Analog Forecasting with Dynamics-Adapted Kernels

Zhizhen Zhao
Department of Mathematics and Center for Atmosphere and Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
E-mail: jzhao@cims.nyu.edu

Dimitrios Giannakis
Department of Mathematics and Center for Atmosphere and Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
E-mail: dimitris@cims.nyu.edu

Abstract. Analog forecasting is a non-parametric technique introduced by Lorenz in 1969 which predicts the evolution of states of a dynamical system (or observables defined on the states) by following the evolution of the sample in a historical record of observations which most closely resembles the current initial data. Here, we introduce a suite of forecasting methods which improve traditional analog forecasting by combining ideas from state-space reconstruction for dynamical systems and kernel methods developed in harmonic analysis and machine learning. The first improvement is to augment the dimension of the initial data using Takens’ delay-coordinate maps to recover information in the initial data lost through partial observations. Then, instead of using Euclidean distances between the states, weighted ensembles of analogs are constructed according to similarity kernels in delay-coordinate space, featuring an explicit dependence on the dynamical vector field generating the data. The eigenvalues and eigenfunctions of associated kernel operators define diffusion coordinates and diffusion distances on the data manifold, improving the identification of skillful analogs compared to the Euclidean distance. Moreover, forecasts based on kernel-weighted ensembles have significantly higher skill than the conventional approach following a single analog. We discuss alternative approaches for kernel-weighted ensemble forecasting based on the Nyström method for out-of-sample extension of functions, as well as multiscale methods based on Laplacian pyramids. We illustrate these techniques in applications to forecasting in a low-order deterministic model for atmospheric dynamics with chaotic metastability, and interannual-scale forecasting in the North Pacific sector of a comprehensive climate model.

Keywords: Analog forecasting, kernel methods, delay-coordinate maps, out-of-sample extension
1. Introduction

In many branches of science and engineering, advances in modeling capabilities and the proliferation of experimental and observational data have spurred the development of new mathematical and computational techniques for predicting the dynamical behavior of complex nonlinear systems with imperfect models and partial knowledge of the initial data. The Earth’s climate system is a classical example of such a complex and partially observed system coupling physical processes from the atmosphere, ocean, and land over a wide range of spatial and temporal scales. Currently, as well as in the foreseeable future \([27]\), the equations of motion for the full climate system cannot be integrated via direct numerical simulation and must be parameterized to account for unresolved scales, while in some cases the equations of motion are only partially known \([10]\). A topic of significant current interest is how to make accurate and reliable predictions of large-scale climate patterns over timescales ranging from weeks to decades, using imperfect models initialized with the incomplete information about the current state of the system provided by observational networks \([44, 40, 20]\).

One way of making initial-value forecasts in complex systems is through large-scale numerical simulation. For example, in climate science, regional weather models and general circulation models (GCMs) are used extensively to make forecasts over lead times spanning a few hours (in weather forecasting) to several decades (in long-term climate projections). Broadly speaking, these numerical models are based on a dynamical core for the atmosphere and/or ocean derived from the Navier-Stokes equations on a rotating sphere, which is coupled to a complex network of physics models describing processes such as radiative energy transfer and phase transitions. In an operational environment, the models are coupled with a data assimilation system providing initial conditions from the (incomplete) data available from sensors such as satellites, weather stations, and buoys. The significant increases in forecast skill for the weather and climate over the past decades is due to advances in both the forward models and data assimilation systems \([44]\).

Distinct from large-scale numerical models are parametric and non-parametric empirical low-order methods \([35, 21]\). These techniques do not rely on a large-scale forward model in the forecast step. Rather, predictions are made using a model trained on a reference time series (which may be obtained from historical observations of nature, or a large-scale numerical model), and initialized with the currently observed data from nature. The empirical models generally make coarser forecasts and may also underperform large-scale models in terms of innovation, but their strengths lie in their low computational cost and ability to be trained using data from nature. Low computational cost is desirable in scenarios where real-time or near real-time forecasts are required but the cost of frequent large-scale model runs is prohibitive \([59]\), and in applications where low-order process models are embedded in large-scale models as subgrid-scale parameterization schemes \([24]\). Equally important, low-order empirical models can be trained using data acquired from nature, therefore avoiding some of the dynamical model errors present in large-scale forward models. For instance, a major barrier in extending the skill of numerical weather forecasts at timescales greater than about
Analog Forecasting with Dynamics-Adapted Kernels

A week is the poor dynamical representation of organized tropical convection [41]. As a result, low-order empirical models frequently outperform their large-scale counterparts in predicting climate phenomena which are the outcome of convective organization, including the Indian monsoon [60] and the Madden-Julian oscillation [37].

1.1. Analog forecasting

In this paper, we will focus on methods inspired by a non-parametric empirical forecasting technique called analog forecasting. This technique was introduced by Lorentz in 1969 [38] as a method for predicting the time evolution of observables in dynamical systems based on a historical record of training data. In the initialization stage of analog forecasting, one identifies an analog, i.e., the state in the historical record which most closely resembles the current initial data. Then, in the forecast step, the historical evolution of that state is followed for the desired lead time, and the observable of interest is predicted based on its value on the analog. Unlike prediction using GCMs and parametric low-order models, the analog approach makes no assumptions about the dynamics of the system being examined. Thus, if the historical record of observations comes from nature, the method avoids the dynamical model errors of parametric models essentially by construction, allowing the deployment of the method to real-world problems where parametric models are known to perform poorly [58, 60]. Analog forecasting is also useful in situations where the reference time series itself is the output of numerical models. For instance, in [11] analogs are used to assess the long-range internal predictability properties of GCMs. Nonlinear forecasting based on libraries of past histories has also been used in ecology [53].

Three major factors influencing the efficacy of analog forecasting are (i) the identification of “skillful” analogs in the training data; (ii) the choice of forecast observable (predictand); (iii) model error in the training data. Geometrically, the ability to identify skillful analogs amounts to being able to identify subsets of the training data whose dynamical evolution will shadow the future time evolution for the given initial data. In general, the required shadowing accuracy to achieve a given forecast accuracy depends on both the dynamical system and the observable, as different observables have different predictability limits. For instance, the predictability of the full state vector is limited by the maximal Lyapunov exponent of the dynamical system, and a trivial observable equal to a constant is predictable at arbitrarily long lead times. Indeed, an important problem in data-driven forecasting techniques is to identify observables which are physically meaningful but also predictable. The properties of the prediction observable also affect the reliability of analog forecasts in situations where the training data have model error. In particular, in some situations it may be preferable (e.g., when the time span of the available data from nature is short), or necessary (e.g., when the predictand is not directly observable) to use a training dataset generated by an imperfect model, despite that analog forecasting cleverly avoids dynamical model error if the training data are acquired for nature.

In real-world applications, the problem of analog identification is compounded by
the fact that the initial data are often high-dimensional and incomplete (i.e., insufficient to uniquely determine the future dynamical evolution of the system). Moreover, highly predictable observables are not known a priori. As a concrete example, consider climate forecasting in the North Pacific Ocean given observations of sea surface temperature (SST) over a spatial grid. In this example, which will be studied in Section 4.2, the dimension of the initial data is equal to the number of spatial gridpoints where SST measurements are taken. For the spatial resolutions typical of current-generation observational networks and GCMs ($\lesssim 1^\circ$) the number of North Pacific gridpoints is $O(10^4)$. However, knowledge of SST at a single instance of time is not sufficient to determine its future evolution, as many of the essential degrees of freedom of the atmosphere and the ocean (e.g., the vertically-resolved circulation, temperature, and density fields) are not represented in SST snapshots. As a result, similar SST snapshots may be produced by states lying far apart on the attractor, and identifying analogs solely on the basis of similarity of SST snapshots is likely to lead to poor prediction skill.

Regarding the choice of prediction observable, considerable efforts are being made by oceanographers to identify patterns which characterize the variability of the ocean on timescales of several years. In the North Pacific, two such prominent patterns are the Pacific decadal oscillation (PDO) \cite{39} and the North Pacific Gyre oscillation (NPGO) \cite{25}, which are conventionally obtained by applying principal components analysis (PCA) to seasonally-detrended and low-pass filtered SST data. In general, observables measuring the PDO and NPGO activity [e.g., the principal components (PCs) corresponding to these patterns] are subjective, in the sense that they are defined through a data analysis algorithm applied to a finite dataset. PCA-type algorithms, in particular, are known to have poor skill in extracting dynamically significant patterns in certain nonlinear dynamical systems \cite{2, 22}, motivating the development of alternative data-driven approaches for extracting observables for empirical forecasting.

1.2. Contributions of this work

The central theme of this work is that kernel methods, in conjunction with state-space reconstruction methods for dynamical systems, can lead to significant improvements in the traditional formulation of analog forecasting, benefiting both the identification of skillful analogs, as well as the construction of intrinsic slow modes for prediction. To motivate our approach, we note that a generic feature of many dynamical systems of interest in science and engineering is that, after initial transients have decayed, the system evolves on low-dimensional subsets of phase space (attractors) \cite{36, 26}. These nonlinear structures are embedded in the high-dimensional ambient data space in potentially complicated ways, meaning that the best-fitting linear hyperplane is likely to be of significantly higher dimension than the intrinsic dimension of the data. Kernel algorithms \cite{3, 15, 16, 57, 47, 9} are well-suited to handle nonlinear geometrical structures of data such as manifolds. Here, our objective is to improve the traditional analog forecasting using these techniques, adapted to deal with
data generated by dynamical systems [30, 7, 28].

In particular, kernel operators, suitably normalized to account for a potentially non-uniform sampling density of the data [15], can be made to converge to Laplace-Beltrami operators associated with a Riemannian metric on the nonlinear manifold sampled by the data, and that metric depends on the kernel [9]. In this context, kernels can therefore be thought of as inducing a Riemannian geometry, sometimes referred to as diffusion geometry, to the training data. Having in mind that the skill of analog forecasting depends on the ability to geometrically identify good analogs in the training data, we propose to use the properties of a kernel-induced geometry to improve the efficacy of this step. In particular, following [30, 31, 7, 28], our approach is to take advantage of a special structure present in data generated by dynamical systems, namely the time ordering of the samples which is the outcome of the dynamics.

The time-ordering of the data samples has been utilized at least since the early 1980s in state-space reconstruction methods [55, 12, 50]. In these works, it was established that under generic conditions one can recover the topology of the attractor of a dynamical system from time-ordered partial observations by embedding the data into a higher-dimensional space consisting of lagged sequences of observations over a temporal window. At the very least, when the initial data are incomplete, this suggests to perform analog forecasting using Euclidean distances or similarity kernels defined in delay-coordinate space. Besides topological aspects, however, delay-coordinate mapping also influences the geometry of the data irrespective of whether or not the initial data is complete. That is, pairwise distances on the data manifold in delay coordinate space depend not only on “snapshots” at single time instances, but on dynamical trajectories (“videos”). In [7], it was shown that adding delays increasingly biases the kernel-induced metric towards the most stable subspace of the Lyapunov metric, enhancing the kernel’s timescale separation capability. This approach was also used heuristically in earlier work on so-called nonlinear Laplacian spectral analysis (NLSA) algorithms [30, 31], where kernels in delay-coordinate space were found to have significantly higher skill in recovering patterns analogous to the PDO and NPGO in North Pacific data from a GCM. Moreover, a set of spatial patterns recovered via NLSA were found to yield Galerkin-reduced models reproducing faithfully the chaotic regime transitions in a low-order model for the atmosphere [22] where PCA is known to fail dramatically. The low-order atmospheric model and the high-dimensional North Pacific GCM data will both be employed here as applications of our kernel analog forecasting techniques. Moreover, we will study an example with model error in the training data simulated by treating North Pacific SST output from a GCM operating at a given atmospheric resolution as “nature” and the corresponding output obtained from a model with a coarser atmosphere as training data from an imperfect model.

Besides delay-coordinate mappings, the kernel used in NLSA features an additional dependence on the time-tendency of the data evaluated through finite differences of the data in time. In [28], it was shown that the finite-differenced data provide an empirical
approximation of the dynamical vector field on the attractor generating the dynamics. That work also established that by incorporating a suitable directional dependence, biasing the kernel towards states which are mutually aligned with the dynamical flow, one gains additional invariance properties and ability to extract slow dynamical time scales. These so-called cone kernels will be employed here as our preferred similarity kernel to identify skillful analogs in the training data, and we will also use Laplace-Beltrami eigenfunctions associated with cone kernels to construct slow observables with high predictability.

With any kernel method, the ability to perform analog forecasts hinges on the ability to compute low-dimensional coordinates associated with the kernel, and to evaluate observables at points lying outside the training dataset. Here, we carry out these tasks using two popular out-of-sample extension techniques, namely the Nyström method [43, 5, 16] and Laplacian pyramids [47]. The Nyström method performs particularly well when the prediction observable is itself a kernel eigenfunction (which will be the case in the North Pacific SST application), but may become ill-conditioned for more general observables. In those situations, the multiscale procedure for out-of-sample extension in Laplacian pyramids is preferable. Both the Nyström method and Laplacian pyramids formulate out-of-sample extension as a weighted sum of the function values on the training dataset, with weights determined empirically through the kernel similarity function evaluated for the out-of-sample datum and each point in the training data. These weights naturally lead to weighted ensembles of analogs with significantly higher forecast skill than what is possible with single analogs.

This paper is organized as follows. In Section 2, we describe the mathematical framework for kernel-based analog forecasting, focusing on kernels designed for dimension reduction and feature extraction in dynamical systems and out-of-sample extension methods. We present our analog forecasting techniques in Section 3. In Section 4, we discuss applications of these techniques to forecasting in a low-order model for the atmosphere and the North Pacific sector of comprehensive climate models. We conclude in Section 5.

2. Mathematical framework

We consider that we observe samples of a $d$-dimensional signal $z_i = z(t_i)$ sampled uniformly at times $\{t_i\}_{i=1}^N$ with $t_i = (i - 1) \delta t$. The signal is generated by a dynamical system, and observations of $z_i$ at a single instance of time or over a finite time interval will serve as initial data to forecast a quantity of interest (a prediction observable) $f$. We are interested in a partial-observations scenario where $z_i$ provides incomplete knowledge of the state of the system in phase space. For instance, in the ocean application in Section 4.2, $z_i$ will be the SST field measured at $d$ spatial gridpoints in the North Pacific ocean, but observations of $z_i$ alone are not sufficient to uniquely determine the system state and its future time evolution. In such scenarios, it is natural to apply state-space reconstruction methods [55, 12, 50], and use time-lagged trajectories of $z_i$ (rather than snapshots) as initial data, since the data becomes more Markovian under delay-coordinate maps.
In general, the prediction observable can be objectively defined for the system, or it can be constructed through a data analysis technique. In the low-order atmospheric model studied in Section 4.1, \( f \) will be a vector-valued observable consisting of the components of the state vector that capture metastable regime transitions—this is an example of an objectively-defined observable. On the other hand, in Section 4.2 the goal will be to predict the time-evolution of low-frequency modes of variability representing the PDO and NPGO in the North Pacific Ocean, and these modes will be constructed by processing the data via kernel algorithms. Desirable features of such data-driven observables is to evolve on slow intrinsic dynamical timescales, and be stable for out-of-sample extension. For both types of observables, our goal is to compute the kernel-induced affinity of the given initial data, \( y \), to the training dataset (analog identification), and make predictions \( f(y, \tau) \) at lead time \( \tau \) by a weighted sum of the values of the observable on the training dataset shifted forward by the desired lead time. The methods employed here for computing the analog weights are based on kernel methods for out-of-sample extension [16, 47].

Following a discussion of delay-coordinate maps (Section 2.1), in Section 2.2, we introduce the dynamics-adapted kernels used here for both analog identification and construction of prediction observables. Subsequently, in Section 2.4, we describe two methods for extending functions on high-dimensional ambient spaces, which will be used in the prediction strategies developed in Section 3. Additional details on delay-coordinate maps and dynamics-adapted kernels can be found in [32, 28].

2.1. Delay-coordinate maps

Delay-coordinate maps [55, 12, 50] is a powerful tool to recover topological features of attractors of dynamical systems lost through partial observations. Fixing an integer parameter \( q \) (the number of delays), each sample \( z_i \) in \( d \)-dimensional ambient data space with \( i \geq q \) is mapped to the point \( x_{i-q+1} \) in a space of dimension \( m = dq \) (hereafter, called delay-coordinate space) given by the lagged sequence

\[
  z_i \mapsto x_{i-q+1} = (z_i, z_{i-1}, \ldots, z_{i-(q-1)}).
\]  

According to a theorem of Takens [55], which was generalized in [50] to systems with fractal attractors under prevalent sets of observation functions, the image of \( z_i \) in delay-coordinate space lies with high probability on a manifold \( \mathcal{M} \), which is in one-to-one correspondence with the attractor of the dynamical system generating the data, provided that the dimension \( m \) is sufficiently large. Thus, because knowledge of the position on the attractor is sufficient to determine the future evolution of the system, the time series \( \{x_i\}_{i=1}^{N-q+1} \) becomes Markovian even if \( \{z_i\}_{i=1}^{N} \) is non-Markovian. Since analog forecasting relies on following the evolution of observables on such time series starting from appropriate initial conditions (analogs), it is natural to identify those initial conditions using the lagged sequences \( x_i \) rather than \( z_i \). This approach should be beneficial in both traditional analog forecasting with respect to Euclidean distances and the kernel-based methods developed here.
Besides topological effects, time-lagged embedding introduces a new geometry to the dataset, irrespective of whether or not the observations are complete. In particular, because each point in delay-coordinate space corresponds to the segment of dynamical evolution in (2.1), distances in that space depend on the dynamical system generating the data. The work in [7] showed that adding suitably-weighted delays increasingly biases the Riemannian metric of $\mathcal{M}$ towards an invariant subspace associated with the most stable Lyapunov direction of the dynamical system (specifically, the most stable Oseledets subspace associated with the Lyapunov metric [45]). As has been demonstrated in a number of applications [30, 7, 13, 54], the geometry of the data in delay-coordinate space significantly enhances the ability to extract intrinsic dynamical timescales with kernel methods. Thus, it may be desirable to perform delay-coordinate mapping even if the observations are full to take advantage of this property. In what follows, we will consider by default that the delay-coordinate mapping in (2.1) has been applied to the training and initial data used for analog forecasting, noting that the special case $q = 1$ corresponds to no embedding.

2.2. Dynamics-adapted kernels

In the context of this work, a kernel is a function $K_\epsilon : \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R}_+$ which maps pairs of points in the $m$-dimensional data space to non-negative numbers, and exhibits fast (exponential) decay away from the diagonal controlled by a bandwidth parameter $\epsilon > 0$ [9]. A standard choice commonly used in machine learning and harmonic analysis is the radial Gaussian kernel [3, 17, 33, 15, 4], $K_\epsilon(y, x) = e^{-\|y-x\|^2/\epsilon}$. In general, one can think of $K_\epsilon(y, x)$ as a measure of similarity between the data points.

As $\epsilon$ becomes small, the kernel becomes localized to small neighborhoods of the manifold $\mathcal{M} \subset \mathbb{R}^m$ sampled by the data. Formally, we will consider that $\mathcal{M}$ is a smooth, compact manifold, but in real-world problems this assumption is likely to only be valid on a coarse scale. Among the major recent advances in machine learning and harmonic analysis has been to establish connections between operators for data analysis constructed through such localizing kernels with classical differential operators on manifolds. In particular, the asymptotic behavior of the radial Gaussian kernel in the limit of large data and small bandwidth ($\epsilon \to 0$) has been studied extensively in the literature [33, 57, 15, 51].

Consider the integral operator

$$G_\epsilon f(y) = \int_{\mathcal{M}} K_\epsilon(y, x)f(x) \, d\mu(x). \quad (2.2)$$

Here, $f$ is a sufficiently smooth scalar- or vector-valued function on $\mathcal{M}$ (an observable), and $d\mu$ a volume form on $\mathcal{M}$. Typically, $d\mu$ is chosen as the volume form that $\mathcal{M}$ inherits from its embedding in $\mathbb{R}^m$, but can also be set to a kernel-dependent volume form simplifying somewhat the small-$\epsilon$ asymptotics [28]. In what follows we will follow the latter approach, but either choice leads to the same limit operator in (2.4) as $\epsilon \to 0$ if a suitable normalization procedure [15, 9] is applied to $G_\epsilon$. The point $y$ can be any point in data space, but if restricted
Analog Forecasting with Dynamics-Adapted Kernels

In what follows, we will employ both of these aspects of $G_\epsilon$ and other related kernel operators.

In practical applications with time-ordered data, manifold integrals such as (2.2) are approximated by discrete Monte Carlo sums of the form

$$\hat{G}_\epsilon f(y) = \sum_{i=1}^{n} K_\epsilon(y, x_i) f(x_i),$$

(2.3)

where $n = N - q + 1$ is the available number of samples after delay-coordinate mapping. Assuming that the discrete time series $x_i$ is ergodic, the pathwise sums will converge to manifold integrals but with a bias that depends on sampling distribution of the data. Here, the sampling distribution is the equilibrium measure of the dynamics, and in the limit of large data ($n \to \infty$) we have $\hat{G}_\epsilon f(y) \to G_\epsilon(f \rho)(y)$, where $\rho$ is the density of the equilibrium measure relative to $d\mu$ [9]. However, biases of this type can be corrected by applying a normalization procedure to $G_\epsilon$ (described in Section 2.3 ahead). In particular, starting from an exponentially decaying kernel, one can construct an operator, $P_\epsilon$, whose leading-order behavior is related to the Laplace-Beltrami operator associated with a Riemannian metric on $\mathcal{M}$ that depends on the kernel. Specifically, it is possible to arrange that

$$P_\epsilon = I + \epsilon \Delta + O(\epsilon^2),$$

(2.4)

where $\Delta$ is the kernel-induced Laplace Beltrami operator, and $(P_\epsilon - I)/\epsilon \to \Delta$ as $\epsilon \to 0$. Because Laplace-Beltrami operators are in one-to-one correspondence with Riemannian metrics on manifolds [49], one can think of the kernel as inducing to the dataset a new Riemannian geometry, and use the properties of that geometry to study the asymptotic behavior of kernel operators such as $P_\epsilon$ for data analysis tasks. Note that $P_\epsilon$ can be approximated from finite datasets via discrete sums analogous to (2.3).

In the context of analog forecasting, the pertinent tasks that will be accomplished through $P_\epsilon$ are (i) identification of analogs in the training data $x_i$ given initial data $y$; (ii) assignment of weights for ensemble forecasts based on multiple analogs; (iii) construction of intrinsically slow observables. Thus, we are interested in working with kernels such that the corresponding induced geometry facilitates these tasks. Intuitively, the induced geometry should be able to choose neighborhoods of the training data which are stable under dynamical evolution; i.e., if we start with points within an initially spherical ball in the kernel geometry, the radius of that ball should vary weakly under dynamical evolution. Moreover, because the dynamical system operates in an abstract phase space independently of the function that produces the observed data $x_i$ (the observation modality), we would like the kernel geometry to have strong invariance properties under general invertible transformations of the data. Note that changes in observation modality can lead to arbitrarily large changes in the induced geometry with respect to the radial Gaussian kernel. Similar considerations apply in
the construction of data-driven observables, and in fact we will see that the kernel-induced
genometry naturally leads to data-driven forecast observables through the eigenfunctions of
$P$. 

Several kernels for dynamical systems have been proposed in the literature [52, 30, 7, 56, 28, 8] and could be employed in the analog forecasting techniques developed here in
different scenarios. In this paper, as a concrete example meeting the objectives outlined
above, we work with the family of “cone kernels” introduced in [28]. A novel aspect of these
kernels (which will turn out to be beneficial for analog forecasting) is that they feature an
explicit dependence on the time tendency of the data, estimated through finite differences
in time. For instance, $v = (x(t) - x(t - \delta t))/\delta t$ corresponds to a first-order backward finite-
difference approximation for the time derivative of the data, but higher-order and/or central
or forward schemes can equally be used. In [28], a geometrical interpretation was given for
these quantities in terms of the vector field $V$ on the attractor generating the dynamics. As
a result, incorporating $v$ in kernels allows one to bias the Riemannian metric of the data in
a way that depends explicitly on the generator.

In cone kernels, that dependence enters through the norm of $V$, estimated through the
norm $\|v\|$ in ambient data space, and the angle between $V$ and the relative displacement
vector $\omega = y(t') - x(t)$ of time-indexed data samples, $y(t'), x(t) \in \mathbb{R}^m$. The cosine
of that angle is estimated through $\cos \eta = v \cdot \omega / (\|v\| \|\omega\|)$, and similarly we compute
$w = (y(t') - y(t' - \delta t))/\delta t$ and $\cos \theta = w \cdot \omega / (\|w\| \|\omega\|)$. Introducing a parameter $\zeta \in [0, 1)$
(in addition to the kernel bandwidth $\epsilon$), we define the cone kernel by

$$K_\epsilon(y(t'), x(t)) = e^{-A(y(t'), x(t))/\epsilon}, \quad A(y(t'), x(t)) = \frac{\|\omega\|^2}{\|w\| \|v\|} \left[ (1 - \zeta \cos^2 \theta)(1 - \zeta \cos^2 \eta) \right]^{1/2}. \quad (2.5)$$

As $\zeta$ approaches 1, this kernel assigns higher affinity to data samples whose relative
displacement vector is aligned with either $v$ and/or $w$; i.e, it assigns strong affinity to pairs
of samples whose relative displacement vector lies within a narrow cone aligned with the
dynamical vector field. This leads to distance contraction along the dynamical flow, such that
the new geometry will be well-adapted to identifying analogs and intrinsic slow timescales.
In particular, it can be shown [28] that in the induced geometry of cone kernels the norm
$\|u\|^2_K$ of tangent vectors on the manifold is related to the ambient-space norm $\|u\|$ through
the expression

$$\|u\|^2_K = \frac{1}{\|V\|^2} \left( \|u\|^2 - \zeta \frac{(V \cdot u)^2}{\|V\|^2} \right). \quad (2.6)$$

That is, for $\zeta > 1$ the metric preferentially contracts tangent vectors aligned with $V$ having
$(V \cdot u)^2/\|V\|^2 \approx \|u\|^2$, and the amount of length contraction becomes arbitrarily large as
$\zeta \to 1$. As a result, in the cone-kernel geometry with $\zeta \approx 1$, small spherical balls centered on
the initial data $y$ will preferentially contain samples in the training dataset lying along a tube
of integral curves of the dynamical vector field containing the integral curve passing through
$y$. The time evolution of these samples is therefore expected to be representative of the time
Analog Forecasting with Dynamics-Adapted Kernels

11

11

evolution of \( y \); i.e., the samples identified in the cone-kernel geometry are expected to be
good analogs. Note that this construction is stable under changes in observation modality, in
the sense that the analogs identified via cone kernels applied to datasets related by invertible
transformations will be similar even if the corresponding ambient-space metrics [i.e., the dot
product and corresponding norm in (2.6)] are Lipschitz far. In Section 2.3, we will see that
the cone kernel geometry is also beneficial for constructing slow intrinsic observables through
the corresponding Laplace-Beltrami eigenfunctions.

We also note that as a result of the scaling by \( 1/\|w\|\|v\| \) cone kernels acquire a scale
invariance which is useful for computing affinities for heterogeneous datasets consisting of
measurements with different physical dimension (units) \[13\]. The outcome of this scaling
factor in the induced geometry is a conformal scaling by the inverse squared norm of the
dynamical vector field which can be seen in (2.6). Heuristically, the result of this conformal
factor is to preferentially increase the resolution (“zoom in”) at the regions of the data
manifold where the phase space velocity \( \|V\| \) is small (and the sampling density is large),
and decrease the resolution when \( \|V\| \) is large. Large \( \|V\| \) is expected to occur during
metastable regime transitions, which will be a prominent feature of the atmospheric model
studied in Section 4.1. In \[30\], this system was studied using as a kernel the special case in (2.5)
with \( \zeta = 0 \), where it was found that the \( 1/\|w\|\|v\| \) scaling is highly beneficial to
recover the intermittent regime transitions in Galerkin-reduced models.

2.3. Accessing the induced geometry with eigenfunctions

In situations where one has no explicit knowledge of the dynamical system generating the
data and/or the corresponding manifold \( \mathcal{M} \), aspects of the kernel-induced geometry can be
accessed through eigenfunctions of kernel operators, \( P_\epsilon \), derived from \( G_\epsilon \) in (2.2). These
eigenfunctions can be used as nonlinear dimension-reduction coordinates as is common
practice in harmonic analysis and machine learning \[3, 15\], but they also form orthonormal
function bases on \( \mathcal{M} \) to construct data-driven observables \[30, 32\].

To construct \( P_\epsilon \), we employ the normalization procedure first introduced in the diffusion
maps algorithm \[15\] and further studied in \[9\]. This procedure involves a sequence of so-
called right and left normalizations of \( G_\epsilon \) designed to remove the influence of the sampling
density \( \rho \) in the limit of large data. In our construction, we set the volume form \( d\mu \) in the
definition of \( G_\epsilon \) in (2.2) equal to the Riemannian measure of the metric associated with the
cone kernels. This volume form is related to the volume form \( d\bar{\theta} \) inherited by the embedding
of \( \mathcal{M} \) into the ambient data space through a scaling, \( d\mu = d\bar{\mu}/\|V\|^{\dim \mathcal{M}} \), that depends on
the data-space norm of the dynamical vector field and the dimension of the manifold \[28\].
We then define

\[
P_\epsilon f = \frac{H_\epsilon f}{H_\epsilon 1}, \quad H_\epsilon f = G_\epsilon (f \rho / \rho_\epsilon), \quad \rho_\epsilon = G_\epsilon \rho,
\]

where \( 1 \) is the function equal to 1 at every point in \( \mathcal{M} \).

A key property of \( P_\epsilon \) is that it can be approximated through Monte-Carlo sums over
Analog Forecasting with Dynamics-Adapted Kernels

ergodic trajectories on the manifold. Specifically, noting that the discrete kernel sum
$$\sum_{i=1}^{n} K_{\epsilon}(y, x_i) = \hat{G}_{\epsilon} 1$$
converges to \( G_{\epsilon} \rho(y) := \rho_{\epsilon}(y) \), \( P_{\epsilon} \) can be approximated by
$$\hat{P}_{\epsilon} = \frac{\hat{H}_{\epsilon} f}{\hat{H}_{\epsilon} 1}, \quad \hat{H}_{\epsilon} f = \hat{G}_{\epsilon} (f / q_{\epsilon}), \quad q_{\epsilon} = \hat{G}_{\epsilon} 1. \quad (2.8)$$

Equivalently, the operators in (2.7) and (2.8) can be expressed as kernel integral operators,
$$P_{\epsilon} f(y) = \int_{\mathcal{M}} p_{\epsilon}(y, x) f(x) d\mu(x), \quad H_{\epsilon} f(y) = \int_{\mathcal{M}} K'_{\epsilon}(y, x) f(x) d\mu(x),$$
$$\hat{P}_{\epsilon} f(y) = \sum_{i=1}^{n} \hat{p}_{\epsilon}(y, x_i) f(x_i), \quad \hat{H}_{\epsilon} f(y) = \sum_{i=1}^{n} \hat{K}'_{\epsilon}(y, x_i) f(x_i), \quad (2.9)$$

where
$$p_{\epsilon}(y, x) = \frac{K'_{\epsilon}(y, x)}{H_{\epsilon} 1(y)}, \quad K'_{\epsilon}(y, x) = \frac{K_{\epsilon}(y, x)}{\rho_{\epsilon}(x)}, \quad \hat{p}_{\epsilon}(y, x_i) = \frac{\hat{K}'_{\epsilon}(y, x_i)}{\hat{H}_{\epsilon} 1(y)}, \quad \hat{K}'_{\epsilon}(y, x_i) = \frac{K_{\epsilon}(y, x_i)}{q_{\epsilon}(x_i)}.$$ 

It is evident from the last equations that the operations to obtain \( K'_{\epsilon} \) and \( p_{\epsilon} \) correspond to right and left normalizations, respectively (similarly for \( \hat{K}'_{\epsilon} \) and \( \hat{p}_{\epsilon} \)). Note also that (up to proportionality constants) \( \rho_{\epsilon} \) and \( q_{\epsilon} \) are kernel estimates of the sampling density on \( \mathcal{M} \) with respect to \( d\mu \). Moreover, \( P_{\epsilon} \) is an averaging operator since \( p_{\epsilon}(y, x) > 0 \) and \( P_{\epsilon} 1 = 1 \) (similarly for \( \hat{P}_{\epsilon} \)).

It can be shown that the operator \( P_{\epsilon} \) admits the asymptotic expansion in (2.4) for all non-singular sampling densities \( \rho \) [28], and as a result \( (P_{\epsilon} - I)/\epsilon \) converges to the Laplace-Beltrami operator \( \Delta \) associated with the Riemannian metric induced by the cone kernels. Correspondingly, up to a proportionality constant, \( \Delta \) can be approximated from data by the discrete operator \( \hat{P}_{\epsilon} - I \). The matrix \( P = [\hat{p}_{\epsilon}(x_i, x_j)]_{i,j=1}^{n} \) is a Markov matrix over the training dataset representing \( \hat{P}_{\epsilon} \). In what follows, we drop the circumflex from \( \hat{P}_{\epsilon} \) for notational simplicity, so that \( P_{\epsilon} f \) will be given by either (2.7) or (2.8) in a continuous or discrete setting, respectively.

Consider now the eigenvalue problem associated with \( P_{\epsilon} \),
$$P_{\epsilon} \phi_i = \lambda_i \phi_i, \quad i \in \{0, 1, \ldots\}, \quad (2.10)$$
where \( \lambda_i \) are real eigenvalues and \( \phi_i \) the corresponding eigenfunctions. By convention, we set \( \lambda_0 = 1 \) and \( \phi_0 = \text{const.} \), and order the eigenvalues such that \( \lambda_0 > |\lambda_1| \geq |\lambda_2| \cdots \geq 0 \). Note that the reality of the eigenvalues is a consequence of the self-adjointness of \( P_{\epsilon} \) in a suitable inner product as discussed below. Moreover, as \( \epsilon \to 0 \), the eigenfunctions of \( P_{\epsilon} \) converge to the eigenfunctions of \( \Delta \). In this limit, \( P_{\epsilon} \) becomes positive semidefinite, but it may have negative eigenvalues for \( \epsilon > 0 \). To take advantage of the efficiency and stability of eigenvalue solvers for symmetric matrices, it is convenient to obtain \( \phi_i \) by first computing the eigenvectors \( \phi_i \) of the symmetric matrix \( S = D^{1/2} P D^{-1/2} \), where \( D = \text{diag}[\hat{H}_{\epsilon} 1(x_i)]_{i=1}^{n} \), and then evaluating \( \phi_i = D^{-1/2} \phi_i \).
A popular approach in harmonic analysis and machine learning is to construct mappings of $\mathcal{M}$ into Euclidean spaces using the eigenfunctions of $P_\epsilon$ as coordinates. Specifically, for a parameter $s > 0$, one defines the family of maps $\Phi_{s,l} : \mathcal{M} \mapsto \mathbb{R}^l$ given by

$$\Phi_{s,l}(x) = r(s)(\lambda_s^1\phi_1(x), \ldots, \lambda_s^l\phi_l(x)),$$

were $r(s)$ is a scaling function for the coordinates. The parameter $s$ can be interpreted as the diffusion time for a diffusion process on the manifold with transition probability kernel $p_\epsilon(y, x)$, but note that this diffusion process and $s$ are in general not related to the dynamical system and the physical time $t$, respectively.

Allowing $l$ to go to infinity, $\Phi_{s,\infty} \equiv \Phi_s$ becomes a map of $\mathcal{M}$ into the Hilbert space $\ell^2$ of real sequences $\{\alpha_1, \alpha_2, \ldots\}$ with $\sum_{i=1}^{\infty}|\alpha_i|^2 < \infty$. In [6] it was shown that if the $\phi_i$ are Laplace-Beltrami eigenfunctions [i.e., $\epsilon \to 0$ in (2.10)], then as $s \to 0$ and with the choice $r(s) = s^{(\dim \mathcal{M} + 2)/4}$ and under certain conditions on sectional curvature, $\Phi_s$ is an embedding of $\mathcal{M}$ into $\ell^2$; i.e., it is invertible on its image, and has a continuous derivative of full rank. Moreover, that embedding is isometric with respect to the Riemannian metric associated with the Laplace-Beltrami operator and the eigenfunctions. Thus, it is possible to recover the kernel-induced Riemannian geometry of the data to arbitrarily high precision using the $\phi_i$ as coordinates. This result was improved in [46], where it was shown that the number of required eigenfunctions is actually finite. In practical data-driven applications, one introduces a threshold $\delta \in (0, 1)$, and chooses $l$ as the largest integer for which $|\lambda_l^\alpha| > \delta$. Moreover, $r(s)$ can be set to a constant as it can be absorbed in a renormalization of the coordinates. The leading-$l$ Laplace-Beltrami eigenfunctions are then approximated by eigenfunctions of the discrete operator in (2.8), giving for each state $x_i$ in the training data an empirical reduced representation in $\mathbb{R}^l$. The Euclidean distance in the image space of $\Phi_{s,l}$,

$$D_{s,l}(y, x) = \|\Phi_{s,l}(y) - \Phi_{s,l}(x)\|,$$

is called diffusion distance [15]. In Section 3.2, we will use diffusion distances for analog identification.

Besides their role as dimension-reduction coordinates, the eigenfunctions in (2.10) form an orthonormal basis for a Hilbert space $L^2(\mathcal{M}, \mu_\epsilon)$ of scalar-valued functions on $\mathcal{M}$ with inner product

$$\langle f_1, f_2 \rangle_{\mu_\epsilon} = \int_{\mathcal{M}} f_1(x)f_2(x) \, d\mu_\epsilon(x), \quad d\mu_\epsilon(x) = \frac{H_{\epsilon,1}(x)\rho(x)}{\rho_\epsilon(x)} \, d\mu(x).$$

In this inner product, the action of $P_\epsilon$ on functions can be expressed in terms of a symmetric kernel, $\sigma_\epsilon(y, x) = \sigma_\epsilon(x, y)$, such that

$$P_\epsilon f(y) = \int_{\mathcal{M}} \sigma_\epsilon(y, x) f(x) \, d\mu_\epsilon(x), \quad \sigma_\epsilon(y, x) = \frac{K_\epsilon(y, x)}{H_{\epsilon,1}(y)H_{\epsilon,1}(x)}.$$

It then follows that $P_\epsilon$ is self-adjoint on $L^2(\mathcal{M}, \mu_\epsilon)$, so that the eigenvalues in (2.10) are real, and the eigenfunctions can be made orthonormal. Note that $d\mu_\epsilon \to d\mu$ as $\epsilon \to 0$, and
the Riemannian inner product of \( L^2(\mathcal{M}, \mu) \) is recovered. By the spectral theorem [see (3.7) ahead], the diffusion distance for \( s = 1 \) and \( l \to \infty \), can be obtained from the \( L^2(\mathcal{M}, \mu) \) distance between the transition probability kernels \( p_\epsilon(y, \cdot) \) and \( p_\epsilon(x, \cdot) \) viewed as functions on the manifold, i.e.,

\[
\lim_{l \to \infty} D_{1,l}(y, x) := D(y, x) = \|p_\epsilon(y, \cdot) - p_\epsilon(x, \cdot)\|_\mu. \tag{2.14}
\]

In the discrete case, \( \hat{P}_\epsilon \) from (2.8) is self-adjoint with respect to the inner product

\[
\langle f_1, f_2 \rangle_{\pi_\epsilon} = \sum_{i=1}^{n} f_1(x_i) f_2(x_i) \pi_\epsilon(x_i), \quad \pi_\epsilon(x_i) = \frac{\hat{H}_\epsilon 1(x_i)}{q_\epsilon(x_i)} = \sum_{j=1}^{n} \frac{K_\epsilon(x_i, x_j)}{q_\epsilon(x_i)q_\epsilon(x_j)},
\]

and we also have

\[
\hat{P}_\epsilon f(y) = \sum_{i=1}^{n} \hat{\sigma}_\epsilon(y, x_i) \pi_\epsilon(x_i), \quad \hat{\sigma}_\epsilon(y, x_i) = \frac{K_\epsilon(y, x_i)}{\hat{H}_\epsilon 1(y)\hat{H}_\epsilon 1(x_i)}.
\]

Note that as the number of samples tends to infinity, \( \pi_\epsilon(x_i) \) converges (up to a proportionality constant) to \( H_\epsilon 1(x_i)/\rho_\epsilon(x_i) \) so that \( \sum_{i=1}^{n} f(x_i) \pi_\epsilon(x_i) \to \int_{\mathcal{M}} f(x) \, d\mu_\epsilon(x) \). Thus, in both the continuous and discrete settings we may expand observables as \( f = \sum_i c_i \phi_i \), with \( c_i = \langle \phi_i, f \rangle \) and \( \|\phi_i\| = 1 \). The constants \( c_i \) can be thought of as Fourier coefficients with respect to the orthonormal basis \( \{ \phi_i \} \). We say that \( f \) has bandwidth \( b \) if the coefficients \( c_i \) all vanish for \( i > b \).

In fact, it turns out that band-limited observables can be interpreted as having bounded roughness on the manifold [32]. In the present context, a natural measure of roughness of function is the Dirichlet form,

\[
E(f) = \langle \text{grad } f, \text{grad } f \rangle_\mu, \tag{2.15}
\]

defined using the Hodge inner product, \( \langle \cdot, \cdot \rangle_\mu \), for vector fields on \( \mathcal{M} \) and the gradient operator, \( \text{grad} \) mapping functions to vector fields. Note that both \( \langle \cdot, \cdot \rangle \) and \( \text{grad} \) depend strongly on the kernel-induced Riemannian metric. Let \( \lambda'_i = (1 - \lambda_i)/\epsilon \). As \( \epsilon \to 0 \), these quantities converge to the eigenvalues of the Laplace-Beltrami operator \( \Delta \), satisfying the relation \( \lambda'_i = E(\phi_i)/\|\phi_i\|_\mu \). Thus, \( \lambda'_i \) measures the roughness of the corresponding eigenfunction \( \phi_i \), and for a function with bandwidth \( b \), \( E(f)/\|f\|^2 \leq \lambda'_b \). Such functions with small \( b \) behave stably under out-of-sample extension [16], and are also expected to have favorable predictability properties.

The constructions introduced above, namely the dimension-reduction coordinates in (2.11), the diffusion distance in (2.12), and the Dirichlet form in (2.15) depend crucially on the properties of the kernel-induced geometry. For the cone kernels in (2.5), the induced geometry has the local distance function given by (2.6), which contracts lengths along the dynamical flow as described in Section 2.2. In fact, it is possible to show that a
consequence of this length contraction is that, as the cone kernel parameter $\zeta$ approaches 1, the corresponding Laplace-Beltrami operator admits the asymptotic expansion $[28]$, 

$$
\Delta f = -\frac{\text{div}_\mu(V(f)V)}{1-\zeta} + O((1-\zeta)^0). \tag{2.16}
$$

In the above equation, $\text{div}_\mu$ is the divergence operator on vector fields associated with the kernel-induced volume form, $d\theta$, and $V(f)$ represents the action on the dynamical vector field $V$ on the function $f$ [i.e., $V(f)$ is the directional derivative of $f$ along $V$]. Note that a general Laplace-Beltrami operator $\Delta f = -\text{div grad} f$ (evaluated, e.g., through a radial Gaussian kernel) depends on the full gradient of $f$, which generally depends on the observation map for the dynamical system through the ambient-space induced metric. On the other hand, the directional derivative $V(f)$ is intrinsic to the dynamical system, and is independent of the observation modality. Asymptotically, as $\zeta \to 1$, the only dependence of the Laplace-Beltrami operator associated with cone kernels on the observation map enters through the volume form $d\theta$ in the divergence operator. Thus, observables constructed through cone kernel eigenfunctions are invariant for equivalence classes of observation maps preserving $d\mu$. In addition, it also follows from (2.16) that $\zeta \to 1$ cone kernels induce a Dirichlet form, $E(f) = \langle V(f), V(f) \rangle$, that assigns large roughness to functions that exhibit strong variability along the dynamical flow. Equivalently, observables which are strongly band-limited with respect to cone kernels [i.e., lie in the subspace of $L^2(\mathcal{M}, \mu)$ spanned by the leading-few eigenfunctions] are expected to vary slowly in the course of dynamical evolution. Such observables are good candidates for revealing intrinsic slow dynamical timescales, and are also expected to have favorable predictability properties. In Section 3, our approach will be to use individual cone kernel eigenfunctions as data-driven forecast observables.

### 2.4. Out-of-sample extension schemes

Let $M = \{x_i\}_{i=1}^n \subset \mathcal{M}$ be the training dataset. Broadly speaking, the goal of out-of-sample extension schemes is to extend scalar- or vector-valued functions defined on $M$ to functions defined in the whole of data space $\mathbb{R}^m$. As a concrete example, let $f$ be a scalar-valued function on $M$, and let $\tilde{f} : \mathbb{R}^m \mapsto \mathbb{R}$ be its out-of-sample extension. For simplicity of exposition, we consider that $f$ is scalar-valued; out-of-sample extension of vector-valued functions can be carried out by applying the methods described below component wise, but note that there may be additional benefits using so-called multitask kernels simultaneously enforcing smoothness on $\mathcal{M}$ and the image space of $f$ [14].

In general, desirable properties of $\tilde{f}$ are as follows. (i) Restricted to $\mathcal{M} \subset \mathbb{R}^m$, $\tilde{f}$ should be a well-behaved function; e.g., $\tilde{f}|_{\mathcal{M}}$ should lie in a function space such as $L^2(\mathcal{M}, \mu_e)$ introduced in Section 2.3. (ii) More broadly, $\tilde{f}$ should be well-behaved with respect to a notion of smoothness defined in the whole of $\mathbb{R}^m$. (iii) $\tilde{f}$ should be consistent on $M$, i.e., $\tilde{f}(x_i) = f(x_i)$ for all $x_i \in M$. (iv) $\tilde{f}$ should be consistent on $\mathcal{M}$ as the number of samples $n$ increases; i.e., if $f$ is given by some function $g$ on $\mathcal{M}$ restricted to $M$, then $\tilde{f}$ should converge
to \( g \) as \( n \to \infty \) and \( \epsilon \to 0 \). In what follows, we outline two kernel-based methods for out-of-sample extension, called geometric harmonics [16] and Laplacian pyramids [47], that attempt to meet these objectives. Geometric harmonics is based on the Nyström method [43] for out-of-sample extension of functions, and is better suited for observables with few leading nonzero coefficients in the eigenfunction basis. On the other hand, Laplacian Pyramids [47] a multiscale iterative approach, and is better-conditioned for more general observables which are not expressible in terms of a tightly truncated eigenfunction expansion. Both of these schemes will be used in the development of the prediction strategies in Section 3 and the applications in Section 4.

2.4.1. Geometric harmonics. This algorithm employs the Nyström method [43] to extend functions defined on \( \mathcal{M} \) (or the finite dataset \( M \subset \mathcal{M} \)) to functions in a reproducing kernel Hilbert space (RKHS), \( \mathcal{H} \), on \( \mathbb{R}^m \). The RKHS construction is introduced to induce smoothness on the extended function, and also to ensure that the extended function is concentrated near the training dataset. The Nyström extension method is strongly related to Gaussian process regression [48], which is an extension method in the field of statistical inference. In the field of geostatistics, this extension method is better known as kriging [34].

The central ingredient in the construction of \( \mathcal{H} \) is a symmetric positive semidefinite kernel \( \bar{K}_\epsilon : \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R} \) defined on data space. Specifically, we require that (i) \( \bar{K}_\epsilon(y, z) = \bar{K}_\epsilon(z, y) \) for all \( y, z \in \mathbb{R}^m \); (ii) for any \( l \geq 1 \) and any sequence \( \alpha = (\alpha_1, \ldots, \alpha_l) \) of real numbers and points \( y_1, \ldots, y_l \) in \( \mathbb{R}^m \) we have \( \sum_{i,j=1}^{l} \alpha_i \alpha_j \bar{K}_\epsilon(y_i, y_j) \geq 0 \). In the present context, a natural choice for \( \bar{K}_\epsilon \) is

\[
\bar{K}_\epsilon(y, z) = \int_{\mathcal{M}} \sigma_\epsilon(y, x) \sigma_\epsilon(x, z) d\mu_\epsilon(x),
\]

where \( \sigma_\epsilon \) is the symmetric kernel defined in (2.13). To verify that \( \bar{K}_\epsilon \) is positive semidefinite, fix \( l \) and \( \{ y_i \}_{i=1}^l \), and consider the linear operator \( w : \mathbb{R}^l \mapsto L^2(\mathcal{M}, \mu_\epsilon) \) such that \( w(\alpha)(x) = \sum_{i=1}^l \alpha_i \sigma_\epsilon(y_i, x) \). With this definition, we have \( \sum_{i,j=1}^{l} \alpha_i \alpha_j \bar{K}_\epsilon(y_i, y_j) = \langle w(\alpha), w(\alpha) \rangle_{\mu_\epsilon} \), and the latter quantity is non-negative. Note that in applications we replace \( \sigma_\epsilon \) in (2.17) by the kernel \( \sigma_\epsilon \) constructed from discrete data, and we also replace \( \int_{\mathcal{M}} f(x) d\mu_\epsilon(x) \) by \( \sum_{i=1}^{n} f(x_i) \pi_\epsilon(x_i) \).

For simplicity of exposition, in what follows we present the geometric harmonics scheme in the continuous framework, with the understanding that a discrete formulation can be obtained at each step using the results of Section 2.3.

According to the Moore-Aronszajn theorem [1], there exists a unique RKHS \( \mathcal{H} \) of scalar functions on \( \mathbb{R}^m \) associated with \( \bar{K}_\epsilon \) with inner product \( \langle \cdot, \cdot \rangle_{\mathcal{H}} \), such that (i) for almost every \( y \in \mathbb{R}^m \), the function \( \bar{K}_\epsilon(y, \cdot) \) is in \( \mathcal{H} \); (ii) for every \( f \in \mathcal{H} \) and \( y \in \mathbb{R}^m \), point evaluation can be expressed in terms of the bounded linear operator

\[
f \mapsto f(y) = \langle f, \bar{K}_\epsilon(y, \cdot) \rangle_{\mathcal{H}}.
\]

This last equation expresses the reproducing property of \( \mathcal{H} \), and it follows from this property that \( \langle \bar{K}_\epsilon(y, \cdot), \bar{K}_\epsilon(z, \cdot) \rangle_{\mathcal{H}} = \bar{K}_\epsilon(y, z) \). More generally, \( \mathcal{H} \) consists of functions of the form
\( f = \sum_{i=1}^{\infty} \alpha_i \tilde{K}_\epsilon(y_i, \cdot) \), with \( \sum_{i,j=1}^{\infty} \alpha_i \alpha_j \tilde{K}_\epsilon(y_i, y_j) < \infty \). The inner product between \( f \) and another function \( g = \sum_{i=1}^{\infty} \beta_i \tilde{K}_\epsilon(z_i, \cdot) \) is \( \langle f, g \rangle_{\mathcal{H}} = \sum_{i,j=0}^{\infty} \alpha_i \beta_j \tilde{K}_\epsilon(y_i, z_j) \).

Next, consider the linear operator \( Q_\epsilon : L^2(\mathcal{M}, \mu_\epsilon) \mapsto \mathcal{H} \) taking functions on the manifold to functions in the RKHS via the kernel integral

\[
Q_\epsilon f(y) = \int_{\mathcal{M}} \tilde{K}_\epsilon(y, x) f(x) \, d\mu_\epsilon(x). \tag{2.18}
\]

An important property of the kernel \( \tilde{K}_\epsilon \) as defined in (2.17) is that the eigenfunctions \( \phi_i \) of \( P_\epsilon \) in (2.10) are also eigenfunctions of \( Q_\epsilon \), restricted to \( \mathcal{M} \) with non-negative eigenvalues \( \lambda_i^2 \). That is, for \( x \in \mathcal{M} \), we have \( Q_\epsilon \phi_i(x) = \lambda_i^2 \phi_i(x) \). This result in conjunction with the reproducing property leads to a relationship between inner products on \( L^2(\mathcal{M}, \mu_\epsilon) \) and \( \mathcal{H} \) involving eigenfunctions, namely \( \langle \phi_i, f \rangle_{\mu_\epsilon} = \lambda_i^2 \langle \phi_i, f \rangle_{\mathcal{H}} \). Therefore, for two functions \( f, g \in L^2(\mathcal{M}, \mu_\epsilon) \) with \( f = \sum_{i=0}^{\infty} \alpha_i \phi_i \), \( g = \sum_{i=0}^{\infty} \beta_i \phi_i \) and \( \| \phi_i \|_{\mu_\epsilon} = 1 \) we obtain

\[
\langle f, g \rangle_{\mathcal{H}} = \sum_{i=0}^{\infty} \frac{\alpha_i \beta_i}{\lambda_i^2}. \tag{2.19}
\]

Note that while \( Q_\epsilon \) maps functions on the manifold to functions on the ambient data space, the mapping is not consistent on \( \mathcal{M} \); i.e., in general, \( Q_\epsilon f(x) \neq f(x) \). On the other hand, as one can verify (e.g., [16]), the adjoint \( Q_\epsilon^* : \mathcal{H} \mapsto L^2(\mathcal{M}, \mu_\epsilon) \) has the property \( Q_\epsilon^* f(x) = f(x) \) for every \( f \in \mathcal{H} \) and almost every \( x \in \mathcal{M} \). Thus, \( Q_\epsilon^* \) is a restriction operator on \( \mathcal{M} \).

Next, we discuss how the reproducing property leads naturally to an extension operator which is also consistent on the manifold. In particular, it follows from (2.19) that the subspace \( \mathcal{H}' \) of \( L^2(\mathcal{M}, \mu_\epsilon) \) consisting of functions \( f = \sum_{i=0}^{\infty} \alpha_i \phi_i \) with \( \sum_{i=0}^{\infty} \alpha_i^2 / \lambda_i^2 < \infty \) is also a subspace of \( \mathcal{H} \). For such functions, we define the out-of-sample extension operator \( A_\epsilon : \mathcal{H}' \mapsto \mathcal{H} \) such that \( A_\epsilon f = \bar{f} \), with

\[
\bar{f}(y) = \langle f, \tilde{K}_\epsilon(y, \cdot) \rangle_{\mathcal{H}} = \sum_{i=0}^{\infty} \frac{\langle f, \phi_i \rangle_{\mu_\epsilon} \langle \phi_i, \tilde{K}_\epsilon(y, \cdot) \rangle_{\mu_\epsilon}}{\lambda_i} = \sum_{i=0}^{\infty} \frac{\langle f, \phi_i \rangle_{\mu_\epsilon} P_\epsilon \phi_i(y)}{\lambda_i}. \tag{2.20}
\]

The procedure in (2.20) is called Nyström extension. The intuition behind this approach is that to extend \( f \) we first compute its Fourier coefficients \( \langle f, \phi_i \rangle_{\mu_\epsilon} \) in the \( \{ \phi_i \} \) eigenfunction basis of \( L^2(\mathcal{M}, \mu_\epsilon) \), and then sum the extensions \( P_\epsilon \phi_i(y) \) of the individual eigenfunctions weighted by the Fourier coefficients and the corresponding eigenvalues \( \lambda_i \). Note that consistency on the manifold, \( \bar{f}(x) = f(x) \) for \( x \in \mathcal{M} \), follows immediately from the reproducing property. In the discrete setting, the continuous operator \( P_\epsilon \) in (2.20) is replaced by its discrete counterpart, \( \tilde{P}_\epsilon \), and all \( \langle \cdot, \cdot \rangle_{\mu_\epsilon} \) inner products are replaced by \( \langle \cdot, \cdot \rangle_{\mu_\epsilon} \).

In [16], the extended eigenfunctions \( \tilde{\phi}_i = A_\epsilon \phi_i \) are called geometric harmonics. These functions have the special property \( \tilde{\phi}_i = \lambda_i P_\epsilon \phi_i \), meaning that, up to a proportionality constant, the pure eigenfunctions can be consistently extended via \( P_\epsilon \) (or \( Q_\epsilon \)). Moreover, in [16] it is also shown that the geometric harmonics are stationary points of a Rayleigh quotient, \( c_{\mathcal{M}}(F) = \| Q_\epsilon F \|_{\mu_\epsilon} / \| F \|_{\mathcal{H}} \), expressing the degree of concentration of a function
$F \in \mathcal{H}$ on the manifold. That is, $\mathcal{H}_l = \text{span}\{\tilde{\phi}_i\}_{i=0}^l$ can be thought of as the $(l+1)$-dimensional subspace of $\mathcal{H}$ whose elements are most strongly concentrated on $\mathcal{M}$, and the $\tilde{\phi}_i$ are orthonormal basis functions for that subspace.

Inspecting (2.20) it becomes apparent that due to the division by the eigenvalues the Nyström extension scheme becomes increasingly ill-conditioned as the bandwidth of $f$ increases. To alleviate this poor conditioning, in [16] $A_\epsilon$ is replaced by a truncated operator,

$$A_{\epsilon,l} f(y) = \sum_{i=0}^l \frac{\langle f, \phi_i \rangle_\mu P_\epsilon \phi_i(y)}{\lambda_i}.$$  

This operator improves the conditioning of the extension scheme based on $A_\epsilon$, but at the expense of some loss of precision on $\mathcal{M}$. The loss of precision can be assessed by computing $\|f - \Pi_{\epsilon,l} f\|_\mu$, where $\Pi_{\epsilon,l}$ is the orthogonal projector onto the subspace of $L^2(\mathcal{M}, \mu_\epsilon)$ spanned by $\{\phi_i\}_{i=0}^l$. Due to these issues, the Nyström method is generally better suited for observables which lie in the span of the leading-few eigenfunctions. For this reason, in the experiments of Section 4 we will use (2.20) for analog forecasting of data-driven observables, constructed as pure eigenfunctions with large corresponding eigenvalues.

### 2.4.2. Laplacian pyramids

Laplacian pyramids [47] is an iterative, multiscale out-of-sample extension scheme which is based on a sequence of smoothing operators $P_{\epsilon_0}, P_{\epsilon_1}, \ldots$ from (2.7) with progressively smaller kernel bandwidths $\epsilon_l$. While this scheme gives up exact consistency on the manifold, it gains significantly in stability compared to schemes based on Nyström-type operators such as (2.20). In particular, Laplacian pyramids is not restricted to observables which are band-limited in the eigenfunction basis. In the previous work [47], a radial Gaussian kernel was used. Here, we apply the Laplacian pyramids technique using the normalized smoothing operators in (2.7), which are based in turn on the cone kernel (2.5). In Section 4, this will be our method of choice for analog forecasting objective observables, such as the components of the state vector.

The Laplacian pyramids construction produces an approximation $f_l$ of the function $f$ desired to be extended having the form

$$f_l = f_0 + g_1 + \ldots + g_l.$$  

(2.21)

In the above, $f_0$ is a kernel estimate of $f$ obtained through a coarse smoothing operator $P_{\epsilon_0}$, and the functions $g_l$ are computed iteratively by applying $P_{\epsilon_l}$ to the residual of the approximation $f_{l-1}$ until a desired accuracy is achieved. The smoothed representation $f_l$ is then extended away from the training dataset by application of $P_{\epsilon_0}, \ldots, P_{\epsilon_l}$ to the corresponding terms in (2.21).

In more detail, at the initialization stage of the algorithm a coarse representation of $f$ is generated by computing $f_0 = P_{\epsilon_0} f$ for a pre-defined bandwidth $\epsilon_0$, and the residual $d_0 = f - f_0$ is also computed. An admissible error parameter, $\varepsilon$, and a refinement schedule for the bandwidth parameters $\epsilon_l$ are also specified. In [47], the bandwidth parameters
are decreased via dyadic subdivision, $\epsilon_t = \epsilon/2^t$, and this approach is also used here. In the ensuing iterative refinement, the residual $d_t = f - f_{t-1}$ at step $i$ is computed, and if $\|d_t\|_{\mu_i} < \epsilon$ the iterations are terminated. Otherwise, the residual is used to generate a more detailed representation of $f$ by carrying out the smoothing operation $g_t = P_\epsilon d_{t-1}$ and setting $f_t = f_{t-1} + g_t$. This iterative procedure involving both smoothing and refinement is repeated until the desired accuracy on the training data is reached. Finally, in the extension step, a function $\bar{f}$ defined in the full ambient data space is constructed by applying the smoothing operators to the corresponding terms in the smoothed representation of $f$; i.e.,

$$\bar{f}(y) = B_t f(y) := P_{\epsilon_0} f_0(y) + \sum_{i=1}^{l} P_{\epsilon_i} g_i(y),$$  \hspace{1cm} \text{(2.22)}$$

where $B_t$ is a linear operator taking functions on the training dataset to functions on the ambient data space. In general, extended functions via Laplacian pyramids are not equal to their values on the training dataset $\mathcal{M}$ [cf. the Nyström method in (2.20)]. Instead, the method is asymptotically consistent as $\epsilon_0 \to 0$, since $P_{\epsilon_0} f(y) \to f(y)$ and $P_{\epsilon_i} g_i(y) \to 0$ for $y \in \mathcal{M}$.

As an alternative to (2.22), in Sections 3 and 4 we will perform out-of-sample extension using averaging operators $\hat{P}_{\epsilon,k}$ based on kernels truncated to $k$ nearest neighbors. To construct these operators, we first identify the $k$-th nearest-neighbor set $R_{y,k} = \{x_{i_j}\}_{j=1}^{k}$ of the point $y$ in data space with respect to the kernel $K_\epsilon$, such that $x_{i_1} = \arg\max_{x_j \in \mathcal{M}} K_\epsilon(y, x_j)$ and $x_{i_j} = \arg\max_{x_j \in \mathcal{M}\setminus R_{y,j-1}} K_\epsilon(y, x_j)$ for $j > 1$. We then define the operator $\hat{P}_{\epsilon,k}$ using the transition probabilities $\hat{p}_{\epsilon,k} \epsilon$ obtained from the same expressions as (2.9), but replacing the kernel $K_\epsilon$ by $K_{\epsilon,k}$, where $K_{\epsilon,k}(y, x_i) = K_\epsilon(y, x_i)$ if $x_i \in R_{y,k}$, and $K_{\epsilon,k}(y, x_i) = 0$ otherwise. An asymptotic analysis of $\hat{P}_{\epsilon,k}$ is beyond the scope of this work, but we note the reference [57], where appropriately truncated kernels are shown to converge to the same limit operators as their non-truncated counterparts. Of course, the properties of $\hat{P}_\epsilon$ and $\hat{P}_{\epsilon,k}$ will differ at finite bandwidth $\epsilon$, and in Section 4 we will present examples where the highest skill of the kernel-based forecasting methods developed in Section 3 occurs for tightly-truncated kernels with $k \ll n$.

3. Prediction strategies

In this Section, we employ the mathematical framework described in Section 2 to develop empirical forecasting strategies. These techniques are inspired by Lorenz’s analog method [38], so we begin in Section 3.1 with an overview of the conventional formulation of this approach based on Euclidean distances. In Section 3.2, we reformulate analog forecasting using diffusion distances associated with dynamics-adapted kernels. Then, in Section 3.3 we propose a scheme for weighted ensemble forecasting based on out-of-sample extension operators. We conclude this section with a discussion of forecast skill metrics (Section 3.4), emphasizing the differences between data-driven observables defined on the training dataset and objective observables defined on the full data space.
3.1. Conventional analog forecasting

Broadly speaking, analog forecasting produces a forecast \( f(y, \tau) \) of some observable \( f \) of a dynamical system at forecast lead time \( \tau \) given initial data \( y \). Assuming that \( \tau \) is some non-negative integer multiple \( r \) of the sampling interval \( \delta t \), the necessary ingredients needed to carry out this task are (i) a training time series \( M' = \{x_{i}\}_{i=1}^{n+r} \) from Section 2 with \( x_i = x((i-1)\delta t) \) lying in the same space as \( y \); (ii) the corresponding time series of the observable values, \( f_i = f(t_i) \). In a perfect-model scenario, the training dataset \( M' \) is generated by the same dynamical system as the dynamical system being forecast—this would be the case if \( x_i \) are observations of nature. In an imperfect-model scenario, the \( x_i \) have systematic biases compared to the time evolution of \( y \). This situation arises when \( M' \) is the result of an integration of a numerical model with dynamical errors relative to nature. However, as stated in Section 1.1, it may be desirable to use a long integration from a numerical model to improve the sampling density of state space despite model errors. Moreover, in some cases, the pairs \( (x_i, f_i) \) may not be observationally accessible. A notable example from climate science where this issue is encountered is forecasting and estimation of the interior circulation and thermal structure of the ocean. These variables are not accessible via remote sensing from satellites, and are only sparsely known from in situ sensors such as gliders and floats.

In the standard formulation of analog forecasting, the initial data are “snapshots” taken at single time instances. Here, we generally consider that \( y \) and \( x_i \) are lagged time series of the form in (2.1) consisting of \( q \) samples, noting that conventional case of analog forecasting with snapshot initial data corresponds to the special case \( q = 1 \). If \( q \) is sufficiently large (and under mild conditions on the dynamical system, the observation function, and the sampling interval), the data points \( x_i \) are in one-to-one correspondence with points on the attractor [50], and therefore we can consider the observed values \( f_i \) as the result of a function on the attractor, i.e., with some abuse of notation, \( f_i = f(t_i) = f(x_i) \). In real-world applications this assumption is likely to be valid only at a coarse scale, but adding delays generally improves the smoothness of \( f_i \) as function values on \( M \), and also make the time series \( \{x_i\} \) more Markovian. Moreover, as discussed in Section 2, time-lagged embedding enhances timescale separation capability with kernel methods. Hereafter, we will formally assume that the \( f_i \) are values of a smooth function on \( M \).

Let \( S_\tau \) be the time-\( \tau \) shift operator on time series such that \( S_\tau f(x(t_i)) = f(x(t_i + \tau)) \). Analog forecasting is a two-step procedure which involves finding the sample in the subset \( M = \{x_i\}_{i=1}^{n} \) of the training data via lying closest to \( y \) with respect to the Euclidean distance, and then time-advancing that sample with the shift operator to produce a forecast. That is, one first identifies the analog \( x_i \) of \( y \) via

\[
i = \arg\min_{j \in \{1,...,n\}} \|x_j - y\|, \tag{3.1}
\]

and then evaluates the forecast

\[
f(y, \tau) = S_\tau f(x(t_i)). \tag{3.2}
\]
Note that, by ergodicity, as the number of samples $n$ tends to infinity $S_n f(x(t_i))$ converges (in the weak sense) to a function $S_T f$ on $\mathcal{M}$. In deterministic dynamical systems that function is unique, but if stochasticity is present $S_n f$ is a random function that will change from realization to realization.

We see from (3.1) and (3.2) that analog forecasting is a fully non-parametric approach, relying solely on the geometric identification of nearest neighbors in the training data and empirical time shifts of time series. These properties make the method especially attractive in situations where one does not have knowledge of the equations of motion, or an appropriate parametric model is not available. For instance, in the ocean application in Section 4.2 construction of parametric regression models for the low-frequency modes is especially hard, and linear autoregressive models have worse skill than a trivial persistence forecast [19]. Nevertheless, despite that analog forecasting cleverly avoids dynamical model errors, it suffers from its inability to generate innovations. This motivates future work on hybrid analog-parametric models, taking advantage of both the fidelity of analog forecasting and innovation of parametric models.

3.2. Analog forecasting with diffusion distances

It is evident from (3.1) that the skill of analog forecasting depends strongly on the distance function used to identify analogs. At the very least, with partial observations, neighborhoods in the space of initial data do not necessarily correspond to neighborhoods on the attractor, meaning that the nearest neighbors of the initial data $y$ can have very different dynamical trajectories among themselves and also relative to the future evolution of $y$. In such situations (which we discuss with experiments in Section 4.1), it is highly beneficial to use delay-coordinate maps to recover missing degrees of freedom on the attractor.

If sufficiently many delays are used, then in the limit of large data ($n \to \infty$) and in the absence of model error any distance function in data space will suffice to produce a forecast of skill limited only by the predictability limits of the dynamical system. However, in realistic scenarios involving datasets with finite sample counts, the analogs identified in finite-sized balls centered on $y$ will be sensitive to the choice of distance function. Intuitively, using a distance function with respect to which nearby points have similar time tendency should improve analog forecasting skill at least in the short to medium term. In Section 2.2, we saw that cone kernels induce a geometry that contracts local distances between points which are mutually aligned with the dynamical flow [see (2.6)], and we propose to use that geometry for analog identification.

In this approach, we represent the cone-kernel geometry in a coarse scale by the diffusion coordinates $\Phi_{s,l}$ in (2.11). Specifically, we identify analogs using the diffusion distance $D_l(y, x_i)$ in (2.12) computed from the leading-few eigenfunctions. To compute this diffusion distance, we first evaluate the eigenfunction values, $\tilde{\phi}_l(y)$, at the initial data using either the Nyström method in (2.20) or the Laplacian pyramids in (2.22), and then compute $D_{s,l}(y, x_i) = \|\Phi_{s,l}(y) - \Phi_{s,l}(x_i)\|$, where $\tilde{\Phi}_{s,l}(y) = (\lambda_{s,1}^{l} \tilde{\phi}_1(y), \ldots, \lambda_{s,l}^{l} \tilde{\phi}_l(y))$. We then replace
the analog-identification step of conventional analog forecasting in (3.1) by

$$i = \arg\min_{j \in \{1, \ldots, n\}} D_{s,l}(y, x_i).$$

(3.3)

The operation with the shift map in (3.2) used to produce the forecast at lead time $\tau$ remains the same. As an alternative to (3.3), it is also possible to identify a single analog directly from the kernel affinity, i.e.,

$$i = \arg\min_{j \in \{1, \ldots, n\}} K_{\epsilon}(y, x_i).$$

(3.4)

This approach is computationally cheaper and involves fewer parameters than (3.3), but does not have a natural geometrical interpretation if the sampling density on the manifold is non-uniform relative to the Riemannian volume form $d\mu$. Our experiments in Section 4 ahead demonstrate that these two approaches for analog identification both improve prediction skill compared to the traditional analog method in (3.1) based on Euclidean distances, especially in high-dimensional data spaces. In low-dimensional problems with dense sampling, (3.3) and (3.4) have similar performance (and provide a modest improvement compared to Euclidean distances), but in high-dimensional settings analog identification via diffusion distances was found to be the best performer among the single-analog methods studied here.

### 3.3. Kernel-weighted ensemble forecasts

The methods described in Sections 3.1 and 3.2 both use a single state in the past to predict the evolution of an observable. However, there is no reason that other states in the training data should not carry useful predictive information for the observable of interest. We now describe a method that leverages this information through weighted ensembles of the training data with weights determined using the out-of-sample extension schemes of Section 2.4.

First, note that the single-analog forecasts in (3.2) can also be expressed as an expectation value over the training dataset with respect to a Dirac measure centered on $x(t_i)$ from (3.1) and (3.3). That is, we can write

$$f(y, \tau) = \mathbb{E}_{\rho_y} S_\tau f = \sum_{j=1}^{n} f(x(t_j + \tau)) \rho_y(x(t_j)),$$

where $\rho_y$ is a $y$-dependent probability density in the training data, given in this case by $\rho_y(x_j) = \delta_{ij}$. In the limit of large data, the discrete Monte-Carlo sum converges to a continuous manifold integral,

$$f(y, \tau) = \mathbb{E}_\rho S_\tau f = \int_{\mathcal{M}} S_\tau f(x) \rho(x) d\mu_\epsilon(x),$$

with $\rho(x_j) = \delta_{x_i}(x)$. Notice now that the last two equations have structural similarities with the out-of-sample extension equations (2.20) and (2.22), both of which are expressed in terms of averaging operators. More specifically, the action of the operator $P_\epsilon$ in (2.7) and its
discrete counterpart $\hat{P}_\epsilon$ in (2.8) can be expressed as $P_\epsilon f(y) = \mathbb{E}_{p_\epsilon(y,\cdot)} f$ and $\hat{P}_\epsilon f(y) = \mathbb{E}_{\hat{p}_\epsilon(y,\cdot)} f$, respectively, where $p_\epsilon$ and $\hat{p}_\epsilon$ are the corresponding transition probability kernels. In other words, the action of these operators can be described in terms of weighted sums over the training data with weights given by $p_\epsilon$ or $\hat{p}_\epsilon$.

With this observation, we formulate kernel-weighted ensemble forecasting schemes using $p_\epsilon(y,\cdot)$ and $\hat{p}_\epsilon(y,\cdot)$ to replace the Dirac $\delta$-measures used above. In particular, we put forward two schemes, where the ensemble weights are chosen according to either the Nyström method or the multiscale Laplacian pyramids approach. Specifically, in the Nyström case we set

$$f(y, \tau) = \sum_{i=1}^{l} \frac{\langle f, \phi_i \rangle_{\mu}}{\lambda_i} \mathbb{E}_{p_\epsilon(y,\cdot)} S_\tau \phi_i,$$

(3.5)

and in the Laplacian pyramids case we define,

$$f(y, \tau) = \mathbb{E}_{p_\epsilon(y,\cdot)} S_\tau f_0 + \sum_{i=1}^{l} \mathbb{E}_{p_\epsilon} S_\tau g_i.$$

(3.6)

Note that at lead time $\tau = 0$ these expressions are consistent with the corresponding out-of-sample extension equations, i.e., (2.20) and (2.22), respectively. Moreover, (3.5) and (3.6) are written down using the continuous inner products and transition probability kernels, but as usual the discrete counterparts of these equations can be obtained by replacing $\langle \cdot, \cdot \rangle_{\mu}$ and $p_\epsilon$ by $\langle \cdot, \cdot \rangle_{\epsilon}$ and $\hat{p}_\epsilon$, respectively.

Qualitatively, the ensemble weights $p_\epsilon(y, x_i)$ in (3.5) and (3.6) measure how similar the initial data $y$ are to the samples in the training dataset. To see how this similarity measure is related to the kernel-induced geometry, consider the spectral expansion

$$p_\epsilon(y, x) = 1 + \sum_{i=1}^{\infty} \lambda_i \bar{\phi}_i(y) \phi_i(x),$$

(3.7)

where $\bar{\phi}_i(y)$ is the extension of eigenfunction $\phi_i$ evaluated at $y$. For $\epsilon$ small-enough so that the eigenvalues $\lambda_i$ are all non-negative, it follows that the nonlinear similarity measure $p_\epsilon(y, x)$ is given up to an additive constant by the dot product between the feature vectors $\bar{\Phi}_{1/2}(x) = (\lambda_1^{1/2} \bar{\phi}_1(x), \lambda_2^{1/2} \bar{\phi}_2(x), \ldots)$ and $\bar{\Phi}_{1/2}(y) = (\lambda_1^{1/2} \bar{\phi}_1(y), \lambda_2^{1/2} \bar{\phi}(y), \ldots)$ of the data into the infinite-dimensional sequence space $\ell^2$. This is an example of the so-called “kernel trick” in machine learning, where a nonlinear problem is converted into a linear problem by embedding the data into a higher-dimensional feature space. In the asymptotic limit $\epsilon \to 0$, the embedding becomes isometric with respect to the kernel-induced Riemannian metric [in this case, the cone-kernel metric with local length function given by (2.6)], and $p_\epsilon(y, x)$ becomes equal to the dot product of the feature vectors adapted to the kernel geometry. In Section 4, we will demonstrate that this approach can provide significantly higher forecast skill compared to forecasting with a single analog in either Euclidean or diffusion distances.
3.4. Error metrics

In Section 4, we will assess the methods described in Sections 3.1–3.3 in hindcast experiments against test datasets consisting of sequences of initial data \( N = \{ y_i \}_{i=1}^{\hat{n}+k} \) with \( y_i = y((i-1)\delta t) \). We perform these tests using two distinct types of forecast observables. Observables of the first type are those naturally defined on the test dataset independently of the model, i.e., for an observable \( f \), we have the ground truth-values \( F_i = f(y_i) \). Examples of such observables are the full state vector, \( F_i = y_i \), or some of its components as it will be the case in Section 4.1. Observables of the second type are data-driven observables, constructed through a data analysis algorithm applied to the training data. For example, in Section 4.2 the forecast observables will be the leading low-frequency eigenfunctions of North Pacific SST recovered by cone kernels. For data-driven observables, there is no objective notion of ground truth on the test data, but a natural reference value is given by the out-of-sample extension of \( f \) on \( N \), i.e., \( F_i = Af(y_i) \), where \( A \) is an out-of-sample extension operator [e.g., 2.20 or 2.22]. Another possibility is to compute a surrogate observable \( \tilde{f} \) by applying the data analysis algorithm directly to the test data, and treat \( F_i = \tilde{f}(y_i) \) as the ground truth. In all cases, the forecast error at lead time \( \tau = k\delta t \) for initial data \( y_i \) is \( \varepsilon_i(\tau) = f(y_i, \tau) - F_i + k \).

Based on these definitions for the ground truth and the corresponding forecast errors, we compute two time-averaged skill scores, namely the root mean squared error (RMSE) and pattern correlation (PC) scores. These are given by

\[
\text{RMSE}(\tau) = \sqrt{\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \varepsilon_i^2(\tau)}, \quad \text{PC}(\tau) = \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \frac{(f(y_i, \tau) - \bar{f})(F_i + k - \bar{F})}{\sigma_f \sigma_F},
\]

where \( \bar{f} = \sum_{i=1}^{\hat{n}} f(y_i, \tau)/\hat{n} \), \( \bar{F} = \sum_{i=1}^{\hat{n}} F_i/\hat{n} \), \( \sigma_f^2 = \sum_{i=1}^{\hat{n}} (f(y_i, \tau) - \bar{f})^2/\hat{n} \), and \( \sigma_F^2 = \sum_{i=1}^{\hat{n}} (F_i + k - \bar{F})^2/\hat{n} \). If \( M \) and \( N \) both sample the same manifold (i.e., there is no model error in the training data), the notions of ground truth based on \( Af \) and \( \tilde{f} \) and the corresponding skill scores will converge as the number of training and test samples tend to infinity. At finite sample counts, the forecast skill scores relative to \( \tilde{f} \) will generally be larger than those based on \( Af \). In applications with model error in the training data, the ground truth based on \( \tilde{f} \) may become problematic since the same data analysis algorithm applied to \( M \) and \( N \) may not yield comparable observables. For completeness, in Section 4.2 we present forecast skill results using both approaches.

4. Experimental results

4.1. A chaotic intermittent low-order atmosphere model

We illustrate several key aspects of the forecasting methods described in Section 3, including Takens’ time-lagged embedding, cone kernels, and kernel-weighted ensemble forecasts, through experiments on a six-dimensional low-order model of the atmosphere featuring chaotic metastability [23, 22]. Specifically, setting \( u(t) = (u_1(t), \ldots, u_6(t)) \in \mathbb{R}^6 \), we
consider a deterministic dynamical system of the form \( \dot{u}(t) = V(u(t)) \), where the vector field, \( V(u) = (V_1(u), \ldots, V_6(u)) \), is given by

\[
\begin{align*}
V_1(u) &= \gamma^*_1 u_3 - C(u_1 - u^*_1), \\
V_2(u) &= - (\alpha_1 u_1 - \beta_1) u_3 - C u_2 - \delta_1 u_4 u_6, \\
V_3(u) &= (\alpha_1 u_1 - \beta_1) u_2 - \gamma_1 u_1 - C u_3 + \delta_1 u_4 u_5, \\
V_4(u) &= \gamma^*_2 u_6 - C(u_4 - u^*_4) + \epsilon(u_2 u_6 - u_3 u_5), \\
V_5(u) &= - (\alpha_2 u_1 - \beta_2) u_6 - C u_5 - \delta_2 u_4 u_3, \\
V_6(u) &= (\alpha_2 u_1 - \beta_2) u_5 - \gamma_2 u_4 - C u_6 + \delta_2 u_4 u_2.
\end{align*}
\]

Here, \( \alpha_i, \beta_i, \gamma_i, \gamma^*_i, \epsilon, C, \) and \( u^* \) are model parameters. This model is derived by a low-order truncation of the streamfunction in the barotropic vorticity equation in a channel in the presence of topography and a zonal (east–west) forcing profile. In particular, the state-vector components \( u_i \) are expansion coefficients in the spatial Fourier basis for the channel. Physically, this model provides a coarse approximation of atmospheric dynamics in the presence of orography and a zonal background flow—it is a simplified model for the polar jet stream interacting with continental mountain ranges, such as the mountain ranges of North America. Observationally, it is known that such flows exhibit metastable transitions between so-called zonal and blocked states, where the jet propagates relatively unimpeded, or is significantly blocked by the presence of topography, respectively. With an appropriate choice of parameters [22], the six-dimensional model exhibits a qualitatively similar behavior, and the two regimes are manifested by distinct regions in the \((u_1, u_4)\) plane (see Figure 4.1). In the spatial domain, these components of the state vector represent purely zonal flow, with no variation in the meridional direction, reflecting the interesting regime switching between zonal and blocked states.

In what follows, we apply the techniques of Section 3 to predict the evolution of the observable \( f = (u_1, u_4) \) representing the regimes, using as initial data either the full state vector, \( y = u \), or a vector formed by applying delay-coordinate maps to the projected
state vector \( z = (u_1, u_4) \) in accordance with (2.1). We note that forecasting with reduced parametric models is particularly challenging in this example. For instance, in [22] it was found that even a modest dimension reduction of the full model to four degrees of freedom by Galerkin projection onto the PCA basis produces models that either decay to fixed points, or become locked on orbits of unrealistically high temporal regularity, failing to reproduce chaotic regime transitions. On the other hand, Galerkin projection onto a basis derived by eigenfunctions from the kernel in (2.5) with \( \zeta = 0 \) was able to exhibit chaotic regime transitions using only three dimensions [30]. This suggests that the patterns derived from Laplace-Beltrami eigenfunctions are more intrinsic to the dynamical system than the linear PCA modes.

Our training dataset, \( M \), consists of \( n = 10^4 \) samples taken every \( \delta t = 1 \) day (d) from an equilibrated simulation of the full dynamical system performed using a Runge-Kutta method. A scatterplot of the components \((u_1, u_4)\) of the state vector in the training data is depicted in Figure 4.1. To generate the test dataset \( N \), we changed the initial conditions and ran another equilibrated simulation with \( \tilde{n} = 5000 \) data points.

First, we discuss hindcast experiments using the full state vector as initial data. In these experiments we did not perform delay-coordinate mapping as the initial data are complete [i.e., we set \( q = 1 \) in (2.1)]. Therefore, the only dependence of the kernel (2.5) in the dynamics is through the finite-difference estimates, \( v \) and \( w \), of the dynamical vector field \( V \), which were evaluated here using a first-order backward scheme. Figure 4.2 shows sample forecast trajectories and the RMSE and PC skill scores from (3.8) obtained via the empirical forecasting methods described in Section 3; namely, conventional analog forecasting with Euclidean distances, analog forecasting with diffusion distances and kernel affinities, and kernel-weighted ensemble forecasting with Laplacian-pyramid weights. Note that we do not use the Nyström approach in (3.5) as \( f \) is not a band-limited observable in the eigenfunction basis.

For the analog forecasts with diffusion distances, we used the cone kernel with \( \zeta = 0.995 \), corresponding to strong influence of the angular terms associated with the dynamical flow. The bandwidth parameter was set to \( \epsilon = 1 \). We computed the diffusion distances via (2.14) using a transition probability kernel \( \hat{p}_{\epsilon,k} \) truncated at \( k = 10 \) nearest neighbors (see Section 2.4.2). For the affinity-based analog forecasts we again used the cone kernel with \( \zeta = 0.995 \), and also examined the case with \( \zeta = 0 \) corresponding to no influence of the angular terms. We find that the analogs chosen with respect to the \( \zeta = 0 \) kernel affinities via (3.4) have essentially identical skill to the canonical analog forecast, but there is a marked improvement for the cone kernel with \( \zeta = 0.995 \). The improvement of skill is reflected in the improved RMSE and PC results in Figures 4.2(b) and 4.2(c), and but is more prominent in the forecast time series in Figure 4.2(a). There, the analog based on Euclidean distance deviates from the ground truth for \( u_1(\tau) \) after \( \tau \approx 360 \) d, but the affinity-based analog with \( \zeta = 0.995 \) tracks the true trajectory up to \( \tau \approx 480 \) d. We therefore see that choosing analogs with respect to an affinity measure that preferentially selects samples with similar time
Figure 4.2: Hindcast experiments for the low-order atmospheric model with complete initial data. (a) Ground-truth and predicted time series for $u_1$ via conventional Euclidean-distance analogs, diffusion-distance analogs with cone kernel parameter $\zeta = 0.995$, affinity-based analogs with $\zeta = 0$ and 0.995, and weighted-ensemble forecast via Laplacian pyramids. (b,c) Root mean squared error (RMSE) and pattern correlation (PC) scores, respectively, for the forecast models in (a).

tendency (as $\zeta \approx 1$ cone kernels do) leads to improvement of skill. Note that in this densely-sampled, low-dimensional example the analog forecasts with $\zeta = 0.995$ diffusion distances have marginally poorer skill than their affinity-based counterparts (but they still perform better than Euclidean distances). In the experiments of Section 4.2 with high-dimensional data, we observe that the diffusion distances give higher prediction skill than kernel affinities.

The most significant improvement of skill, however, takes place for the kernel-weighted ensemble forecasts made via the Laplacian pyramids approach in (3.6). Here, we used $\zeta = 0.995$, and truncated the kernel $\hat{p}_{k_1,k}$ at $k = 10$ nearest neighbors. The initial bandwidth
parameter $\epsilon_0$ was set to the median of the distances between the $k$ nearest neighbors. The error tolerance for Laplacian pyramids was $\varepsilon = 0.05$ and $l = 2$ iterations were used for extension. In Figure 4.2, the ensemble mean forecast accurately tracks the true trajectory to at least 480 d, the RMSE is lower at both intermediate and long times ($\tau \gtrsim 60$ d), and the PC values exceed those from other methods for up to 300 d leads. To summarize, in these experiments the cone-kernel geometry with $\zeta \approx 1$ (which produces neighborhoods aligned with the dynamical flow), improves the identification of analogs, and moreover there are further benefits from using multiple samples in the training data through weighted ensembles constructed with these kernels.

Next, to mimic a real-world scenario where only partial observations are available, we perform an experiment where only two coordinates, $z_i = (u_1(t_i), u_4(t_i))$, of the full state vector are observed. Following Section 2.1, we form the vectors of initial data $x_i$ by applying delay-coordinate mapping to $z_i$ with $q$ lags. In an operational setting, this would correspond to predicting the state of the atmosphere given its observed evolution over $q$ days in the past. The results from hindcast experiments with $q = 1$ and $q = 20$ [i.e., no delay coordinate mapping and delay coordinate mapping over a physical time window $(q - 1)\delta t = 19$ d, respectively] and the same forecasting techniques used in the fully-observed case are shown in Figure 4.3.

Without Takens’ time-lagged embedding, the analog forecasts with partial observations perform significantly worse than the same analog forecasts with complete observations for all forecasting methods. This is consistent with the two-dimensional initial data with $q = 1$ being insufficient to determine neighborhoods on the attractor, and there is evidence that the intrinsic dimension of the attractor in this class of systems is greater than two [22]. On the other hand, using Takens’ time-lagged embedding with $q = 20$ delays leads to nearly equal skill as the case with full observations. Thus, in this relatively low-dimensional example, delay-coordinate mapping is a highly effective way of recovering the necessary dynamical information for empirical forecasting. Similar to the complete-observation experiments, the $\zeta = 0.995$ cone-kernel affinity measure gave the best single-analog forecast, and the weighted ensemble through Laplacian pyramids provides the highest skill across the board.

4.2. Long-range forecasting in the North Pacific sector of a comprehensive climate model

As a high-dimensional application, we study low-frequency SST variability in the North Pacific sector of the coupled climate model CCSM3 [18]. The datasets used in this study are two long equilibrated control integrations available in the public domain (https://www.earthsystemgrid.org) with designations b30.004 and b30.009. Integration b30.004 spans 900 years (y) using a 2.8° nominal resolution for the atmospheric component of the model; b30.009 spans 450 years, but employs a higher, 1.4°, atmosphere resolution. Both integrations employ the same 1° ocean resolutions. In this suite of experiments, we treat the first 400 years of integration b30.004 as the training dataset $M$. We use this dataset both to construct data-driven low-frequency observables for prediction, as well as training data for
empirical forecasting. Two sets of experiments are discussed below, the first of which is a perfect-model experiment using the second half of b30.004 as the test dataset. The second set of experiments introduces model error, using the data from b30.009 as test data from “nature”. Elsewhere [19], we perform hindcast experiments against actual observational data acquired via remote sensing, involving multi-variate datasets of SST and arctic sea ice concentration. The b30.004 and b30.009 datasets were also used in [29], where the dominant spatiotemporal modes of North Pacific SST variability as extracted by NLSA algorithms were compared across different GCMs.

Figure 4.3: Same experiments as in Figure 4.2, but for incomplete initial data, $z = (u_1, u_4)$, reconstructed using the delay-coordinate mapping in (2.1) with $q$ lags. The experiments shown here are for $q = 1$ and $q = 20$, corresponding to no delay-coordinate mapping and delay coordinate mapping over a physical time window of 19 days, respectively.
4.2.1. Recovering low-frequency observables. Following [29, 28], we extract low-frequency SST modes using $q = 24$ delay-coordinate lags (corresponding to a two-year temporal embedding window) to induce timescale separation. Throughout, we work with cone kernels with $\zeta = 0.995$ as these kernels provide better timescale separation in the eigenfunctions [28] than the $\zeta = 0$ kernels used in [29]. The SST snapshots have dimension $d = 6671$ (equal to the number of North Pacific ocean gridpoints in CCSM3), and the dimension of the data vectors $x_i$ after delay-coordinate mapping via (2.1) is $m = 24 \times 6671 = 160,104$. Throughout, we work with a kernel bandwidth parameter $\epsilon = 2.25$, but our results are qualitatively robust for bandwidths in the interval $0.5–5$.

Representative Laplace-Beltrami eigenfunctions computed from the training data with these delay-coordinate and kernel parameters are shown in Figure 4.4. The eigenfunctions separate the temporal variability of the data into qualitatively distinct families, namely periodic [Figure 4.4(a)], low-frequency [Figures 4.4(b,d)], and intermittent [Figure 4.4(c)] modes. The periodic modes closely resemble sinusoids with frequencies given by integer multiples of $1 \text{ y}^{-1}$. These modes represent the harmonics of the seasonal cycle (a prominent source of variability in North Pacific SST), and form doubly-degenerate pairs with a phase offset of $\pi/2$. The low-frequency modes describe oscillations taking place at interannual to decadal timescales, and are characterized by red-noise-like power spectra. The intermittent modes consist of periodic signals at the seasonal-cycle harmonics modulated by low-frequency envelopes. We refer the interested reader to [29, 13, 28] for detailed discussions on the properties and physical significance of these mode families.

In what follows, our focus will be on forecasts of the leading-two low-frequency eigenfunctions, i.e., $\phi_5$ and $\phi_8$. These eigenfunctions carry significant power on interannual to decadal timescales, making them good candidate observables for long-range forecasting, but in addition they are associated with physically meaningful patterns in spatiotemporal reconstructions (e.g., see the online supporting animation in [28]). In particular, $\phi_5$ and $\phi_8$ display the salient features of two prominent low-frequency climate patterns in the North Pacific—the PDO [39] and NPGO [25], respectively. The former is characterized by a horseshoe-like temperature anomaly pattern, developing east of Japan, together with an anomaly of the opposite sign along the west coast of North America. Based on these observations, we select eigenfunctions $\phi_5$ and $\phi_8$ as data-driven observables for empirical forecasting. It is important to note that our kernel construction is essential to achieve timescale separation; in particular, if no Takens embedding is used ($q = 1$) the temporal character of the eigenfunctions becomes corrupted, mixing low-frequency and periodic variability. However, despite their low-frequency character, the PDO and NPGO are particularly challenging to predict with parametric models, and first-order autoregressive models often fail to beat the persistence forecast [19]. In the predictive skill results discussed below we include the persistence forecast as a reference for the forecast skill that can be achieved in these time series with simple autoregressive models.
4.2.2. Forecasting in a perfect-model environment. We begin with an application of the techniques of Section 3 in hindcasts of a test dataset generated by the same model as the model generating the training data—as stated earlier, the test dataset in this set of experiments consists of the 400 years of CCSM3 integration b30.004 following the 400 years of the training data. Figures 4.5 and 4.6 display predictive skill results for the PDO ($\phi_5$) and NPGO ($\phi_8$) eigenfunctions obtained with the methods described in Section 3, as well as with the persistence forecast. The prediction observables in these experiments are pure eigenfunctions, so we performed weighted-ensemble forecasts using both the Nyström method and Laplacian pyramids via (3.5) and (3.6), respectively. Moreover, because the prediction observables are data driven (i.e., they are not objectively defined on the test dataset), the skill scores were evaluated treating as ground truth either the out-of-sample extension of $\phi_5$ and $\phi_8$ (Figure 4.5), or the eigenfunctions $\hat{\phi}_5$ and $\hat{\phi}_8$ evaluated on the test dataset (Figure 4.6),
as discussed in Section 3.4. In Figure 4.5, the skill scores of the Nyström-based ensemble forecasts were computed using ground-truth values also obtained from the Nyström method via (2.20), as this method is consistent on the training dataset for pure eigenfunctions. The other methods were compared with out-of-sample extended eigenfunctions from Laplacian pyramids. Specifically, the $P_\epsilon$ operators in (2.22), were constructed with an initial bandwidth scale $\epsilon_0 = 2.25$ using all data points and without $k$-nearest-neighbor truncation. The same parameters were also used in the weighted-ensemble forecast in (3.6). Note that because there is no model error in these experiments, the training and test data sample the same manifold so the two types of skill scores used in Figures 4.5 and 4.6 are mutually consistent in the large-data limit. However, at finite sample counts, the ground truth convention of Figure 4.6 generally yields lower skill scores than Figure 4.5.

The main finding from these experiments is that the kernel-weighted ensemble forecasts, implemented either with the Nyström method in (3.5) or with Laplacian pyramids in (3.6), are the best performers across the board, substantially improving over single-analog forecasting approaches as well as persistence forecasts. In the case of the PDO, PC scores exceeding 0.5 (a popular lower threshold for “useful” forecasts) persist for 32-month lead times. In contrast, the persistence forecast crosses the PC = 0.5 threshold at 25-month leads, and has decayed to 0.38 at $\tau = 32$ months. Note that for this dataset the persistence forecast performs better than linear autoregressive models [19], which are popular parametric models for the PDO [42]. Improvements over the persistence forecast are even more significant for the NPGO eigenfunction, which is a more rapidly decorrelating variable than the PDO. In this case, the PC = 0.5 crossing time for the kernel-weighted forecasts is 19 months, which is almost a factor of two improvement compared to the 10-month crossing time for the persistence forecast.

Consider now the forecasts made by single-analog methods. In this example with a high-dimensional ambient data space, the analogs selected via the cone-kernel affinity measure using (3.4) have nearly identical skill as the conventional Euclidean-distance analogs (cf. the low-dimensional example in Section 4.1). Therefore, the affinity-based results are not included in Figures 4.5 and 4.6. On the other hand, the analogs selected via diffusion distances [computed here via (2.14) using $l = 30$ eigenfunctions and unit diffusion time $s$] provide some improvement of skill over conventional analogs, mainly in the short to medium term. Note that in the literature it is customary to calibrate the forecast error vs. lead time curves such that the error at $\tau = 0$ vanishes (e.g., [11]), and our skill scores with single analogs will generally appear lower than in those approaches.

Despite the improvement of skill of diffusion-distance analogs over their Euclidean-distance counterparts, it is evident from Figures 4.6 and 4.5 that all single-analog methods are affected by large errors at short lead times. These errors are caused by poor reconstruction accuracy using single analogs. That is, unless the sampling density on the attractor is large, the ground truth at initialization time ($\tau = 0$) may deviate significantly from the estimated value $f(y, 0)$ from the analog. Substantial errors at initialization time typically persist at
Figure 4.5: Forecast skill for the low-frequency eigenfunctions $\phi_5$ and $\phi_8$ representing the PDO and NPGO, respectively, in a perfect-model scenario. The RMSE and PC scores from (3.8) where evaluated for the persistence forecast, single-analog forecasts based on Euclidean distances and diffusion distances, and kernel-weighted ensemble forecasts using the Nyström method and Laplacian pyramids.

least for the short term, and in such cases single-analog methods will have lower short-term skill than methods which are able to accurately reconstruct the forecast observable at initialization, including the persistence forecast. Indeed, as shown in Figures 4.5(a,c) and 4.6(a,c), the single-analog forecasts for the PDO (a slowly-decorrelating observable) fail to beat persistence for both short and long times. Single-analog forecasts of the NPGO, which is a more rapidly decorrelating variable, do outperform persistence at moderate to long lead times [$\tau \gtrsim 10$ months in Figures 4.5(b,d) and 4.6(b,d)], but the short-term skill for this observable is also poor. Note that poor short-term skill with single-analog methods was not an issue in the experiments in Section 4.1 with the low-order atmospheric model (see Figures 4.2 and 4.3). There, the dimension of both the ambient data space and the attractor
was low, and the sampling density was sufficiently high for accurate reconstruction at short leads even with a single analog. In contrast, by appropriately weighing multiple samples in the training data, the kernel-weighted ensemble methods perform well in both low- and high-dimensional ambient spaces, and for both short and long lead times.

4.2.3. Forecasting with model error in the training data To test the robustness of our techniques in the presence of model error in the training data, we now study the forecast skill of the PDO and NPGO in an experiment where the training data come from the first 400 years of the b30.004 CCSM integration, but the test data come from integration b30.009. As mentioned earlier, these two integrations differ in the resolution of the atmospheric component of CCSM. These two datasets therefore have different atmospheric dynamics, which lead in turn to differences in SST variability through the associated atmosphere–ocean momentum and heat fluxes. Because b30.009 has higher atmospheric resolution than
b30.004, in these experiments we view b30.009 as “nature” and b30.004 as an imperfect model. More broadly, this experimental setup is representative of many empirical modeling applications in the physical sciences, where long-term observations of nature are not available for training.

We repeated the experiments described in Section 4.2.2, i.e., the persistence forecast, single-analog forecasts with Euclidean and diffusion distances, and kernel-weighted ensemble forecasts with Nyström and Laplacian-pyramid weights, on the new test dataset. The skill scores from these experiments are shown in Figures 4.7 and 4.8, where the ground truth is given by out-of-sample extension of the PDO and NPGO eigenfunctions from b30.004 to b30.009 and the PDO ($\tilde{\phi}_5$) and NPGO ($\tilde{\phi}_8$) eigenfunctions computed directly on b30.009, respectively. Due to model error, the latter ground-truth values are systematically biased away from the values of the PDO and NPGO on the training dataset, even for infinitely-many training and test samples and zero lead time.

As expected, the skill scores in these experiments with model error in the training data are generally lower than their perfect-model counterparts in Figures 4.5 and 4.6. In the case of the PDO, PC scores exceeding 0.5 persist for 28-month lead times, i.e., 4 months less than in the perfect model experiment. In contrast, the persistence forecast crosses the PC = 0.5 threshold at 24-month leads, and has decayed to 0.42 at $\tau = 28$ months. Comparing Figures 4.6(a,b) and 4.8(a,b), we notice that the initial RMSE at $\tau = 0$ for the PDO increases from about 0.1 to 0.2, and the initial RMSE for the NPGO increases from 0.27 to 0.42. However the kernel-weighted ensemble methods still provide significantly higher skill than persistence, and also outperform single-analog methods using Euclidean or diffusion distances. In the case of the NPGO, the PC = 0.5 crossing time for the ensemble forecasts remains a factor of two higher than the persistence forecast [see Figure 4.7(d)].

5. Conclusions

In this paper, we proposed a family of data-driven, non-parametric forecasting methods for observables of dynamical systems. Inspired by Lorenz’s analog forecasting technique [38], these methods perform forecasting by identifying the most closely matching state to the initial data, or a weighted ensemble of states, in a training dataset comprised of historical observations of nature, or output from a potentially imperfect model. A key ingredient of these techniques is that the selection of analogs is performed not with respect to Euclidean distances in data space, but rather with respect to a modified Riemannian geometry favoring samples in the training data with similar dynamical evolution to the observed data at forecast initialization. Operationally, this geometry is accessed through kernel operators and their associated diffusion distances and eigenfunctions, computed with the diffusion maps algorithm [15] and out-of-sample extension techniques [16, 47]. These kernel operators incorporate empirically accessible aspects of the dynamical system through Takens delay-coordinate maps [30, 7] and finite-difference approximations of the dynamical vector field on the attractor [28], enhancing timescale separation and robustness to changes of observation
modality. Out-of-sample extension operators were also used here in a novel way to create weighted forecast ensembles leveraging the predictive information from multiple samples in the training data.

We demonstrated the efficacy of these schemes in hindcast experiments involving a low-order deterministic model for the atmosphere with chaotic regime metastability, as well as interannual to decadal variability of sea surface temperature (SST) in the North Pacific sector of a comprehensive climate model. We also studied a North Pacific SST experiment with model error in the training data. In all experiments, our proposed weighted-ensemble methods led to significant improvement of forecast skill compared to conventional analog forecasting with Euclidean distances. The weighted-ensemble methods also significantly
outperformed forecasts made with single analogs identified using diffusion distances.

In the experiments with the low-order atmospheric model, we demonstrated that Takens delay-coordinate maps are an effective way of improving forecast skill with partially observed initial data for all methods studied here. In this relatively low-dimensional setting with dense sampling of the attractor, single-analog forecasts with diffusion distances provided a small to moderate improvement prediction skill compared to conventional analogs. On the other hand, forecasts with the weighted-ensemble methods were able to track the metastable regime transitions in this model for longer times, leading to a pronounced improvement of skill especially at moderate to long leads.

In the North Pacific SST applications, we constructed data-driven forecast observables through kernel eigenfunctions representing two prominent interannual patterns of variability in the North Pacific, namely the Pacific decadal oscillation (PDO) \[39\] and the North Pacific gyre oscillation \[25\]. The timescale separation capabilities of the kernels used here

Figure 4.8: Same skill scores as Figure 4.7, but computed using as ground truth the PDO and NPGO eigenfunctions obtained from the test data with model error.
contributed significantly to the clean low-frequency character and favorable predictability properties of the observables, extending the range of useful forecasts \( > 0.5 \) pattern correlation score (PC)] to 2.5 years for the PDO and 1.5 years for the NPGO. As a benchmark, the persistence forecast for these variables (which is more skillful than simple autoregressive models) decays below the 0.5 PC threshold at 2 years (PDO) and 0.8 years (NPGO). Again, kernel-weighted ensembles (as opposed to single analogs) were crucial to obtain these results, but in this high-dimensional environment single-analog methods with diffusion distances also performed markedly better than Euclidean-distance analogs.

Stemming from this work are several future research directions. Here, we focused on forecasting with the kernels developed in \([30, 32]\), but other kernels for dynamical systems have been proposed in the literature \([52, 56, 8]\) and could be employed in our forecasting schemes. Moreover, we have focused on Nyström-type operators and Laplacian pyramids to extend functions, but other interpolation and extrapolation techniques, such as kriging, can be explored. Finally, analog methods are able to avoid the dynamical model errors of parametric models, but have limited ability to generate innovations. As a result, it should be fruitful to develop blended methods taking advantage of both parametric and non-parametric forecast models.

Acknowledgments

The research of Dimitrios Giannakis is partially supported by ONR DRI grant N00014-14-0150 ONR MURI grant 25-74200-F7112. Zhizhen Zhao is supported as a postdoctoral fellow through the last grant. The authors wish to thank Darin Comeau and Andrew Majda for stimulating conversations.

References


Analog Forecasting with Dynamics-Adapted Kernels


