lower dimension than the number of training samples.

**Shift operators.** In order to parallel the construction of the target function  $f_{\tau,\ell}$  from Section 2.2, we would now like to define a Koopman operator on  $L^2(\mu_n)$ . However, an obstruction to this is that, unlike the  $L^2(\mu)$  setting associated with the invariant measure, the composition operator with respect to the dynamical flow does not lift to an operator on equivalence classes of functions on the  $L^2(\mu_n)$  spaces associated with the sampling measures. This is because the flow  $\Phi^\tau : \Omega \rightarrow \Omega$  on state space does not preserve null sets with respect to  $\mu_n$ , meaning that if  $f, f' : \Omega \rightarrow \mathbb{C}$  are measurable functions lying in the same  $L^2(\mu_n)$  equivalence class, their images  $U^\tau f = f \circ \Phi^\tau$  and  $U^\tau f' = f' \circ \Phi^\tau$  may lie in different  $L^2(\mu_n)$  equivalence classes. Nevertheless, for any  $q \in \mathbb{N}_0$ , an analogous notion to the Koopman operator  $U^q \Delta t$  on  $L^2(\mu)$  is provided by the shift operator  $U_n^q : L^2(\mu_n) \rightarrow L^2(\mu_n)$  [64], defined as

$$U_n^q f(\omega_j) = \begin{cases} f(\omega_{j+q}), & j+q \leq n, \\ 0, & \text{otherwise.} \end{cases}$$

Hereafter, we will refer to the  $\mathbb{C}^n$  vector

$$\mathbf{y}_\tau = [y_{1+q}, \dots, y_n, 0, \dots, 0]^T, \quad (17)$$

representing the response  $U_n^q \iota_n Y \in L^2(\mu_n)$  for  $\tau = q \Delta t$ , as the *analog vector*.

**Empirical error minimization.** With these definitions, and assuming throughout that  $\tau = q \Delta t$  for some  $q \in \mathbb{N}_0$ , the empirical generalization error  $\mathcal{E}_{\tau,n} : L^2(\mu_{X,n}) \rightarrow \mathbb{R}$  is given by (cf. (4))

$$\mathcal{E}_{\tau,n}(f) := \|\iota_n f \circ X - U^\tau Y\|_{L^2(\mu_n)}^2,$$

where  $\iota_n : L^2(\mu_n) \rightarrow L^2(\mu_n)$  maps each function in  $L^2(\mu_n)$  to its  $L^2(\mu_n)$  equivalence class. This functional is minimized by a unique element  $Z_{\tau,n} \in L^2(\mu_{X,n})$  analogous to the regression function



$Z_\tau$  from Sections 2.1–2.3. Moreover, we may split the empirical generalization error as

$$\mathcal{E}_{\tau,n}(f) = \mathcal{A}_{\tau,n}(f) + \sigma_{\tau,n},$$

with (cf. (5))

$$\mathcal{A}_{\tau,n}(f) = \| \iota_n f - Z_{\tau,n} \|_{L^2(\mu_{X,n})}^2, \quad \sigma_{\tau,n} = \| Z_{\tau,n} \circ X - U^\tau Y \|_{L^2(\mu_n)}^2.$$

**Empirical hypothesis space.** To construct an empirical target function, we proceed again analogously to the infinite-dimensional case in Sections 2.1–2.3. That is, we seek the minimizer of the empirical excess generalization error  $\mathcal{A}_{\tau,n}(f)$  for  $f$  lying in an  $\ell$ -dimensional empirical hypothesis space  $\mathcal{H}_{\ell,n}$ , which is chosen as a subspace of an ambient RKHS  $\mathcal{K}_n \subset C(\mathcal{X})$  associated with an empirical Mercer kernel  $k_n : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ . Note that we allow the reproducing kernel  $k_n$  to depend on  $n$  in order to be able to take advantage of the variety of normalized kernel algorithms in the literature [65–69]. Given any  $x \in \mathcal{X}$ , we shall refer to the  $\mathbb{C}^n$  vector

$$\mathbf{k}(x) = [k_n(x, x_1), \dots, k_n(x, x_n)]^T,$$

representing the  $L^2(\mu_{X,n})$  equivalence class  $\iota_n k_n(x, \cdot)$  of the kernel section  $k_n(x, \cdot) \in \mathcal{K}_n$  as the *kernel vector*.

Next, because  $\mathcal{K}_n \subseteq L^2(\mu_{X,n})$ , we can consider  $\iota_n : \mathcal{K}_n \rightarrow L^2(\mu_{X,n})$  as a (finite-rank, and thus compact) operator between Hilbert spaces, inducing the self-adjoint integral operator  $G_n := G_{\mu_{X,n}} = \iota_n \iota_n^*$  on  $L^2(\mu_{X,n})$ . The leading  $\ell$  orthonormal eigenvectors  $\phi_{1,n}, \dots, \phi_{\ell,n}$  of  $G_n$ , corresponding to positive eigenvalues  $\lambda_{1,n} \geq \dots \geq \lambda_{\ell,n}$ , respectively, induce the  $\ell$ -dimensional empirical hypothesis space  $\mathcal{H}_{\ell,n} \subseteq \mathcal{K}_n$  given by (cf. (10))

$$\mathcal{H}_{\ell,n} = \text{span}\{\psi_{1,n}, \dots, \psi_{\ell,n}\},$$

where

$$\psi_{i,n} = \iota_n^* \phi_{i,n} / \lambda_{i,n}^{1/2} \quad (18)$$

are orthonormal functions in  $\mathcal{K}_n$ . We then compute the minimizer  $f_{\tau,\ell,n} \in \mathcal{H}_{\ell,n}$  of  $\mathcal{A}_{\tau,n}(f)$  over this hypothesis space, obtaining, in direct analogy to (14),

$$\begin{aligned} f_{\tau,\ell,n}(x) &= T_{\ell,n} Z_{\tau,n} = \sum_{i=1}^{\ell} \frac{\alpha_{i,n}(\tau)}{\lambda_{i,n}} \langle \iota_n k_{n,x}, \phi_{i,n} \rangle_{L^2(\mu_{X,n})} \\ &= \sum_{i=1}^{\ell} \frac{\alpha_{i,n}(\tau)}{\lambda_{i,n}^{1/2}} \psi_{i,n}(x), \end{aligned} \quad (19)$$

where  $\alpha_{i,n}(\tau) = \langle \phi_{i,n}, U_n^q Y \rangle_{L^2(\mu_{X,n})}$ , and  $T_{\ell,n} : L^2(\mu_{X,n}) \rightarrow \mathcal{H}_{\ell,n}$  is the pseudoinverse operator associated with  $\mathcal{H}_{\ell,n}$ . This minimizer constitutes the empirical target function utilized by KAF.

**Definition 9 (Empirical Target Function).** The empirical target function  $f_{\tau,\ell,n}$  at lead time  $\tau$  associated with the response  $Y$  and  $\ell$ -dimensional hypothesis space  $\mathcal{H}_{\ell,n}$  is the minimizer of the empirical excess generalization error functional  $\mathcal{A}_{\tau,n}$  over  $\mathcal{H}_{\ell,n}$ , given by (19).

The expression in (19) can be written more compactly in matrix form using the column vector representations  $\phi_i \in \mathbb{C}^n$  of the  $\phi_{i,n}$ , given by eigenvectors of the  $n \times n$  kernel matrix  $\mathbf{G} = [k_n^*(x_i, x_j)]/n$  representing  $G_n$ , and chosen such that  $\phi_i \cdot \phi_j / n = \delta_{ij}$ . Note, in particular, that the expansion coefficients  $\alpha_{i,n}(\tau)$  are simply equal to the dot products  $\alpha_{i,n}(\tau) = \phi_i \cdot \mathbf{y}_\tau / n$  with the analog vector. Treating the remaining terms in (19) in a similar manner, we arrive at the expression

$$f_{\tau,\ell,n}(x) = \mathbf{k}(x) \mathbf{A}_\ell \mathbf{y}_\tau, \quad \mathbf{A}_\ell = \boldsymbol{\Phi}_\ell \boldsymbol{\Lambda}_\ell^{-1} \boldsymbol{\Phi}_\ell^* / n^2, \quad (20)$$

where  $\mathbf{k}(x)$  is the kernel vector,  $\boldsymbol{\Phi}_\ell$  is the  $n \times \ell$  matrix whose columns consist of the eigenvectors  $\phi_i \in \mathbb{C}^n$ ,  $\boldsymbol{\Lambda}_\ell$  is the  $\ell \times \ell$  diagonal matrix whose diagonal entries consist of  $\lambda_{i,n}$ , and  $*$  denotes

complex-conjugate transpose. This formula expresses the KAF target function as a sesquilinear form  $(\mathbf{k}(x), \mathbf{y}_\tau) \mapsto \mathbf{k}(x)^* \mathbf{A}_\ell \mathbf{y}_\tau$ , mapping pairs of kernel and analog vectors to  $\mathbb{C}$ -valued forecasts. Letting  $\mathbf{V}_\ell = \boldsymbol{\Lambda}_\ell^{-1/2} \boldsymbol{\Phi}_\ell^* / n$ , where  $\mathbf{V}_\ell^* \mathbf{V}_\ell = \boldsymbol{\Lambda}_\ell$ , the empirical target function is reexpressed as

$$f_{\tau,\ell,n}(x) = \mathbf{V}_\ell \mathbf{k}(x) \cdot \mathbf{V}_\ell \mathbf{y}_\tau.$$

This particular form emphasizes that the forecast is the result of taking the inner product of suitably projected kernel vector and equivalently projected analog vectors.

**Remark 10.** In KPCA [58], as well as related manifold learning techniques [70,71,65,72,68,69], eigenvectors of kernel matrices such as  $\phi_i$  above are employed for unsupervised feature extraction. In particular, it is common to use the  $\phi_i$  as coordinate vectors of dimension reduction maps,  $x_j \mapsto (\lambda_1^{-1/2} \phi_{1,n}(x_j), \dots, \lambda_\ell^{-1/2} \phi_{\ell,n}(x_j)) \in \mathbb{C}^\ell$ , mapping potentially high-dimensional covariate data into low-dimensional Euclidean spaces, where geometrical data relationships are revealed. In contrast, KAF/KPCR is a supervised learning technique where the goal is to perform out-of-sample prediction of a random variable (the response observable  $U^\tau Y$ ). A common aspect of the two methods is that they both rely heavily on eigendecompositions of kernel integral operators, but the end goals are fundamentally different. Note, in particular, that in KPCR one seeks to use as many eigenvectors  $\phi_i$  that can be computed from the available training data with tolerable sample error, and the number  $\ell$  of such eigenvectors can be far higher than the dimension of covariate space  $\mathcal{X}$ . For instance, in the L63 examples in Fig. 2 and Section 6.2 we use  $\ell = O(10^3)$ , which clearly does not serve as a “dimension reduction” map for the 3- or 1-dimensional covariate spaces employed there.

### 3. Error analysis and convergence

The previous section has shown how to calculate both an empirical target function  $f_{\tau,\ell,n}$  (Definition 9) and an ideal target function  $f_{\tau,\ell}$  (Definition 2), corresponding to two different hypothesis spaces,  $\mathcal{H}_{\ell,n}$  and  $\mathcal{H}_\ell$ , as well as two different error functionals,  $\mathcal{E}_{\tau,n}$  and  $\mathcal{E}_\tau$ , respectively. This section addresses the connection between the two functions, with the ultimate goal being that of bounding the error  $\mathcal{E}_\tau(f_{\tau,\ell,n})$  of the empirical target function as much as possible. Among other reasons, the availability of such bounds is useful for assessing the risk of overfitting the training data; that is, the possibility that  $\mathcal{E}_\tau(f_{\tau,\ell,n}) \gg \mathcal{E}_{\tau,n}(f_{\tau,\ell,n})$  for the chosen empirical hypothesis space. Note, in particular, that for a variety of kernels (e.g., strictly positive-definite kernels) it is possible to make  $\mathcal{E}_{\tau,n}(f_{\tau,\ell,n})$  at fixed  $n$  arbitrarily small by increasing  $\ell$ , but this reduction of empirical error eventually leads to an increase of the “true” error  $\mathcal{E}_\tau(f_{\tau,\ell,n})$  with respect to the invariant measure of the dynamics. See Section 6.1 for an illustration of this phenomenon.

The analysis of the error  $\mathcal{E}_\tau(f_{\tau,\ell,n})$  is typically organized into analysis of the error  $\mathcal{E}_\tau(f_{\tau,\ell})$  of the ideal target function (i.e., the excess generalization error), and the difference in error  $\mathcal{E}_\tau(f_{\tau,\ell,n}) - \mathcal{E}_\tau(f_{\tau,\ell})$ , denoted by  $\mathcal{D}_{\tau,\ell,n}$ , between the empirical and ideal target functions, referred to as the sample error [30]. In other words, error analysis uses the following decomposition:

$$\begin{aligned} \mathcal{E}_\tau(f_{\tau,\ell,n}) &= \mathcal{E}_\tau(f_{\tau,\ell}) + \mathcal{D}_{\tau,\ell,n} \\ &= \mathcal{E}_\tau(f_{\tau,\ell}) + (\mathcal{E}_\tau(f_{\tau,\ell,n}) - \mathcal{E}_\tau(f_{\tau,\ell})). \end{aligned}$$

This section examines in detail these contributions, and establishes sufficient conditions for convergence of the KAF target function to the conditional expectation.



### 3.1. KAF generalization error

The excess generalization error  $\mathcal{A}_\tau(f_{\tau,\ell})$  from (8) of the KAF target function  $f_{\tau,\ell}$  in (14) is given by

$$\mathcal{A}_\tau(f_{\tau,\ell}) = \|\Pi_{H_\ell^\perp} Z_\tau\|_{L^2(\mu_X)}^2 = \sum_{i=\ell+1}^{\infty} |\alpha_i(\tau)|^2,$$

where  $\Pi_{H_\ell^\perp} : L^2(\mu_X) \rightarrow L^2(\mu_X)$  is the orthogonal projection mapping into the orthogonal complement  $H_\ell^\perp$  of the hypothesis space  $H_\ell$  in  $L^2(\mu_X)$ . It follows from the above that  $\mathcal{A}_\tau(f_{\tau,\ell})$  vanishes as  $\ell \rightarrow \infty$  for any  $Z_\tau \in L^2(\mu_X)$ , and thus for any response variable  $U^\tau Y \in L^2(\mu)$ , iff the sequence of projections  $\Pi_{H_\ell^\perp}$  converges pointwise to 0 as  $\ell \rightarrow \infty$  (i.e.,  $\Pi_{H_\ell^\perp} g \rightarrow 0$  for any  $g \in L^2(\mu_X)$ ). By Lemma 5, this happens in turn iff  $K$  is a dense subspace of  $L^2(\mu_X)$ , i.e., iff  $G$  is a strictly positive operator. Since  $\mathcal{A}_\tau(f) = 0$  iff  $f = Z_\tau$ , we obtain the following basic consistency result expressed in terms of a positivity condition on the kernel  $k$ .

**Theorem 11.** *Let  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$  be an  $L^2(\mu_X)$ -strictly-positive kernel with corresponding RKHS  $\mathcal{K}$ . Then, for any response variable  $U^\tau Y \in L^2(\mu)$  and lead time  $\tau \geq 0$ , as  $\ell \rightarrow \infty$ , the target functions  $f_{\tau,\ell}$  from (14) converge to the conditional expectation  $\mathbb{E}[U^\tau Y | X] = X \circ Z_\tau$ , in the sense that  $\lim_{\ell \rightarrow \infty} \|f_{\tau,\ell} - Z_\tau\|_{L^2(\mu_X)}^2 = 0$ .*

Convergence with respect to the (stronger) RKHS norm of  $\mathcal{K}$ , as well as more precise estimates of the  $L^2(\mu_X)$  error, can be obtained under the additional assumption that the regression function  $Z_\tau$  lies in the subspace  $K \subset L^2(\mu_X)$ . In that case,  $Z_\tau$  has a representative  $f_\tau \in \mathcal{K}$ , given by Nyström extension as

$$f_\tau = \tilde{T}Z_\tau = \sum_{i=1}^{\infty} \alpha_i(\tau) \psi_i / \lambda_i^{1/2},$$

where the infinite sum on the right-hand side converges in  $\mathcal{K}$  norm. That is,  $f_\tau$  is given by the  $\mathcal{K}$ -norm limit of the partial sums  $\sum_{i=1}^{\ell} \alpha_i \psi_i / \lambda_i^{1/2}$ . The latter are precisely equal to the target functions  $f_{\tau,\ell}$  from (14), and therefore we conclude that  $\lim_{\ell \rightarrow \infty} \|f_{\tau,\ell} - f_\tau\|_{\mathcal{K}} = 0$ .

To obtain an estimate of  $\mathcal{A}(f_{\tau,\ell})$ , observe that  $K$  coincides with the range of  $G^{1/2}$ , the square root of  $G$ . It then follows that for  $Z_\tau \in K$ , there exists  $W_\tau \in L^2(\mu_X)$  such that  $Z_\tau = G^{1/2}W_\tau$ , which allows the excess generalization error to be rewritten as

$$\mathcal{A}_\tau(f_{\tau,\ell}) = \sum_{i=\ell+1}^{\infty} \lambda_i |\langle \phi_i, W_\tau \rangle_{L^2(\mu_X)}|^2.$$

The Cauchy–Schwarz inequality then yields

$$\mathcal{A}_\tau(f_{\tau,\ell}) \leq \left( \sum_{i=\ell+1}^{\infty} \lambda_i \right) \|W_\tau\|_{L^2(\mu_X)}^2,$$

where  $\sum_{i=\ell+1}^{\infty} \lambda_i < \text{tr } G$  is finite. Thus, in this case we can bound the decay of the excess generalization error by the decay of the tail sum of the eigenvalues of  $G$ .

The study of decay rates of the eigenvalues of an integral operator [73] is an active field of research. In the setting of Mercer kernels and compactly supported probability measures studied here, it can be shown that  $\lambda_i = o(i^{-1})$  for large-enough  $i$  [74, Theorem 2.4], which is consistent with the fact that  $G$  is trace-class. Estimates of the rate of decay of the tail sum are possible under additional regularity conditions on the kernel, including, for example, specialized notions of Lipschitz continuity. In such cases, it is possible to express the decay rate of  $\mathcal{A}_\tau(f_{\tau,\ell})$  as being algebraic, i.e.,  $\mathcal{A}_\tau(f_{\tau,\ell}) \leq C\ell^{-\gamma}$ , for some positive constants  $C$  and  $\gamma$  [75].

### 3.2. KAF sample error

In this section, we will establish that, under natural assumptions on the dynamical system and the reproducing kernels, the difference in error  $\mathcal{D}_{\tau,\ell,n}$  between the empirical and ideal target functions vanishes in the limit of large data,  $n \rightarrow \infty$ . We will do so by establishing a stronger result, namely that  $f_{\tau,\ell,n}$  converges uniformly to  $f_{\tau,\ell}$  in an appropriate compact set containing the supports of  $\mu_X$  and the sampling measures  $\mu_{X,n}$ .

**Basic assumptions for convergence.** Our first assumption is that (i)  $\Omega$  has the structure of a metric space, equipped with its Borel  $\sigma$ -algebra  $\mathcal{F}$ ; (ii)  $\mu$  is a Borel probability measure with compact support  $\Omega_\mu \subseteq \Omega$ ; and (iii) all of  $\Phi^t$ ,  $X$ , and  $Y$  are continuous. Note that, by continuity of  $\Phi^t$ , the Koopman operator  $U^t$  maps continuous functions to continuous functions for all  $t \geq 0$ , preserving the supremum norm of bounded continuous functions in  $C_b(\Omega)$ . See the commutative diagram in Fig. 1 for an illustration of the relationships between the Koopman operator on  $C(\Omega)$  and  $L^2(\mu_X)$ .

Our second assumption pertains to the convergence of the empirical measures  $\mu_n$  underlying the data to the invariant measure. Specifically, we assume that, for the starting state  $\omega_1 \in \Omega$ , the measures  $\mu_n$  converge to  $\mu$  weakly; that is, for every bounded, continuous function  $g : \Omega \rightarrow \mathbb{C}$ ,  $\lim_{n \rightarrow \infty} \int_\Omega g d\mu_n = \int_\Omega g d\mu$ . The weak convergence of  $\mu_n$  to  $\mu$ , in conjunction with the continuity of  $X$ , implies in turn that  $\mu_{X,n}$  converges weakly to  $\mu_X$ , i.e.,  $\lim_{n \rightarrow \infty} \int_{\mathcal{X}} f d\mu_{X,n} = \int_{\mathcal{X}} f d\mu_X$ , for all  $f \in C_b(\mathcal{X})$ .

Our third assumption relates to the existence of a compact set in which both the covariate data  $x_i$  and the support  $\mu_X$  lie. Specifically, for the starting state  $\omega_1 \in \Omega$  underlying the covariate training data, we assume that there exists a compact set  $\mathcal{U} \subseteq \mathcal{X}$  containing  $\mathcal{X}_\mu = \text{supp } \mu_X$ , as well as  $\text{supp } \mu_{X,n} = \{x_1, \dots, x_n\}$  for every  $n \in \mathbb{N}$ . This condition is automatically satisfied if the state space  $\Omega$  is already a compact space (e.g., ergodic dynamics on a torus), and is also satisfied by many systems with appropriate dissipative dynamics. Examples of such systems include ordinary differential equation models on  $\Omega = \mathbb{R}^d$  with quadratic nonlinearities, such as the L63 system [76] studied in Section 6.2, as well as partial differential equation models possessing inertial manifolds [77]. For our purposes, the existence of the compact set  $\mathcal{U}$  allows the (generally distinct) ideal and empirical RKHSs,  $\mathcal{K}(\mathcal{U})$  and  $\mathcal{K}_n(\mathcal{U})$ , respectively, to be viewed as subspaces of the Banach space  $C(\mathcal{U})$ . In the latter space, the relevant notion of convergence is convergence with respect to the uniform norm.

Next, we make an assumption on the convergence of the empirical reproducing kernels  $k_n$  of  $\mathcal{K}_n$  to the reproducing kernel  $k$  of  $\mathcal{K}$ . Specifically, we assume that, as  $n \rightarrow \infty$ ,  $k_n$  converges to  $k$  uniformly on  $\mathcal{U} \times \mathcal{U}$  (i.e., with respect to  $C(\mathcal{U} \times \mathcal{U})$  norm). This assumption is trivially satisfied if one works with data-independent kernels,  $k_n = k$ , and also holds for many classes of normalized kernels, including [71,65–69].

Finally, we assume that the response variable  $U^\tau Y$  is bounded on  $X^{-1}(\mathcal{U})$ , i.e.,  $C_Y = \sup_{\omega \in X^{-1}(\mathcal{U})} |U^\tau Y(\omega)| < \infty$ .

**Physical measures.** We define the basin of  $\mu$  as the maximal set  $M_\mu \subseteq \Omega$  for which the sampling measures  $\mu_n$ , starting from any  $\omega_1 \in M_\mu$ , converge weakly to  $\mu$ . If the dynamics is ergodic (i.e., every invariant set  $S \in \mathcal{F}$  under  $\Phi^t$  for all  $t \in \mathbb{R}$  has either  $\mu(S) = 0$  or  $\mu(S) = 1$ ), then  $\mu$ -a.e.  $\omega_1 \in \Omega$  lies in  $M_\mu$ , and the support  $\Omega_\mu$  lies in the topological closure of  $M_\mu$ . Moreover, for many dynamical systems encountered in applications,  $M_\mu$  can be a significantly “larger” set than  $\Omega_\mu$ . In particular, for systems possessing physical measures [78],  $M_\mu$  has positive measure with respect to an ambient measure on  $\Omega$  (e.g., Lebesgue measure), whereas  $\Omega_\mu$  oftentimes has zero ambient measure (e.g., if  $\Omega_\mu$  is an attractor developing under dissipative dynamics). In such cases, the methods will converge from an experimentally accessible set of initial conditions that can lie outside of  $\Omega_\mu$ .

**Uniform convergence on  $\mathcal{U}$ .** We assume throughout that the basic assumptions stated above hold. For simplicity, we will assume that for the given hypothesis space dimension  $\ell$ , all eigenvalues  $\lambda_1, \dots, \lambda_\ell$  are simple (if this is not the case, the argument presented below can be modified using appropriate projection operators onto eigenspaces of  $G$  and  $G_n$ ).

Since the ideal target function  $f_{\tau,\ell}$  from (14) and the empirical target function (19) are linear combinations of  $\ell < \infty$  continuous functions,  $\psi_i$  and  $\psi_{i,n}$ , respectively, convergence of  $f_{\tau,\ell,n}$  to  $f_\ell$  in  $C(\mathcal{U})$  norm will follow if it can be shown that, as  $n \rightarrow \infty$  and for each  $i \in \{1, \dots, \ell\}$ , (i) the eigenvalues  $\lambda_{i,n}$  converge to  $\lambda_i$ ; (ii) the RKHS functions  $\psi_{i,n}$  converge, up to multiplication by a constant phase factor, to  $\psi_i$  in  $C(\mathcal{U})$  norm; and (iii) each of the expansion coefficients  $\alpha_{i,n}(\tau)$  converges to  $\alpha_i(\tau)$ . The first two of these claims are a consequence of the following lemma, which is based on [66, Theorem 15], [51, Corollary 2], and [79, Theorem 7].

**Lemma 12.** *Under the basic assumptions for convergence, the following hold:*

(i) *For each nonzero eigenvalue  $\lambda_i$  of  $G$ ,  $\lambda_{i,n}$  converges to  $\lambda_i$  as  $n \rightarrow \infty$ .*

(ii) *For every RKHS function  $\psi_i$  corresponding to  $\lambda_i > 0$ , there exist complex numbers  $c_{i,n}$  of unit modulus, such that  $\lim_{n \rightarrow \infty} \|\psi_{i,n} - c_{i,n}\psi_i\|_{C(\mathcal{U})} = 0$ .*

**Remark 13.** In [59], Rosasco et al. approach the problem of establishing spectral convergence of the empirical integral operators  $G_n = \iota_n \iota_n^*$  associated with a fixed ( $n$ -independent) kernel  $k$  by considering the operators  $\tilde{G} = \iota_n^* \iota_n$  acting on the corresponding RKHS,  $\mathcal{K}$ . They show that for i.i.d. training data sampled from  $\mu_X$ , as  $n \rightarrow \infty$  these operators converge in Hilbert–Schmidt norm, and thus in spectrum, to the integral operator  $\tilde{G} = \iota^* \iota$ , and provide explicit rates of convergence. Aside from losing the precise error bounds that the i.i.d. assumption affords, the weaker ergodicity assumption employed in this work could be used to establish convergence in Hilbert–Schmidt norm of the data-driven operators associated with the sampling measures  $\mu_{X,n}$  along orbits of the dynamics, for which the data are not independent. An advantage of this approach is that one does not need to introduce  $C(\mathcal{U})$  as an auxiliary comparison space for the operators  $G_n$  and  $G$  (which act on different spaces). Moreover, convergence of  $\tilde{G}_n$  in Hilbert–Schmidt norm is stronger than the type of operator convergence considered in [66,51,79] (namely, collective compact convergence), which leads to Lemma 12. At the same time, however, a potential limitation of the Hilbert–Schmidt approach is that it requires the existence of a fixed RKHS, and as previously discussed, in many cases the kernels  $k_n$  and corresponding RKHSs  $\mathcal{K}_n$  depend on the training data. For this reason, we have opted to work in the more general setting of Lemma 12 despite a somewhat weaker convergence result, but we should point out that the results of [59] are available as an option when KAF is implemented with data-independent kernels.

Next, let  $\tilde{T}_n : \mathcal{K}_n \rightarrow \mathcal{K}_n$ , be the empirical Nyström extension operator on  $\mathcal{K}_n := \iota_n \mathcal{K}_n$ , defined analogously to  $\tilde{T}$  from Section 2.2. Also, for any probability measure  $\rho : \mathcal{F} \rightarrow [0, 1]$ , let  $\rho(f) = \int_{\mathcal{X}} f d\rho$ , where  $f \in \mathbb{L}^1(\rho)$ . To verify convergence of the expansion coefficients  $\alpha_{i,n}(\tau)$ , note that Lemma 12(ii) implies that for each  $i$  such that  $\lambda_i > 0$ , the continuous representatives of  $\phi_{i,n}$ , given by  $\varphi_{i,n} = \tilde{T}_n \phi_{i,n} = \psi_{i,n}/\lambda_{i,n}^{1/2}$ , converge in  $C(\mathcal{U})$  norm and up to phase to the continuous representative  $\varphi_i = \tilde{T} \phi_i = \psi_i/\lambda_i^{1/2}$  of  $\phi_i$ . Moreover, because the products  $\alpha_{i,n}(\tau)\psi_{i,n}$  are invariant under multiplication of  $\phi_{i,n}$  by a constant phase factor, without loss of generality, we may assume that the  $c_{i,n}$  in Lemma 12 are all equal

to 1. Then, for any  $\tau = q \Delta t$  with  $q \in \mathbb{N}_0$ , we have

$$\begin{aligned} \alpha_{i,n}(\tau) &= \mu_n((\phi_{i,n}^* \circ X)(U_n^q \iota_n Y)) \\ &= \frac{1}{n} \sum_{j=1}^{n-q} \varphi_{i,n}^*(x_j) U^\tau Y(\omega_j) \\ &= \frac{n-q}{n} \mu_{n-q}((\varphi_{i,n}^* \circ X) U^\tau Y), \end{aligned}$$

and defining  $\tilde{\mu}_n = (n-q)n^{-1}\mu_n$ , it follows that

$$\begin{aligned} |\alpha_{i,n}(\tau) - \alpha_i(\tau)| &= |\mu_n((\phi_{i,n} \circ X) U_n^q \iota_n Y) - \mu((\phi_i \circ X) U^\tau Y)| \\ &= |\tilde{\mu}_n((\varphi_{i,n} \circ X) U^\tau Y) - \mu((\varphi_i \circ X) U^\tau Y)| \\ &\leq |\tilde{\mu}_n((\varphi_{i,n} - \varphi_i) \circ X) U^\tau Y| \\ &\quad + |(\tilde{\mu}_n - \mu)((\varphi_i \circ X) U^\tau Y)|, \\ &\leq C_Y \|\varphi_{i,n} - \varphi_i\|_{C(\mathcal{U})} \\ &\quad + |(\tilde{\mu}_n - \mu)((\varphi_i \circ X) U^\tau Y)|. \end{aligned}$$

In the last line above, the first term converges to 0 by uniform convergence of  $\varphi_{i,n}$  to  $\varphi_i$  on  $\mathcal{U}$ , and the second term by weak convergence of  $\mu_n$  to  $\mu$ , so we conclude that  $\alpha_{i,n}(\tau)$  converges to  $\alpha_i(\tau)$ . Moreover, by continuity of the dynamics and covariate and response variables, the convergence is uniform with respect to  $\tau$  lying in compact sets.

We summarize the main results of Sections 3.1 and 3.2 in the following theorem:

**Theorem 14.** *Under the basic assumptions for convergence, for every lead time  $\tau = q \Delta t$ ,  $q \in \mathbb{N}_0$ , and hypothesis space dimension  $\ell$  such that  $\lambda_\ell > 0$ , the KAF target function  $f_{\tau,\ell,n} \in \mathcal{K}_n$  converges as  $n \rightarrow \infty$  to the ideal target function  $f_{\tau,\ell} \in \mathcal{K}$ , uniformly on  $\mathcal{U}$ . Moreover, if the reproducing kernel  $k$  of  $\mathcal{K}$  is  $L^2(\mu_X)$ -strictly-positive-definite, then by Theorem 11,  $f_{\tau,\ell,n}$  converges to the regression function  $Z_\tau$  associated with the conditional expectation,  $\mathbb{E}[U^\tau Y | X] = Z_\tau \circ X$ , in the sense of the iterated limit*

$$\lim_{\ell \rightarrow \infty} \lim_{n \rightarrow \infty} f_{\tau,\ell,n} = \lim_{\ell \rightarrow \infty} f_{\tau,\ell} = Z_\tau.$$

Here, the  $n \rightarrow \infty$  and  $\ell \rightarrow \infty$  limits are taken in  $C(\mathcal{U})$  and  $L^2(\mu_X)$  norm, respectively. Moreover, the convergence is uniform with respect to  $\tau$  lying in compact sets.

### 3.3. Mixing and loss of predictability

Before closing Section 3, we discuss some aspects of the long-time behavior of the conditional expectation and the KAF target functions in the presence of mixing dynamics, which will be useful in our interpretation of the L63 experiments in Section 6.2. First, we recall that the measure-theoretic definition of mixing [e.g., 80] can be equivalently stated as the condition that for any  $g, h \in L^2(\mu)$ ,

$$\lim_{\tau \rightarrow \infty} \langle U^{\tau*} g, h \rangle_{L^2(\mu)} = \left( \int_{\Omega} g^* d\mu \right) \left( \int_{\Omega} h d\mu \right).$$

Thus, under mixing dynamics, inner products of the form  $\langle U^{\tau*} g, h \rangle_{L^2(\mu)}$ , which can be thought of as temporal cross-correlation functions, converge to constants equal to products of the expectation values  $\mathbb{E}[g^*] = \int_{\Omega} g^* d\mu$  and  $\mathbb{E}[h] = \int_{\Omega} h d\mu$ . Using the projection representation of the conditional expectation in (3), it then follows that

$$\begin{aligned} \lim_{\tau \rightarrow \infty} \langle g, \mathbb{E}[U^\tau Y | X] \rangle_{L^2(\mu)} &= \lim_{\tau \rightarrow \infty} \langle g, \Pi_X U^\tau Y \rangle_{L^2(\mu)} \\ &= \lim_{\tau \rightarrow \infty} \langle U^{\tau*} \Pi_X g, Y \rangle_{L^2(\mu)} \\ &= \mathbb{E}[(\Pi_X g)^*] \mathbb{E}[Y]. \end{aligned}$$

Therefore, because  $g$  in the above is arbitrary, and  $\Pi_X$  leaves constant functions invariant, we conclude that  $\mathbb{E}[U^\tau Y | X]$  converges weakly to a constant function equal to  $\mathbb{E}[Y]$ , i.e.,

$$\lim_{\tau \rightarrow \infty} \langle g, \mathbb{E}[U^\tau Y | X] - \mathbb{E}[Y]1_\Omega \rangle_{L^2(\mu)} = \lim_{\tau \rightarrow \infty} \langle g, \Pi_X(U^\tau Y - \mathbb{E}[Y]1_\Omega) \rangle_{L^2(\mu)} = 0,$$

where  $1_\Omega$  is the function on  $\Omega$  equal everywhere to 1. We interpret this behavior as a loss of predictability due to mixing dynamics.

Observe now that the  $L^2(\mu)$  element  $f_{\tau,\ell} \circ X$ , where  $f_{\tau,\ell}$  is the ideal target function, can be expressed as  $\Pi_{\tilde{H}_\ell} U^\tau Y$ , where  $\Pi_{\tilde{H}_\ell}$  is the orthogonal projection on  $L^2(\mu)$  mapping into the pullback  $\tilde{H}_\ell = H_\ell \circ X$  of the  $\ell$ -dimensional hypothesis space  $H_\ell$  into  $L^2_\mu(\mu)$ . If  $H_\ell$  contains constant functions, then it follows from similar arguments as above, in conjunction with the fact that  $\tilde{H}_\ell$  is finite-dimensional, that as  $\tau \rightarrow \infty$ ,  $f_{\tau,\ell}$  converges in  $L^2(\mu)$  norm (and not merely weakly) to a constant equal to  $\mathbb{E}[Y]$ . We will discuss practical ways of ensuring that  $\tilde{H}_\ell$  always contains constant functions, ensuring in turn this type of long-term statistical consistency with the infinite-dimensional case, using Markov-normalized reproducing kernels in Section 5.

With regard now to the empirical target function, since the  $n \rightarrow \infty$  convergence of  $f_{\tau,\ell,n}$  to  $f_{\tau,\ell}$  may not be uniform with respect to  $\tau \in \mathbb{R}$ , we cannot use this result to make a statement about the relation between  $f_{\tau,\ell,n}$  and  $\mathbb{E}[Y]$  as  $\tau \rightarrow \infty$ . Nevertheless, it is still possible to ensure (through Markov normalization of the kernel) that, at fixed  $n$ ,  $f_{\tau,\ell,n} \circ X$  lies in a finite-dimensional subspace of  $L^2(\mu)$  containing constant functions. In that case, for large-enough  $n$ , and long-enough, but bounded,  $\tau$ , we can expect  $f_{\tau,\ell,n}$  to be an approximately constant function equal to  $\mathbb{E}[Y]$ .

**Remark 15.** Time series prediction techniques can generally be categorized as being *direct* or *iterated* methods [5]. The KAF target function  $f_{\tau,\ell,n}$  from (19) provides direct prediction, in the sense that every lead time  $\tau$  has a distinct associated forecast function, which is evaluated once at the given initial data to yield a prediction. In contrast, in iterated prediction, one sets a timestep  $\Delta t > 0$ , and constructs a function  $g : \mathcal{X} \rightarrow \mathcal{X}$  that propagates the covariate over that timestep. Then, to obtain prediction with lead time  $\tau = q \Delta t$ ,  $q \in \mathbb{N}$ , the function  $g$  is iteratively evaluated  $q$  times,  $x_{j+1} = g(x_j)$ , using the covariate  $x \in \mathcal{X}$  observed at forecast initialization as the initial condition  $x_0 = x$ , and the result is fed into a function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ , yielding a forecast  $y = f(x_q)$  of the response. Direct methods are more general than iterated methods, as the latter are generally based on some type of Markovianity assumption for the training data. Indeed, in Theorem 14, the convergence of  $f_{\tau,\ell,n}$  to the  $L^2$ -optimal conditional expectation was established without invoking any assumption about the dynamics in covariate space. In contrast, iterated methods effectively construct a surrogate dynamical model on  $\mathcal{X}$ , which if successful, may provide access to long-term statistics in ways that are not possible by direct methods. For example, the time series  $(q \Delta t) \mapsto (f \circ g^q)(x)$  produced by an iterated method need not converge to a constant as  $q \rightarrow \infty$ , but may exhibit variance and higher-order statistical moments resembling those induced by the invariant measure. See [81] for examples of iterated models using radial basis functions.

## 4. Extensions

This section shows how the KAF/KPCR learning framework presented thus far can shed light on other aspects of the kernel approach other than using leading principal components (eigenfunctions) to approximate the conditional expectation of observables. The first extension (Section 4.1) describes how KRR may be

characterized as resulting from the same variational problem as that of KPCR, albeit with a nonlinear, rather than linear, hypothesis space. The second extension (Section 4.2) shows how KAF can be implemented using a class of non-symmetric kernels. The third extension (Section 4.3) shows how quantities other than the conditional expectation, such as the conditional probability and estimates of the forecast error, may also be approximated, and what their utility may be in practical problems.

### 4.1. Kernel ridge regression (KRR)

In KRR [24], the hypothesis space is a closed ball of radius  $R$  in the RKHS  $\mathcal{K}$ ;

$$\mathcal{H}_R = \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq R\}.$$

Note that  $\mathcal{H}_R$  is not a linear subspace of  $\mathcal{K}$ , and thus the projection  $\Pi_{\mathcal{H}_R} : L^2(\mu_X) \rightarrow L^2(\mu_X)$  mapping into  $H_R = \iota_{\mathcal{H}_R}$  is a nonlinear operator. Although representations for this particular nonlinear operator are known [30], those for the inverse  $\iota_{\mathcal{H}_R}^{-1}$  are generally intractable. However, by using Lagrange multipliers, optimization over  $\mathcal{H}_R$  may be transformed into a linear problem. In particular, the problem of minimizing  $\|f - Z_\tau\|_{L^2(\mu_X)}$  such that  $\|f\|_{\mathcal{K}} \leq R$  is a constrained optimization problem for which there exists a parameter  $\eta_R > 0$ , dependent on  $R$ , such that the penalized optimization problem

$$\min_{f \in \mathcal{K}} \|f - Z_\tau\|_{L^2(\mu_X)}^2 + \eta_R \|f\|_{\mathcal{K}}^2$$

is an equivalent formulation. The solution to this problem is known to be [30]

$$f_{\tau,R} = \iota^*(G + \eta_R \text{Id})^{-1} Z_\tau.$$

The empirical solution  $f_{\tau,R,n} \in \mathcal{K}_n$ , meanwhile, is given by

$$f_{\tau,R,n}(x) = \mathbf{k}(x)^*(\mathbf{G} + \eta_R \mathbf{I})^{-1} \mathbf{y}_\tau, \quad (21)$$

where  $\mathbf{G}$ ,  $\mathbf{k}$ , and  $\mathbf{y}_\tau$  are the kernel matrix, kernel vector, and analog vector from Section 2.4. As with KPCR, the KRR target function  $f_{\tau,R}$  also converges in mean square to the conditional expectation, in the sense that, as the regularization parameter  $\eta_R$  is decreased to zero,  $\|f_{\tau,R} - Z_\tau\|_{L^2(\mu_X)}$  converges to 0 if the kernel  $k$  is  $L^2(\mu_X)$ -strictly-positive (cf. Theorem 11). Moreover, under the assumptions stated in Section 3.2,  $f_{\tau,R,n}$  converges to  $f_{\tau,R}$  in  $C(\mathcal{U})$  norm, as  $n \rightarrow \infty$ , so that an analog of Theorem 14 holds for  $f_{\tau,R,n}$ .

Note that, unlike KPCR, the KRR estimator in (21) does not require eigendecomposition of  $\mathbf{G}$ , and only depends on kernel values (thus, it can be thought of as making direct use of the “kernel trick”; see Section 2.3). Still, the standard implementation of KRR relies on a computationally expensive full inversion of the kernel matrix, whose eigenvalues are perturbed away from zero by some regularizing parameter  $\eta_R$ . A hybrid approach is to employ a low-rank approximation as in KPCR, while perturbing the eigenvalues away from zero as in KRR. With the notation of (20), this leads to the target function

$$f_{\tau,\ell,R,n}(x) = \mathbf{k}(x)^T \Phi_\ell(\Lambda_\ell + \eta_R \mathbf{I})^{-1} \Phi_\ell^* / n^2.$$

All of the KPCR, KRR, and hybrid estimators approximate the conditional expectation when the parameters are sufficiently relaxed, but the rates of convergence may differ. In general, KRR is useful when insensitivity to noise is desired, but it can be computationally expensive as it involves full matrix inversion. KPCR, on the other hand, can converge very rapidly when it turns out that the regression function lies in the leading eigenspaces of  $G$ .



#### 4.2. Non-symmetric kernels

The KAF formulation presented in Sections 2 and 3 makes heavy use of RKHSs, and is therefore restricted to Hermitian, positive-definite kernels. Yet, many popular kernel-based algorithms utilize non-symmetric kernels, typically constructed by normalization of symmetric kernels. Examples include normalized graph Laplacians [66] and Markov kernels approximating heat kernels on manifolds [71,65,68,69]. We now describe an extension of KAF to a class of non-symmetric kernels, whose corresponding integral operators on  $L^2(\mu)$  are related to integral operators associated with symmetric positive-definite kernels by similarity transformations. We will see that target functions can still be constructed using Nyström-type extensions into RKHSs, with extension operators derived from integral operators induced by non-symmetric kernels.

Specifically, let  $w : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a continuous, positive-definite bivariate function on covariate space, not necessarily symmetric, such that

$$d(x)w(x, x') = d(x')w(x', x), \quad \forall x, x' \in \mathcal{X}, \quad (22)$$

for a continuous, strictly positive function  $d : \mathcal{X} \rightarrow \mathbb{R}$ . Notice the similarity between (22) and the detailed balance relation in reversible Markov chains. We have:

**Proposition 16.** *With notation as above, and if (22) holds, the function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  with*

$$k(x, x') = w(x, x')/d(x')$$

*is a continuous, symmetric, positive-definite kernel. Moreover, denoting the corresponding RKHS, inclusion map, and Nyström operator by  $\mathcal{K}$ ,  $\iota : \mathcal{K} \rightarrow L^2(\rho)$ ,  $\tilde{T} = \iota^+$ , respectively, where  $\rho$  is any compactly supported finite Borel measure on  $\mathcal{X}$ , the following hold.*

(i) *The integral operator*

$$W : f \mapsto \int_{\mathcal{X}} w(\cdot, x)f(x)d\rho(x)$$

*is well-defined as a bounded operator  $W : L^2(\rho) \rightarrow \mathcal{K}$ .*

(ii) *The integral operator  $J = \iota W$  is a trace-class operator on  $L^2(\mu_X)$ , with real eigenvalues  $\eta_1 \geq \eta_2 \geq \dots \searrow 0^+$ . Moreover, there exist a Riesz basis  $\{\xi_1, \xi_2, \dots\}$  of  $L^2(\mu_X)$  and a dual basis  $\{\xi'_1, \xi'_2, \dots\}$  with  $\langle \xi'_i, \xi_j \rangle_{L^2(\mu_X)} = \delta_{ij}$  consisting of eigenfunctions of  $J$  and  $J^*$ , respectively, i.e.,*

$$J\xi_j = \eta_j\xi_j, \quad J^*\xi'_j = \eta_j\xi'_j.$$

(iii) *The domain  $D(J^+)$  of the Moore–Penrose pseudoinverse of  $J$  is a dense subspace of  $D(\tilde{T})$ , and on this subspace the Nyström operator  $\tilde{T}$  takes the form  $\tilde{T}|_{D(J^+)} = WJ^+$ . As a result, for every  $f \in D(\tilde{T})$  we have*

$$\tilde{T}f = \sum_{i: \eta_i > 0} \frac{\langle \xi'_i, f \rangle_{L^2(\mu_X)}}{\eta_i^{1/2}} \vartheta_i, \quad (23)$$

where  $\vartheta_i = W\xi_i/\eta_i^{1/2}$  are orthogonal functions in  $\mathcal{K}$ .

**Proof.** See Appendix A.2.  $\square$

Proposition 16 was inspired by [59, Section 4], where an auxiliary RKHS analogous to  $\mathcal{K}$  was used to establish spectral convergence for a class of non-symmetric graph Laplacian operators. Here, our perspective is somewhat different as we introduce  $\mathcal{K}$  as a consequence of the detailed balance condition in (22), rather than assuming its existence a priori and deducing from it a non-symmetric kernel such as  $w$  (as done in [59]). Moreover, our objective here is not to establish spectral convergence

of integral operators acting on  $\mathcal{K}$  (since our kernels are typically data-dependent; see Remark 13), but rather to identify an RKHS whose corresponding Nyström operator is computable using integral operators associated with non-symmetric kernels. Indeed, similarly to [59], a key aspect of Proposition 16 is that the Nyström operator for  $\mathcal{K}$  is constructed using integral operators associated with  $w$  and their pseudoinverses, without invoking the operator  $\iota^*$  associated with the reproducing kernel  $k$  of  $\mathcal{K}$ . See also [48] for further applications of non-symmetric kernels to forecasting.

The eigenvalues  $\eta_i$  appearing in Proposition 16 are different from the eigenvalues  $\lambda_i$  of the integral operator  $G = \iota^*$  associated with  $k$ . Instead, as can be directly verified, the  $\eta_i$  coincide with the eigenvalues of a self-adjoint integral operator  $\hat{G}$  on  $L^2(\rho)$  associated with a different reproducing kernel than  $k$ , to which  $J$  is related by a similarity transformation. Specifically, observe that under detailed balance, the kernel  $\hat{k} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  given by

$$\begin{aligned} \hat{k}(x, x') &= d^{1/2}(x)w(x, x')d^{-1/2}(x') \\ &= d^{1/2}(x)k(x, x')d^{1/2}(x') \end{aligned} \quad (24)$$

is also positive-definite, and has an associated RKHS  $\hat{\mathcal{K}}$  with restriction operator  $\hat{\iota} : \hat{\mathcal{K}} \rightarrow L^2(\rho)$  and self-adjoint integral operator  $\hat{G} = \hat{\iota}^* : L^2(\rho) \rightarrow L^2(\rho)$ . Letting  $M_d : L^2(\rho) \rightarrow L^2(\rho)$  be the bounded multiplication operator by the continuous function  $d$ , i.e.,  $M_d f = df$ , we can express  $J$  as a similarity transformation of  $\hat{G}$ , namely

$$J = M_d^{-1/2} \hat{G} M_d^{1/2}, \quad (25)$$

where  $M_d^{1/2}$  and  $M_d^{-1/2}$  are both bounded multiplication operators by the strict positivity and continuity of  $d$  and compactness of the support of  $\mu_X$ . It then follows that  $J$  and  $G$  have the same eigenvalues  $\eta_i$ , and we can construct the biorthonormal bases  $\xi_i$  and  $\xi'_i$  by first computing orthonormal eigenvectors  $\hat{G}_i \hat{\phi}_i = \eta_i \hat{\phi}_i$  (taking advantage of specialized solvers for self-adjoint operators), and setting

$$\xi_i = d^{-1/2} \hat{\phi}_i, \quad \xi'_i = d^{1/2} \hat{\phi}_i. \quad (26)$$

Similarly, we have

$$\vartheta_i = d^{-1/2} \hat{\psi}_i, \quad \hat{\psi}_i = \hat{\iota}^* \hat{\phi}_i / \eta_i^{1/2}. \quad (27)$$

Note that  $G = M_d^{-1/2} \hat{G} M_d^{-1/2}$ , so apart from special cases,  $G$  and  $\hat{G}$  are not similar operators.

As in the case of symmetric kernels, when forecasting with the class of non-symmetric kernels considered in this section we construct target functions analogously to (12) and (19) by truncating the expansion in (23) to  $\ell$  eigenfunctions.

#### 4.3. Conditional variance and conditional probability

In forecasting applications, it is important to be able to perform uncertainty quantification; that is, estimate the error of the target function. Moreover, besides point forecasts of a given response variable, it is oftentimes of interest to predict the probability of occurrence of events defined in terms of the response meeting certain criteria (e.g., exceeding a specified threshold). We now discuss how KAF techniques can be employed to carry out these tasks.

First, regarding error estimation, consider the  $L^2(\mu)$  observable  $\beta_\tau = |U^\tau Y - \mathbb{E}[U^\tau Y | X]|^2$ ,

which measures the square error of the conditional expectation (and thus, the optimal target function  $Z_\tau$ ). The conditional expectation of  $\beta_\tau$  with respect to  $X$  is known as the conditional variance,

$$\text{var}[U^\tau Y | X] = \mathbb{E}[\beta_\tau | X],$$



and satisfies

$$\int_{\Omega} \text{var}[U^{\tau} Y | X] d\mu = \sigma_{\tau}$$

by construction. Thus, the conditional variance is equal in expectation to the intrinsic error from (6), providing an unbiased estimator of the square forecast error. Moreover, being a conditional expectation,  $\text{var}[U^{\tau} | X]$  is expressible as the pullback of a unique element  $W_{\tau} \in L^2(\mu_X)$ , such that  $\text{var}[U^{\tau} | X] = W_{\tau} \circ X$ , which can be empirically approximated using KAF. Specifically, applying KAF to the function  $\beta_{\tau, \ell, n} = |U^{\tau} Y - f_{\tau, \ell, n} \circ X|^2$  leads to the estimator  $s_{\tau, \ell, n} \in \mathcal{H}_{\ell, n}$  of  $W_{\tau}$  given by

$$s_{\tau, \ell, n} = \mathbf{k}(x)^* \mathbf{A}_{\ell} \beta_{\tau, \ell, n},$$

where  $\beta_{\tau, \ell, n} = [\beta_{\tau, \ell, n}(\omega_1), \dots, \beta_{\tau, \ell, n}(\omega_n)]^T$  is a column vector in  $\mathbb{C}^n$  containing the values of  $\beta_{\tau, \ell, n}$  on the training states  $\omega_i$  (cf. the analog vector  $\mathbf{y}_{\tau}$  in (17)). Because, as with many projection methods,  $s_{\tau, \ell, n}(x)$  is not guaranteed to be non-negative, in practice we perform error estimation using

$$\varepsilon_{\tau, \ell, n}(x) = |s_{\tau, \ell, n}(x)|^{1/2}. \quad (28)$$

The function  $\varepsilon_{\tau, \ell, n} \circ X$  then converges in the limit of large data to  $\sqrt{\text{var}[U^{\tau} Y | X]}$ , analogously to the convergence of  $f_{\tau, n, \ell} \circ X$  to  $\mathbb{E}[U^{\tau} Y | X]$  in Theorem 14. More generally, note that for any function  $\Gamma : \mathbb{C} \rightarrow \mathbb{C}$ , such that  $\Gamma \circ Y$  lies in  $L^2(\mu)$ , the conditional expectation  $\mathbb{E}[U^{\tau}(\Gamma \circ Y) | X]$  is approximated by

$$g_{\tau, \ell, n} = \mathbf{k}(x)^* \mathbf{A}_{\ell} \Gamma(\mathbf{y}_{\tau}), \quad (29)$$

where  $\Gamma(\mathbf{y}_{\tau})$  is the column vector in  $\mathbb{C}^n$  obtained by element-wise application of  $\Gamma$  to the analog vector  $\mathbf{y}_{\tau}$ .

Next, turning to approximations for conditional probability, let  $\Theta \in \mathcal{F}$  be an event (i.e., a measurable subset of  $\Omega$ ), defined through certain conditions on  $Y(\omega)$  being met. For instance, in the forecasting of rare or extreme events, one might employ a formulation such as

$$\Theta = \{\omega \in \Omega : Y(\omega) > \theta\}, \quad (30)$$

where  $\theta$  is a large threshold parameter.

Every event  $\Theta$  has an associated indicator function  $\chi_{\Theta} \in L^2(\mu)$ , evolving under the action of the Koopman operator as

$$U^{\tau} \chi_{\Theta} = \chi_{\Theta_{\tau}}, \quad \Theta_{\tau} = A^{-\tau}(\Theta).$$

Note, in particular, that every point lying initially in  $\Theta_{\tau}$  will be mapped into  $\Theta$  after dynamical evolution over time  $\tau$ . The conditional expectation

$$\mathbb{P}[\Theta_{\tau} | X] := \mathbb{E}[U^{\tau} \chi_{\Theta} | X]$$

then gives the conditional probability for  $\Theta$  to occur at lead time  $\tau$  given  $X$ . In the context of KAF, approximations for conditional probability are obtained by setting  $\Gamma$  in (29) to be the indicator function of the set  $Y(\Theta) \subset \mathbb{C}$ , leading to the target function

$$\tilde{g}_{\tau, \ell, n} = \mathbf{k}(x)^T \mathbf{A}_{\ell} \chi_{Y(\Theta)}(\mathbf{y}_{\tau}).$$

Because  $\tilde{g}_{\tau, \ell, n}$  is not guaranteed to take values in the interval  $[0, 1]$ , in order to obtain meaningful forecasts of conditional probability we threshold it, leading to the estimator

$$g_{\tau, \ell, n}(x) = \begin{cases} 1, & g_{\tau, \ell, n}(x) > 1, \\ g_{\tau, \ell, n}(x), & 0 < g_{\tau, \ell, n}(x) \leq 1, \\ 0, & g_{\tau, \ell, n}(x) \leq 0, \end{cases} \quad (31)$$

where  $g_{\tau, \ell, n} \circ X$  approximates  $\mathbb{P}[\Theta_{\tau} | X]$  analogously to Theorem 14. For example, for the event in (30),  $g_{\tau, \ell, n}(x)$  estimates the likelihood that  $Y$  will exceed  $\theta$  at lead time  $\tau$ , given the covariate value  $x \in \mathcal{X}$ .

## 5. Choice of kernel

In this section, we discuss practical guidelines for choosing the kernel  $k_n$  on covariate space  $\mathcal{X}$  employed in KAF.

### 5.1. Strictly positive-definite kernels

As a general guideline, in order to ensure that the empirical target function  $f_{\tau, \ell, n}$  from Definition 9 converges to the regression function  $Z_{\tau}$  from Definition 1 for an arbitrary response variable  $Y$  and lead time  $\tau$  (i.e., Theorem 14 holds), the empirical kernels  $k_n$  should converge, as  $n \rightarrow \infty$ , to an  $L^2(\mu_X)$ -strictly-positive kernel  $k$ , uniformly on the compact set  $\mathcal{U} \subseteq \mathcal{X}$ . Because every Mercer kernel which is strictly positive-definite on the support of a compactly supported Borel probability measure  $\rho$  is  $L^2(\rho)$ -strictly-positive (see Section 2.3), a convenient way of ensuring  $L^2(\mu_X)$ -strict-positivity of  $k$  is to work with empirical kernels  $k_n$  whose restrictions on  $\text{supp } \mu_{X, n} = \{x_1, \dots, x_n\}$  are strictly positive-definite for every  $n \in \mathbb{N}$ . For example, in the case  $\mathcal{X} = \mathbb{R}^m$ , it is known that radial Gaussian kernels are strictly positive-definite on the whole of  $\mathcal{X}$  [82]. Therefore, one can work with

$$k_n(x, x') = k(x, x') = e^{-\|x - x'\|^2 / \epsilon} \quad (32)$$

for some positive bandwidth parameter  $\epsilon$ , and the conditions of Theorem 14 will be satisfied.

The radial Gaussian kernel in (32) will be employed in the circle example in Section 6.1. It is an instance of a local kernel [69] of the form

$$k(x, x') = h(L(x, x')/\epsilon), \quad (33)$$

where  $L$  is a continuous, positive, symmetric function on  $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , and  $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  a strictly positive, bounded, continuous, shape function with rapid (faster-than-polynomial) decay at infinity. In the case of (32), we have  $L(x, x') = \|x - x'\|^2$  and  $h(u) = e^{-u}$ . See Ref. [83] for additional examples of kernels commonly used in machine learning applications.

**Remark 17.** On the finite-dimensional linear covariate space  $\mathcal{X} = \mathbb{R}^d$ , the covariance kernel,  $k(x, x') = x \cdot x'$ , which is employed in the proper orthogonal decomposition [84,85] and linear inverse modeling techniques [86], is not  $L^2(\mu_X)$ -strictly-positive. Indeed, one can verify that for this choice of kernel, the corresponding integral operators  $G_n$  and  $G$  are of at most rank  $d$  [87, Section 9], thereby bounding the dimension of the hypothesis spaces  $\mathcal{H}_{\ell, n}$  and  $\mathcal{H}_{\ell}$  by  $d$ . Thus, unless  $Z_{\tau}$  happens to lie in the span of the leading  $d$  eigenfunctions of  $G$  (which are, in this case, linear functions on  $\mathcal{X}$ ), the empirical target function  $f_{\tau, \ell, n}$  will fail to converge to  $Z_{\tau}$ .

### 5.2. Variable-bandwidth, Markov kernels

Next, we discuss two modifications of the radial Gaussian kernel on  $\mathbb{R}^d$ , which can play a fairly substantial role in improving the robustness of the hypothesis space, particularly for data with strong contrasts in sampling density (e.g., the L63 example in Section 6.2).

**Variable bandwidth.** Our first modification is to introduce a strictly-positive, continuous bandwidth function  $r_n : \mathcal{X} \rightarrow \mathbb{R}$ , turning (32) into a variable-bandwidth Gaussian kernel, viz.

$$k_n(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\epsilon r_n(x) r_n(x')}\right). \quad (34)$$

Intuitively, the role of  $r_n$  is to correct for variations in the sampling density of the data in covariate space. In particular, for a

well conditioned kernel integral operator  $G_n$ , the number of datapoints lying within radius  $O(\epsilon^{1/2})$  balls centered at each datapoint should not exhibit significant variations across the dataset, yet, the standard radial Gaussian kernel from (32) has no mechanism for preventing this from happening. For appropriately chosen  $r_n$ , the variable-bandwidth kernel in (34) can, in effect, vary the radii of these balls to help improve conditioning.

The different bandwidth functions proposed in the literature include near-neighbor distances [88] and kernel density estimates [68]. In the numerical experiments of Section 6.2, we will employ the latter approach, defining

$$r_n(x) = q_n^{-1/\tilde{m}}(x), \quad q_n(x) = \frac{1}{(\pi\tilde{\epsilon})^{\tilde{m}/2}} \int_{\mathcal{X}} e^{-\|x-x'\|^2/\tilde{\epsilon}} d\mu_{X,n}(x'). \quad (35)$$

Here,  $\tilde{\epsilon}$  a positive bandwidth parameter (different from  $\epsilon$  in (34)), and  $\tilde{m}$  a positive parameter approximating the dimension of the support  $\mathcal{X}_\mu$ . The parameters  $\epsilon$ ,  $\tilde{\epsilon}$ , and  $\tilde{m}$  are all determined from the data automatically. See [64,52] for descriptions of this procedure, including pseudocode [52, Algorithm 1].

It can be shown [52] that if  $\mathcal{X}_\mu$  has the structure of a smooth manifold, with a Riemannian metric inherited from its embedding in  $\mathcal{X} = \mathbb{R}^d$ , the bandwidth functions  $r_n$  in (35) induce a conformal change of metric, such that, in the new geometry, the measure  $\mu_X$  has uniform density relative to the Riemannian volume measure. That is, the conformal change of metric can be thought of as “balancing out” variations of the sampling density relative to the ambient-space metric, thus improving robustness to sampling errors. It should be noted that while here we do not assume that  $\mathcal{X}_\mu$  has manifold structure, the balancing effect of the bandwidth functions is still expected to take place.

**Symmetric Markov normalization.** Our second modification of the radial Gaussian kernel is to normalize it to a  $L^2(\mu_{X,n})$ -strictly-positive Markov-ergodic kernel using the normalization procedure introduced in [67]. This involves first computing the strictly positive, continuous functions

$$\begin{aligned} u_n(x) &= \int_{\mathcal{X}} k_n(x, x') d\mu_{X,n}(x'), \\ v_n(x) &= \int_{\mathcal{X}} \frac{k_n(x, x')}{u_n(x')} d\mu_{X,n}(x'), \end{aligned} \quad (36)$$

and then defining the Markov kernel  $p_n : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , with

$$p_n(x, x') = \int_{\mathcal{X}} \frac{k_n(x, x'') k_n(x'', x')}{u_n(x) v_n(x'') u_n(x')} d\mu_{X,n}(x''). \quad (37)$$

It can be readily verified that with this definition  $p_n$  acquires the Markov property,  $\int_{\mathcal{X}} p_n(x, x') d\mu_{X,n}(x') = 1$ , for all  $x \in \mathcal{X}$ . Moreover, it can be shown that if  $k_n$  is strictly positive-definite on  $\text{supp}(\mu_{X,n})$  then so is  $p_n$  [89, Lemma 12]. It can further be shown [79] that as  $n \rightarrow \infty$ ,  $p_n$  converges in  $C(\mathcal{U})$  norm to an  $L^2(\mu_X)$ -strictly-positive Markov kernel  $p : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  (given by an analogous formula to (37)), so that the spectral convergence results in Lemma 12 hold with  $k_n$  and  $k$  replaced by  $p_n$  and  $p$ , respectively.

In the context of KAF, a useful property of Markov kernels is that the associated integral operators  $G_n$  and  $G$  have the top eigenvalue  $\lambda_{1,n} = \lambda_1 = 1$  with a constant corresponding eigenfunction. This implies, in particular, that the corresponding RKHSs,  $\mathcal{K}_n$  and  $\mathcal{K}$ , respectively, always contain constant functions, and thus can naturally capture the mean of the response variable  $U^T Y$ . The eigenfunctions corresponding to  $< 1$  eigenvalues can then be thought of as capturing progressively finer-scale features of  $U^T Y$ , which are orthogonal to the mean. An illustration of this behavior is provided in Figs. 4 and 5.

In fact, in many ways, an RKHS  $\mathcal{K}$  with a Markov-ergodic reproducing kernel resembles a Sobolev space associated with

a heat kernel on a manifold. Specifically, using the Nyström extension operator, one can define a Dirichlet energy functional on the dense subspace  $K = \iota\mathcal{K}$  of  $L^2(\mu_X)$  that assigns a measure of roughness of functions analogous to the Dirichlet energy in Sobolev spaces. See [89] for additional discussion on this topic. Appendix B of that reference also contains pseudocode for computing the eigenfunctions  $\phi_{i,n}$  and associated RKHS functions  $\psi_{i,n}$  for the kernel in (37), which complements the pseudocode in Table 1 of this paper.

**Non-symmetric normalizations.** While the class of symmetric Markov kernels in (37) is attractive due to its direct correspondence with RKHSs, in a variety of learning applications, including spectral clustering [66] and approximation of heat operators on manifolds [65,72,69], it is significantly more common to employ normalizations leading to non-symmetric kernels. As a popular example, we mention here the diffusion maps algorithm [65], which is based on the class of Markov kernels  $p_n : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  with

$$\begin{aligned} p_n(x, x') &= \frac{\kappa(x, x')}{v_n(x) u_n^\alpha(x')}, \\ u_n(x) &= \int_{\mathcal{X}} \kappa(x, x') d\mu_{X,n}(x'), \\ v_n(x) &= \int_{\mathcal{X}} \frac{\kappa(x, x')}{u_n^\alpha(x')} d\mu_{X,n}(x'). \end{aligned}$$

Here,  $\kappa$  is a continuous, symmetric, strictly positive, positive-definite kernel (e.g., the radial Gaussian kernel from (32)),  $u_n$  is the normalization function from (36), and  $\alpha$  is a real parameter (typically set to 0, 1/2, or 1). One can verify that the kernel  $p_n$  just defined satisfies the detailed balance condition in (22) for  $w = p_n$  and  $d = v_n/u_n^\alpha$ , and thus can be employed for KAF as described in Section 4. An advantage of these kernels over the kernels in (37) is that they do not require integration with respect to  $x''$  in their definition, thus avoiding a source of sampling error. A disadvantage is that one needs to keep track of a biorthogonal pair of bases,  $\{\xi_i\}$  and  $\{\xi'_i\}$ , as opposed to a single orthonormal basis  $\{\phi_i\}$  in the symmetric case.

### 5.3. Kernels based on delay-coordinate maps

Delay-coordinate maps is a technique originally introduced for empirical state space reconstruction of partially observed dynamical systems [90,91], which employs the time ordering of the covariate data to embed it in a higher-dimensional space. Assuming that the dynamics  $\Phi^t$  is invertible, and choosing a positive integer parameter  $q$  (the number of delays), we define the covariate map  $X_q : \Omega \rightarrow \mathcal{X}^q$  such that

$$X_q(\omega) = (X(\omega), X(\Phi^{-\Delta t}(\omega)), \dots, X(\Phi^{-(q-1)\Delta t}(\omega))). \quad (38)$$

Note that  $X_q(\omega)$  can be empirically evaluated without explicit knowledge of the dynamical flow  $\Phi^t$  on state space, so-long as time-ordered covariate data  $X(\Phi^{-j,\Delta t}(\omega))$  are available over the temporal window  $[-(q-1)\Delta t, 0]$  at the sampling interval  $\Delta t$ .

Intuitively, this approach should increase the information content of the covariate map, since the time ordering of the data is a manifestation of the underlying dynamical flow. This intuition has been made mathematically precise in a number of “embeddology” results for flows in finite-dimensional state spaces [92], as well as classes of partial differential equation models [93]. These studies have established that under natural assumptions,  $X_q$  becomes an injective map for sufficiently large  $q$ , even if the raw covariate map  $X$  is not injective. In that case, the temporal evolution of the covariate  $X_q(\omega)$  on the support  $\mu \circ X_q^{-1}$  of the invariant measure on delay-coordinate space becomes a homeomorphic copy of the dynamics on the support  $\Omega_\mu$ , with optimal potential

predictability. While rigorously verifying the appropriate embedding conditions in practice is oftentimes difficult, it is generally expected that including delays can recover at least some of the state space degrees of freedom lost due to non-injectivity of  $X$ , leading to more skillful forecasts. Indeed, delay-coordinate maps have been employed in a number of parametric [81] and nonparametric [2,4,5] forecasting methodologies, and have also been found useful for extraction of coherent features from time series data [94–98,47,51,52].

In the context of KAF, we generally expect the conditional expectation  $\mathbb{E}[U^T Y \mid X_q]$  approximated by the algorithm to exhibit smaller intrinsic error (see (6)) with increasing  $q$ , and thus smaller total error for appropriately constructed hypothesis spaces. One way of constructing these spaces is to employ the strictly positive-definite, Markovian kernels described in Sections 5.1 and 5.2, replacing the “snapshot” data in  $\mathcal{X}$  with delay-embedded sequences in  $\mathcal{X}^q$ . For instance, a natural analog of the local kernel in (33) is  $k^{(q)} : \mathcal{X}^q \times \mathcal{X}^q \rightarrow \mathbb{R}$ , where

$$k^{(q)}(\tilde{x}, \tilde{x}') = h(L_q(\tilde{x}, \tilde{x}')/\epsilon). \quad (39)$$

Here,  $L_q$  is the symmetric function

$$L_q(\tilde{x}, \tilde{x}') = \frac{1}{q} \sum_{j=0}^{q-1} L(x_j, x'_j),$$

induced on  $\mathcal{X}^q$  from  $L$ , where  $\tilde{x} = (x_1, \dots, x_q)$  and  $\tilde{x}' = (x'_1, \dots, x'_q)$ . The kernel  $k^{(q)}$  can then be normalized via the symmetric or non-symmetric normalization procedures described in Section 5.2 to yield a Markov kernel. See, e.g., [12,14,11,15] for applications of KAF with Gaussian kernels on delay-coordinate space.

Before closing this section, we should point out that while beneficial from the point of view of topological state-space reconstruction, delay-coordinate maps with large numbers of delays may face potential limitations from the point of view of spectral characteristics of the underlying dynamical system. In particular, observe that the pullback  $L_{\Omega,q} : \Omega \times \Omega \rightarrow \mathbb{R}_+$  of  $L_q$  on state space, i.e.,  $L_{\Omega,q}(\omega, \omega') = L_q(X_q(\omega), X_q(\omega'))$  has the structure of an ergodic average of the continuous function  $L_{\Omega}(\omega, \omega') = L(X(\omega), X(\omega'))$  under the product dynamical flow  $\Phi^t \times \Phi^t$ , viz.

$$L_{\Omega,q}(\omega, \omega') = \frac{1}{q} \sum_{j=0}^{q-1} L_{\Omega}(\omega_{-j}, \omega'_{-j}),$$

where  $\omega_j = \Phi^{j\Delta t}(\omega)$  and  $\omega'_j = \Phi^{j\Delta t}(\omega')$ . As a result, by the Birkhoff pointwise ergodic theorem [80], as  $q \rightarrow \infty$ ,  $L_{\Omega,q}$  converges  $\mu \times \mu$ -almost everywhere to a function  $L_{\Omega,\infty} \in L^2(\mu \times \mu)$ . Further, it can be shown [51] that in the same limit, the kernel integral operators  $G_{\Omega,q}$  on  $L^2(\mu)$  associated with the pullback kernel  $k^{(q)} : \Omega \times \Omega \rightarrow \mathbb{R}$  induced by (39),  $k^{(q)}_{\Omega}(\omega, \omega') = k^{(q)}(X(\omega), X(\omega'))$ , converge in operator norm, and thus in spectrum, to a Hilbert–Schmidt integral operator  $G_{\Omega,\infty} : L^2(\mu) \rightarrow L^2(\mu)$  associated with the kernel

$$k_{\Omega,\infty}(\omega, \omega') = h(L_{\Omega,\infty}(\omega, \omega')/\epsilon).$$

Now, by invariance of Birkhoff averages, at any time  $t \in \mathbb{R}$  the kernel  $k_{\Omega,\infty}$  is invariant under the Koopman operator  $U^t \otimes U^t : L^2(\mu \times \mu) \rightarrow L^2(\mu \times \mu)$  for the product system, and the latter implies that  $U^t$  and  $G_{\Omega,\infty}$  are commuting operators [51]. As a result, every eigenspace of  $G_{\Omega,\infty}$  at nonzero eigenvalue (which is finite-dimensional by compactness of this operator) is a finite union of Koopman eigenspaces. The latter implies, in particular, that the nullspace of  $G_{\Omega,\infty}$  must necessarily contain the Koopman-invariant subspace  $L^2(\mu)$  associated with the continuous spectrum of  $U^t$  (see [99,39,51] for precise definitions

of this subspace). If it now happens that the subspace  $L^2_{\mathcal{X}}(\mu)$  where the conditional expectation  $\mathbb{E}[U^T Y \mid X]$  lies has a nonzero intersection with the continuous-spectrum subspace, as would typically be the case in systems with sufficiently complex (mixing) dynamics, then the hypothesis spaces associated with  $k_{\Omega,\infty}$  will fail to be dense in  $L^2_{\mathcal{X}}(\mu)$ , and thus the KAF target functions may fail to converge to  $\mathbb{E}[U^T Y \mid X]$ . Since the empirical integral operators with large numbers of delays are spectrally close to  $G_{\Omega,\infty}$ , this behavior indicates that there may be situations where increasing  $q$  beyond a limit may be detrimental to forecast skill.

## 6. Applications

We present two examples to illustrate how to build a kernel forecasting function, as well as some basic properties of convergence to the conditional expectation. See Table 1 for a summary of the algorithmic steps involved in KAF.

### 6.1. Circle rotation

Our first example is periodic flow on the circle,  $\Omega = S^1$ . Expressed in terms of canonical angle coordinates  $\omega \in [0, 2\pi)$ , the dynamical flow map  $\Phi^t$  takes the form of a translation,

$$\Phi^t(\omega) = \omega + \alpha t \pmod{2\pi}, \quad \alpha \in \mathbb{R},$$

with a period of  $2\pi/\alpha$ , exhibiting a unique ergodic invariant Borel probability measure  $\mu$ , equal to a normalized Lebesgue measure. As covariate and response spaces, we choose  $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ , and we prescribe covariate and response maps,  $X$  and  $Y$ , respectively, given by simple trigonometric functions as follows:

$$X(\omega) = \cos(\omega), \quad Y(\omega) = \sin(\omega).$$

Under this setup, we have  $U^T Y(\omega) = \sin(\omega + \alpha\tau)$ , and the conditional expectation  $\mathbb{E}[U^T Y \mid X = x]$  is the average of  $U^T Y$  at the two angles for which  $X(\omega + \alpha\tau) = x$ ; specifically,

$$Z_{\tau}(x) = \frac{\sin(\arccos(x) + \alpha\tau) + \sin(-\arccos(x) + \alpha\tau)}{2} = x \sin(\alpha\tau).$$

The intrinsic error  $\sigma_{\tau}$  may then be computed directly as

$$\sigma_{\tau} = \|Y\|_{L^2(\mu)}^2 \cos^2(\alpha\tau), \quad \|Y\|_{L^2(\mu)} = 1/\sqrt{2}.$$

Observe that the intrinsic error is maximal (and equal to the squared  $L^2(\mu)$  norm of the response variable) when  $\tau = q\pi/\alpha$ , and minimal (and equal to zero) when  $\alpha\tau = (2q+1)\pi/2$ , where  $q$  is any integer.

The pushforward measure  $\mu_X$  in covariate space is supported on the interval  $[-1, 1] \subset \mathbb{R}$ , where it has the density

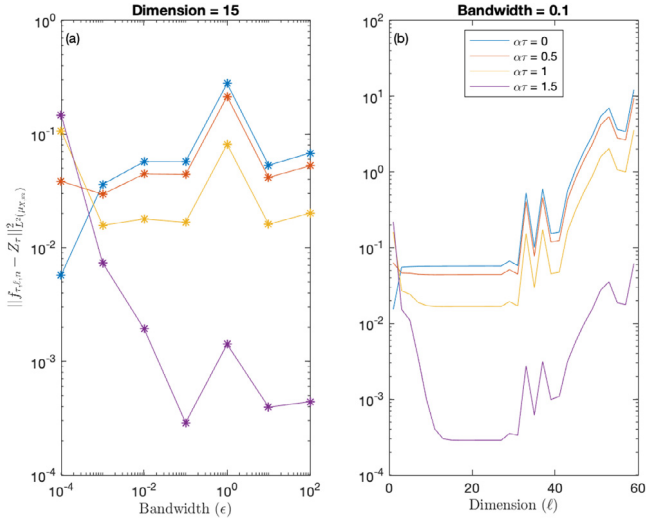
$$\varrho(x) = \frac{d\mu_X(x)}{d\text{Leb}} = \frac{1}{2\pi\sqrt{1-x^2}}$$

relative to Lebesgue measure. Note that  $\varrho(x)$  diverges at the boundary points  $x = \pm 1$ , but nevertheless lies in the  $\mathbb{L}^1$  space associated with the Lebesgue measure. Given a kernel  $k : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  meeting the conditions of Section 3, the eigenvalue problem for the associated integral operator  $G$  then becomes

$$\int_{-1}^1 k(x, x') \phi_i(x') \varrho(x') dx' = \lambda_i \phi_i(x).$$

A closed, analytic expression for this eigenvalue problem is not known for arbitrary choices of kernel  $k$ . Instead, using the radial Gaussian kernel from (32), we numerically solve the eigenvalue problem for the data-driven operator  $G_n$ , constructed from a sequence  $x_1, \dots, x_n \in \mathbb{R}$  of covariate points obtained from an underlying dynamical trajectory  $\omega_1, \dots, \omega_n \in S^1$ , as described in Section 2.4. Using the corresponding sequence  $y_1, \dots, y_n \in \mathbb{R}$  of





**Fig. 3.** Empirical excess generalization error  $\tilde{\mathcal{A}}_{\tau,m}(f_{\tau,\ell,n})$  of KAF target functions  $f_{\tau,\ell,n}$  for periodic flow on the circle with angular frequency  $\alpha = \sqrt{2}$ , obtained using the radial Gaussian kernel and a training dataset of  $n = 1000$  samples. The empirical excess generalization error is plotted for the lead times  $\tau \in \{0, 1/2\alpha, 1/\alpha, 3/2\alpha\}$ , using  $m = 10,000$  test points. In Panel (a), the dimension of the hypothesis space is fixed at  $\ell = 15$ , and the bandwidths are varied to take on values  $\epsilon \in \{10^{-4}, 10^{-3}, \dots, 10^2\}$ . In Panel (b), the bandwidth is fixed at  $\epsilon = 0.1$ , and the dimensions take on all odd integers from 1 to 59. For reference, note that the squared  $L^2(\mu)$  norm of the response variable  $U^\tau Y$  is equal to  $1/2$ .

response variables, we then build the KAF target function  $f_{\tau,\ell,n}$  via (19).

Here, we set the frequency  $\alpha = \sqrt{2}$ , and employ a training dataset of  $n = 1000$  samples, taken at an interval of  $\Delta t = 2\pi/100$  time units. Note that  $\Delta t$  is rationally independent from the rotation period, which ensures that the discrete-time map  $\Phi^{\Delta t}$  provides an ergodic sampling of  $\mu$ . Using this dataset, we have computed  $f_{\tau,\ell,n}$  for lead times  $\tau = q \Delta t/\alpha$ , with  $q$  an integer in the interval  $[0, 24]$ . To assess forecast skill, one can compute

the mean square error (MSE)

$$\tilde{\mathcal{E}}_{\tau,m}(f_{\tau,\ell,n}) = \frac{1}{m} \sum_{j=1}^m |f_{\tau,\ell,n}(\tilde{x}_j) - U^\tau Y(\tilde{\omega}_j)|^2$$

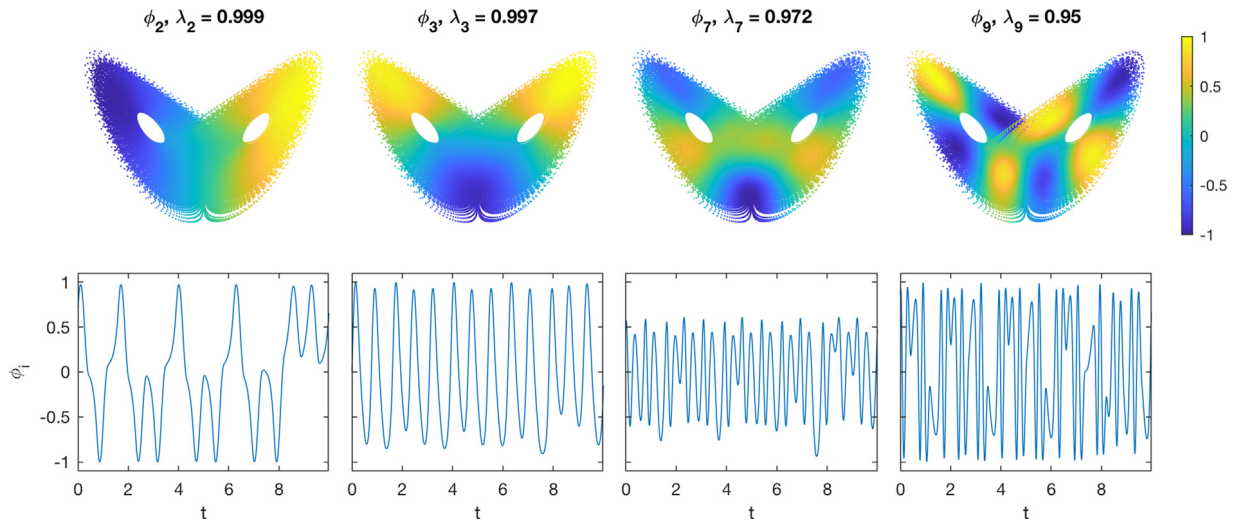
on a verification dataset  $\tilde{\omega}_j = \Phi^{(j-1)\Delta t}(\tilde{\omega}_1)$ ,  $\tilde{x}_j = X(\tilde{\omega}_j)$ . Since the intrinsic error happens to be analytically expressible for this problem, we report the empirical excess generalization error

$$\tilde{\mathcal{A}}_{\tau,m}(f_{\tau,\ell,n}) = \frac{1}{m} \sum_{j=1}^m |f_{\tau,\ell,n}(\tilde{x}_j) - Z_\tau(\tilde{x}_j)|^2.$$

The empirical MSE and excess generalization error approximate the true MSE and excess generalization error,  $\mathcal{E}_\tau(f_{\tau,\ell,n})$  and  $\mathcal{A}_\tau(f_{\tau,\ell,n})$ , respectively, and converge to these quantities as  $m \rightarrow \infty$ . Here, we employ a verification dataset of  $m = 10,000$  samples, starting from a state  $\tilde{\omega}_1$  chosen randomly and uniformly on the interval  $[0, 2\pi)$  (so that  $\omega_1 - \tilde{\omega}_1$  and  $\Delta t$  are rationally independent with probability 1).

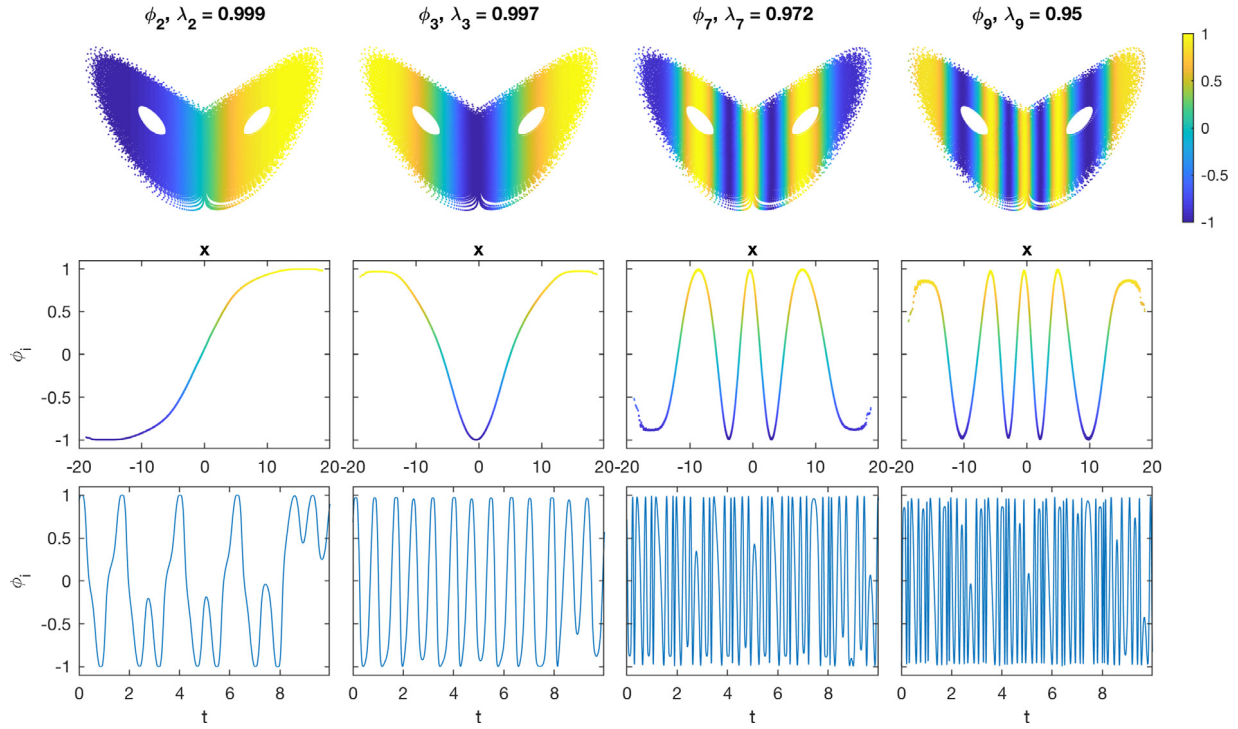
Fig. 3 shows the absolute value of empirical excess generalization error, plotted against the bandwidth parameter  $\epsilon$  of the Gaussian kernel and the hypothesis space dimension  $\ell$  for representative lead times  $\tau$  in the range  $[0, \pi/(2\alpha)]$ . In Fig. 3(a),  $\ell$  is kept fixed at 15, and  $\epsilon$  varies logarithmically in the interval  $[10^{-4}, 10^2]$ . The results show agreement between several different choices of  $\epsilon$  in some regimes of  $\tau$ , but also notable discrepancy in the region where the intrinsic error is already very small (i.e., when  $\alpha\tau$  is close to  $\pi/2$ ). In such a regime, the less sensitive kernels of large bandwidth are better able to capture that the generalization error is close to 0. In general, the  $\tilde{\mathcal{A}}_{\tau,m}$  values in Fig. 3(a) lie approximately in the interval  $[10^{-4}, 3 \times 10^{-1}]$ , which corresponds to approximately  $2 \times 10^{-4}$  to  $5 \times 10^{-1}$  multiples of the squared  $L^2(\mu)$  norm of the covariate variable.

Fig. 3(b) shows the behavior of empirical excess generalization error at fixed  $\epsilon = 0.1$  and representative values of  $\ell$  in the range 1 to 60. Employing just the first eigenfunction performs best for  $\alpha\tau = 0$ , but employing more eigenfunctions is better for larger values of  $\alpha\tau$ . Most notable, however, is the characteristic bias-variance tradeoff as  $\ell$  increases, with a valley of optimal values of  $\ell$  between 10 and 30. For instance, at  $\alpha\tau = 1.5$ , the error decreases from  $\simeq 2 \times 10^{-1}$  for  $\ell = 1$  to a minimal value



**Fig. 4.** Representative data-driven eigenfunctions  $\phi_{i,n}$  and their corresponding eigenvalues  $\lambda_{i,n}$ , computed from the fully observed L63 system. Top: Scatterplots of the eigenfunction values  $\phi_{i,n}(x_j)$  on the covariate training data  $x_j = \omega_j \in \mathbb{R}^3$ . Bottom: Eigenfunction time series  $t_j \mapsto \phi_{i,n}(x_j)$  over a portion of the training dataset spanning 10 natural time units. Notice that, despite the fact that the L63 attractor is not a Riemannian manifold, the eigenfunctions qualitatively resemble a generalized Fourier basis associated with a heat kernel. That is, as  $\lambda_{i,n}$  decreases,  $\phi_{i,n}$  exhibits increasingly small-scale oscillatory behavior, allowing one to represent functions of increasingly fine structure through eigenfunction expansions.





**Fig. 5.** As in Fig. 4, but for eigenfunctions computed from the partially observed L63 system. Top: Scatterplots of the eigenfunction values  $\phi_{i,n}(x_j)$  on the L63 states  $\omega_j$  underlying the covariate data  $x_j \in \mathbb{R}$ . Middle: Eigenfunction values  $\phi_{i,n}(x_j)$  versus  $x_j$ . Coloring is as in the top row. Bottom: Eigenfunction time series  $t_j \mapsto \phi_{i,n}(x_j)$  over a portion of the training dataset spanning 10 natural time units. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of  $\simeq 2 \times 10^{-4}$  for  $\ell = 20$ , but then increases for larger  $\ell$  to  $\simeq 10^{-2}$  values. This is a manifestation of the fact that the true error  $\mathcal{E}_\tau(f_{\tau,\ell,n})$  may increase with  $\ell$  at fixed  $\tau$  and  $n$ , even though the empirical error  $\mathcal{E}_{\tau,n}(f_{\tau,\ell,n})$  is always a non-increasing function of  $\ell$ .

## 6.2. Lorenz 63 system

In the L63 system [55], the state space is  $\Omega = \mathbb{R}^3$ . The dynamical flow  $\Phi^t(\omega_0)$  starting from  $\omega_0 \in \mathbb{R}^3$  is given by solution of the initial-value problem

$$\dot{\omega}(t) = \vec{V}(\omega(t)), \quad \omega(0) = \omega_0,$$

where  $\vec{V} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  is the smooth vector field with components  $(V^1, V^2, V^3)$  at  $\omega = (\omega^1, \omega^2, \omega^3)$  given by  $V^1 = \sigma(\omega^2 - \omega^1)$ ,  $V^2 = \omega^1(\mu - \omega^3)$ , and  $V^3 = \omega^1\omega^2 - \beta\omega^3$ . Here,  $\beta$ ,  $\mu$ , and  $\sigma$  are real parameters, which we set to the classical values  $\beta = 8/3$ ,  $\mu = 28$ , and  $\sigma = 10$ . For this choice of parameters, the L63 system is rigorously known to have a compact attractor  $\Omega_\mu \subset \mathbb{R}^3$  [100] with fractal dimension  $\approx 2.06$  [101], supporting a physical invariant measure  $\mu$  with a single positive Lyapunov exponent  $\Lambda \approx 0.91$  [102]. Due to dissipative dynamics, the attractor is contained within absorbing balls [76], ensuring the existence of the compact set  $\mathcal{U} \subseteq \mathcal{X}$  in covariate space. In light of these facts, all of the assumptions on the dynamical system made in Section 3.2 rigorously hold. The L63 system is also rigorously known to be mixing [103], and thus exhibits the loss of long-term predictability discussed in Section 3.3.

In the experiments that follow, we shall let  $\mathcal{Y} = \mathbb{R}$ , and let the response variable  $Y : \Omega \rightarrow \mathcal{Y}$  pick out one of the state vector components, i.e., if  $\omega = (\omega^1, \omega^2, \omega^3) \in \mathbb{R}^3$ , then  $Y(\omega) = \omega^i$ , for either  $i = 1, 2, 3$ . To illustrate the conditional probability framework discussed in Section 4.3, for each such response variable we will consider the event  $\Theta = \{\omega \in \Omega : Y(\omega) > \theta\}$ , where  $\theta$  is an empirical mean of  $Y$  computed from the training data. That

is, we will use KAF to estimate the conditional probability that the components of the state vector exceed their mean values. As for the covariate variable  $X$ , we will consider two cases, namely, full observations,  $\mathcal{X} = \mathbb{R}^3$  and  $X = \text{Id}$ , and a partially observed setup with  $\mathcal{X} = \mathbb{R}$  and  $X(\omega) = \omega^1$ . Consequently, in the partially observed setup  $\mathbb{E}[U^\tau Y | X]$  represents the conditional expectation of the  $i$ th component of  $\Phi^\tau(\omega)$ , given the first component of  $\omega$ .

All experiments use covariate data  $x_j = X(\omega_j)$  and response data  $y_{\tau,j} = U^\tau Y(\omega_j)$  generated from the same underlying trajectory  $\omega_1, \dots, \omega_n \in \mathbb{R}^3$ , with  $\omega_j = \Phi^{(j-1)\Delta t}(\omega_1)$ . The trajectory  $\omega_j$  was numerically generated in Matlab using the ode45 solver, starting from an arbitrary initial condition  $\omega_0$  and waiting for a long spinup time before collecting the first sample  $\omega_1$ . We nominally work with a training dataset consisting of  $n = 64,000$  samples, taken at a sampling interval  $\Delta t$  equal to 0.01 natural time units (i.e., about 1/100 of the characteristic Lyapunov timescale  $1/\Lambda$  of the system). Additional experiments with dataset sizes ranging from  $n = 640$  to  $n = 512,000$  and/or a longer sampling interval of  $\Delta t = 0.1$  were also conducted to investigate the performance of KAF relative to the  $(n, \Delta t) = (64,000, 0.01)$  baseline case.

To assess forecast skill, we use empirical root mean square error (RMSE) metrics computed from an independent verification dataset as in the circle example of Section 6.1. Specifically, the RMSE of the target function  $f_{\tau,\ell,n}$  at lead time  $\tau$  is given by  $\sqrt{\tilde{\mathcal{E}}_{\tau,m}(f_{\tau,\ell,n})}$ , where the verification dataset has the same number of  $m = 64,000$  samples as the training dataset, and was obtained via a similar spinup procedure starting from a different initial condition. Note that aside from the covariate and response data in the training phase, and the covariate data in the verification phase, no other information about the system state and/or dynamics was provided to the KAF algorithm.

The first step in the KAF pipeline is to compute the kernel eigenfunctions  $\phi_{i,n}$ , whose corresponding RKHS functions  $\psi_{i,n}$  form orthonormal bases for the hypothesis spaces  $\mathcal{H}_{n,\ell}$ . For that,

we employ the variable-bandwidth, Markov-normalized kernels from (37), with automatically tuned bandwidth and dimension parameters (see Section 5). Representative eigenfunctions for the fully- and partially-observed systems are displayed in Figs. 4 and 5, respectively, in both scatterplot and time series form. There, it is evident that the eigenfunctions behave like a generalized Fourier basis on the support of the measure  $\mu_X$ , with eigenfunctions corresponding to smaller eigenvalues allowing to resolve functions of increasingly smaller-scale variability on the L63 attractor. Notice, in particular, that in the partially observed example with  $\mathcal{X} = \mathbb{R}$ , the  $\phi_{i,n}$  are increasingly oscillatory, orthogonal functions on the real line, which pull back to  $\mathcal{G}$ -measurable functions on the attractor in  $\mathbb{R}^3$  with no variability in the  $\omega^2$  and  $\omega^3$  coordinates. It is precisely such a lack of variability that contributes to degraded forecast skill when faced with non-injective covariate functions.

Next, using the eigenfunctions and the response data in the training phase, we construct the empirical target functions  $f_{\tau,\ell,n}$  from (19). We also compute the error estimators  $\varepsilon_{\tau,\ell,n}$  from (28), which we use to place “error bars” around our forecasts of the form  $f_{\tau,\ell,n}(x) \pm \varepsilon_{\tau,\ell,n}(x)$ . According to Section 4.3, for an unbiased error estimate, the RMS value of  $\varepsilon_{\tau,\ell,n}$  in the verification phase should be close to the actual RMS forecast error. We use  $\ell = 3000$  and 1000 eigenfunctions for the fully observed and partially observed setups, respectively.

**Forecasting state vector components.** Fig. 2 shows prediction results for the three components of the state vector for the fully observed and partially observed systems, together with error estimates based on  $\varepsilon_{\tau,\ell,n}$ . We show representative forecast trajectories starting from an arbitrary initial condition in the verification dataset, as well as aggregate RMSE scores as a function of lead time, normalized by empirical standard deviation (i.e., the  $L^2(\mu_n)$  norm of  $Y - \int_{\Omega} Y d\mu_n$ ). Starting from the fully observed examples, the RMSE of all three state vector components  $\omega^i$  exhibits an initial exponential-like increase from near-zero values for  $\tau \lesssim 0.5 \simeq 0.5/\Lambda$ . This period is followed by an intermediate-time regime with more gradual RMSE increase and noticeable oscillatory behavior, until convergence to the equilibrium standard deviation (normalized RMSE  $\simeq 1$ ) at late times,  $\tau \gtrsim 4$ .

Examining the individual forecast trajectories, it is evident that the late-time convergence of the RMSE to a near-constant values is a manifestation of the trajectories converging to the mean,  $\mathbb{E}[Y]$ . The numerical results are therefore consistent with the theoretically expected late-time behavior of KAF in the presence of mixing dynamics, discussed in Section 3.3. It is also evident from Fig. 2 that the error estimators  $\varepsilon_{\tau,\ell,n}$  provide useful uncertainty quantification. That is, the error bars derived from these quantities envelop, for the most part, the true trajectories, and their RMS values agree well with the forecast RMSE.

Overall, in the fully-observed experiments,  $\omega^3$  is the most predictable state vector component (likely due to symmetry of the L63 equations), followed by  $\omega^1$  and  $\omega^2$  which are nearly equally predictable (again due to symmetry). If one were to set a normalized RMSE value of 0.6 as a threshold for loss of skill,  $\omega^3$  would remain predictable out to  $\simeq 3$  natural time units (i.e.,  $\simeq 3$  Lyapunov timescales), whereas  $\omega^1$  and  $\omega^2$  would remain predictable out to  $\tau \simeq 2$ . Setting that threshold to 0.8 increases the predictability horizon of  $\omega^3$  and  $\omega^1/\omega^2$  to  $\tau \simeq 5$  and 2.75, respectively.

Turning now to the Fig. 2 results for the partially observed system, it is clear that the act of observing  $\omega^1$  only in the covariate space bears a significant impact on forecast skill, particularly for  $\omega^2$  and  $\omega^3$ . Indeed, for these state vector components, the non-injectivity of the covariate function means that the normalized RMSE can be significant even at  $\tau = 0$ , without ever dropping

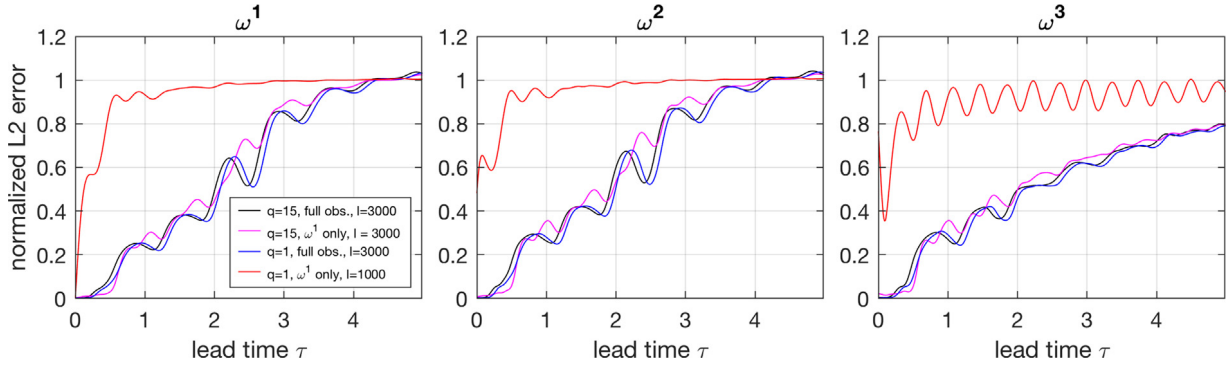
below  $\lesssim 0.4$ . Yet, even though the method cannot overcome the intrinsic error of this observational setup, it is nevertheless capable of providing adequate uncertainty quantification, as manifested by the ability of the estimated error bars to envelop the true trajectories (with the notable exception of certain extremal points) and the close agreement between the RMS values of  $\varepsilon_{\tau,\ell,n}$  and the forecast RMSE.

**Delay-coordinate maps.** As a demonstration of the efficacy of delay-coordinate maps in recovering forecast skill lost due to forecast observations, in Fig. 6 we compare the RMSE scores from the fully and partially observed experiments in Fig. 2 with their counterparts obtained by including  $q = 15$  delays to the respective covariate maps. Specifically, we construct delay-coordinate maps via (38) based on either  $X(\omega) = (\omega^1, \omega^2, \omega^3)$  or  $X(\omega) = \omega^1$ , and build KAF models using the same class of kernels, sampling intervals, and training dataset sizes as the experiments without delays. In both cases we use the same number of eigenfunctions as the fully observed case with no delays,  $\ell = 3000$ . Note that the delay-coordinate experiments based on full observations are interesting despite the fact that there is no potential predictability to be gained—this is because incorporating delays can introduce strong colinearities in the training data, increasing the likelihood of overfits. In addition, the L63 system falls squarely in the class of mixing dynamical systems discussed in Section 5.3, where incorporating delays in the kernel can suppress the nonzero eigenvalues of the corresponding integral operators, making them prone to sampling errors.

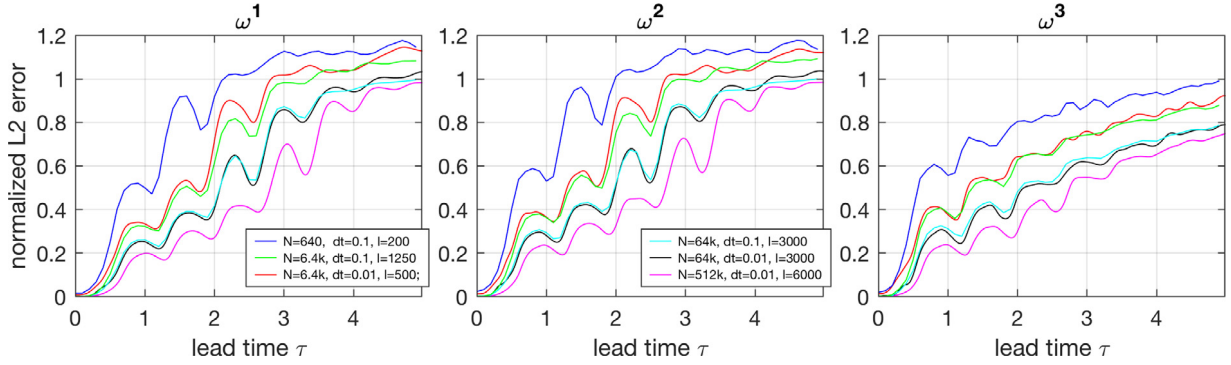
As is evident in Fig. 6, adding delays results in a considerable increase of skill for the partially observed system, with RMSE scores in generally good agreement with the fully observed systems. Interestingly, the RMSE values for short-term forecasts with  $0.2 \lesssim \tau \lesssim 0.5$  appear to be smaller than the fully observed experiments (either with or without delays), though longer-term forecasts exhibit intervals (e.g.,  $\tau \in [2.2, 2.5]$ ) where the RMSE of the partially observed system with delays is noticeably higher. Meanwhile, the fully observed system with delays exhibits very comparable skill as the system without delays, despite the issues associated with colinearity and mixing dynamics mentioned above.

**Sensitivity analysis.** The results displayed in Fig. 2 were obtained using a fairly dense sampling of the L63 attractor, corresponding to  $n \Delta t = 640$  natural time units, or, approximately 800 oscillations assuming a characteristic oscillatory timescale of 0.8. Moreover, the sampling interval  $\Delta t = 0.01$  was short compared to the oscillation and Lyapunov timescales of the system. To assess the performance of KAF in environments with shorter and/or less frequently sampled training data, we have performed a suite of forecasting experiments with full observations ( $X = \text{Id}$ ) that differ from our nominal setup with  $n = 64,000$  and  $\Delta t = 0.01$  by various modifications of the number of training samples  $n$  and sampling interval  $\Delta t$ . RMSE results from these experiments are depicted in Fig. 7.

Before delving to a description of these results, let us recall that, as with any supervised learning technique, KAF strives for a balance between generalization error (the difference between the ideal target function  $f_{\tau,\ell}$  and the regression function) and sample error (the difference between  $f_{\tau,\ell}$  and the empirical target function  $f_{\tau,\ell,n}$ ). This balance is attained by controlling the number of eigenfunctions (principal components)  $\ell$  employed, and the generalization error is a decreasing function of  $\ell$ . On the other hand, at fixed  $n$ , the sample error is generally an increasing function of  $\ell$ . As the sampling provided by the training data becomes poorer (by decreasing  $n$  and/or  $\Delta t$ ), the values of  $\ell$  achieving that balance become smaller, generally resulting to a decrease of forecast skill.



**Fig. 6.** Normalized RMSE versus lead time plots for KAF applied to the state vector components of the L63 system, illustrating the effects of incorporating delays in the covariate function. The black and red lines show the RMSE scores for the fully and partially observed cases from Fig. 2, respectively, using  $n = 64,000$  training samples and  $\ell = 3000$  (full observations) and  $\ell = 1000$  (partial observations) eigenfunctions. The blue and magenta lines show RMSE scores obtained by incorporating  $q = 15$  delays to the full and partial observation maps, respectively. The generally high consistency of the  $q = 15$  results with those of the fully observed experiments without delays illustrate that (i) delay-coordinate maps successfully recover information loss due to partial observations; and (ii) KAF behaves stably in delay-spaces with potentially poor conditioning due to colinearity of delay coordinates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 7.** Normalized RMSE versus lead time plots for KAF applied to the state vector components of the L63 system, under full observations and for a variety of numbers of training samples ( $n$ ), sampling intervals ( $\Delta t$ ), and hypothesis space dimensions ( $\ell$ ). The case with  $n = 64,000$  and  $\Delta t = 0.01$ , shown here in a black line, is identical to the fully observed case from Fig. 2. The experiments shown in colored lines correspond to modifications of the sample number and/or sampling frequency relative to the  $n = 64,000$ ,  $\Delta t = 0.01$  baseline case. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The RMSE results in Fig. 7 are for values of  $\ell$  chosen on the basis of yielding good skill over the full range  $\tau \in [0, 5]$  of lead times examined. It should be noted that we did not perform an exhaustive search to select these values, as we found that the dependence of skill on  $\ell$  exhibits plateau behavior, analogously to the circle example in Fig. 3(b). We should also point out that in an “operational” environment one would typically select different values of  $\ell$  for each lead time  $\tau$  so as to minimize RMSE. In particular, that for a mixing dynamical system such as L63, as  $\tau$  increases the conditional expectation  $\mathbb{E}[U^\tau Y | X]$  weak-converges to a constant, indicating that for the class of Markov kernels employed in this work (where the top eigenspace is spanned by constant functions) smaller  $\ell$  values may be warranted at large  $\tau$ .

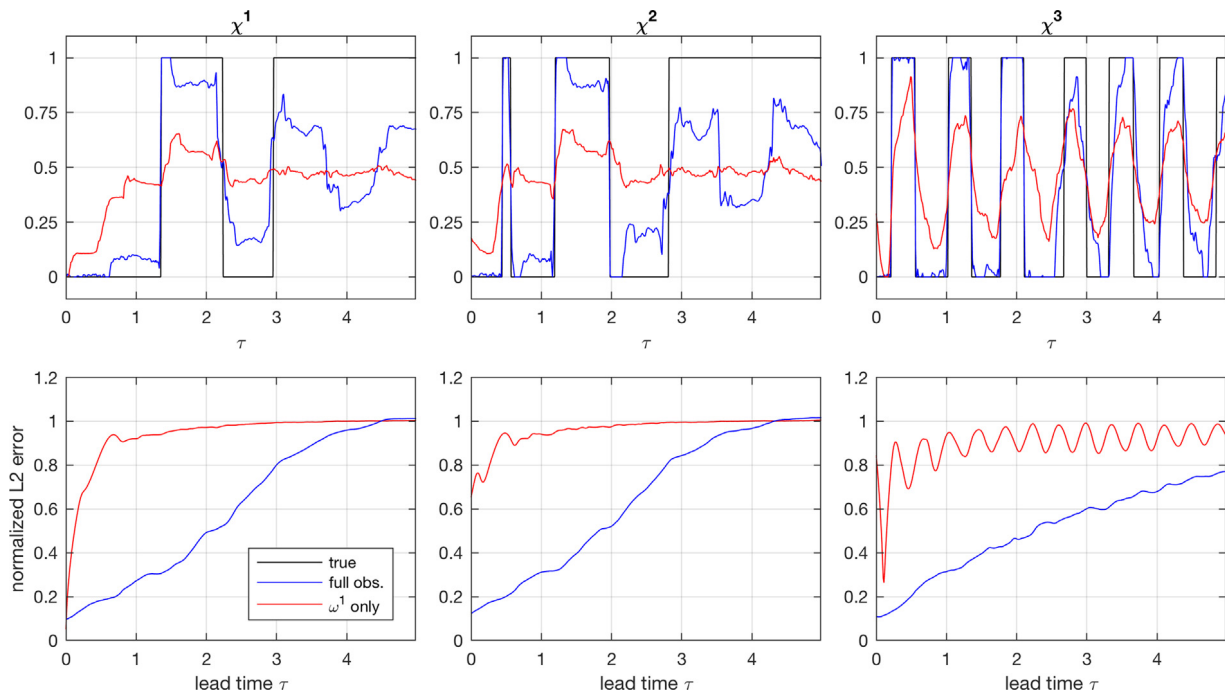
With these considerations in mind, we now turn to the results in Fig. 7. First, note that increasing the sampling interval by a factor of 10 to  $\Delta t = 0.1$ , while keeping the number of training samples and eigenfunctions fixed to the nominal values from Fig. 2,  $(n, \ell) = (64,000, 3000)$ , imparts little change to forecast skill. This suggests that the convergence of the leading  $\ell$  eigenspaces of the empirical integral operators  $G_n$  employed in KAF is largely unaffected by this reduction of sampling frequency. On the other hand, a tenfold reduction of the number of samples to  $n = 6400$ , using either  $\Delta t = 0.01$  or  $\Delta t = 0.1$ , does impart a noticeable reduction of skill, as we are forced to work with smaller numbers of eigenfunctions,  $\ell = 500$  and  $1250$  for  $\Delta t = 0.01$  and  $0.1$ , respectively. Nevertheless, at least over

short to moderate lead times,  $\tau \lesssim 2$ , the reduction of skill compared to the  $n = 64,000$  cases is fairly modest. For instance, using again a 0.6 value of normalized RMSE as a useful-skill threshold, the forecasts of the  $\omega^1$  and  $\omega^2$  variables based on the  $n = 6400$  remain useful out to  $\tau \simeq 1.8$  versus  $\tau \simeq 2.1$  for  $n = 64,000$ . Larger discrepancies are observed at longer leads,  $\tau \gtrsim 2$ , as well as for the  $\omega^3$  observable which loses about 1 natural time unit of predictability horizon based on the 0.6 normalized RMSE threshold. Reducing the dataset size by another order of magnitude to  $n = 640$  (using a sampling interval of  $\Delta t = 0.1$  and  $\ell = 200$  eigenfunctions), the reduction of skill is, as might be expected, more noticeable, bringing down the predictability horizon for  $\omega^1$  and  $\omega^2$  to  $\tau \simeq 1.1$  natural time units. Still, despite this reduction of skill, the ability to control the complexity of the forecast function by controlling the number of eigenfunctions allows KAF to behave stably in sparsely sampled environments.

In Fig. 7, we also show RMSE results for a larger dataset consisting of  $n = 512,000$  samples taken at a  $\Delta t = 0.01$  sampling interval. Using  $\ell = 6000$  eigenfunctions, this larger dataset is seen to provide a noticeable improvement of skill over the  $n = 64,000$  benchmark, particularly for  $\tau \gtrsim 1$  leads. For instance, the  $n = 512,000$  setup maintains lower than 0.6 normalized RMSE values out to  $\tau \simeq 2.7$ , which represents a  $\simeq 30\%$  increase over the  $n = 64,000$  case.

**Forecasting conditional probability.** As our final numerical results, we show in Fig. 8 trajectory and RMSE results for prediction of





**Fig. 8.** As in Fig. 2, but for prediction of the characteristic functions  $\chi^1$ ,  $\chi^2$ , and  $\chi^3$ , representing the L63 states whose components  $\omega^1$ ,  $\omega^2$ , and  $\omega^3$ , respectively, are greater than their mean. The initial condition in the top panels is the same as in the top panels of Fig. 2.

the characteristic functions  $\chi^i$ , corresponding to the conditional probabilities for  $\omega^i$  to take greater-than-average values. These forecasts were obtained using the empirical target functions  $g_{\tau,\ell,n}$  from (31), constructed using the same parameter values as the state vector forecasts in Fig. 2 based on  $f_{\tau,\ell,n}$ . Compared to forecasts of the state vector components, forecasts of characteristic functions are hampered by the fact that a characteristic function has discontinuities (apart from trivial cases), potentially inducing Gibbs oscillations in approximations by finite linear combinations of RKHS functions. Such oscillations may in turn induce overshoots outside the interval  $[0, 1]$ , necessitating the use of thresholding in (31). Evidence of active thresholding can be seen in Fig. 8, particularly at early times ( $\tau \lesssim 0.5$ ) where the relative RMSE is significantly larger than the corresponding state vector results in Fig. 2. Despite that, the conditional probability results are broadly consistent with their state vector counterparts. That is,  $\chi^1/\chi^2$  and  $\chi^3$  remain predictable out to 2–3 and 4–5 natural time units, similarly to  $\omega^1/\omega^2$  and  $\omega^3$ , respectively, and as expected, the fully observed forecasts fare substantially better than the partially observed ones. It is worthwhile noting that unlike  $\omega^i$ , the fully observed  $\chi^i$  forecasts have not converged to near-constant values at the end of the examined prediction intervals (i.e., at  $\tau = 5$ ).

## 7. Conclusions

As shown throughout this paper, the theoretical underpinning of the kernel approach to forecasting is that it approximates the conditional expectation of observables, in the sense of minimizing mean square forecast error. The extent to which approximating the conditional expectation is one of the better ways of producing forecasts depends on the specific dynamical system and the goals of the forecaster, but is a natural and common target in many applications. Although there are many ways of achieving this approximation, we have shown in this article that the kernel approach is a distinguished such method both theoretically, given the central importance of Hilbert space theory for both kernels and the conditional expectation, and practically, as the

computation requires only eigenfunction computation and matrix multiplication.

In contrast to the usual expositions of kernel methods that present kernels and RKHSs as an axiomatic starting point, we have followed a different order in which the appearance of kernels arises naturally within a learning framework (with hypothesis spaces as a prominent object) and with error minimization as a starting point. It is this perspective on kernel based forecasting, i.e. one that deemphasizes kernels in favor of conditional expectation, that is the main contribution of this paper. Additionally, we presented and proved the results that the empirically obtained kernel forecasting function approaches, in the limit of large data, the ideal kernel forecasting function, which itself approaches the true conditional expectation of the response variable acted upon by the Koopman operator as more principal components (kernel eigenfunctions) are utilized. Precise estimates on the rate of convergence is an active area of research and one that depends on specific aspects of both the dynamical system and the employed kernels.

One of the advantages of an understanding of kernel forecasting based on conditional expectation, so we have argued, is that it enables the computation of a host of related quantities, including conditional probability and conditional variance. The former can be used to handle the binary classification problem that arises when trying to detect extreme or rare events. The latter, meanwhile, is instrumental in providing more informative forecasts that detail the level of uncertainty involved. Another benefit of the statistical learning framework is that it shows the connection between the two most common kernel methods, KPCR (of which KAF is an example) and KRR; in particular, they both follow the same variational logic, but the former is based on a choice of a linear hypothesis space whereas the latter uses a nonlinear one. Although KRR may be a simpler algorithm to implement, and may be more accurate in the presence of noise, KPCR can converge much more rapidly when the predictand happens to lie in the space spanned by the leading principal components. We have also shown how KAF can be implemented using a class of non-symmetric kernels satisfying a detailed-balance condition, as well as kernels based on delay-coordinate maps.



Applications of KAF to two low-dimensional dynamical systems were presented for the sake of illustration. The first system, periodic flow on a circle, is in fact not a system for which conditional expectation is a good estimate of forecasts, at least when the abscissa is the only quantity on which the forecast is conditioned. Nevertheless, we demonstrate that KAF implemented with a radial Gaussian kernel converges to this conditional expectation quite rapidly, and that the dependence of its error on the number of principal components follows a U-shaped curve that is characteristic of the classic bias-variance trade-off of statistical learning. The second system, the L63 system, exhibits a number of the hallmark challenges in forecasting of complex systems, including invariant measures supported on non-smooth sets (fractal attractors) and mixing dynamics. Despite these challenges, we saw that KAF, implemented with a judiciously chosen variable-bandwidth, Markov-normalized Gaussian kernel, successfully predicts the state vector components, as well as their associated conditional probabilities to take greater-than average values. As expected, conditioning on the full state produces better forecasts than conditioning on just partial observations of the state, but in both cases the method yielded adequate uncertainty quantification through estimates of the conditional variance. It was also found that incorporating a sufficient number of delays leads to recovery of most of the forecast skill lost due to partial observations. The L63 example also demonstrates that forecasts based on partial conditioning are better for some choices of response variables than others. In particular, as expected from symmetry considerations, the first coordinate of the state vector has greater predictive value for the third coordinate than for the second coordinate.

There are two chief challenges in utilizing kernel methods in real-world applications. The first is an appropriate choice of response, as well as a covariate variable with sufficiently rich predictive value. The second challenge, particularly when dealing with high-dimensional covariate spaces, is a choice of kernel such that as much of the dynamical features of interest can be characterized by as few of the leading principal components as possible. In general, the response and covariate are selected with the certainty that there is close association between the two, but with the precise nature of the correspondence being either unknown, or intractable to reproduce analytically or numerically. In real-world applications, this issue is further compounded by the fact that the response space is oftentimes multi-dimensional. While in this paper we did not directly address this situation, it is natural to consider extensions of KAF to the setting of vector-valued response functions using operator-valued kernel techniques [104,105] for multi-task learning. Another potential direction for future research is to establish connections between KAF and RKHS embeddings of probability distributions [106,63,107]. As for the choice of kernel, recent approaches for learning kernels targeted to specific response functions [108] could potentially provide effective ways of ensuring that the response is well-captured by the leading eigenspaces of the corresponding integral operator, thus improving forecast skill. The main goal of this paper has been to clarify the theoretical justification for utilizing kernels in forecasting observables of dynamical systems, so that the forecaster can focus on the remaining problem of leveraging specific scientific knowledge of the system into optimal choices of response, covariate, and kernel.

### CRedit authorship contribution statement

**Romeo Alexander:** Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing.  
**Dimitrios Giannakis:** Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing, Supervision.

### Acknowledgments

Dimitrios Giannakis acknowledges support by ONR YIP grant N00014-16-1-2649, NSF grants DMS-1521775 and 1842538, and DARPA grant HR0011-16-C-0116. Romeo Alexander was supported as a Ph.D. student from the first NSF grant and the DARPA grant. The authors would like to thank Suddhasattwa Das, Krithika Manohar, and Andrew Stuart for fruitful conversations. In addition, they would like to thank the Editor and two anonymous Reviewers for constructive comments which have led to improvements of the manuscript. Dimitrios Giannakis 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.

### Appendix. Definitions and technical results

#### A.1. Moore–Penrose pseudoinverse

We state below the definition of the pseudoinverse of a linear map between Hilbert spaces [57].

**Definition 18** (Moore–Penrose pseudoinverse). Let  $H_1$  and  $H_2$  be Hilbert spaces over the complex numbers, and  $A : D(A) \rightarrow H_2$  a closed linear map with dense domain  $D(A) \subseteq H_1$ . Then, a densely defined operator  $A^+ : D(A^+) \rightarrow H_1$  with domain  $D(A^+) \subseteq H_2$  is said to be a Moore–Penrose pseudoinverse of  $A$  if (i)  $\ker A^+ = \text{ran } A^\perp$ ; (ii)  $\text{ran } A^+ = \ker A^\perp$ ; and (iii)  $AA^+f = f$  for all  $f \in \text{ran } A$ .

If  $A^+$  in the above definition exists, then it is closed and unique. Moreover, if  $A$  has closed range, then  $A^+$  always exists, and is a bounded operator with  $D(A^+) = H_2$ . If  $A$  is bounded, we can express  $A^+$  on the potentially restricted domain  $D((AA^*)^+) \subseteq D(A^+)$  of the pseudoinverse of the self-adjoint operator  $AA^*$  through a formula with a direct counterpart in finite-dimensional linear algebra, viz.

$$A^+f = A^*(AA^*)^+f, \quad \forall f \in D((AA^*)^+). \quad (\text{A.1})$$

#### A.2. Proof of Proposition 16

First, the symmetry and positive-definiteness of  $k$  follow directly from its definition and the detailed-balance condition in (22). In particular, since  $d$  is strictly positive, (24) implies that  $k$  is positive-definite if and only if  $\hat{k}$  is positive-definite, and the latter is indeed positive-definite since it is related to the positive-definite kernel  $w$  by a similarity transformation.

Next, the integral operator  $W$  is equal to  $\iota^*M_d$ , and because  $M_d$  is a bounded, invertible operator with bounded inverse,  $W$  is well-defined on  $L^2(\rho)$  and  $\text{ran } W = \text{ran } \iota^* \subseteq \mathcal{K}$ , proving Claim (i).

Turning to Claim (ii), the fact that  $J$  is a trace-class non-negative operator with real eigenvalues follows from its relation to  $G$  (which has all of these properties by positive-definiteness and continuity of  $\hat{k}$ ) via the similarity transformation in (25). In addition, the existence of the biorthonormal Riesz bases  $\{\xi_1, \xi_2, \dots\}$  and  $\{\xi'_1, \xi'_2, \dots\}$  follows from (26) in conjunction with continuity of  $d^{1/2}$  and  $d^{-1/2}$  (and thus boundedness of these functions on the compact support of  $\rho$ ).

Finally, to prove Claim (iii) note first that by Definition 18,

$$\begin{aligned} D(J^+) &= \text{ran } J \oplus \ker J^* = \text{ran } (\iota W) + \ker (W^* \iota^*) \\ &= \text{ran } (\iota^* M_d) + \ker (M_d \iota^*) = \text{ran } (\iota^*) \oplus \ker (\iota^*) \\ &= \text{ran } G \oplus \ker G = D(G^+), \end{aligned}$$

where the second equality in the second line follows from the fact that  $M_d$  is a bounded, invertible operator with bounded inverse. Moreover,

$$\text{ran } G = \text{ran}(\iota^*) \subseteq \text{ran } \iota,$$

and because  $\text{ran } \iota^* \subseteq \ker \iota^\perp$ , we have

$$\overline{\text{ran } G} = \ker G = \ker \iota^* = \overline{\text{ran } \iota}. \quad (\text{A.2})$$

It therefore follows that  $\text{ran } G$  is a dense subspace of  $\text{ran } \iota$ . Now, by Lemma 4 and (A.2),

$$D(\tilde{T}^+) = \text{ran } \iota \oplus \ker \iota^* = \text{ran } \iota \oplus \ker G,$$

so we conclude that  $D(J^+) = D(G^+)$  is a dense subspace of  $D(\tilde{T})$ , as claimed. Moreover,

$$\begin{aligned} \tilde{T}|_{D(J^+)} &= \tilde{T}|_{D(G^+)} = \iota^* G^+ \\ &= \iota^* (J M_d^{-1})^+ = \iota^* M_d J^+ = W J^+. \end{aligned}$$

The expression for  $\tilde{T}f$  in (23) follows from the result just proved, the definitions of  $\xi_i$ ,  $\xi'_i$ , and  $\theta_i$  in (26) and (27), and the fact that  $\tilde{T}$  is a closed operator. We then verify that the  $\vartheta_i$  are indeed orthogonal, viz.

$$\begin{aligned} \langle \vartheta_i, \vartheta_j \rangle_{\mathcal{K}} &= \frac{1}{\sqrt{\eta_i \eta_j}} \langle W M_d^{-1/2} \hat{\phi}_i, W M_d^{-1/2} \hat{\phi}_j \rangle_{\mathcal{K}} \\ &= \frac{1}{\sqrt{\eta_i \eta_j}} \langle \iota^* M_d^{1/2} \hat{\phi}_i, \iota^* M_d^{1/2} \hat{\phi}_j \rangle_{\mathcal{K}} \\ &= \frac{1}{\sqrt{\eta_i \eta_j}} \langle \hat{\phi}_i, M_d^{1/2} \iota^* M_d^{1/2} \hat{\phi}_j \rangle_{L^2(\rho)} \\ &= \frac{1}{\sqrt{\eta_i \eta_j}} \langle \hat{\phi}_i, \hat{G}_\rho \hat{\phi}_j \rangle_{L^2(\rho)} = \delta_{ij}. \end{aligned}$$

This completes the proof of Claim (iii) and Proposition 16.

## References

- [1] E. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues, *J. Atmos. Sci.* 26 (1969) 636–646, [http://dx.doi.org/10.1175/1520-0469\(1969\)26<636:aparbn>2.0.co;2](http://dx.doi.org/10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2).
- [2] J.D. Farmer, J.J. Sidorowich, Predicting chaotic time series, *Phys. Rev. Lett.* 59 (8) (1987) 845–848, <http://dx.doi.org/10.1103/PhysRevLett.59.845>.
- [3] M. Casdagli, Nonlinear prediction of chaotic time series, *Physica D* 35 (3) (1989) 335–356, [http://dx.doi.org/10.1016/0167-2789\(89\)90074-2](http://dx.doi.org/10.1016/0167-2789(89)90074-2).
- [4] G. Sugihara, R.M. May, Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, *Nature* 344 (1990) 734–741, <http://dx.doi.org/10.1038/344734a0>.
- [5] T. Sauer, Time series prediction by using delay coordinate embedding, in: N.A. Gerhsenfeld A.S. Weigend (Ed.), *Time Series Prediction: Forecasting the Future and Understanding the Past*, in: SFI Studies in the Sciences of Complexity, vol. 15, Addison-Wesley, 1993, pp. 175–193.
- [6] J. Fan, Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer Science & Business Media, 2008.
- [7] Z. Zhao, D. Giannakis, Analog forecasting with dynamics-adapted kernels, *Nonlinearity* 29 (2016) 2888–2939, <http://dx.doi.org/10.1088/0951-7715/29/9/2888>.
- [8] L. Delle Monache, T. Nipen, Y. Liu, G. Roux, R. Stull, Kalman filter and analog schemes to postprocess numerical weather predictions, *Mon. Weather Rev.* 139 (11) (2011) 3554–3570, <http://dx.doi.org/10.1175/2011MWR3653.1>.
- [9] A. Atencia, I. Zawadzki, A comparison of two techniques for generating ensembles. Part II: Analogs selection and comparison of techniques, *Mon. Weather Rev.* 143 (2015) 2890–2908, <http://dx.doi.org/10.1175/MWR-D-14-00342.1>.
- [10] H. Van den Dool, *Empirical Methods in Short-Term Climate Prediction*, Oxford University Press, Oxford, 2006.
- [11] R. Alexander, Z. Zhao, E. Szekeley, D. Giannakis, Kernel analog forecasting of tropical intraseasonal oscillations, *J. Atmos. Sci.* 74 (2017) 1321–1342, <http://dx.doi.org/10.1175/JAS-D-16-0147>.
- [12] D. Comeau, Z. Zhao, D. Giannakis, A.J. Majda, Data-driven prediction strategies for low-frequency patterns of North Pacific climate variability, *Clim. Dyn.* 48 (5–6) (2017) 1855–1872, <http://dx.doi.org/10.1007/s00382-016-3177-5>.
- [13] H. Ding, M. Newman, M.A. Alexander, A.T. Wittenberg, Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs, *J. Clim.* 31 (2018) 5437–5459, <http://dx.doi.org/10.1175/JCLI-D-17-0661.1>.
- [14] D. Comeau, D. Giannakis, Z. Zhao, A.J. Majda, Predicting regional and pan-arctic sea ice anomalies with kernel analog forecasting, *Clim. Dyn.* 52 (9–10) (2019) 5507–5525, <http://dx.doi.org/10.1007/s00382-018-4459-x>.
- [15] X. Wang, J. Slawinska, D. Giannakis, Extended-range statistical ENSO prediction through operator-theoretic techniques for nonlinear dynamics, *Sci. Rep.* 10 (2020) 2636, <http://dx.doi.org/10.1038/s41598-020-59128-7>.
- [16] C. Voyant, M.L. Nivet, C. Paoli, M. Muselli, G. Notton, Meteorological time series forecasting based on MLP modelling using heterogeneous transfer functions, *J. Phys. Conf. Ser.* 574 (1) (2015) 012064, <http://dx.doi.org/10.1088/1742-6596/574/1/012064>.
- [17] K. Chakraborty, K. Mehrotra, C.K. Mohan, S. Ranka, Forecasting the behavior of multivariate time series using neural networks, *Neural Netw.* 5 (6) (1992) 961–970, [http://dx.doi.org/10.1016/S0893-6080\(05\)80092-9](http://dx.doi.org/10.1016/S0893-6080(05)80092-9).
- [18] N.K. Ahmed, A.F. Atiya, N.E. Gayar, H. El-Shishiny, An empirical comparison of machine learning models for time series forecasting, *Econ. Rev.* 29 (5–6) (2010) 594–621, <http://dx.doi.org/10.1016/j.strusafe.2015.05.001>.
- [19] M.A. Aizerman, Theoretical foundations of the potential function method in pattern recognition learning, *Autom. Remote Control* 25 (1964) 821–837.
- [20] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ACM, 1992, pp. 144–152.
- [21] T. Hofmann, B. Schölkopf, A.J. Smola, Kernel methods in machine learning, *Ann. Statist.* 36 (3) (2008) 1171–1220, <http://dx.doi.org/10.1214/009053607000000677>.
- [22] I. Steinwart, A. Christmann, *Support Vector Machines*, Springer, 2008.
- [23] H. Drucker, C.J. Burges, L. Kaufman, A.J. Smola, V. Vapnik, Support vector regression machines, in: *Adv. Neural Inf. Process. Syst.*, 1997, pp. 155–161.
- [24] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, in: *ICML-1998 Proceedings of the 15th International Conference on Machine Learning* (04 1999).
- [25] R. Rosipal, L.J. Trejo, A. Cichocki, Kernel Principal Component Regression with EM Approach to Nonlinear Principal Components Extraction, *Tech. Rep. 12*, University of Paisley, Scotland, 2000.
- [26] K.-R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, V. Vapnik, Predicting time series with support vector machines, in: *International Conference on Artificial Neural Networks*, Springer, 1997, pp. 999–1004.
- [27] P. Exterkate, P.J. Groenen, C. Heij, D. van Dijk, Nonlinear forecasting with many predictors using kernel ridge regression, *Int. J. Forecast.* 32 (3) (2016) 736–753, <http://dx.doi.org/10.2139/ssrn.1738192>.
- [28] F.E. Tay, L. Cao, Application of support vector machines in financial time series forecasting, *Omega* 29 (4) (2001) 309–317, [http://dx.doi.org/10.1016/S0305-0483\(01\)00026-3](http://dx.doi.org/10.1016/S0305-0483(01)00026-3).
- [29] V.N. Vapnik, An overview of statistical learning theory, *IEEE Trans. Neural Netw.* 10 (5) (1999) 988–999, <http://dx.doi.org/10.1109/72.788640>.
- [30] F. Cucker, D. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, in: *Cambridge Monographs on Applied and Computational Mathematics*, Cambridge University Press, 2007.
- [31] V. Kuznetsov, M. Mohri, Learning theory and algorithms for forecasting non-stationary time series, in: *Adv. Neural Inf. Process. Syst.*, 2015, pp. 541–549.
- [32] V. Kuznetsov, M. Mohri, Time series prediction and online learning, in: *Conference on Learning Theory*, 2016, pp. 1190–1213.
- [33] V. Baladi, Positive Transfer Operators and Decay of Correlations, in: *Advanced Series in Nonlinear Dynamics*, vol. 16, World scientific, Singapore, 2000.
- [34] T. Eisner, B. Farkas, M. Haase, R. Nagel, *Operator Theoretic Aspects of Ergodic Theory*, Vol. 272, Springer, 2015.
- [35] M. Budisić, R. Mohr, I. Mezić, Applied Koopmanism, *Chaos* 22 (2012) 047510, <http://dx.doi.org/10.1063/1.4772195>.
- [36] B.O. Koopman, Hamiltonian systems and transformation in Hilbert space, *Proc. Natl. Acad. Sci.* 17 (5) (1931) 315–318, <http://dx.doi.org/10.1073/pnas.17.5.315>.
- [37] M. Dellnitz, O. Junge, On the approximation of complicated dynamical behavior, *SIAM J. Numer. Anal.* 36 (1999) 491, <http://dx.doi.org/10.1137/S0036142996313002>.
- [38] I. Mezić, A. Banaszuk, Comparison of systems with complex behavior, *Physica D* 197 (2004) 101–133, <http://dx.doi.org/10.1016/j.physd.2004.06.015>.
- [39] I. Mezić, Spectral properties of dynamical systems, model reduction and decompositions, *Nonlinear Dynam.* 41 (2005) 309–325, <http://dx.doi.org/10.1007/s11071-005-2824-x>.
- [40] C.W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, D.S. Henningson, Spectral analysis of nonlinear flows, *J. Fluid Mech.* 641 (2009) 115–127, <http://dx.doi.org/10.1017/S0022112009992059>.
- [41] P.J. Schmid, Dynamic mode decomposition of numerical and experimental data, *J. Fluid Mech.* 656 (2010) 5–28, <http://dx.doi.org/10.1017/S0022112010001217>.

- [42] G. Froyland, O. Junge, P. Koltai, Estimating long-term behavior of flows without trajectory integration: The infinitesimal generator approach, *SIAM J. Numer. Anal.* 51 (1) (2013) 223–247, <http://dx.doi.org/10.1137/110819986>.
- [43] G. Froyland, G.A. Gottwald, A. Hammerlindl, A computational method to extract macroscopic variables and their dynamics in multiscale systems, *SIAM J. Appl. Dyn. Syst.* 13 (4) (2014) 1816–1846, <http://dx.doi.org/10.1137/130943637>.
- [44] D. Giannakis, J. Slawinska, Z. Zhao, Spatiotemporal feature extraction with data-driven Koopman operators, in: *J. Mach. Learn. Res. Proceedings*, vol. 44, 2015, pp. 103–115.
- [45] M.O. Williams, I.G. Kevrekidis, C.W. Rowley, A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition, *J. Nonlinear Sci.* 25 (6) (2015) 1307–1346, <http://dx.doi.org/10.1007/s00332-015-9258-5>.
- [46] S. Klus, C. Koltai, P. Scütte, On the numerical approximation of the Perron-Frobenius and Koopman operator, *J. Comput. Dyn.* 3 (1) (2016) 51–79, <http://dx.doi.org/10.3934/jcd.2016003>.
- [47] S.L. Brunton, B.W. Brunton, J.L. Proctor, E. Kaiser, J.N. Kutz, Chaos as an intermittently forced linear system, *Nature Commun.* 8 (19) (2017) <http://dx.doi.org/10.1038/s41467-017-00030-8>.
- [48] J. Harlim, H. Yang, Diffusion forecasting model with basis functions from QR-decomposition, *J. Nonlinear Sci.* 28 (2018) 847–872, <http://dx.doi.org/10.1007/s00332-017-9430-1>.
- [49] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, F. Noé, Data-driven model reduction and transfer operator approximation, *J. Nonlinear Sci.* 28 (2018) 985–1010, <http://dx.doi.org/10.1007/s00332-017-9437-7>.
- [50] M. Korda, M. Putinar, I. Mezić, Data-driven spectral analysis of the Koopman operator, *Appl. Comput. Harmon. Anal.* 48 (2) (2020) 599–629, <http://dx.doi.org/10.1016/j.acha.2018.08.002>.
- [51] S. Das, D. Giannakis, Delay-coordinate maps and the spectra of Koopman operators, *J. Stat. Phys.* 175 (6) (2019) 1107–1145, <http://dx.doi.org/10.1007/s10955-019-02272-w>.
- [52] D. Giannakis, Data-driven spectral decomposition and forecasting of ergodic dynamical systems, *Appl. Comput. Harmon. Anal.* 62 (2) (2019) 338–396, <http://dx.doi.org/10.1016/j.acha.2017.09.001>.
- [53] J.C. Ferreira, V.A. Menegatto, Positive definiteness reproducing kernel Hilbert spaces, and beyond, *Ann. Funct. Anal.* 4 (2013) 64–88, <http://dx.doi.org/10.15352/afa/1399899838>.
- [54] V.I. Paulsen, M. Raghupathi, An introduction to the theory of reproducing kernel Hilbert spaces, in: *Cambridge Studies in Advanced Mathematics*, Cambridge University Press, Cambridge, 2016, <http://dx.doi.org/10.1017/CBO9781316219232>.
- [55] E.N. Lorenz, Deterministic nonperiodic flow, *J. Atmos. Sci.* 20 (1963) 130–141, [http://dx.doi.org/10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](http://dx.doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2).
- [56] O. Kallenberg, Foundations of modern probability, in: *Probability and Its Applications*, Springer, New York, 1997, <http://dx.doi.org/10.1007/b98838>.
- [57] F.J. Beutler, W.L. Root, The operator pseudoinverse in control and systems identification, in: M. Zuhair Nashed (Ed.), *Generalized Inverses and Applications*, University of Wisconsin-Madison, Academic Press, 1973, pp. 397–494, <http://dx.doi.org/10.1016/B978-0-12-514250-2.50014-7>.
- [58] B. Schölkopf, A. Smola, K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Comput.* 10 (1998) 1299–1319, <http://dx.doi.org/10.1162/089976698300017467>.
- [59] L. Rosasco, M. Belkin, E. De Vito, On learning with integral operators, *J. Mach. Learn. Res.* 11 (2010) 905–934, <http://dx.doi.org/10.5555/1756006.1756036>.
- [60] R.R. Coifman, S. Lafon, Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions, *Appl. Comput. Harmon. Anal.* 21 (2006) 31–52, <https://doi.org/10.1016/j.acha.2005.07.005>.
- [61] A. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* 63 (1950) 337–404, <http://dx.doi.org/10.1090/S0002-9947-1950-0051437-7>.
- [62] C.M. Brislawn, Traceable integral kernels on countable generated measure spaces, *Pacific. J. Math.* 150 (2) (1991) 229–240.
- [63] B.K. Sriperumbudur, K. Fukumizu, G.R. Lanckriet, Universality, characteristic kernels and RKHS embedding of measures, *J. Mach. Learn. Res.* 12 (2011) 2389–2410.
- [64] T. Berry, D. Giannakis, J. Harlim, Nonparametric forecasting of low-dimensional dynamical systems, *Phys. Rev. E* 91 (2015) 032915, <http://dx.doi.org/10.1103/PhysRevE.91.032915>.
- [65] R.R. Coifman, S. Lafon, Diffusion maps, *Appl. Comput. Harmon. Anal.* 21 (2006) 5–30, <http://dx.doi.org/10.1016/j.acha.2006.04.006>.
- [66] U. von Luxburg, M. Belkin, O. Bousquet, Consistency of spectral clustering, *Ann. Statist.* 26 (2) (2008) 555–586, <http://dx.doi.org/10.1214/009053607000000640>.
- [67] R. Coifman, M. Hirn, Bi-stochastic kernels via asymmetric affinity functions, *Appl. Comput. Harmon. Anal.* 35 (1) (2013) 177–180, <http://dx.doi.org/10.1016/j.acha.2013.01.001>.
- [68] T. Berry, J. Harlim, Variable bandwidth diffusion kernels, *Appl. Comput. Harmon. Anal.* 40 (1) (2016) 68–96, <http://dx.doi.org/10.1016/j.acha.2015.01.001>.
- [69] T. Berry, T. Sauer, Local kernels and the geometric structure of data, *Appl. Comput. Harmon. Anal.* 40 (3) (2016) 439–469, <http://dx.doi.org/10.1016/j.acha.2015.03.002>.
- [70] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.* 15 (2003) 1373–1396, <http://dx.doi.org/10.1162/089976603321780317>.
- [71] M. Hein, J.-Y. Audibert, U. von Luxburg, From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians, in: *Learning Theory*, in: *Lecture Notes in Computer Science*, vol. 3559, Springer, Berlin, 2005, pp. 470–485, [http://dx.doi.org/10.1007/11503415\\_32](http://dx.doi.org/10.1007/11503415_32).
- [72] A. Singer, From graph to manifold Laplacian: The convergence rate, *J. Appl. Comput. Harmon. Anal.* 21 (2006) 128–134, <http://dx.doi.org/10.1016/j.acha.2006.03.004>.
- [73] H. König, Eigenvalue distribution of compact operators, in: *Operator Theory: Advances and Applications*, vol. 16, Birkhäuser, Basel, 1986.
- [74] J.C. Ferreira, V.A. Menegatto, Eigenvalues of integral operators defined by smooth positive definite kernels, *Integral Equations Operator Theory* 64 (2009) 61–81, <http://dx.doi.org/10.1007/s00020-009-1680-3>.
- [75] J. Ferreira, V.A. Menegatto, Eigenvalue decay rates for positive integral operators, *Ann. Mat. Pura Appl.* 192 (6) (2013) 1025–1041, <http://dx.doi.org/10.1007/s10231-012-0256-z>.
- [76] K. Law, A. Shukla, A.M. Stuart, Analysis of the 3DVAR filter for the partially observed Lorenz'63 model, *Discrete Contin. Dyn. Syst.* 34 (3) (2013) 1061–10178, <http://dx.doi.org/10.3934/dcds.2014.34.1061>.
- [77] P. Constantin, C. Foias, B. Nicolaenko, R. Témam, *Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations*, Springer, 1989, <http://dx.doi.org/10.1007/978-1-4612-3506-4>.
- [78] L.-S. Young, What are SRB measures and which dynamical systems have them? *J. Stat. Phys.* 108 (2002) 733–754, <http://dx.doi.org/10.1023/A:1019762724717>.
- [79] D. Giannakis, A. Ourmazd, J. Slawinska, Z. Zhao, Spatiotemporal pattern extraction by spectral analysis of vector-valued observables, *J. Nonlinear Sci.* 29 (5) (2019) 2385–2445, <http://dx.doi.org/10.1007/s00332-019-09548-1>.
- [80] P. Walters, An introduction to ergodic theory, in: *Graduate Texts in Mathematics*, vol. 79, Springer-Verlag, New York, 1981.
- [81] M. Small, K. Judd, Comparisons of new nonlinear modeling techniques with applications to infant respiration, *Physica D* 117 (1998) 283–298, [http://dx.doi.org/10.1016/S0167-2789\(97\)00311-4](http://dx.doi.org/10.1016/S0167-2789(97)00311-4).
- [82] C.A. Micchelli, Interpolation of scattered data: Distance matrices and conditionally positive definite functions, *Constr. Approx.* 2 (1986) 11–22, <http://dx.doi.org/10.1007/BF01893414>.
- [83] M.C. Genton, Classes of kernels for machine learning: A statistics perspective, *J. Mach. Learn. Res.* 2 (2001) 299–312.
- [84] D.D. Kosambi, Statistics in function space, *J. Indian Math. Soc.* 7 (1943) 76–88.
- [85] P. Holmes, J.L. Lumley, G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, 1996.
- [86] C. Penland, Random forcing and forecasting using principal oscillation pattern analysis, *Mon. Weather Rev.* 117 (10) (1989) 2165–2185, [http://dx.doi.org/10.1175/1520-0493\(1989\)117<2165:RFAFUP>2.0.CO;2](http://dx.doi.org/10.1175/1520-0493(1989)117<2165:RFAFUP>2.0.CO;2).
- [87] S. Das, D. Giannakis, Koopman spectra in reproducing kernel Hilbert spaces, 2018, [arXiv:1801.07799](https://arxiv.org/abs/1801.07799).
- [88] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: *Advances in Neural Information Processing Systems*, vol. 17, 2004, pp. 1601–1608.
- [89] S. Das, D. Giannakis, J. Slawinska, Reproducing kernel Hilbert space compactification of unitary evolution groups, 2019, [arXiv:1808.01515](https://arxiv.org/abs/1808.01515).
- [90] N.H. Packard, et al., Geometry from a time series, *Phys. Rev. Lett.* 45 (1980) 712–716, <http://dx.doi.org/10.1103/physrevlett.45.712>.
- [91] F. Takens, Detecting strange attractors in turbulence, in: *Lecture Notes in Mathematics*, vol. 898, Springer, Berlin, 1981, pp. 366–381, <http://dx.doi.org/10.1007/bfb0091924>.
- [92] T. Sauer, J.A. Yorke, M. Casdagli, Embedology, *J. Stat. Phys.* 65 (3–4) (1991) 579–616, <http://dx.doi.org/10.1007/bf01053745>.
- [93] J.C. Robinson, A topological delay embedding theorem for infinite-dimensional dynamical systems, *Nonlinearity* 18 (5) (2005) 2135–2143, <http://dx.doi.org/10.1088/0951-7715/18/5/013>.
- [94] D.S. Broomhead, G.P. King, Extracting qualitative dynamics from experimental data, *Physica D* 20 (2–3) (1986) 217–236, [http://dx.doi.org/10.1016/0167-2789\(86\)90031-x](http://dx.doi.org/10.1016/0167-2789(86)90031-x).
- [95] R. Vautard, M. Ghil, Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Physica D* 35 (1989) 395–424, [http://dx.doi.org/10.1016/0167-2789\(89\)90077-8](http://dx.doi.org/10.1016/0167-2789(89)90077-8).

- [96] D. Giannakis, A.J. Majda, Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability, *Proc. Natl. Acad. Sci.* 109 (7) (2012) 2222–2227, <http://dx.doi.org/10.1073/pnas.1118984109>.
- [97] T. Berry, R. Cressman, Z. Gregurić-Ferenčak, T. Sauer, Time-scale separation from diffusion-mapped delay coordinates, *SIAM J. Appl. Dyn. Syst.* 12 (2013) 618–649, <http://dx.doi.org/10.1137/12088183x>.
- [98] H. Arbabi, I. Mezić, Ergodic theory dynamic mode decomposition and computation of spectral properties of the Koopman operator, *SIAM J. Appl. Dyn. Syst.* 16 (4) (2017) 2096–2126, <http://dx.doi.org/10.1137/17M1125236>.
- [99] P.R. Halmos, *Lectures on Ergodic Theory*, American Mathematical Society, Providence, 1956.
- [100] W. Tucker, The Lorenz attractor exists, *C. R. Acad. Sci., Paris I* 328 (1999) 1197–1202.
- [101] M.J. McGuinness, The fractal dimension of the Lorenz attractor, *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 262 (1968) 413–458, <http://dx.doi.org/10.1098/rsta.1968.0001>.
- [102] J.C. Sprott, *Chaos and Time-Series Analysis*, Oxford University Press, Oxford, 2003.
- [103] S. Luzzatto, I. Melbourne, F. Paccaut, The Lorenz attractor is mixing, *Comm. Math. Phys.* 260 (2) (2005) 393–401, <http://dx.doi.org/10.1007/s00220-005-1411-9>.
- [104] C.A. Micchelli, M. Pontil, On learning vector-valued functions, *Neural Comput.* 17 (1) (2005) 177–204, <http://dx.doi.org/10.1162/0899766052530802>.
- [105] J. Slawinska, A. Ourmazd, D. Giannakis, A new approach to signal processing of spatiotemporal data, in: *IEEE Statistical Signal Processing Workshop*, Freiburg, Germany, 2018, pp. 338–342. <http://dx.doi.org/10.1109/SSP.2018.8450704>.
- [106] L. Song, J. Huang, A. Smola, K. Fukumizu, Hilbert space embeddings of conditional distributions with applications to dynamical systems, in: *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009, pp. 961–968.
- [107] S. Klus, I. Schuster, K. Muandet, Eigendecomposition of transfer operators in reproducing kernel Hilbert spaces, *J. Nonlinear Sci.* 30 (2019) 283–315, <http://dx.doi.org/10.1007/s00332-019-09574-z>.
- [108] O. Owahdi, G.R. Yoo, Kernel flows: From learning kernels from data into the abyss, *J. Comput. Phys.* 389 (2019) 22–47, <http://dx.doi.org/10.1016/j.jcp.2019.03.040>.