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

Hassan A. Kingravi
Pindrop Security
hkingravi@pindropsecurity.com

Harshal Maske & Girish Chowdhary
MAE, Oklahoma State University
{harshal.maske,girish.chowdhary}@okstate.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 a dynamical systems layer over temporal evolution of the weights of a kernel model is a valid approach to spatiotemporal modeling that does not necessarily require the design of complex nonstationary kernels. Furthermore, we show that such a predictive model can be utilized to determine sensing locations that guarantee that the hidden state of the predictive model can be recovered with very few measurements. The approach is validated on real-world datasets.

1 Introduction

Modeling of large-scale stochastic phenomena with both spatial and temporal (spatiotemporal) evolution is a fundamental problem in the applied sciences [1, 7]. While spatiotemporal phenomena have been traditionally modeled using first-principles approaches like PDEs, data-driven models have gained more attention in the machine learning and statistics communities in recent years [4]. Kernel methods represent a class of well-studied and powerful nonparametric methods for inference in spatial domains [19], and have been successfully applied to spatiotemporal modeling [4, 17]; recent techniques focus on nonstationary covariance kernel design and associated hyperparameter learning algorithms. The Process Convolution with Local Smoothing Kernels (PCLSK) approach [8] captures nonstationary structure by allowing the kernel to vary across the input space, and can utilize a Gaussian process (GP) framework, as shown in [6, 12, 14]. Apart from directly modeling the covariance function using additional latent GPs, another class of methods, henceforth known as Latent Extension of Input Space (LEIS), map the nonstationary process into a latent space, in which the problem becomes approximately stationary [13, 18]. Both approaches are nonconvex and require MCMC or other sophisticated optimization methods for solution. They also scale poorly because typically data is retained across both space and time.

The geostatistics literature has many examples of the dynamical spatiotemporal modeling approach, where the focus is on finding good dynamical transition models on the linear combination of weights in a parameterized model, an example of which is the kriged (kernel) Kalman filter [4, 11, 16]. The advantage here is that when the spatial and temporal dynamics are hierarchically separated and if linear transition models are used, the learning problem becomes more tractable, and long-term, periodic behavior can be inferred. In this case, complex nonstationary kernels are often not necessary. The approach presented in this paper aligns closely with this vein of work. The main difference is that we view the problem from the more abstract viewpoint of constructing a Bayesian observer in a reproducing kernel Hilbert space. 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 2) quantifying the minimum number of measurements required to estimate the state of functional evolution. In particular, if feedback is allowed, monitoring (state recovery) and prediction (filtering) can be made more efficient than other nonstationary kernel methods.

2 Kernel Observers

We focus on predictive inference of a time-varying stochastic process, whose mean f evolves 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, the prototypical example of which is the *abstract evolution equation*, i.e. ODEs on Banach spaces [2]. To make this approach computationally realizable, we restrict the sequence f_τ to lie in a reproducing kernel Hilbert space (RKHS) [17]. Here $k : \Omega \times \Omega \to \mathbb{R}$ is a positive-definite Mercer kernel on a domain Ω that 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 proposed model assumes spatiotemporal evolution in the input domain corresponds to temporal evolution of the mixing weights of a kernel model alone in the functional domain: this separation 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. In this case, the method is functionally equivalent to a kernel Kalman filter (KKF), with the main contributions being our approximate formulation (see below) and the minimization of the number of measurements required to infer the latent state. Both nonstationary kernel methods and KKFs can be derived using Bayesian inference: the main difference is that nonstationary kernel methods are truly nonparametric methods in that the time indices τ are part of the observations, and thus the kernel model must retain them for training. On the other hand, KKFs decouple the time and space variables, and thus can learn long-term patterns without including time in the kernel function, at the cost of missing subtler local time-dependent correlations[3].

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 $\mathbb{R}^N$. To avoid working in dual space and have the parameters grow with the data, we work with an approximate feature map $\widehat{\psi}(x) := [\, \widehat{\psi}_1(x) \; \cdots \; \widehat{\psi}_M(x) \,]$ to an approximate feature space $\widehat{\mathcal{H}}$. Typical examples of such maps include random Fourier features [15], FastFood [10], A la Carte [22], and the Nyström method [20]. Here we use the dictionary of atoms approach as follows: let Ω be compact. Given points $\mathcal{C} = \{c_1, \ldots, c_M\}$, $c_i \in \Omega$, define the 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, where $\widehat{\psi}_i(x) := k(x, c_i)$. In the approximate space case, we replace the transition operator $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ in (1) by $\widehat{\mathcal{A}} : \widehat{\mathcal{H}} \to \widehat{\mathcal{H}}$. The finite-dimensional evolution equations approximating (1) in approximate dual form are

$$w_{\tau+1} = \widehat{A} w_\tau + \eta_\tau, \quad y_\tau = K w_\tau + \zeta_\tau, \tag{2}$$

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 $K_{(i)} = [\, \widehat{\psi}_1(x_i) \; \widehat{\psi}_2(x_i) \; \cdots \; \widehat{\psi}_M(x_i) \,]$. In systems-theoretic language, the matrix acts as a measurement operator. To the best of our knowledge, this approximate formulation of the kernel Kalman filter is new: the paper [16] formulates the filter in dual space, and [11] makes no connection to RKHSs. We define the *observability matrix* as $\mathcal{O}_\Upsilon = [\, (K\widehat{A}^{\tau_1})^T \; \cdots \; (K\widehat{A}^{\tau_L})^T \,]$ where $\Upsilon = \{\tau_1, \ldots, \tau_L\}$ are the set of instances τ_i when we apply the operators K_{τ_i}. A linear system is said to be *observable* if $\mathrm{Rank}\mathcal{O}_\Upsilon = M$) for $\Upsilon = \{0, 1, \ldots, M-1\}$ [23]. Observability guarantees two critical facts: firstly, it guarantees that the state w_0 can be recovered exactly from a finite series of measurements $\{y_{\tau_1}, \ldots, y_{\tau_L}\}$; in particular, defining $y_\Upsilon = \left[y_{\tau_1}^T, \cdots, y_{\tau_L}^T\right]^T$, 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. We are now in a position to formally state the spatiotemporal modeling and inference problem considered: given a spatiotemporally evolving system modeled using (2), 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**.

2.1 Main Theoretical Result

We take a geometric approach for the choice of sampling locations for inferring w_τ using the Jordan decomposition of $\widehat{A}$. Let r be the number of unique eigenvalues of $\widehat{A}$, and let $\gamma(\lambda_i)$ denote the geometric multiplicity of eigenvalue λ_i. Then the *cyclic index* of $\widehat{A}$ is defined as $\ell = \max_{1 \leq i \leq r} \gamma(\lambda_i)$[21]. To state the main theorem, we need the following technical condition:

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 $K_{(i)} := [\, \widehat{\psi}_1(x_i) \; \cdots \; \widehat{\psi}_M(x_i) \,]$, define $\mathcal{I}_{(i)} := \{\iota_1^{(i)}, \iota_2^{(i)}, \ldots, \iota_{M_i}^{(i)}\}$ 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.

Random sampling is a simple but effective way to generate shaded matrices with minimal effort.

Theorem 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 (2). 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 $\Lambda = [\, \Lambda_1 \; \cdots \; \Lambda_O \,]$ may have repeated eigenvalues. Let ℓ be the cyclic index of $\widehat{A}$. Define $\mathbf{K} = [\, K^{(1)^T} \; \ldots \; K^{(\ell)^T} \,]^T$ 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 (2) is observable if Υ has distinct values, and $|\Upsilon| \geq M$.*

For a detailed explication of these theoretical results, see section 2.2 in [9]: for the proof of Theorem 1, see the proof of Proposition 2.3. This theorem lends many interesting insights for the modeling of spatiotemporal phenomena: of these, two particularly fascinating ones are a) functions with complex dynamics (with a small cyclic index) can be recovered with less sensor placements than functions with simpler dynamics, and b) for monitoring, *the number of sensor placements are essentially independent of the dimensionality M*, but depend rather on the cyclic index of $\widehat{A}$.

3 Experimental Results

3.1 Comparison With Nonstationary Kernel Methods

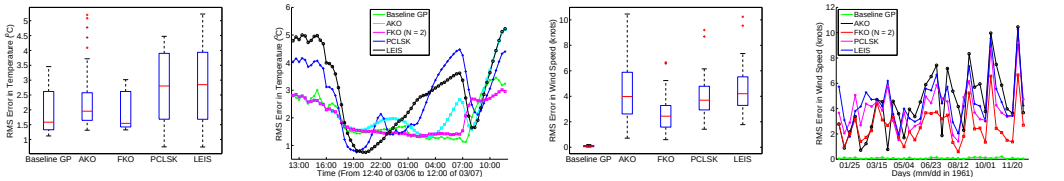
We compare two variants of the kernel observer on two real-world datasets (**Intel** Berkeley and **Irish** Wind) with three competing techniques (see §1): a) PCLSK (with the method in [6]), b) LEIS (with the method in [13]), and c) a baseline GP, which is a sparse GP model trained using all of the data available per time step. Recall from §2 that the kernel observer is a Bayesian state estimator utilizing the dynamics-measurement pair $(K, \widehat{A})$. The first variant of the kernel observer, called *autonomous* and denoted by AKO, is feedforward only (i.e. K is not utilized), and is a measure of the modeling fidelity of (2). The second variant, called *feedback*, and denoted by FKO, utilizes K. 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 [5]); 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 $\widehat{A}$. In the prediction step, in AKO, $\widehat{A}$ is used to propagate w_τ forward to make predictions with no feedback, and in FKO, a Kalman filter with $N \geq \ell$ is used to propagate w_τ forward to make predictions.

The Intel Berkeley dataset consists of temperature data from wireless sensors. The training data is taken from 00:20 hrs on March 6th 2004 at intervals of 20 minutes, and testing was performed over unseen data beginning 12:40 hrs of the same day. Out of 50 sensor 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 1a and as a time-series in Figure 1b. 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 AKO outperforms PCLSK and LEIS overall, and FKO with $N = 2$ significantly outperforms all other methods. The second dataset is the Irish wind dataset, which consists of daily average wind speed (in knots = 0.542 m/s) data collected from year 1961 to 1978 at 12 meteorological stations in the Republic of Ireland. The prediction error results are presented in box-plot form in Figure 1c and as a time-series in Figure 1d. Again, the cyclic index of $\widehat{A}$ was determined to be 2. In this case, AKO's performance is comparable to PCLSK and LEIS, while the kernel observer with feedback with $N = 2$ again outperforms all other methods. The kernel observer is an order of magnitude faster than the competitors: e.g. on **Intel**, the total training and prediction times for PCLSK, LEIS, and FKO are 121.4 sec, 43.8 sec, and 2.1 sec respectively.

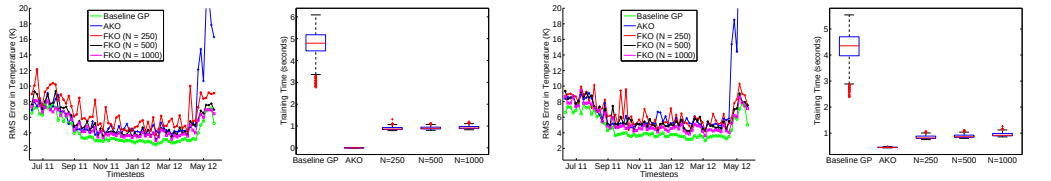
3.2 Prediction of Global Ocean Surface Temperature

We analyzed the feasibility of our approach on a large dataset from the National Oceanographic Data Center: the 4 km AVHRR Pathfinder project, which is a satellite monitoring global ocean surface temperature. This dataset has 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 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 AKO and FKO models, we followed the same procedure outlined in Section 3.1, but with $\mathcal{C} = \{c_1, \ldots, c_M\}$, $c_j \in \mathbb{R}^2$, $|\mathcal{C}| = 300$. The Kalman filter for FKO used $N \in \{250, 500, 1000\}$ at random locations to track the system state given a random initial condition w_0. The observers are compared to a baseline GP model trained on approximately $400,000$ measurements per day, to get a fair comparison. Figures 2a and 2c compare the autonomous and feedback approach with $1,000$ samples to the baseline GP; here, it can be seen that AKO does well in the beginning (beating FKO_{250}), 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 FKO does well throughout. Figures 2a and 2c show a comparison of the speedup with the baseline. These results demonstrate that our approach is a good fit for practical problems with large amounts of data. Future work will explore image analysis and kernels on more general domains.



(a) **Intel** Error (boxplot) (b) **Intel** Error (time-series) (c) **Irish** Error (boxplot) (d) **Irish** Error (time-series)

Figure 1: Comparison of kernel observer to PCLSK and LEIS methods.



(a) Error (day) (b) Estimation time (day) (c) Error (night) (d) Estimation time (night)

Figure 2: Performance of the kernel observer over AVVHR satellite 2012 data.

References

[1] Tim P Barnett, David W Pierce, and Reiner Schnur. Detection of anthropogenic climate change in the world's oceans. *Science*, 292(5515):270–274, 2001.

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

[3] Noel Cressie and Christopher K Wikle. Space-time kalman filter. *Encyclopedia of environmetrics*.

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

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

[6] 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.

[7] Matthew J Heaton, Matthias Katzfuss, Shahla Ramachandar, Kathryn Pedings, Eric Gilleland, Elizabeth Mannshardt-Shamseldin, and Richard L Smith. Spatio-temporal models for large-scale indicators of extreme weather. *Environmetrics*, 22(3):294–303, 2011.

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

[9] Hassan A Kingravi, Harshal Maske, and Girish Chowdhary. Kernel controllers: A systems-theoretic approach for data-driven modeling and control of spatiotemporally evolving processes. *arXiv preprint arXiv:1508.02086*, 2015.

[10] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the International Conference on Machine Learning*, 2013.

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

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

[13] 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.

[14] 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.

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

[16] Liva Ralaivola and Florence d'Alché Buc. Time series filtering, smoothing and learning using the kernel kalman filter. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 3, pages 1449–1454. IEEE.

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

[18] 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.

[19] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

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

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

[22] Zichao Yang, Andrew Wilson, Alex Smola, and Le Song. {A la Carte–Learning Fast Kernels}. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 1098–1106, 2015.

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