Kernel Observers: Systems-Theoretic Modeling and Inference of Spatiotemporally Varying Processes

Hassan Kingravi
Pindrop Security
Atlanta, GA 30308
hkingravi@pindropsecurity.com

Harshal Maske and Girish Chowdhary
University of Illinois at Urbana Champaign
Urbana, IL 61801
hmaske2@illinois.edu, girishc@illinois.edu

Abstract

We consider the problem of estimating the latent state of a spatiotemporally evolving continuous function using very few sensor measurements. We show that layering a dynamical systems prior over temporal evolution of weights of a kernel model is a valid approach to spatiotemporal modeling, and that it does not require the design of complex nonstationary kernels. Furthermore, we show that such a differentially constrained predictive model can be utilized to determine sensing locations that guarantee that the hidden state of the phenomena can be recovered with very few measurements. We provide sufficient conditions on the number and spatial location of samples required to guarantee state recovery, and provide a lower bound on the minimum number of samples required to robustly infer the hidden states. Our approach outperforms existing methods in numerical experiments.

1 Introduction

Modeling of large-scale stochastic phenomena with both spatial and temporal (spatiotemporal) evolution is a fundamental problem in the applied sciences and social networks. The spatial and temporal evolution in such domains is constrained by stochastic partial differential equations, whose structure and parameters may be time-varying and unknown. While modeling spatiotemporal phenomena has traditionally been the province of the field of geostatistics, it has in recent years gained more attention in the machine learning community [2]. The data-driven models developed through machine learning techniques provide a way to capture complex spatiotemporal phenomena that are not easily modeled by first-principles alone, such as stochastic partial differential equations.

In the machine learning community, kernel methods represent a class of extremely well-studied and powerful methods for inference in spatial domains; in these techniques, correlations between the input variables are encoded through a covariance kernel, and the model is formed through a linear weighted combination of the kernels [14]. In recent years, kernel methods have been applied to spatiotemporal modeling with varying degrees of success [2, 14]. Many recent techniques in spatiotemporal modeling have focused on nonstationary covariance kernel design and associated hyperparameter learning algorithms [4, 7, 12]. The main benefit of careful design of covariance kernels over approaches that simply include time as an additional input variable is that they can account for intricate spatiotemporal couplings. However, there are two key challenges with these approaches: the first is ensuring the scalability of the model to large scale phenomena, which manifests due to the fact that the hyperparameter optimization problem is not convex in general, leading to methods that are difficult to implement, susceptible to local minima, and can become computationally intractable for large datasets. In addition to the challenge of modeling spatiotemporally varying processes, we are interested in addressing the second very important, and widely unaddressed challenge: given a predictive model of the spatiotemporal phenomena, how can the current latent state of the phenomena be estimated using as few sensor measurements as possible? This is called

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

the *monitoring problem*. Monitoring a spatiotemporal phenomenon is concerned with estimating its current state, predicting its future evolution, and inferring the initial conditions utilizing limited sensor measurements. The key challenges here manifest due to the fact that it is typically infeasible or expensive to deploy sensors at a large scale across vast spatial domains. To minimize the number of sensors deployed, a predictive data-driven model of the spatiotemporal evolution could be learned from historic datasets or through remote sensing (e.g. satellite, radar) datasets. Then, to monitor the phenomenon, the key problem would boil down to reliably and quickly estimate the evolving latent state of the phenomena utilizing measurements from very few sampling locations.

In this paper, we present an alternative perspective on solving the spatiotemporal monitoring problem that brings together kernel-based modeling, systems theory, and Bayesian filtering. Our main contributions are two-fold: first, we demonstrate that spatiotemporal functional evolution can be modeled using stationary kernels with a linear dynamical systems layer on their mixing weights. In other words, the model proposed here posits *differential constraints*, embodied as a linear dynamical system, on the spatiotemporal evolution of a kernel based models, such as Gaussian Processes. This approach does not necessarily require the design of complex spatiotemporal kernels, and can accommodate positive-definite kernels on any domain on which it's possible to define them, which includes non-Euclidean domains such as Riemannian manifolds, strings, graphs and images [6]. Second, we show that the model can be utilized to determine sensing locations that guarantee that the hidden states of functional evolution can be estimated using a Bayesian state-estimator with very few measurements. We provide sufficient conditions on the number and location of sensor measurements required and prove non-conservative lower bounds on the minimum number of sampling locations. The validity of the presented model and sensing techniques is corroborated using synthetic and large real datasets.

1.1 Related Work

There is a large body of literature on spatiotemporal modeling in geostatistics where specific process dependent kernels can be used [17, 2]. From the machine learning perspective, a naive approach is to utilize both spatial and temporal variables as inputs to a Mercer kernel [10]. However, this technique leads to an ever-growing kernel dictionary. Furthermore, constraining the dictionary size or utilizing a moving window will occlude learning of long-term patterns. Periodic or nonstationary covariance functions and nonlinear transformations have been proposed to address this issue [7, 14]. Work focusing on nonseparable and nonstationary covariance kernels seeks to design kernels optimized for environment-specific dynamics, and to tune their hyperparameters in local regions of the input space. Seminal work in [5] proposes a process convolution approach for space-time modeling. This model captures nonstationary structure by allowing the convolution kernel to vary across the input space. This approach can be extended to a class of nonstationary covariance functions, thereby allowing the use of a Gaussian process (GP) framework, as shown in [9]. However, since this model's hyperparameters are inferred using MCMC integration, its application has been limited to smaller datasets. To overcome this limitation, [12] proposes to use the mean estimates of a second isotropic GP (defined over latent length scales) to parameterize the nonstationary covariances. Finally, [4] considers nonistropic variation across different dimension of input space for the second GP as opposed to isotropic variation by [12]. Issues with this line of approach include the nonconvexity of the hyperparameter optimization problem and the fact that selection of an appropriate nonstationary covariance function for the task at hand is a nontrivial design decision (as noted in [16]).

Apart from directly modeling the covariance function using additional latent GPs, there exist several other approaches for specifying nonstationary GP models. One approach maps the nonstationary spatial process into a latent space, in which the problem becomes approximately stationary [15]. Along similar lines, [11] extends the input space by adding latent variables, which allows the model to capture nonstationarity in original space. Both these approaches require MCMC sampling for inference, and as such are subject to the limitations mentioned in the preceding paragraph.

Geostatistics approach that finds dynamical transition models on the linear combination of weights of a parameterized model [2, 8] is advantageous when the spatial and temporal dynamics are hierarchically separated leading to a convex learning problem. As a result complex nonstationary kernels are often not necessary (although they can be accommodated). The approach presented in this paper aligns closely with this vein of work. A system theoretic study of this viewpoint enables the fundamental contributions of the paper, which are 1) allowing for inference on more general domains with a larger class of basis functions than those typically considered in the geostatistics community, and

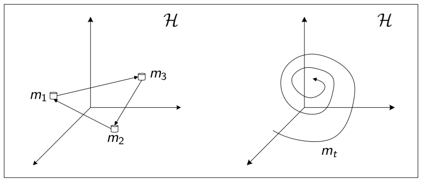
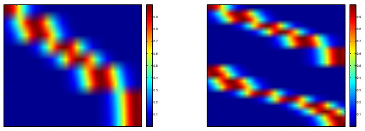


Figure 1: Two types of Hilbert space evolutions. Left: discrete switches in RKHS $\mathcal{H}$; Right: smooth evolution in $\mathcal{H}$.



(a) 1-shaded (Definition 1) (b) 2-shaded (Eq. (4))

Figure 2: Shaded observation matrices for dictionary of atoms.

2) quantifying the minimum number of measurements required to estimate the state of functional evolution.

It should be noted that the contribution of the paper concerning sensor placement is to provide sufficient conditions for monitoring rather than optimization of the placement locations, hence a comparison with these approaches is not considered in the experiments.

2 Kernel Observers

This section outlines our modeling framework and presents theoretical results associated with the number of sampling locations required for monitoring functional evolution.

2.1 Problem Formulation

We focus on predictive inference of a time-varying stochastic process, whose mean f evolves temporally as $f_{\tau+1} \sim \mathbb{F}(f_\tau, \eta_\tau)$, where $\mathbb{F}$ is a distribution varying with time τ and exogenous inputs η. Our approach builds on the fact that in several cases, temporal evolution can be hierarchically separated from spatial functional evolution. A classical and quite general example of this is the *abstract evolution equation* (AEO), which can be defined as the evolution of a function u embedded in a Banach space $\mathcal{B}$: $\dot{u}(t) = \mathcal{L}u(t)$, subject to $u(0) = u_0$, and $\mathcal{L} : \mathcal{B} \to \mathcal{B}$ determines spatiotemporal transitions of $u \in \mathcal{B}$ [1]. This model of spatiotemporal evolution is very general (AEOs, for example, model many PDEs), but working in Banach spaces can be computationally taxing. A simple way to make the approach computationally realizable is to place restrictions on $\mathcal{B}$: in particular, we restrict the sequence f_τ to lie in a reproducing kernel Hilbert space (RKHS), the theory of which provides powerful tools for generating flexible classes of functions with relative ease [14]. In a kernel-based model, $k : \Omega \times \Omega \to \mathbb{R}$ is a positive-definite Mercer kernel on a domain Ω that models the covariance between any two points in the input space, and implies the existence of a smooth map $\psi : \Omega \to \mathcal{H}$, where $\mathcal{H}$ is an RKHS with the property $k(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$. The key insight behind the proposed model is that spatiotemporal evolution in the input domain corresponds to temporal evolution of the mixing weights of a kernel model alone in the functional domain. Therefore, f_τ can be modeled by tracing the evolution of its mean embedded in a RKHS using switched ordinary differential equations (ODE) when the evolution is continuous, or switched difference equations when it is discrete (Figure 1). The advantage of this approach is that it allows us to utilize powerful ideas from systems theory for deriving necessary and sufficient conditions for spatiotemporal monitoring. In this paper, we restrict our attention to the class of functional evolutions $\mathbb{F}$ defined by linear Markovian transitions in an RKHS. While extension to the nonlinear case is possible (and non-trivial), it is not pursued in this paper to help ease the exposition of the key ideas. The class of linear transitions in RKHS is rich enough to model many real-world datasets, as suggested by our experiments.

Let $y \in \mathbb{R}^N$ be the measurements of the function available from N sensors, $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ be a linear transition operator in the RKHS $\mathcal{H}$, and $\mathcal{K} : \mathcal{H} \to \mathbb{R}^N$ be a linear measurement operator. The model for the functional evolution and measurement studied in this paper is:

$$f_{\tau+1} = \mathcal{A} f_\tau + \eta_\tau, \quad y_\tau = \mathcal{K} f_\tau + \zeta_\tau, \tag{1}$$

where η_τ is a zero-mean stochastic process in $\mathcal{H}$, and ζ_τ is a Wiener process in $\mathbb{R}^N$. Classical treatments of kernel methods emphasize that for most kernels, the feature map ψ is unknown, and possibly infinite-dimensional; this forces practioners to work in the dual space of $\mathcal{H}$, whose dimensionality is the number of samples in the dataset being modeled. This conventional wisdom precludes the use of kernel methods for most tasks involving modern datasets, which may have millions and sometimes billions of samples [13]. An alternative is to work with an approximate feature map $\widehat{\psi}(x) := \left[\, \widehat{\psi}_1(x) \;\cdots\; \widehat{\psi}_M(x) \,\right]$ to an approximate feature space $\widehat{\mathcal{H}}$, with the property that for every element $f \in \mathcal{H}$, $\exists \widehat{f} \in \widehat{\mathcal{H}}$ and an $\epsilon > 0$ s.t. $\|f - \widehat{f}\| < \epsilon$ for an appropriate function norm. A few such approximations are listed below.

Dictionary of atoms Let Ω be compact. Given points $\mathcal{C} = \{c_1, \ldots, c_M\}$, $c_i \in \Omega$, we have a dictionary of atoms $\mathcal{F}^{\mathcal{C}} = \{\psi(c_1), \cdots, \psi(c_M)\}$, $\psi(c_i) \in \mathcal{H}$, the span of which is a strict subspace $\widehat{\mathcal{H}}$ of the RKHS $\mathcal{H}$ generated by the kernel. Here,

$$\widehat{\psi}_i(x) := \langle \psi(x), \psi(c_i) \rangle_{\mathcal{H}} = k(x, c_i) \tag{2}$$

Low-rank approximations Let Ω be compact, let $\mathcal{C} = \{c_1, \ldots, c_M\}$, $c_i \in \Omega$, and let $K \in \mathbb{R}^{M \times M}$, $K_{ij} := k(c_i, c_j)$ be the Gram matrix computed from $\mathcal{C}$. This matrix can be diagonalized to compute approximations $(\widehat{\lambda}_i, \widehat{\phi}_i(x))$ of the eigenvalues and eigenfunctions $(\lambda_i, \phi_i(x))$ of the kernel [18]. These spectral quantities can then be used to compute $\widehat{\psi}_i(x) := \sqrt{\widehat{\lambda}_i}\widehat{\phi}_i(x)$.

Random Fourier features Let $\Omega \subset \mathbb{R}^n$ be compact, and let $k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$ be the Gaussian RBF kernel. Then random Fourier features approximate the kernel feature map as $\widehat{\psi}_\omega : \Omega \to \widehat{\mathcal{H}}$, where ω is a sample from the Fourier transform of $k(x, y)$, with the property that $k(x, y) = \mathbb{E}_\omega[\langle \widehat{\psi}_\omega(x), \widehat{\psi}_\omega(y) \rangle_{\widehat{\mathcal{H}}}]$ [13]. In this case, if $V \in \mathbb{R}^{M/2 \times n}$ is a random matrix representing the sample ω, then $\widehat{\psi}_i(x) := \left[\, \frac{1}{\sqrt{M}} \sin([Vx]_i), \frac{1}{\sqrt{M}} \cos([Vx]_i) \,\right]$

In the approximate space case, we replace the transition operator $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ in (1) by $\widehat{\mathcal{A}} : \widehat{\mathcal{H}} \to \widehat{\mathcal{H}}$. This approximate regime, which trades off the flexibility of a truly nonparametric approach for computational realizability, still allows for the representation of rich phenomena, as will be seen in the sequel. The finite-dimensional evolution equations approximating (1) in dual form are

$$w_{\tau+1} = \widehat{A} w_\tau + \eta_\tau, \quad y_\tau = K w_\tau + \zeta_\tau, \tag{3}$$

where we have matrices $\widehat{A} \in \mathbb{R}^{M \times M}$, $K \in \mathbb{R}^{N \times M}$, the vectors $w_\tau \in \mathbb{R}^M$, and where we have slightly abused notation to let y_τ, η_τ and ζ_τ denote their $\widehat{\mathcal{H}}$ counterparts. Here K is the matrix whose rows are of the form $K_{(i)} = \widehat{\Psi}(x_i) = \left[\, \widehat{\psi}_1(x_i) \; \widehat{\psi}_2(x_i) \;\cdots\; \widehat{\psi}_M(x_i) \,\right]$. In systems-theoretic language, each row of K corresponds to a *measurement* at a particular location, and the matrix itself acts as a measurement operator. We define the *generalized observability matrix* [20] as $\mathcal{O}_\Upsilon = \begin{bmatrix} K\widehat{A}^{\tau_1} \\ \cdots \\ K\widehat{A}^{\tau_L} \end{bmatrix}$ where $\Upsilon = \{\tau_1, \ldots, \tau_L\}$ are the set of instances τ_i when we apply the operator K. A linear system is said to be *observable* if $\mathcal{O}_\Upsilon$ has full column rank (i.e. $\mathrm{Rank}\mathcal{O}_\Upsilon = M$) for $\Upsilon = \{0, 1, \ldots, M-1\}$ [20]. Observability guarantees two critical facts: firstly, it guarantees that the state w_0 can be recovered exactly from a finite series of measurements $\{y_{\tau_1}, y_{\tau_2}, \ldots, y_{\tau_L}\}$; in particular, defining $y_\Upsilon = \left[y_{\tau_1}^T, y_{\tau_2}^T, \cdots, y_{\tau_L^T}\right]^T$, we have that $y_\Upsilon = \mathcal{O}_\Upsilon w_0$. Secondly, it guarantees that a feedback based *observer* can be designed such that the estimate of w_τ, denoted by $\widehat{w}_\tau$, converges exponentially fast to w_τ in the limit of samples. Note that all our theoretical results assume $\widehat{A}$ is available: while we perform system identification in the experiments (Section 3.3), it is not the focus of the paper.

We are now in a position to formally state the spatiotemporal modeling and inference problem considered: given a spatiotemporally evolving system modeled using (3), choose a set of N sensing locations such that even with $N \ll M$, the functional evolution of the spatiotemporal model can be estimated (which corresponds to *monitoring*) and can be predicted robustly (which corresponds to *Bayesian filtering*). Our approach to solve this problem relies on the design of the measurement operator K so that the pair $(K, \widehat{A})$ is observable: any Bayesian state estimator (e.g. a Kalman filter) utilizing this pair is denoted as a **kernel observer** [1]. We will leverage the spectral decomposition of $\widehat{A}$ for this task. (refer section 1 in supplementary for details on spectral decomposition)

2.2 Main Results

In this section, we prove results concerning the observability of spatiotemporally varying functions modeled by the functional evolution and measurement equations (3) formulated in Section 2.1. In

[1] In the case where no measurements are taken, for the sake of consistency, we denote the state estimator as an **autonomous kernel observer**, despite this being somewhat of an oxymoron.

particular, observability of the system states implies that we can recover the current state of the spatiotemporally varying function using a small number of sampling locations N, which allows us to 1) track the function, and 2) predict its evolution forward in time. We work with the approximation $\widehat{\mathcal{H}} \approx \mathcal{H}$: given M basis functions, this implies that the dual space of $\widehat{\mathcal{H}}$ is $\mathbb{R}^M$. Proposition 1 shows that if $\widehat{A}$ has a full-rank Jordan decomposition, the observation matrix K meeting a condition called *shadedness* (Definition 1) is sufficient for the system to be observable. Proposition 2 provides a lower bound on the number of sampling locations required for observability which holds for any $\widehat{A}$. Proposition 3 constructively shows the existence of an abstract measurement map $\widetilde{K}$ achieving this lower bound. Finally, since the measurement map does not have the structure of a kernel matrix, a slightly weaker sufficient condition for the observability of any $\widehat{A}$ is in Theorem 1. Proof of all claims are in the supplementary material.

Definition 1. ***(Shaded Observation Matrix)*** *Given $k : \Omega \times \Omega \to \mathbb{R}$ positive-definite on a domain Ω, let $\{\widehat{\psi}_1(x), \ldots, \widehat{\psi}_M(x)\}$ be the set of bases generating an approximate feature map $\widehat{\psi} : \Omega \to \widehat{\mathcal{H}}$, and let $\mathcal{X} = \{x_1, \ldots, x_N\}$, $x_i \in \Omega$. Let $K \in \mathbb{R}^{N \times M}$ be the observation matrix, where $K_{ij} := \widehat{\psi}_j(x_i)$. For each row $K_{(i)} := \left[\, \widehat{\psi}_1(x_i) \cdots \widehat{\psi}_M(x_i) \,\right]$, define the set $\mathcal{I}_{(i)} := \{\iota_1^{(i)}, \iota_2^{(i)}, \ldots, \iota_{M_i}^{(i)}\}$ to be the indices in the observation matrix row i which are nonzero. Then if $\bigcup_{i \in \{1,\ldots,N\}} \mathcal{I}^{(i)} = \{1, 2, \ldots, M\}$, we denote K as a* shaded observation matrix *(see Figure 2a).*

This definition seems quite abstract, so the following remark considers a more concrete example.

Remark 1. *let $\widehat{\psi}$ be generated by the dictionary given by $\mathcal{C} = \{c_1, \ldots, c_M\}$, $c_i \in \Omega$. Note that since $\widehat{\psi}_j(x_i) = \langle \psi(x_i), \psi(c_j) \rangle_{\mathcal{H}} = k(x_i, c_j)$, K is the kernel matrix between $\mathcal{X}$ and $\mathcal{C}$. For the kernel matrix to be shaded thus implies that there does not exist an atom $\psi(c_j)$ such that the projections $\langle \psi(x_i), \psi(c_j) \rangle_{\mathcal{H}}$ vanish for all x_i, $1 \leq i \leq N$. Intuitively, the shadedness property requires that the sensor locations x_i are privy to information propagating from every c_j. As an example, note that, in principle, for the Gaussian kernel, a single row generates a shaded kernel matrix*[2].

Proposition 1. *Given $k : \Omega \times \Omega \to \mathbb{R}$ positive-definite on a domain Ω, let $\{\widehat{\psi}_1(x), \ldots, \widehat{\psi}_M(x)\}$ be the set of bases generating an approximate feature map $\widehat{\psi} : \Omega \to \widehat{\mathcal{H}}$, and let $\mathcal{X} = \{x_1, \ldots, x_N\}$, $x_i \in \Omega$. Consider the discrete linear system on $\widehat{\mathcal{H}}$ given by the evolution and measurement equations (3). Suppose that a full-rank Jordan decomposition of $\widehat{A} \in \mathbb{R}^{M \times M}$ of the form $\widehat{A} = P \Lambda P^{-1}$ exists, where $\Lambda = \left[\, \Lambda_1 \cdots \Lambda_O \,\right]$, and there are no repeated eigenvalues. Then, given a set of time instances $\Upsilon = \{\tau_1, \tau_2, \ldots, \tau_L\}$, and a set of sampling locations $\mathcal{X} = \{x_1, \ldots, x_N\}$, the system (3) is observable if the observation matrix K_{ij} is shaded according to Definition 1, Υ has distinct values, and $|\Upsilon| \geq M$.*

When the eigenvalues of the system matrix are repeated, it is not enough for K to be shaded. In the next proposition, we take a geometric approach and utilize the rational canonical form of $\widehat{A}$ to obtain a lower bound on the number of sampling locations required. Let r be the number of unique eigenvalues of $\widehat{A}$, and let γ_{λ_i} denote the geometric multiplicity of eigenvalue λ_i. Then the *cyclic index* of $\widehat{A}$ is defined as $\ell = \max_{1 \leq i \leq r} \gamma_{\lambda_i}$[19] (see supplementary section 1 for details).

Proposition 2. *Suppose that the conditions in Proposition 1 hold, with the relaxation that the Jordan blocks $\left[\, \Lambda_1 \cdots \Lambda_O \,\right]$ may have repeated eigenvalues (i.e. $\exists \Lambda_i$ and Λ_j s.t. $\lambda_i = \lambda_j$). Then there exist kernels $k(x, y)$ such that the lower bound ℓ on the number of sampling locations N is given by the cyclic index of $\widehat{A}$.*

Section 2 in supplementary gives a concrete example to build intuition regarding this lower bound. We now show how to construct a matrix $\widetilde{K}$ corresponding to the lower bound ℓ.

Proposition 3. *Given the conditions stated in Proposition 2, it is possible to construct a measurement map $\widetilde{K} \in \mathbb{R}^{\ell \times M}$ for the system given by (3), such that the pair $(\widetilde{K}, \widehat{A})$ is observable.*

The construction provided in the proof of Proposition 3 is utilized in Algorithm 1, which uses the rational canonical structure of $\widehat{A}$ to generate a series of vectors $v_i \in \mathbb{R}^M$, whose iterations

[2] However, in this case, the matrix can have many entries that are extremely close to zero, and will probably be very ill-conditioned.

Algorithm 1 Measurement Map $\widetilde{K}$

```
Input: $\widehat{A} \in \mathbb{R}^{M \times M}$
Compute Rational canonical form, such that $C = Q^{-1}\widehat{A}^T Q$. Set $C_0 := C$, and $M_0 := M$.
for $i = 1$ to $\ell$ do
    Obtain MP $\alpha_i(\lambda)$ of $C_{i-1}$. This returns associated indices $\mathcal{J}^{(i)} \subset \{1, 2, \dots, M_{i-1}\}$.
    Construct vector $v_i \in \mathbb{R}^M$ such that $\xi_{v_i}(\lambda) = \alpha_i(\lambda)$ .
    Use indices $\{1, 2, \dots, M_{i-1}\} \setminus \mathcal{J}^{(i)}$ to select matrix $C_i$. Set $M_i := |\{1, 2, \dots, M_{i-1}\} \setminus \mathcal{J}^{(i)}|$
end for
Compute $\mathring{K} = [v_1^T, v_2^T, ..., v_\ell^T]^T$
Output: $\widetilde{K} = \mathring{K} Q^{-1}$
```

$\{v_1, \dots, \widehat{A}^{m_1-1}v_1, \dots, v_\ell, \dots, \widehat{A}^{m_\ell-1}v_\ell\}$ generate a basis for $\mathbb{R}^M$. Unfortunately, the measurement map $\widetilde{K}$, being an abstract construction unrelated to the kernel, does not directly select $\mathcal{X}$. We will show how to use the measurement map to guide a search for $\mathcal{X}$ in Remark 2. For now, we state a sufficient condition for observability of a general system.

Theorem 1. *Suppose that the conditions in Proposition 1 hold, with the relaxation that the Jordan blocks $[\Lambda_1 \quad \cdots \quad \Lambda_O]$ may have repeated eigenvalues. Let ℓ be the cyclic index of $\widehat{A}$. Define*

$$\mathbf{K} = [\, K^{(1)^T} \ \dots \ K^{(\ell)^T} \,]^T \tag{4}$$

*as the ℓ-*shaded matrix *which consists of ℓ shaded matrices with the property that any subset of ℓ columns in the matrix are linearly independent from each other. Then system (3) is observable if Υ has distinct values, and $|\Upsilon| \geq M$.*

While Theorem 1 is a quite general result, the condition that any ℓ columns of $\mathbf{K}$ be linearly independent is a very stringent condition. One scenario where this condition can be met with minimal measurements is in the case when the feature map $\widehat{\psi}(x)$ is generated by a dictionary of atoms with the Gaussian RBF kernel evaluated at sampling locations $\{x_1, \dots, x_N\}$ according to (2), where $x_i \in \Omega \subset \mathbb{R}^d$, and x_i are sampled from a non-degenerate probability distribution on Ω such as the uniform distribution. For a semi-deterministic approach, when the dynamics matrix $\widehat{A}$ is block-diagonal, we can utilize a simple heuristic:

Remark 2. *Let Ω be compact, $\mathcal{C} = \{c_1, \dots, c_M\}$, $c_i \in \Omega$, and let the approximate feature map be defined by (2). Consider the system (3) with $\widehat{A} = \Lambda$, and let $\Upsilon = \{0, 1, \dots, M-1\}$. Then the measurement map $\widetilde{K}$'s values lie in $\{0, 1\}$; in particular, each row $\widetilde{K}^{(j)}$, $j \in \{1, \dots, \ell\}$, corresponds to a subspace $\widehat{\mathcal{H}}_j$, generated by a subset of centers $\mathcal{C}^{(j)} \subset \mathcal{C}$. Generate samples $x_i^{(j)}$ to create a kernel matrix $K^{(j)}$ that is shaded only with respect to centers $\mathcal{C}^{(j)}$. Once this is done, move on to the next subspace $\widehat{\mathcal{H}}_{j+1}$. When all ℓ rows of $\widetilde{K}$ are accounted for, construct the matrix $\mathbf{K}$ as in (4). Then the resulting system $(\mathbf{K}, \widehat{A})$ is observable.*

This heuristic is formalized in Algorithm 2 in the supplementary. Note that in practice, the matrix $\widehat{A}$ needs to be inferred from measurements of the process f_τ. If no assumptions are placed on $\widehat{A}$, it's clear that at least M sensors are required for the system identification phase. Future work will study the precise conditions under which system identification is possible with less than M sensors.

3 Experimental Results

3.1 Sampling Locations for Synthetic Data Sets

The goal of this experiment is to investigate the dependency of the observability of system (3) on the shaded observation matrix and the lower bound presented in Proposition 2. The domain is fixed on the interval $\Omega = [0, 2\pi]$. First, we pick sets of points $\mathcal{C}^{(\iota)} = \{c_1, \dots, c_{M_\iota}\}$, $c_j \in \Omega$, $M = 50$, and construct a dynamics matrix $A = \Lambda \in \mathbb{R}^{M \times M}$, with cyclic index 5. We pick the RBF kernel $k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$, $\sigma = 0.02$. Generating samples $\mathcal{X} = \{x_1, \dots, x_N\}$, $x_i \in \Omega$ randomly, we compute the ℓ-shaded property and observability for this system. Figure 3a shows how shadedness is a necessary condition for observability, validating Proposition 1: the slight gap between shadedness and observability here can be explained due to numerical issues in computing the rank of $\mathcal{O}_\Upsilon$. Next, we again pick $M = 50$, but for a system with a cyclic index $\ell = 18$. We constructed the measurement map $\widetilde{K}$ using Algorithm 1, and the heuristic in Remark 2 (Algorithm 2 in the supplementary) as well as random sampling to generate the sampling locations $\mathcal{X}$. These results are presented in Figure 3b. The plot for random sampling has been averaged over 100 runs. It is evident from the plot that observability cannot be achieved for a number of samples $N < \ell$. Clearly, the heuristic presented outperforms random sampling; note however, that our intent is not to compare the heuristic against random sampling, but to show that the lower bound ℓ provides decisive guidelines for selecting the number of samples while using the computationally efficient random approach.

3.2 Comparison With Nonstationary Kernel Methods on Real-World Data

We use two real-world datasets to evaluate and compare the kernel observer with the two different lines of approach for non-stationary kernels discussed in Section 1.1. For the Process Convolution with Local Smoothing Kernel (PCLSK) and Latent Extension of Input Space (LEIS) approaches, we compare with NOSTILL-GP [4] and [11] respectively, on the Intel Berkeley and Irish Wind datasets.

Model inference for the kernel observer involved three steps: 1) picking the Gaussian RBF kernel $k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$, a search for the ideal σ is performed for a sparse Gaussian Process model (with a fixed basis vector set $\mathcal{C}$ selected using the method in [3]. For the data set discussed in this section, the number of basis vectors were equal to the number of sensing locations in the training set, with the domain for input set defined over $\mathbb{R}^2$; 2) having obtained σ, Gaussian process inference is used to generate weight vectors for each time-step in the training set, resulting in the sequence $w_\tau, \tau \in \{1, \ldots, T\}$; 3) matrix least-squares is applied to this sequence to infer $\widehat{A}$ (Algorithm 3 in the supplementary). For prediction in the autonomous setup, $\widehat{A}$ is used to propagate the state w_τ forward to make predictions with no feedback, and in the observer setup, a Kalman filter (Algorithm 4 in the supplementary) with N determined using Proposition 2, and locations picked randomly, is used to propagate w_τ forward to make predictions. We also compare with a baseline GP (denoted by 'original GP'), which is the sparse GP model trained using all of the available data.

Our first dataset, the Intel Berkeley research lab temperature data, consists of 50 wireless temperature sensors in indoor laboratory region spanning 40.5 meters in length and 31 meters in width[3]. Training data consists of temperature data on March 6th 2004 at intervals of 20 minutes (beginning 00:20 hrs) which totals to 72 timesteps. Testing is performed over another 72 timesteps beginning 12:20 hrs of the same day. Out of 50 locations, we uniformly selected 25 locations each for training and testing purposes. Results of the prediction error are shown in box-plot form in Figure 4a and as a time-series in Figure 4b, note that 'Auto' refers to autonomous set up. Here, the cyclic index of $\widehat{A}$ was determined to be 2, so N was set to 2 for the kernel observer with feedback. Note that here, even the autonomous kernel observer outperforms PCLSK and LEIS overall, and the kernel observer with feedback with $N = 2$ significantly outperforms all other methods, which is why we did not include results with $N > 2$.

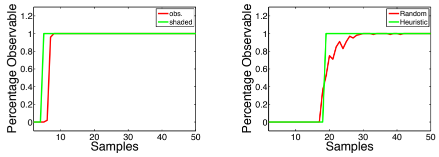
The second dataset is the Irish wind dataset, consisting of daily average wind speed data collected from year 1961 to 1978 at 12 meteorological stations in the Republic of Ireland[4]. The prediction error is in box-plot form in Figure 5a and as a time-series in Figure 5b. Again, the cyclic index of $\widehat{A}$ was determined to be 2. In this case, the autonomous kernel observer's performance is comparable to PCLSK and LEIS, while the kernel observer with feedback with $N = 2$ again outperforms all other methods. Section 3 in supplementary reports the total training and prediction times associated with PCLSK, LEIS, and the kernel observer. We observed that, 1) the kernel observer is an order of magnitude faster, and 2) even for small sets, competing methods did not scale well.

3.3 Prediction of Global Ocean Surface Temperature

We analyzed the feasibility of our approach on a very large dataset from the National Oceanographic Data Center: the 4 km AVHRR Pathfinder project, which is a satellite monitoring global ocean surface temperature (see figure 6a). This dataset is challenging, with measurements at over 37 million possible coordinates, but with only around 3-4 million measurements available per day, leading to a lot of missing data. The goal was to learn the day and night temperature models on data from the year 2011, and then to monitor thereafter for 2012. Success in monitoring would demonstrate two

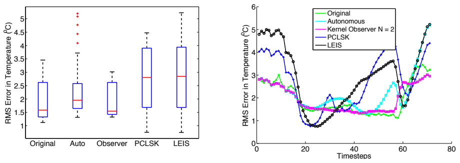
[3] `http://db.csail.mit.edu/labdata/labdata.html`

[4] `http://lib.stat.cmu.edu/datasets/wind.desc`



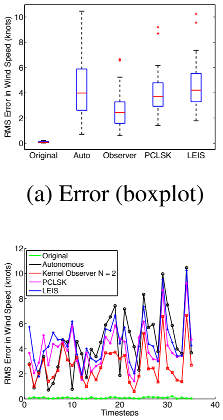
(a) Shaded vs. observability (b) Heuristic vs. random

Figure 3: Kernel observability results.



(a) Error (boxplot) (b) Error (time-series)

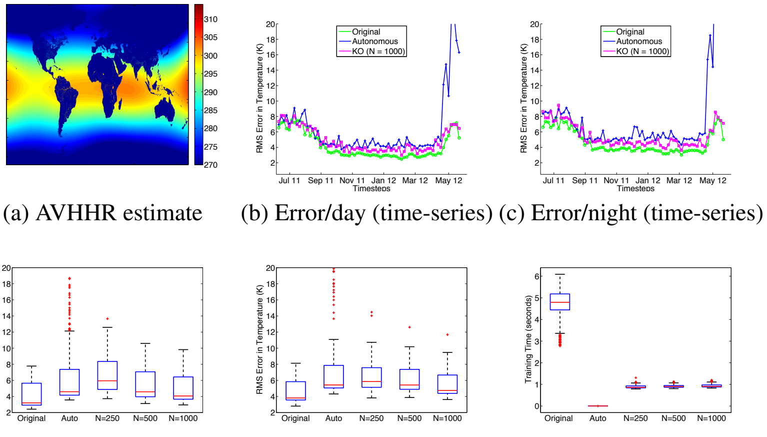
Figure 4: Comparison of kernel observer to PCLSK and LEIS methods on Intel Berkeley dataset.



(a) Error (boxplot)

(b) Error (boxplot)

Figure 5: Irish Wind



(a) AVHHR estimate (b) Error/day (time-series) (c) Error/night (time-series)

(d) Error (day) (e) Error (night) (f) Estimation time (day)

Figure 6: Performance of the kernel observer over AVVHR satellite 2012 data with different numbers of observation locations.

things: 1) the modeling process can capture spatiotemporal trends that generalize across years, and 2) the observer framework allows us to infer the state using a number of measurements that are an order of magnitude fewer than available. Note that due to the size of the dataset and the high computational requirements of the nonstationary kernel methods, a comparison with them was not pursued. To build the autonomous kernel observer and general kernel observer models, we followed the same procedure outlined in Section 3.2, but with $\mathcal{C} = \{c_1, \dots, c_M\}$, $c_j \in \mathbb{R}^2$, $|\mathcal{C}| = 300$. The Kalman filter for the general kernel observer model used $N \in \{250, 500, 1000\}$ at random locations to track the system state given a random initial condition w_0. As a fair baseline, the observers are compared to training a sparse GP model (labeled 'original') on approximately $400,000$ measurements per day. Figures 6b and 6c compare the autonomous and feedback approach with $1,000$ samples to the baseline GP; here, it can be seen that the autonomous does well in the beginning, but then incurs an unacceptable amount of error when the time series goes into 2012, i.e. where the model has not seen any training data, whereas KO does well throughout. Figures 6d and 6e show a comparison of the RMS error of estimated values from the real data. This figure shows the trend of the observer getting better and better state estimates as a function of the number of sensing locations N [5]. Time required for kernel observer is much lesser than retraining the model every time step, see figure 6f.

4 Conclusions

This paper presented a new approach to the problem of monitoring complex spatiotemporally evolving phenomena with limited sensors. Unlike most Neural Network or Kernel based models, the presented approach inherently incorporates differential constraints on the spatiotemporal evolution of the mixing weights of a kernel model. In addition to providing an elegant and efficient model, the main benefit of the inclusion of the differential constraint in the model synthesis is that it allowed the derivation of fundamental results concerning the minimum number of sampling locations required and the

[5] Note that we checked the performance of training a GP with only $1,000$ samples as a control, but the average error was about 10 Kelvins, i.e. much worse than KO.

identification of correlations in the spatiotemporal evolution by building upon the rich literature in systems theory. These results are non-conservative, and as such provide direct guidance in ensuring robust real-world predictive inference with distributed sensor networks.

Acknowledgment: This work was supported by AFOSR grant #FA9550-15-1-0146.

References

[1] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[2] Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2011.

[3] Lehel Csatö and Manfred Opper. Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668, 2002.

[4] Sahil Garg, Amarjeet Singh, and Fabio Ramos. Learning non-stationary space-time models for environmental monitoring. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.*, 2012.

[5] David Higdon. A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5(2):173–190, 1998.

[6] Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel Methods on Riemannian Manifolds with Gaussian RBF Kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[7] Chunsheng Ma. Nonstationary covariance functions that model space–time interactions. *Statistics & Probability Letters*, 61(4):411–419, 2003.

[8] Kanti V Mardia, Colin Goodall, Edwin J Redfern, and Francisco J Alonso. The kriged kalman filter. *Test*, 7(2):217–282, 1998.

[9] C Paciorek and M Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in neural information processing systems*, 16:273–280, 2004.

[10] Fernando Pérez-Cruz, Steven Van Vaerenbergh, Juan José Murillo-Fuentes, Miguel Lázaro-Gredilla, and Ignacio Santamaria. Gaussian processes for nonlinear signal processing: An overview of recent advances. *Signal Processing Magazine, IEEE*, 30(4):40–50, 2013.

[11] Tobias Pfingsten, Malte Kuss, and Carl Edward Rasmussen. Nonstationary gaussian process regression using a latent extension of the input space. *URL http://www. kyb. mpg. de/˜ tpfingst*, 2006.

[12] Christian Plagemann, Kristian Kersting, and Wolfram Burgard. Nonstationary gaussian process regression using point estimates of local smoothness. In *Machine learning and knowledge discovery in databases*, pages 204–219. Springer, 2008.

[13] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184, 2007.

[14] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, December 2005.

[15] Alexandra M Schmidt and Anthony O'Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758, 2003.

[16] Amarjeet Singh, Fabio Ramos, H Durrant-Whyte, and William J Kaiser. Modeling and decision making in spatio-temporal processes for environmental surveillance. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 5490–5497. IEEE, 2010.

[17] Christopher K Wikle. A kernel-based spectral model for non-gaussian spatio-temporal processes. *Statistical Modelling*, 2(4):299–314, 2002.

[18] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *NIPS*, pages 682–688, 2001.

[19] W Murray Wonham. *Linear multivariable control*. Springer, 1974.

[20] Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, NJ, 1996.

Kernel Observers: Supplementary Material

Hassan Kingravi
Pindrop Security
Atlanta, GA 30308
hkingravi@pindropsecurity.com

Harshal Maske and Girish Chowdhary
Department of Agriculture and Biological Engineering
University of Illinois at Urbana Champaign
Urbana, IL 61801
hmaske2@illinois.edu, girishc@illinois.edu

1 Preliminaries on Rational Canonical Structure

We take a geometric approach towards the choice of sampling locations for inferring w_τ in (3). To do so, we utilize the Frobenius canonical form and Jordan canonical decomposition of $\mathcal{A}$ [2]. We use the notation $\mathcal{V}$, with $\dim(\mathcal{V}) = M$, to emphasize the fact that these theorems hold for any finite-dimensional vector space. The linear operator $\mathcal{A} : \mathcal{V} \to \mathcal{V}$ has a characteristic polynomial $\pi(\lambda)$ such that $\pi(\mathcal{A}) = 0$ by the Cayley-Hamilton theorem. The minimal polynomial (MP) of $\mathcal{A}$ is the monic polynomial $\alpha(\lambda)$ of least degree (denoted by $\deg(\cdot)$) such that $\alpha(\lambda) = a_0 + a_1\lambda + \cdots + \lambda^{\deg(\alpha)} = 0$, and $\alpha(\mathcal{A}) = a_0 I + a_1\mathcal{A} + \cdots + \mathcal{A}^{\deg(\alpha)} = 0$. The MP is unique and divides $\pi(\lambda)$, so that $\deg(\alpha) \le \deg(\pi)$. The MP of a vector $v \in \mathcal{V}$ *relative to* $\mathcal{A}$ is the unique monic polynomial ξ_v of least degree such that $\xi_v(\mathcal{A})v = a_0 v + a_1\mathcal{A}v + \cdots + \mathcal{A}^{\deg(\alpha)}v = 0$. If $\deg(\alpha) = M$, then $\mathcal{A}$ is said to be *cyclic* and there exists $v \in \mathcal{V}$, such that the vectors $\{v, \mathcal{A}v, \ldots, \mathcal{A}^{M-1}v\}$ form a basis for $\mathcal{V}$; this is the same as saying that the pair $(v^T, \mathcal{A}^T)$ is observable.

A subspace $\mathcal{V}_\mathcal{S} \subset \mathcal{V}$ s.t. $\mathcal{A}\mathcal{V}_\mathcal{S} \subset \mathcal{V}_\mathcal{S}$ is $\mathcal{A}$-*cyclic* if $\mathcal{A}_{|\mathcal{V}_\mathcal{S}}$, the restriction of $\mathcal{A}$ to the subspace $\mathcal{V}_\mathcal{S}$, is cyclic. If $\alpha(\lambda)$ is the minimal polynomial of $\mathcal{A}$ and $\deg(\alpha) = m < M$, $\exists\ v \in \mathcal{V}$ such that $\{v, \mathcal{A}v, \ldots, \mathcal{A}^{m-1}v\}$ span an m-dimensional $\mathcal{A}$-cyclic subspace $\mathcal{V}_\mathcal{S}$, with v being the *cyclic generator* of $\mathcal{V}_\mathcal{S}$. The subspace $\mathcal{V}_\mathcal{S}$ decomposes $\mathcal{V}$ relative to $\mathcal{A}$. By the rational (or Frobenius) canonical structure theorem, $\mathcal{A}$ can be successively decomposed into subspaces $\mathcal{V}_i \subset \mathcal{V}$, $i \in \{1, \ldots, \ell\}$, s.t. $\mathcal{V} = \mathcal{V}_1 \oplus ... \oplus \mathcal{V}_\ell$, $\mathcal{A}\mathcal{V}_i \subset \mathcal{V}_i$, and $\mathcal{A}_{|\mathcal{V}_i}, i \in \{1, \ldots, \ell\}$, are cyclic[1]. The integer ℓ is unique and is called the *cyclic index of* $\mathcal{A}$. One of our main results is to show that the cyclic index is a lower bound on the number of measurements required to reconstruct w_τ (see Proposition 3 and Algorithm 1).

Recall also that for any matrix $\mathcal{A} \in \mathbb{R}^{M\times M}$, $\exists P \in \mathbb{R}^{M\times M}$ invertible such that $\mathcal{A} = P\Lambda P^{-1}$, where Λ is a unique block diagonal matrix with Jordan blocks with λ_i along the diagonal. If all the eigenvalues λ_i are nonzero and real, we say the matrix has a *full-rank Jordan decomposition*.

2 Discussion of Theoretical Results

The systems-theoretic approach taken in this paper reveals something rather surprising: functions with complex dynamics (with a small cyclic index) can be recovered with less sensor placements than functions with simpler dynamics. Although seemingly counterintuitive, it becomes clear that this is because complex dynamics, which are characterized by a lower geometric multiplicity of the

[1] In general, the subspaces $\mathcal{V}_i$ are not unique for a fixed $\mathcal{A}$.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

eigenvalues, ensure that the orbit $\boldsymbol{\Theta} := \{\widehat{A}w_\tau\}_{\tau\in\Upsilon}$ traverses a greater portion of $\mathbb{R}^M \equiv \widehat{\mathcal{H}}$ and thus that fewer sensors can recover more geometric information. On the other hand, in 'simpler' functional evolution, $\boldsymbol{\Theta}$ evolves along strict subspaces of $\mathbb{R}^M$, and so more independent sensors are required to infer the same amount of information. Recall that cyclic index $\ell > 1$ implies that there exist spaces $\mathcal{V}_i$ s.t. $\mathbb{R}^M = \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_\ell$, which induces the decomposition $\widehat{\mathcal{H}} = \widehat{\mathcal{H}}_1 \oplus \cdots \oplus \widehat{\mathcal{H}}_\ell$. As the simplest nontrivial case, consider $\widehat{\psi}(x)$ defined as in (2), where $\mathcal{C} = \{c_1, c_2\}$, $c_i \in \Omega$, and pick one sensor location $x_1 \in \Omega$. Let (3) be given by $\widehat{A} = \left[\begin{smallmatrix} \lambda & 0 \\ 0 & \lambda \end{smallmatrix}\right]$, $\lambda \in \mathbb{R}$ and $|\lambda| < 1$, and let the system be deterministic (i.e. $\eta_\tau, \zeta_\tau = 0$). Here, $\ell = 2$, because there exists no $v \in \mathbb{R}^2$ s.t. $\mathrm{span}\{v, \widehat{A}v\} = \mathbb{R}^2$. For any initial condition w_0 we get a discrete sequence $\{w_0, \lambda w_0, \lambda^2 w_0, \dots\}$ going to zero along the 1-dimensional subspace generated by w_0 (i.e. w_0 is an eigenvector of $\widehat{A}$). Let the set of time instances Υ be given by $\Upsilon = \{0, 1\}$, and consider a shaded matrix $K = \left[\begin{smallmatrix} k_{11} & k_{12} \end{smallmatrix}\right]$: then the observability matrix is given by $\mathcal{O}_\Upsilon = \left[\begin{smallmatrix} K^T & (K\widehat{A})^T \end{smallmatrix}\right]^T = \left[\begin{smallmatrix} k_{11} & k_{12} \\ \lambda k_{11} & \lambda k_{12} \end{smallmatrix}\right] = \left[\begin{smallmatrix} \mathbf{k}^T \\ \lambda\mathbf{k}^T \end{smallmatrix}\right]$, which is obviously rank-deficient, and where $\left[\begin{smallmatrix} k_{11} & k_{12} \end{smallmatrix}\right]^T := \mathbf{k}$. Intuitively, we have that $\mathcal{O}_\Upsilon w_0 = \left[\begin{smallmatrix} \langle \mathbf{k}, w_0\rangle_{\mathbb{R}^2} \\ \lambda\langle \mathbf{k}, w_0\rangle_{\mathbb{R}^2} \end{smallmatrix}\right]$, which implies that $\mathbf{k}$ doesn't have enough geometric information to recover the initial state. Contrast this with the case when $\widehat{A} = \left[\begin{smallmatrix} \lambda & 1 \\ 0 & \lambda \end{smallmatrix}\right]$; here, $\ell = 1$, and $\mathcal{O}_\Upsilon = \left[\begin{smallmatrix} k_{11} & k_{12} \\ \lambda k_{11} & k_{11}+\lambda k_{12} \end{smallmatrix}\right]$, which is full rank for a shaded kernel matrix K, and hence leads to observability. This fundamental insight is be gained from considering dynamical evolutions in the structure of the model.

Another point to note is that since the collection of bases $\{\widehat{\psi}_i(x)\}_{i=1}^M$ determines the richness of the function space $\widehat{\mathcal{H}} \approx \mathcal{H}$ we operate in, it determines the fidelity of the model approximation to the true time-varying function. As a consequence, observability of the system in $\widehat{\mathcal{H}}$ refers to the best possible approximation in $\widehat{\mathcal{H}}$. The greater the number of bases, the higher the dimensionality, which results in greater model fidelity, but which may require a much greater number of measurements for state recovery. This is where the lower bounds presented in the paper are particularly useful, because they show that for functional evolutions corresponding to certain $\widehat{A}$, *the number of sensor placements are essentially independent of the dimensionality M*, but depend rather on the cyclic index of $\widehat{A}$.

3 Training and prediction times for section 3.2 of main Article

Table 1: Total training and prediction times for Figs. 4 and 5

	Intel Berkeley	**Irish Wind**
Data Size (*bases-timesteps*)	25-72	12-36
Kernel Observer PCLSK LEIS	2.1 sec 121.4 sec 43.8 sec	0.1 sec 7.0 sec 2.8 sec

4 Proofs of Main Theorems

Definition 1. ***(Shaded Observation Matrix)*** *Given $k : \Omega \times \Omega \to \mathbb{R}$ positive-definite on a domain Ω, let $\{\widehat{\psi}_1(x), \dots, \widehat{\psi}_M(x)\}$ be the set of bases generating an approximate feature map $\widehat{\psi} : \Omega \to \widehat{\mathcal{H}}$, and let $\mathcal{X} = \{x_1, \dots, x_N\}$, $x_i \in \Omega$. Let $K \in \mathbb{R}^{N\times M}$ be the observation matrix, where $K_{ij} := \widehat{\psi}_j(x_i)$. For each row $K_{(i)} := \left[\begin{smallmatrix} \widehat{\psi}_1(x_i) & \cdots & \widehat{\psi}_M(x_i) \end{smallmatrix}\right]$, define the set $\mathcal{I}_{(i)} := \{\iota_1^{(i)}, \iota_2^{(i)}, \dots, \iota_{M_i}^{(i)}\}$ to be the indices in the observation matrix row i which are nonzero. Then if $\bigcup_{i\in\{1,\dots,N\}} \mathcal{I}^{(i)} = \{1, 2, \dots, M\}$, we denote K as a* shaded observation matrix *(see Figure 2a).*

This definition seems quite abstract, so the following remark considers a more concrete example.

Remark 1. *let $\widehat{\psi}$ be generated by the dictionary given by $\mathcal{C} = \{c_1, \dots, c_M\}$, $c_i \in \Omega$. Note that since $\widehat{\psi}_j(x_i) = \langle \psi(x_i), \psi(c_j)\rangle_{\mathcal{H}} = k(x_i, c_j)$, K is the kernel matrix between $\mathcal{X}$ and $\mathcal{C}$. For the kernel* *matrix to be shaded thus implies that there does not exist an atom $\psi(c_j)$ such that the projections $\langle \psi(x_i), \psi(c_j) \rangle_{\mathcal{H}}$ vanish for all x_i, $1 \le i \le N$. Intuitively, the shadedness property requires that the sensor locations x_i are privy to information propagating from every c_j. As an example, note that, in principle, for the Gaussian kernel, a single row generates a shaded kernel matrix*[2].

Proposition 1. *Given $k : \Omega \times \Omega \to \mathbb{R}$ positive-definite on a domain Ω, let $\{\widehat{\psi}_1(x), \dots, \widehat{\psi}_M(x)\}$ be the set of bases generating an approximate feature map $\widehat{\psi} : \Omega \to \widehat{\mathcal{H}}$, and let $\mathcal{X} = \{x_1, \dots, x_N\}$, $x_i \in \Omega$. Consider the discrete linear system on $\widehat{\mathcal{H}}$ given by the evolution and measurement equations (3). Suppose that a full-rank Jordan decomposition of $\widehat{A} \in \mathbb{R}^{M \times M}$ of the form $\widehat{A} = P \Lambda P^{-1}$ exists, where $\Lambda = \left[\begin{smallmatrix} \Lambda_1 & \cdots & \Lambda_O \end{smallmatrix}\right]$, and there are no repeated eigenvalues. Then, given a set of time instances $\Upsilon = \{\tau_1, \tau_2, \dots, \tau_L\}$, and a set of sampling locations $\mathcal{X} = \{x_1, \dots, x_N\}$, the system (3) is observable if the observation matrix K_{ij} is shaded according to Definition 1, Υ has distinct values, and $|\Upsilon| \ge M$.*

Proof. To begin, consider a system where $\widehat{A} = \Lambda$, with Jordan blocks $\{\Lambda_1, \Lambda_2, \dots, \Lambda_O\}$ along the diagonal. Then $\widehat{A}^{\tau_i} = \operatorname{diag}([\Lambda_1^{\tau_i} \quad \Lambda_2^{\tau_i} \quad \cdots \quad \Lambda_O^{\tau_i}])$. We have that

$$\begin{aligned}
\mathcal{O}_\Upsilon &= \begin{bmatrix} K\widehat{A}^{\tau_1} \\ \cdots \\ K\widehat{A}^{\tau_L} \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} K & \cdots & K \end{bmatrix}}_{\widehat{\mathbf{K}} \in \mathbb{R}^{N \times ML}} \underbrace{\left[\begin{array}{cccc}
\Lambda_1^{\tau_1} & 0 & \cdots & 0 \\
0 & \Lambda_2^{\tau_1} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \Lambda_O^{\tau_1} \\ \hline
\vdots & \vdots & \ddots & \vdots \\ \hline
\Lambda_1^{\tau_L} & 0 & \cdots & 0 \\
0 & \Lambda_2^{\tau_L} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \Lambda_O^{\tau_L}
\end{array}\right]}_{\widehat{\mathbf{A}} \in \mathbb{R}^{ML \times M}}.
\end{aligned}$$

Recall that a matrix's rank is preserved under a product with an invertible matrix. Design a matrix $U \in \mathbb{R}^{N \times N}$ s.t. $\breve{K} := UK$ is a matrix with one row vector of nonzeros, and all of the remaining rows as zeros. Then $\operatorname{rank}(\widehat{\mathbf{K}}\widehat{\mathbf{A}}) = \operatorname{rank}(U\widehat{\mathbf{K}}\widehat{\mathbf{A}})$. Therefore, we have that

$$\begin{aligned}
\breve{K}\widehat{A}^{\tau_j} &= \begin{bmatrix} \breve{K}_{(1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \widehat{A}^{\tau_j} \\
&= \begin{bmatrix}
k_{11}\lambda_1^{\tau_j} & \binom{\tau_j}{1}\lambda_1^{\tau_j - 1} + k_{12}\lambda_1^{\tau_j} & \cdots & k_{1M}\lambda_O^{\tau_j} \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & 0 \\
0 & 0 & \cdots & 0
\end{bmatrix}.
\end{aligned}$$

Therefore, following some more elementary row operations encoded by $V \in \mathbb{R}^{ML \times ML}$, we get that

$$\begin{aligned}
V \begin{bmatrix} \breve{K} & \cdots & \breve{K} \end{bmatrix} \begin{bmatrix} \widehat{A}^{\tau_1} \\ \vdots \\ \widehat{A}^{\tau_L} \end{bmatrix} &= \begin{bmatrix}
\tilde{k}_{11}\lambda_1^{\tau_1} & \cdots & \tilde{k}_{1M}\lambda_O^{\tau_1} \\
\tilde{k}_{11}\lambda_1^{\tau_2} & \cdots & \tilde{k}_{1M}\lambda_O^{\tau_2} \\
\vdots & \ddots & 0 \\
\tilde{k}_{11}\lambda_1^{\tau_L} & \cdots & \tilde{k}_{1M}\lambda_O^{\tau_L} \\
\mathbf{0} & \cdots & \mathbf{0}
\end{bmatrix} \\
&= \begin{bmatrix} \mathbf{\Phi} \\ \widehat{\mathbf{0}} \end{bmatrix}.
\end{aligned}$$

[2] However, in this case, the matrix can have many entries that are extremely close to zero, and will probably be very ill-conditioned.

If the individual entries $\tilde{k}_{1i}$ are nonzero, and the Jordan block diagonals have nonzero eigenvalues, the columns of $\mathbf{\Phi}$ become linearly independent. Therefore, if $L \geq M$, the column rank of $\mathcal{O}_\Upsilon$ is M, which results in an observable system.

To extend this proof to matrices $\widehat{A} = P\Lambda P^{-1}$, note that

$$\begin{aligned}\mathcal{O}_\Upsilon &= \begin{bmatrix} K\widehat{A}^{\tau_1} \\ \cdots \\ K\widehat{A}^{\tau_L} \end{bmatrix} \\ &= \begin{bmatrix} KP\Lambda^{\tau_1}P^{-1} \\ \cdots \\ KP\Lambda^{\tau_L}P^{-1}. \end{bmatrix} \\ &= \begin{bmatrix} K & \cdots & K \end{bmatrix} \boldsymbol{P}\boldsymbol{\Lambda}^t\boldsymbol{P}^{-1},\end{aligned}$$

where $\boldsymbol{P} \in \mathbb{R}^{ML\times ML}$, $\boldsymbol{\Lambda}^t \in \mathbb{R}^{ML\times ML}$, and $\boldsymbol{P^{-1}} \in \mathbb{R}^{ML\times ML}$ are the block diagonal matrices associated with the system. Since $\boldsymbol{P}$ is an invertible matrix, the conclusions about the column rank drawn before still hold, and the system is observable. □

When the eigenvalues of the system matrix are repeated, it is not enough for K to be shaded. The next proposition proves a lower bound on the number of sampling locations required.

Proposition 2. *Suppose that the conditions in Proposition 1 hold, with the relaxation that the Jordan blocks* $\left[\, \Lambda_1 \cdots \Lambda_O \,\right]$ *may have repeated eigenvalues (i.e.* $\exists \Lambda_i$ *and* Λ_j *s.t.* $\lambda_i = \lambda_j$*). Then there exist kernels* $k(x, y)$ *such that the lower bound* ℓ *on the number of sampling locations* N *is given by the cyclic index of* $\widehat{A}$*.*

Proof. We first prove the lower bound. WLOG, let $\mathbf{K}$ have $\ell - 1$ fully shaded, linearly independent rows, and write it as

$$\mathbf{K} = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1M} \\ \vdots & \vdots & \cdots & \vdots \\ k_{(\ell-1)1} & k_{(\ell-1)2} & \cdots & k_{(\ell-1)M} \end{bmatrix}.$$

Since the cyclic index is ℓ, this implies that at least one eigenvalue, say λ, has ℓ Jordan blocks. Define indices $j_1, j_2, \ldots, j_\ell \in \{1, 2, \ldots, M\}$ as the columns corresponding to the leading entries of the ℓ Jordan blocks corresponding to λ. WLOG, let $j_1 = 1$. Using ideas similar to the last proof, we can write the observability matrix as

$$\mathcal{O}_\Upsilon := \begin{bmatrix} k_{11}\lambda^{\tau_1} & \cdots & k_{1j_\ell}\lambda^{\tau_1} & \cdots \\ \vdots & \ddots & \vdots & \ddots \\ k_{11}\lambda^{\tau_L} & k_{1j_\ell}\lambda^{\tau_L} & \cdots & \\ \vdots & \ddots & \vdots & \ddots \\ k_{(\ell-1)1}\lambda^{\tau_1}\cdots & k_{(\ell-1)j_\ell}\lambda^{\tau_1} & \cdots & \\ \vdots & \ddots & \vdots & \ddots \\ k_{(\ell-1)1}\lambda^{\tau_L} & \cdots & k_{(\ell-1)j_\ell}\lambda^{\tau_L} & \cdots \end{bmatrix}.$$

Define $\boldsymbol{\lambda} := [\lambda^{\tau_1} \quad \lambda^{\tau_2} \quad \cdots \lambda^{\tau_L}]^T$. Then the above matrix becomes

$$\mathcal{O}_\Upsilon := \begin{bmatrix} k_{11}\boldsymbol{\lambda} & \cdots & k_{1j_2}\boldsymbol{\lambda} & \cdots & k_{1j_\ell}\boldsymbol{\lambda} & \cdots \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\ k_{(\ell-1)1}\boldsymbol{\lambda} & \cdots & k_{(\ell-1)j_2}\boldsymbol{\lambda} & \cdots & k_{(\ell-1)j_\ell}\boldsymbol{\lambda} & \cdots \end{bmatrix}.$$

We need to show that one of the columns above can be written in terms of the others. This is equivalent to solving the linear system

$$\begin{bmatrix} k_{1j_1} \\ k_{2j_1} \\ \vdots \\ k_{(\ell-1)j_1} \end{bmatrix} = \begin{bmatrix} k_{1j_2} & \cdots & k_{1j_\ell} \\ k_{2j_2} & \cdots & k_{2j_\ell} \\ \vdots & \ddots & \vdots \\ k_{(\ell-1)j_2} & \cdots & k_{(\ell-1)j_\ell} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{(\ell-1)} \end{bmatrix}.$$

Suppose the kernel matrix on the RHS is generated from the Gaussian kernel. From [1], it's known that every principal minor of a Gaussian kernel matrix is invertible, which implies that $\mathcal{O}_\Upsilon$ cannot be observable. □

Section 2 gives a concrete example to build intuition regarding this lower bound. We now show how to construct a matrix $\widetilde{K}$ corresponding to the lower bound ℓ.

Proposition 3. *Given the conditions stated in Proposition 2, it is possible to construct a measurement map $\widetilde{K} \in \mathbb{R}^{\ell \times M}$ for the system given by (3), such that the pair $(\widetilde{K}, \widehat{A})$ is observable.*

Proof. The construction of the measurement map $\widetilde{K}$ is based on the rational canonical structure of $\widehat{A}^T$ (discussed in section 1), which decomposes $\mathcal{V}$ into $\widehat{A}^T$-cyclic direct summands such that $\mathcal{V} = \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_\ell$, where ℓ is the cyclic index of $\widehat{A}$ as defined in Proposition 2. Let ξ_v be the minimal polynomial (m.p.) of v (relative to $\widehat{A}^T$): it is then the unique monic polynomial of least degree such that $\xi_v(\widehat{A}^T)v = 0$. Let $\alpha_1(\lambda)$ be the m.p. of $\widehat{A}^T_{|\mathcal{V}_1}$: then $\deg(\alpha_1(\lambda)) < M$. By the rational canonical structure theorem [2], there exists a vector $\widehat{v}_1$, such that $\xi_{v_1}(\lambda) = \alpha_1(\lambda)$. Similarly there exists a vector $\widehat{v}_2$, such that $\xi_{v_2}(\lambda) = \alpha_2(\lambda)$, where $\alpha_2(\lambda)$, is the minimal polynomial of $\widehat{A}^T_{|\mathcal{V}_2}$ and so on. Thus we can obtain ℓ such vectors that form the measurement map $\widetilde{K} = [\widehat{v}_1, \widehat{v}_2, \cdots, \widehat{v}_\ell]^T$. Construction of these vectors $\widehat{v}_i$, can be simplified by first performing the Jordan decomposition as $\widehat{A}^T = P\Lambda P^{-1}$. Then the vectors $\widetilde{v}_i$, $i \in \ell$ for Λ, can be constructed such that the entries corresponding to the leading entries of Jordan blocks of $\Lambda_{|\mathcal{V}_i}$ are nonzero. Such a construction ensures that the m.p. of vector $\widetilde{v}_i$ w.r.t $\Lambda_{|\mathcal{V}_i}$, is also the corresponding m.p. of $\Lambda_{|\mathcal{V}_i}$. Hence the required map is given by $\widetilde{K} = [\widetilde{v}_1, \widetilde{v}_2, \ldots, \widetilde{v}_\ell]^T P^{-1}$. □

The construction provided in the proof of Proposition 3 is utilized in Algorithm 1, which uses the rational canonical structure of $\widehat{A}$ to generate a series of vectors $v_i \in \mathbb{R}^M$, whose iterations $\{v_1, \ldots, \widehat{A}^{m_1-1}v_1, \ldots, v_\ell, \ldots, \widehat{A}^{m_\ell-1}v_\ell\}$ generate a basis for $\mathbb{R}^M$ (see Section 1). Unfortunately, the measurement map $\widetilde{K}$, being an abstract construction unrelated to the kernel, does not directly select $\mathcal{X}$. We will show how to use the measurement map to guide a search for $\mathcal{X}$ in Remark 2. For now, we state a sufficient condition for observability of a general system.

Theorem 1. *Suppose that the conditions in Proposition 1 hold, with the relaxation that the Jordan blocks $[\Lambda_1 \quad \Lambda_2 \quad \cdots \quad \Lambda_O]$ may have repeated eigenvalues. Let ℓ be the cyclic index of $\widehat{A}$. We define*

$$\mathbf{K} = \left[\, {\scriptstyle K^{(1)^T} \ \ldots \ K^{(\ell)^T}} \,\right]^T \tag{1}$$

*as the ℓ-*shaded matrix *which consists of ℓ shaded matrices with the property that any subset of ℓ columns in the matrix are linearly independent from each other. Then system (3) is observable if Υ has distinct values, and $|\Upsilon| \geq M$.*

Proof. A cyclic index of ℓ for this system implies that there exists an eigenvalue λ that's repeated ℓ times. We prove the theorem for repeated eigenvalues of dimension 1: the same statement can be proven for repeated eigenvalues for Jordan blocks using the ideas in the proof of Proposition 1. WLOG, let $\mathbf{K}$ have ℓ fully shaded, linearly independent rows, and, assume that the column indices corresponding to this eigenvalue are $\{1, 2, \ldots, \ell\}$. Define $\boldsymbol{\lambda}_i := [\lambda_i^{\tau_1} \quad \lambda_i^{\tau_2} \quad \cdots \lambda_i^{\tau_L}]^T$. Then

$$\mathcal{O}_\Upsilon := \begin{bmatrix} k_{11}\boldsymbol{\lambda}_1 & k_{12}\boldsymbol{\lambda}_2 & \cdots & k_{1M}\boldsymbol{\lambda}_M \\ \vdots & \vdots & \ddots & \vdots \\ k_{\ell 1}\boldsymbol{\lambda}_1 & k_{\ell 2}\boldsymbol{\lambda}_2 & \cdots & k_{\ell M}\boldsymbol{\lambda}_M \end{bmatrix}.$$

Let $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \cdots \boldsymbol{\lambda}_\ell := \boldsymbol{\lambda}$. Focusing on these first ℓ columns of this matrix, this implies that we need to find constants $c_1, c_2, \ldots, c_{\ell-1}$ such that

$$\begin{bmatrix} k_{11} \\ \vdots \\ k_{\ell 1} \end{bmatrix} = c_1 \begin{bmatrix} k_{12} \\ \vdots \\ k_{\ell 2} \end{bmatrix} + \cdots + c_{\ell-1} \begin{bmatrix} k_{1\ell} \\ \vdots \\ k_{\ell\ell} \end{bmatrix}.$$

Algorithm 1 Measurement Map $\widetilde{K}$

```
Input: $\widehat{A} \in \mathbb{R}^{M \times M}$
Compute Frobenius canonical form, such that $C = Q^{-1}\widehat{A}^T Q$. Set $C_0 := C$, and $M_0 := M$.
for $i = 1$ to $\ell$ do
    Obtain MP $\alpha_i(\lambda)$ of $C_{i-1}$. This returns associated indices $\mathcal{J}^{(i)} \subset \{1, 2, \ldots, M_{i-1}\}$.
    Construct vector $v_i \in \mathbb{R}^M$ such that $\xi_{v_i}(\lambda) = \alpha_i(\lambda)$ .
    Use indices $\{1, 2, \ldots, M_{i-1}\} \setminus \mathcal{J}^{(i)}$ to select matrix $C_i$. Set $M_i := |\{1, 2, \ldots, M_{i-1}\} \setminus \mathcal{J}^{(i)}|$
end for
Compute $\mathring{K} = [v_1^T, v_2^T, ..., v_\ell^T]^T$
Output: $\widetilde{K} = \mathring{K}Q^{-1}$
```

Algorithm 2 Sampling locations set $\mathcal{X}$

```
Input: $\widehat{A} = C$, lower bound $\ell$
Decompose $C$ to generate invariant subspaces $\widehat{\mathcal{H}}_j$, $j \in \{1, 2, \ldots, \ell\}$ (see section 1)
for $j = 1$ to $\ell$ do
    Obtain centers $\mathcal{C}^{(j)}$ w.r.t subspace $\widehat{\mathcal{H}}_j$,
    Generate samples $x_i^{(j)}$ to create a kernel matrix $K^{(j)}$ that is shaded only with respect to centers $\mathcal{C}^{(j)}$
end for
Output: Sampling locations set $\mathcal{X} = \{x^{(1)}, x^{(2)} \cdots, x^{(l)}\}$.
```

However, these columns are linearly independent by assumption, and thus no such constants exist, implying that $\mathcal{O}_\Upsilon$ is observable. □

While Theorem 1 is a quite general result, the condition that any ℓ columns of $\mathbf{K}$ be linearly independent is a very stringent condition. One scenario where this condition can be met with minimal measurements is in the case when the feature map $\widehat{\psi}(x)$ is generated by a dictionary of atoms with the Gaussian RBF kernel evaluated at sampling locations $\{x_1, \ldots, x_N\}$ according to (2), where $x_i \in \Omega \subset \mathbb{R}^d$, and x_i are sampled from a non-degenerate probability distribution on Ω such as the uniform distribution. For a semi-deterministic approach, when the dynamics matrix $\widehat{A}$ is block-diagonal, we can utilize a simple heuristic:

Remark 2. *Let Ω be compact, $\mathcal{C} = \{c_1, \ldots, c_M\}$, $c_i \in \Omega$, and let the approximate feature map be defined by (2). Consider the system (3) with $\widehat{A} = \Lambda$, and let $\Upsilon = \{0, 1, \ldots, M-1\}$. Then the measurement map $\widetilde{K}$'s values lie in $\{0, 1\}$; in particular, each row $\widetilde{K}^{(j)}$, $j \in \{1, \ldots, \ell\}$, corresponds to a subspace $\widehat{\mathcal{H}}_j$, generated by a subset of centers $\mathcal{C}^{(j)} \subset \mathcal{C}$. Generate samples $x_i^{(j)}$ to create a kernel matrix $K^{(j)}$ that is shaded only with respect to centers $\mathcal{C}^{(j)}$. Once this is done, move on to the next subspace $\widehat{\mathcal{H}}_{j+1}$. When all ℓ rows of $\widetilde{K}$ are accounted for, construct the matrix $\mathbf{K}$ as in (1). Then the resulting system $(\mathbf{K}, \widehat{A})$ is observable.*

This heuristic is formalized in Algorithm 2. Note that in practice, the matrix $\widehat{A}$ needs to be inferred from measurements of the process f_τ. If no assumptions are placed on $\widehat{A}$, it's clear that at least M sensors are required for the system identification phase. Future work will study the precise conditions under which system identification is possible with less than M sensors.

Algorithm 3 Kernel Observer (Transition Learning)

```
Input: Kernel $k$, basis centers $\mathcal{C}$, final time step $T$.
while $\tau \leq T$ do
  1) Sample data $\{y_\tau^i\}_{i=1}^M$ from $f_\tau$.
  2) Estimate $\widehat{w}_\tau$ via standard kernel inference procedure.
  3) Store weights $\widehat{w}_\tau$ in matrix $\mathcal{W} \in \mathbb{R}^{M \times T}$.
end while
To infer $\widehat{A}$, define matrix $\Phi = \mathcal{W}^T \mathcal{W}$. Then:
for $i = 1$ to $M$ do
  At step $i$, solve system
    $\widehat{A}^{(i)} = \left( (\Phi + \lambda I)^{-1} (\mathcal{W}^T \mathcal{W}^{(i)}) \right)^T,$    (2)
  where $\widehat{A}^{(i)}$, and $\mathcal{W}^{(i)}$ are the $i$th columns of $\widehat{A}$ and $\mathcal{W}^{(i)}$ respectively.
end for
Compute the covariance matrix $\widehat{B}$ of the observed weights $\mathcal{W}$.
Output: estimated transition matrix $\widehat{A}$, predictive covariance matrix $\widehat{B}$.
```

Algorithm 4 Kernel Observer (Monitoring and Prediction)

```
Input: Kernel $k$, basis centers $\mathcal{C}$, estimated system matrix $\widehat{A}$, estimated covariance matrix $\widehat{B}$.
Compute Observation Matrix: Compute the cyclic index $\ell$ of $\widehat{A}$, and compute $K$.
Initialize Observer: Use $\widehat{A}$, $\widehat{B}$, and $K$ to initialize a state-observer (e.g. Kalman filter (KF)) on $\widehat{\mathcal{H}}$.
while measurements available do
  1) Sample data $\{y_\tau^i\}_{i=1}^N$ from $f_\tau$.
  2) Propagate KF estimate $\widehat{w}_\tau$ forward to time $\tau+1$, correct using measurement feedback with $\{y_{\tau+1}^i\}_{i=1}^N$.
  3) Output predicted function $\widehat{f}_{\tau+1}$ of KF.
end while
```

References

[1] Charles A Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. In *Approximation Theory and Spline Functions*, pages 143–145. Springer Netherlands, 1984.

[2] W Murray Wonham. *Linear multivariable control*. Springer, 1974.

Table 2: Notations

Notations	
$\mathcal{A}$	Linear transition operator in the RKHS $\mathcal{H}$
$\widehat{A}$	Linear transition operator in the strict subspace $\widehat{\mathcal{H}}$ of RKHS $\mathcal{H}$
$\alpha(\lambda)$	Minimal polynomial
$\alpha_i(\lambda)$	Minimal polynomial w.r.t Λ_{i-1} or $\mathcal{A}_{\vert\mathcal{V}_i}$
$\mathcal{B}$	Banach space
$\mathcal{C}$	Array of basis centers generating a finite-dimensional covering of $\mathcal{H}$
$\mathcal{C}^{(j)}$	Subset of basis centers $\mathcal{C}$
c_i	Basis center, i^{th} element of $\mathcal{C}$
f_τ	Mean of time-varying stochastic process at instant τ
$\mathcal{H}$	Reproducing Kernel Hilbert Space (RKHS)
$\widehat{\mathcal{H}}$	Approximate Reproducing Kernel Hilbert Space
$\widehat{\mathcal{H}}_j$	j^{th} subspace of $\widehat{\mathcal{H}}$
$k(\cdot,\cdot)$	Positive-definite kernel on a domain Ω
$\widetilde{K}$	Measurement map $\in \mathbb{R}^{\ell \times M}$
$\mathring{K}$	Measurement map $\in \mathbb{R}^{\ell \times M}$ corresponding to Jordan normal form Λ
$\mathcal{K}$	Linear measurement operator that maps $\mathcal{H} \to \mathbb{R}^N$
K	Kernel matrix between the data points and basis vectors
ℓ	Lower bound on sampling locations, the cyclic index of a matrix
L	Total number of time instances, at which measurement or samples are taken.
$\mathcal{L}$	Linear operator in Banach space $\mathcal{B}$
M	Number of atoms in $\widehat{\mathcal{H}}$
N	Number of sensing or sampling locations
Λ	Jordan normal form
$\mathcal{O}_\Upsilon$	Observability Matrix
Υ	Set of instances τ_i
P	Similarity transformation matrix
$\psi(\cdot)$	Smooth map $\psi : \Omega \to \mathcal{H}$
τ	Discrete time index
u	a function in Banach space $\mathcal{B}$
$\mathcal{V}$	Linear space
w_τ	Weight vector $\in \mathbb{R}^M$at instant τ
w_0	Initial weight vector
$\widehat{w}_\tau$	Estimate of w_τ
x_i	i^{th} sensing or sampling location
$x_i^{(j)}$	i^{th} sensing or sampling location w.r.t $\mathcal{C}^{(j)} \subset \mathcal{C}$
$x^{(j)}$	Sensing or sampling locations w.r.t $\mathcal{C}^{(j)} \subset \mathcal{C}$
$\mathcal{X}$	Set of sampling or sensing locations