TIME SERIES RECONSTRUCTION VIA MACHINE LEARNING: REVEALING DECADAL VARIABILITY AND INTERMITTENCY IN THE NORTH PACIFIC SECTOR OF A COUPLED CLIMATE MODEL

DIMITRIOS GIANNAKIS* AND ANDREW J. MAJDA*

ABSTRACT. Many processes in atmosphere-ocean science develop multiscale temporal and spatial patterns, with complex underlying dynamics and time-dependent external forcings. Because of the possible advances in our understanding and prediction of climate phenomena, extracting that variability empirically from incomplete observations is a problem of wide contemporary interest. Here, we present a technique for analyzing climatic time series that exploits the geometrical relationships between the observed data points to recover features characteristic of strongly nonlinear dynamics (such as intermittency), which are not accessible to classical Singular Spectrum Analysis (SSA). The method utilizes Laplacian eigenmaps, evaluated after suitable time-lagged embedding, to produce a reduced representation of the observed samples, where standard tools of matrix algebra can be used to perform truncated Singular Value Decomposition despite the nonlinear manifold structure of the data set. As an application, we study the variability of the upper-ocean temperature in the North Pacific sector of a 700-year equilibrated integration of the CCSM3 model. Imposing no a priori assumptions (such as periodicity in the statistics), our machine-learning technique recovers three distinct types of temporal processes: (1) periodic processes, including annual and semiannual cycles; (2) decadal-scale variability with spatial patterns resembling the Pacific Decadal Oscillation; (3) intermittent processes associated with the Kuroshio extension and variations in the strength of the subtropical and subpolar gyres. The latter carry little variance (and are therefore not captured by SSA), yet their dynamical role is expected to be significant.

1. INTRODUCTION

Coupled atmosphere-ocean processes exhibit variability across a broad range of time scales, including seasonal, interannual, and decadal time scales [19, 27, 20, 21, 26]. There is a strong interest among the climate community in extracting physically-meaningful information about this variability using data from models or observations, with the goal of enhancing our understanding of the underlying dynamics, and improving our predictive capabilities.

A classical way of attacking this problem is through Singular Spectrum Analysis (SSA), or one of its variants [28, 3, 18, 15]. Here, a low-rank approximation of a dynamic process is constructed by first embedding a time series of a scalar or multivariate observable in a high-dimensional vector space H (called embedding space) using the method of delays [25, 24, 14], and then performing a truncated singular-value decomposition (SVD) of the matrix X containing the embedded data [8]. In this manner, information about the dynamical process is extracted from the left and right singular vectors of X with the k largest singular values. The left singular vectors form a set of empirical orthogonal functions (EOFs) which, at each instance of time, are weighted by the corresponding principal components (PCs) determined from the right singular vectors to yield a rank-k reconstruction of X.

A potential drawback of this approach is that it is based on minimizing an operator norm which may be unsuitable for the nonlinear processes encountered in atmosphere-ocean science (AOS). Specifically, the PCs are computed by projecting onto the principal axes of the k-dimensional ellipsoid that best fits the data in the least-squares sense. This construction is optimal when the underlying dynamics are linear, but nonlinear processes will in general give rise to a manifold M in embedding space that deviates significantly from an ellipsoidal shape. Physically, a prominent manifestation

*Courant Institute of Mathematical Sciences, New York University, dimitris@cims.nyu.edu, jonjon@cims.nyu.edu.

of this phenomenon is failure to capture via SSA the intermittent patterns arising in turbulent dynamical systems; i.e., temporal processes that carry low variance but play an important dynamical role [13].

Despite their inherently nonlinear character, such data sets possess a natural linear structure, namely the Hilbert space $L^2(M, \mu)$ of square-integrable functions on M with inner product inherited from the volume element μ of M (the Riemannian measure). This space may be thought of as the collection of all possible weights that can be assigned to the data samples when making a reconstruction, i.e., it is analogous to the space spanned by the right singular vectors in SSA [3]. Similarly, the left singular vectors are naturally identified with elements of the dual space H^* to H. Therefore, it is reasonable to develop algorithms that seek to approximate suitably defined maps from $L^2(M, \mu)$ to H^*. Such maps, denoted here by A, have the advantage of being simultaneously linear and compatible with the nonlinear manifold comprised by the data.

In this paper, we advocate that this approach, implemented via algorithms developed in machine learning, can reveal important aspects of complex AOS data sets which are not accessible to standard SSA. Here, an orthonormal basis for $L^2(M, \mu)$ is constructed through eigenfunctions of the Laplace-Beltrami operator on M, computed efficiently via sparse graph-theoretic algorithms [4, 10]. Projecting the data from embedding space H onto these eigenfunctions gives a matrix representation of A, such that the optimal rank-k reconstruction with respect to the natural norm of maps from $L^2(M, g)$ to H^* is given by standard truncated SVD.

We demonstrate the efficacy of the scheme in an analysis of the North Pacific sector of the Community Climate System model version 3 (CCSM3) [12]. Using a 700-year equilibrated data set of the upper 300 m ocean [1, 26, 7], we identify a number of qualitatively-distinct spatiotemporal processes, each with a meaningful physical interpretation. These include the seasonal cycle, semiannual variability, as well as decadal-scale processes resembling the Pacific Decadal Oscillation (PDO).

Besides these modes, which are familiar from SSA, the spectrum of the manifold-based algorithm also contains modes with a strongly intermittent behavior in the temporal domain, characterized by five-year periods of high-amplitude oscillations with annual and semiannual frequencies, separated by periods of quiescence. Spatially, these modes describe enhanced eastward transport in the Kuroshio extension region, as well as retrograde (westward) propagating temperature anomalies and circulation patterns resembling the subpolar and subtropical gyres. The bursting-like behavior of these modes, a hallmark of strongly-nonlinear dynamics, means that they carry little variance of the raw signal (about an order of magnitude less than the seasonal and PDO modes). As a result, they are not captured by linear SSA.

The plan of this paper is as follows. In Section 2 we describe our theoretical framework. In Section 3 we apply this framework to the upper-ocean temperature in the North Pacific sector of CCSM3. We discuss the implications of these results in Section 4, and conclude in Section 5.

2. Theoretical framework

We consider that we have at our disposal samples of a time-series x_t of a d-dimensional climatic variable sampled uniformly with time step δt. Here, $x_t \in \mathbb{R}^d$ is generated by a dynamical system, but observations of x_t alone are not sufficient to uniquely determine the state of the system in phase space; i.e., our observations are incomplete. For instance, in Section 3 ahead, x_t will be a depth-averaged ocean temperature field restricted in the North-Pacific sector of CCSM3. Our objective is to produce a low-rank reconstruction of x_t taking explicitly into account the fact that the underlying trajectory of the dynamical system lies on a nonlinear manifold M in phase space.

The methodology employed here to address this objective consists of five basic steps: (1) embed the observed data in a vector space of dimension greater than d via the method of delays; (2) map the data from embedding space to a set of orthonormal Laplacian eigenfunctions; (3) evaluate a low-rank approximation of the data in reduced coordinates determined through the eigenfunctions; (4) convert the approximated data back to embedding space; (5) project to physical space $\mathbb{R}^d$ to obtain the reconstructed signal. Below, we provide a summary of each step. Details of the procedure will be presented elsewhere. Hereafter, we shall consider that M is compact and smooth, so that a well-defined spectral theory exists [6]. Even though these conditions may not be fulfilled in practice, eventually we will pass to a discrete, graph-theoretic description [9], where smoothness is not an issue.

Step (1) is familiar from the qualitative theory of dynamical systems [23, 25, 24, 14]. Under generic conditions, the image of x_t in embedding space $H = \mathbb{R}^n$ under the delayed-coordinate mapping,

$$x_t \mapsto X_t = (x_t, x_{t-\delta t}, \dots, x_{t-(q-1)\,\delta t}) \tag{1}$$

lies on a manifold which is diffeomorphic to M (i.e., indistinguishable from M from the point of view of differential geometry), provided that the dimension n of H is sufficiently large. Thus, given a sufficiently-long embedding window $\mathit{\Delta} t = (q-1)\,\delta t$, we obtain a representation of the nonlinear manifold underlying our incomplete observations, which can be thought of as a curved hypersurface in Euclidean space. That hypersurface inherits a Riemannian metric g, i.e., an inner product between tangent vectors on M constructed from the canonical inner product of H.

Steps (2) and (3) effectively constitute a generalization of SSA, adapted to nonlinear data sets. Recall that SSA is essentially an SVD decomposition,

$$X = U \mathit{\Sigma} V^T, \tag{2}$$

of the data matrix $X = [X_0, X_{\delta t}, \dots, X_{(s-1)\delta t}]$, dimensioned $n \times s$ for s samples in n-dimensional embedding space. Here, the key observation is that the map in Eq. (1) naturally gives rise to two linear vector spaces, which are analogous to the spaces spanned by left and right singular vectors of X [3]. The first is the space $L^2(M,\mu)$ of square-integrable functions on M, where $\mu = (\det g)^{1/2}$ is the volume element (Riemannian density) induced on M through the embedding $M \mapsto H$. The second space of interest is the dual space H^* of H. The elements of H^* are functionals, mapping observed data points in H to the real numbers.

To see the correspondence with SVD, let f be a function in $L^2(M,\mu)$, z an arbitrary vector in H, and consider the dual vector $h \in H^*$ defined by

$$h(z) = \int_M \mu(X_t) f(X_t) \langle X_t, z \rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is the inner product of H. That is, f assigns a weight proportional to $f(X_t)$ on the dual vector $\langle X_t, \cdot \rangle$, much like the i-th column of V weighs the i-th column of U in Eq. (2). What one gains by phrasing the problem in this manner is a linear map A taking $L^2(M,\mu)$ to H^* via the rule in Eq. (3), viz. $A(f) = h$. Note that this definition is basis-independent. Moreover, unlike the nonlinear manifold M, A is amenable to analysis through the standard tools of linear algebra. In particular, low-rank reconstruction of A is a well-defined notion.

Having the latter as an objective, the role of the Laplacian eigenfunctions in step (2) is to provide an orthonormal basis of $L^2(M,\mu)$, in which the operator norm $\|A\|$ can be straightforwardly computed via the Frobenius norm of its matrix representation [Eq. (6) ahead]. Specifically, it is well known that the eigenfunctions $\{\phi_0, \phi_1, \dots\}$ of the Laplace-Beltrami operator Δ associated with the metric g, defined via $\Delta \phi_i = \lambda_i \phi_i$ (together with appropriate boundary conditions if M has boundaries), lead to an orthogonal decomposition of $L^2(M,\mu)$ into invariant subspaces $\mathit{\Phi}_i$. That is, we have [6]

$$L^2(M,\mu) = \bigoplus_{i=0}^{\infty} \mathit{\Phi}_i \quad \text{with } \mathit{\Phi}_i = \operatorname{span}\{\phi_k : \lambda_k = \lambda_i\}, \tag{4a}$$

$$\int_M \mu(X) f_i(X) f_j(X) = 0 \quad \text{for any } f_i \in \mathit{\Phi}_i,\ f_j \in \mathit{\Phi}_j, \text{ and } j \neq i. \tag{4b}$$

The components of A in this basis are

$$A_{ij} = \int_M \mu(X_t)(X_t)_i \phi_j(X_t), \tag{5}$$

with $(X_t)_i$ the i-th element of X_t, giving the operator norm through

$$\|A\|^2 = \sum_{ij} A_{ij}^2. \tag{6}$$

Equation (5) may be interpreted as a Fourier transform on compact manifolds.

In applications, the Laplace-Beltrami eigenfunctions for a finite data set are computed by replacing the continuous manifold M via a weighted graph G, and solving the eigenproblem of a Markov matrix P defined on G, constructed so that in the continuum limit, $s \to \infty$, the generator of P (the graph Laplacian) converges to Δ [4, 11, 10, 5]. Note that the Markov matrix employed in this procedure is highly sparse, which means that the cost of the eigenvalue problem for (λ_i, ϕ_i) grows linearly with the number of samples.

The least-favorable scaling in the eigenfunction calculation involves the pairwise distance calculation between the data samples in embedding space. This scales quadratically with the number of samples if done with brute force, which is the approach adopted here. However, an $s \log s$ scaling may be realized if the dimension of H is small-enough for approximate kd-tree-based algorithms to operate efficiently [2]. In the present study, all eigenfunction calculations were performed on a desktop workstation. The scalability of this class of algorithms to large problem sizes has been widely demonstrated in the machine learning and data mining literature.

In step (3), a rank-k approximation $\tilde{A}$ of A is evaluated by selecting the first r invariant subspaces in order of increasing λ_i (with $l = \sum_{i=1}^{r} \dim \mathit{\Phi}_i \geq k$), and performing a truncated SVD of the $n \times l$ matrix $\hat{A} = [A_{ij}]_{j \leq l}$. That is, in matrix notation, the nonzero components of $\tilde{A}$ are

$$\tilde{A} = U_k \mathit{\Sigma}_k V_k^T, \tag{7}$$

where $\mathit{\Sigma}_k$ is a $k \times k$ diagonal matrix containing the k-largest singular values σ_i of $\hat{A}$, and U_k and V_k are respectively $n \times k$ and $l \times k$ matrices whose columns are the corresponding left and right singular vectors. The resulting operator is the highest-norm rank-k linear map from $L^2(M, \mu)$ to H^*, whose kernel is the orthogonal complement of $\bigoplus_{i=1}^{r} \mathit{\Phi}_i$ in $L^2(M, \mu)$.

Step (4) involves computing the reconstructed data $\tilde{X}_t$ in embedding space via the inverse transform [cf. Eq. (5)]

$$(\tilde{X}_t)_i = \sum_{j=1}^{l} \tilde{A}_{ij} \phi_j(X_t). \tag{8}$$

Finally, in step (5), $\tilde{X}_t$ is projected to d-dimensional physical space by writing

$$\tilde{X}_t = (\hat{x}_{t,0}, \hat{x}_{t,\delta_t}, \dots, \hat{x}_{t,(q-1)\,\delta_t}), \tag{9}$$

and taking the average,

$$\tilde{x}_t = \sum_{t',\tau : t'-\tau=t} \hat{x}_{t',\tau}/q. \tag{10}$$

Note that if M is embedded as an ellipsoid in H, then a set of (possibly degenerate) Laplace-Beltrami eigenfunctions will give the projections of X_t on the principal axes of the ellipsoid; i.e., the system trajectory $\phi_i(X_t)$ in the eigenfunction-based coordinates will be equivalent to the right singular vectors in SSA.

3. Modes of variability in the North Pacific sector of CCSM3

We apply the method presented above to study variability in the North Pacific sector of CCSM3; specifically, variability of the mean upper 300 m sea temperature field in the 700-year equilibrated control integration used by Teng and Branstator [26] and Branstator and Teng [7] in work on the initial and boundary-value predictability of subsurface temperature in that model. Here, our objective is to diagnose the prominent modes of variability in a time series generated by a coupled general circulation model.

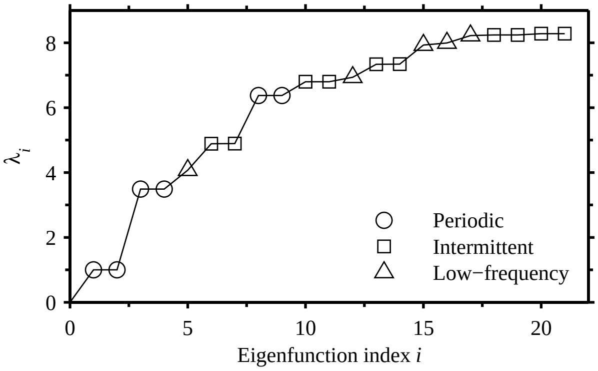

FIGURE 1. Eigenvalues λ_i of the graph Laplacian Δ for the periodic, intermittent, and low-frequency states. Here, we have defined Δ as a positive semidefinite operator, which means that the eigenvalues are non-negative, and obey the ordering $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots$. Moreover, we have normalized the first non-trivial eigenvalue, λ_1, to unity, since multiplication of the λ_i by the same constant can be absorbed by rescaling the Riemannian metric g.

In this analysis, the x_t observable is the mean upper 300 m temperature field sampled every month at $d = 534$ gridpoints (native ocean grid mapped to the model's T42 atmosphere) in the region 20°N–65°N and 120°E–110°W. Throughout, we work with a two-year embedding window; i.e., the dimension of embedding space is $n = d \times 24 = 12{,}816$. For the calculations of the Laplacian eigenvalues and eigenvectors we used the Diffusion Map algorithm of Coifman and Lafon [10].

Figures 1 and 2 show representative eigenvalues and eigenfunctions of the graph Laplacian. Since we are interested in studying temporal evolution processes, we display the eigenfunctions graphically as plots of $\phi_i(X_t)$ versus t, and also show the corresponding Fourier power spectra. Moreover, to study the spatial patterns associated with the eigenfunctions, we have performed temperature field reconstructions by applying the inverse transform in Eq. (8) with $\tilde{A}_{ij}$ replaced by the operator components A_{ij} from Eq. (5) corresponding to each invariant subspace $\mathit{\Phi}_j$. Figure 3 shows reconstructions based on the eigenfunctions of Figure 2.

Carrying out this procedure systematically for several (~ 100) of the eigenfunctions, we find that they fall into three distinct families of periodic, low-frequency, and intermittent modes, described below. Note that embedding [step (1)] is essential to the separability of the eigenfunctions into these processes; the character of the eigenfunctions is mixed if no embedding is performed.

3.1. **Periodic modes.** The periodic modes come in doubly-degenerate pairs (see Figure 1), and have the structure of sinusoidal waves with phase difference $\pi/2$ and frequency equal to integer multiples of 1 year^{-1}. The leading periodic modes, ϕ_1 and ϕ_2, represent the seasonal cycle in the data. In the physical (spatial) domain [Figure 3(b)], these modes generate an annual oscillation of the temperature anomaly, whose amplitude is largest ($\sim$ 1°C) in the western part of the basin ($\sim$ 130°E–160°E) and for latitudes in the range 30°N–45°N. Together with the higher-frequency overtones, the modes in this family are the standard eigenfunctions of the Laplacian on the circle, suggesting that the data manifold M has the geometry of a circle along one of its dimensions.

3.2. **Low-frequency modes.** The low-frequency modes are characterized by high spectral power over interannual to interdecadal timescales, and strongly suppressed power over annual or shorter time scales. As a result, these modes represent the low-frequency variability of the upper ocean, which has been well-studied in the North Pacific sector of CCSM3 [1, 26]. The leading mode in this family [ϕ_5; see Figure 2(b)], gives rise to a typical PDO pattern [Figure 3(c)], where the most prominent basin-scale structure is a horseshoe-like temperature anomaly pattern developing eastward

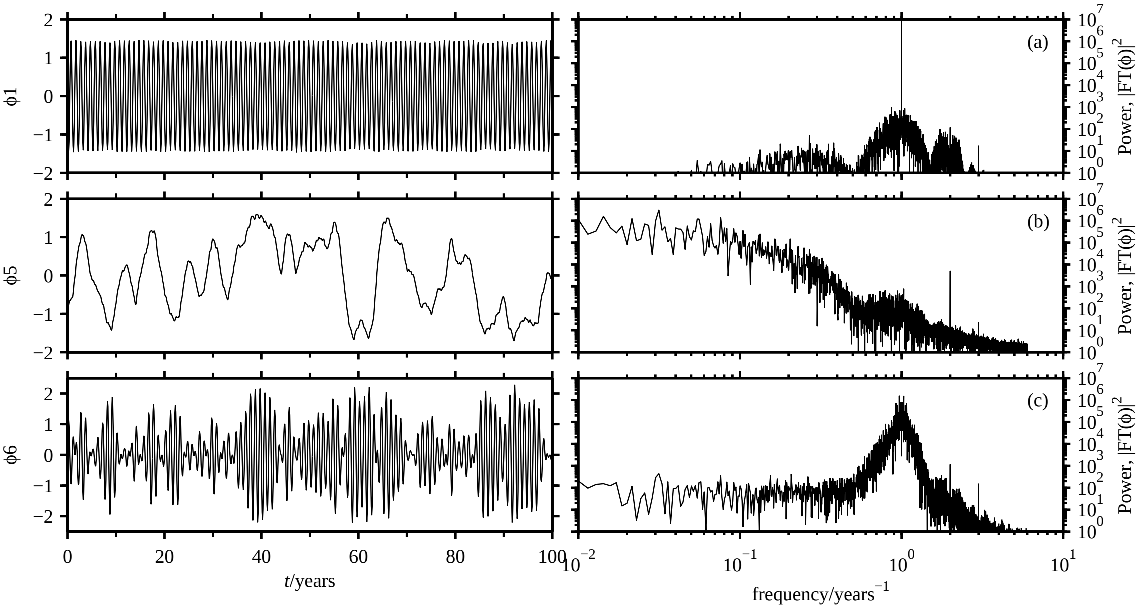

Figure 2. Eigenfunctions of the graph Laplacian corresponding to the eigenvalues from Figure 1 plotted in the temporal (left-hand panels) and frequency domains (right-hand panels). (a) Seasonal eigenfunction ϕ_1. (b) First low-frequency eigenfunction, ϕ_5. (c) First intermittent eigenfunction, ϕ_6.

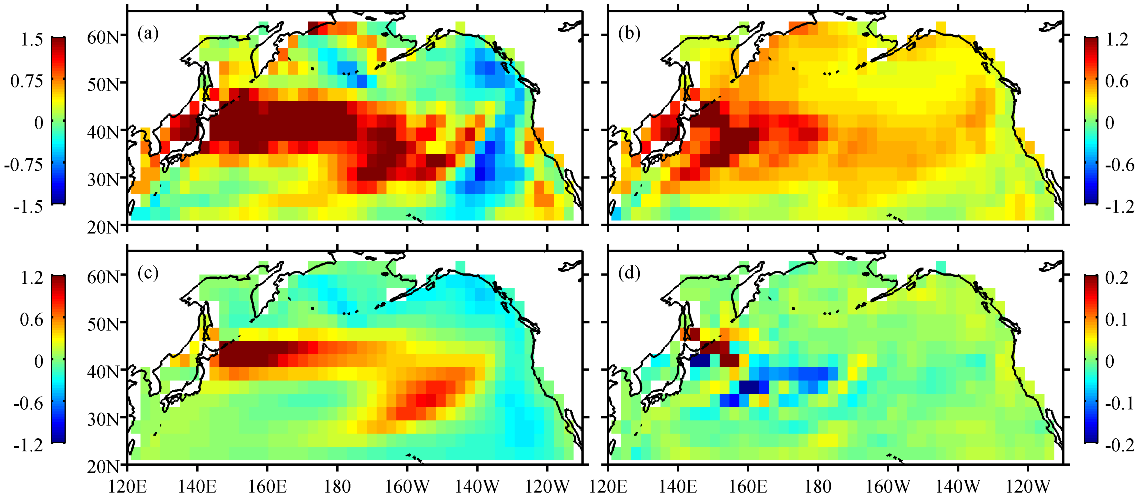

Figure 3. Reconstructions of the upper 300 m temperature anomaly field (annual mean subtracted at each gridpoint). Panel (a) shows the raw data in November of year 91 of Figure 2. Panels (b–d) display reconstructions using (b) the seasonal eigenfunctions, ϕ_1 and ϕ_2; (c) the first low-frequency eigenfunction, ϕ_5, describing the PDO; (d) the first two-fold degenerate set of intermittent eigenfunctions, ϕ_6 and ϕ_7, describing variability of the Kuroshio extension.

along the Kuroshio extension, together with an anomaly of the opposite sign along the west coast of North America. The higher modes in this family gradually develop smaller spatial features and spectral content over shorter time scales than ϕ_5, but have no spectral peaks on annual or shorter timescales.

3.3. **Intermittent modes.** As illustrated in Figure 2(c), the key feature of modes of this family is temporal intermittency, arising out of oscillations at annual or higher frequency, which are modulated by relatively sharp envelopes with a temporal extent in the 2–10-year regime. Like their periodic counterparts, the intermittent modes form nearly degenerate pairs (see Figure 1), and their base frequency of oscillation is an integer multiple of 1 year^{-1}. The resulting Fourier spectrum is dominated by a peak centered at at the base frequency, exhibiting some skewness towards lower frequencies.

In the physical domain, these modes describe processes with relatively fine spatial structure, which are activated during the intermittent bursts, and become quiescent when the amplitude of the envelopes is small. The most physically-recognizable aspect of these processes is enhanced transport along the Kuroshio extension region, shown for the leading-two intermittent modes (ϕ_6 and ϕ_7) in Figure 3(d). This process features sustained eastward propagation of small-scale, ~ 0.2 °C temperature anomalies during the intermittent bursts. The intermittent modes higher in the spectrum also encode rich spatiotemporal patterns, including retrograde (westward) propagating anomalies, and gyre-like patterns resembling the subpolar and subtropical gyres.

4. Discussion

4.1. **Intermittent processes and relation to SSA.** The main result of this analysis, which highlights the importance of taking explicitly into account the nonlinear structure of AOS data sets, is the existence of intermittent patterns of variability in the North Pacific sector of CCSM3, which are not accessible through SSA. This type of variability naturally emerges by studying the properties of individual invariant subspaces $\mathit{\Phi}_i$ of Laplace-Beltrami eigenfunctions on the data manifold (e.g., as done in Figure 3), but in order to produce a more accurate reconstruction, the SVD in Eq. (2) must be applied to combine information from several $\mathit{\Phi}_i$. Here, we apply this procedure to evaluate a rank $k = 30$ reconstruction based on the leading $l = 55$ Laplace-Beltrami eigenfunctions (in order of increasing λ_i), and compare the results with SSA.

As shown in Figure 4, the leading singular values of $\tilde{A}$ from Eq. (7) fall into four distinct families, separated by spectral gaps; viz. $\{\sigma_1, \sigma_2\}$, coupling almost entirely to the annual eigenfunctions, ϕ_1 and ϕ_2; $\{\sigma_3, \dots, \sigma_{12}\}$, dominated by the low-frequency modes in Figure 1 with weak contributions from the intermittent modes; $\{\sigma_{13}, \sigma_{14}\}$; coupling almost entirely to the semiannual modes, ϕ_3 and ϕ_4; $\{\sigma_{15}, \dots, \sigma_{21}\}$, dominated by the intermittent modes with some coupling to the low-frequency modes with high λ_i.

Typical temperature-anomaly patterns associated with these processes are shown in Figure 5. There, the Kuroshio modes of Figure 3(d) become augmented by temperature anomalies developing along the West Coast of North America, and transported westwards at high latitudes or in the sub-tropics. These features, displayed in Figure 5(f), resemble the subpolar and subtropical gyre. The semiannual modes [Figure 5(e)] also exhibit significant amplitude along the West Coast, which is consistent with semiannual variability of the upper ocean associated with the California current [22]. Note that the semiannual modes appear early in the λ_i spectrum of the Laplacian, but their explained variance, as measured by σ_i, is comparatively small. In separate calculations, we have verified that the SVD decomposition of A is qualitatively robust with respect to the number l of Laplacian eigenfunctions used as basis functions for $L^2(M, \mu)$.

A key point brought out by Figures 4 and 5 is that reconstructions based on machine learning are in close agreement with SSA for the annual and low-frequency modes, but intermittent modes have no SSA counterparts. In particular, instead of the qualitatively-distinct families of processes

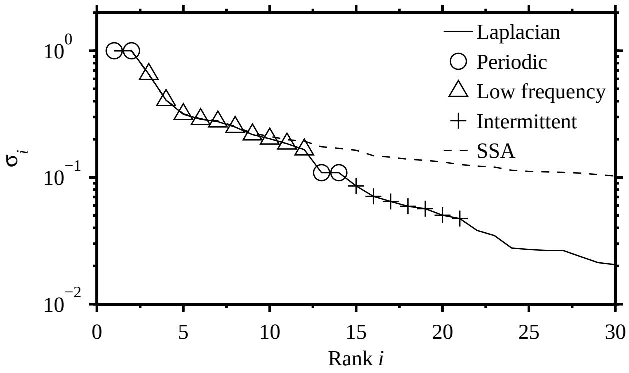

FIGURE 4. Singular values σ_i (normalized so that $\sigma_1 = 1$) evaluated through Laplacian eigenmaps from Eq. (7) (solid line) and SSA (dashed line). The periodic, low-frequency, and intermittent modes indicated here are used in the temperature field reconstructions of Figure 5.

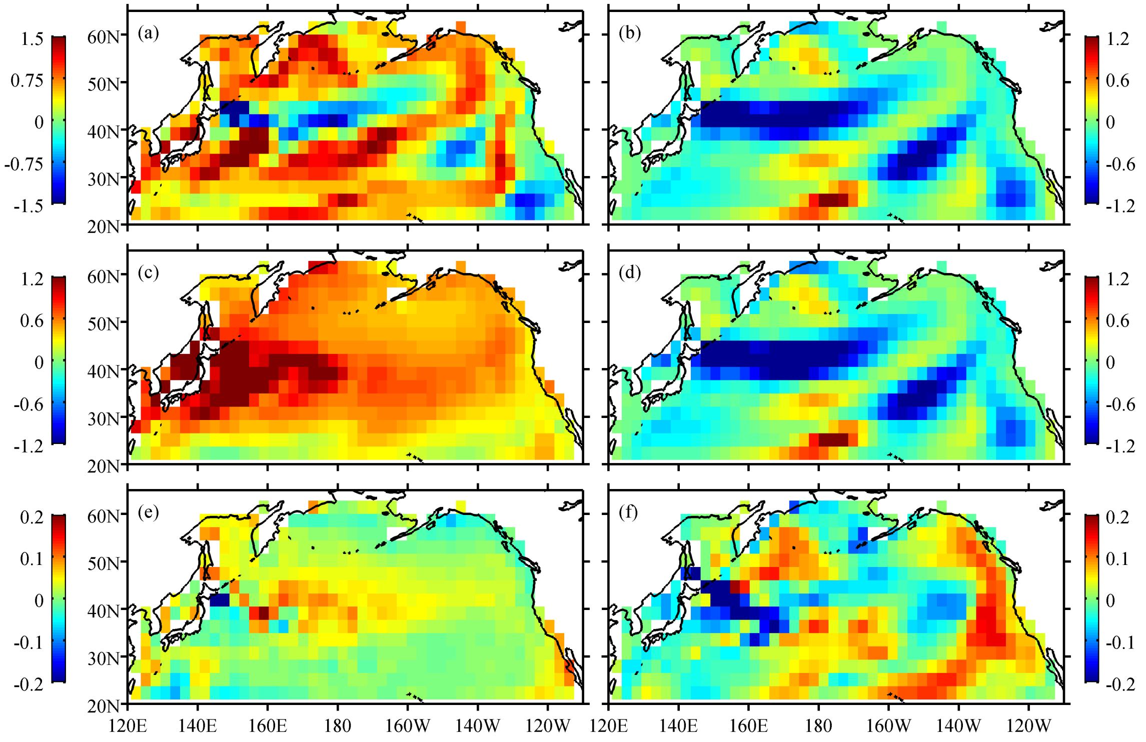

FIGURE 5. Reconstructions of the upper-300 m temperature anomaly field of the 700-year CCSM3 control run through machine learning and SSA. Panel (a) shows the raw data in October of year 45 of Figure 2. Panel (b) displays an SSA reconstruction evaluated using singular vectors (SVs) 3–12 (the low-frequency modes; see Figure 4). Panels (c–f) display reconstructions via Laplacian eigenmaps using (c) the first two SVs of $\tilde{A}$ in Eq. (7) (annual modes); (d) SVs 3–12 (low-frequency modes); (e) SVs 13 and 14 (semiannual modes); (f) SVs 15–21 (intermittent modes).

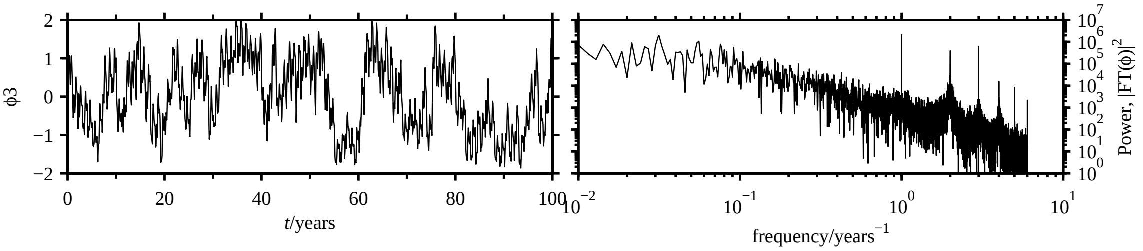

Figure 6. The Laplacian eigenfunction corresponding to the leading "low-frequency" mode evaluated without embedding [cf. Figure (2b)]. Note the pronounced spectral lines with period $\{1, 1/2, 1/3, \dots, 1/6\}$ years.

described above, the SSA spectrum is characterized by a smooth decay involving modes of progressively higher spatiotemporal frequencies, but with no intermittent behavior analogous, e.g., to mode ϕ_6 in Figure 2. Two of the SSA modes exhibit significant semiannual variability, but the frequency content of these modes is not pure, featuring low-frequency beating patterns.

The σ_i values associated with the intermittent modes and, correspondingly, the contributed variance in temperature field reconstructions, is significantly smaller than the periodic or low-frequency modes. However, this is not to say the dynamical significance of these modes is negligible. In fact, intermittent events, carrying low variance, are widely prevalent features of complex dynamical systems [13]. Being able to capture this intrinsically nonlinear behavior constitutes one of the major strengths of the machine-learning based method presented here.

4.2. **The role of lagged embedding.** The embedding in Eq. (1) of the input data x_t in H is essential to the separability of the Laplacian eigenfunctions into distinct families of processes. To illustrate this, in Figure 6 we display the Laplacian eigenfunction that most-closely resembles the PDO mode in Figure 2(b), evaluated without embedding ($q = 1$, $\Delta t = 0$). It is evident from both the temporal and Fourier representations of that eigenfunction that the decadal process recovered in Section 3.2 using a two-year embedding window has been contaminated with high-frequency variability; in particular, prominent spectral lines at integer multiples of 1 year^{-1} down to the maximum frequency of 6/year allowed by the monthly sampling of the data. An even stronger frequency mixing was found to take place in the corresponding temporal SSA modes. In general, representing the dynamical information lost through partial observations via time-lagged embedding, as advocated in the qualitative theory of dynamical systems [23, 25, 8, 24], significantly enhances the quality of time-series reconstructions through either of the machine learning or SSA schemes.

In separate calculations, we have verified that the eigenfunctions separate into periodic, low-frequency, and intermittent processes for embedding windows up to $\Delta t = 10$ years. However, longer embedding windows require more eigenfunctions to produce the same strength of reconstructed signal via Eq. (7).

5. Conclusions

Combining techniques from machine learning and the qualitative theory of dynamical systems, in this work we have presented a scheme for time series reconstruction, which takes explicitly into account the nonlinear geometrical structure of data sets arising in atmosphere-ocean science. Like classical SSA [15], the method presented here utilizes time-lagged embedding and truncated SVD to produce a low-rank reconstruction of time series generated by partial observations of high-dimensional, complex dynamical systems. However, the linear operator used here in the SVD step differs crucially from SSA in that its domain of definition is the Hilbert space of square-integrable functions on the nonlinear manifold M comprised by the data (in a suitable coarse-grained representation via a graph). These functions, analogous to the temporal modes (right singular vectors) in SSA [3], are tailored to the nonlinear geometry of M through its Riemannian measure.

Applying this scheme to the upper-ocean temperature in the North Pacific sector of the CCSM3 model, we find a family of intermittent processes which are not captured by SSA. These processes describe eastward-propagating, small-scale temperature anomalies in the Kuroshio extension region, as well as retrograde-propagating structures at high latitudes and in the subtropics. Moreover, they carry little variance of the raw signal, and display burst-like behavior characteristic of strongly nonlinear dynamics. The remaining identified modes include the familiar PDO pattern of low-frequency variability, as well as annual and semiannual periodic processes.

The nature of the analysis presented here is purely diagnostic. In particular, we have not touched upon the dynamical role of these modes in reproducing the upper ocean dynamics in CCSM3. Here, pertinent open questions are the significance of the intermittent modes in triggering large-scale regime transitions [13], as well as potential improvements of the predictive skill and model error of reduced models utilizing these modes [16, 17]. We plan to study these topics in future work.

Acknowledgments

This work was supported by NSF grant DMS-0456713, ONR DRI grants N25-74200-F6607 and N00014-10-1-0554, and DARPA grants N00014-07-10750 and N00014-08-1-1080. We thank G. Branstator and H. Teng for providing access to the CCSM3 data set used in this analysis.

References

[1] M. Alexander et al. Extratropical atmosphere–ocean variability in CCSM3. *J. Climate*, 19:2496–2525, 2006.

[2] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching. *J. ACM*, 45:891, 1998.

[3] N. Aubry, R. Guyonnet, and R. Lima. Spatiotemporal analysis of complex signals: Theory and applications. *J. Stat. Phys.*, 64(3–4):683–739, 1991.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[5] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *J. Comput. Syst. Sci.*, 74(8):1289–1308, 2008.

[6] P. H. Bérard. *Spectral Geometry: Direct and Inverse Problems*, volume 1207 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1989.

[7] G. Branstator and H. Teng. Two limits of initial-value decadal predictability in a CGCM. *J. Climate*, 23(23):6292–6311, 2010.

[8] D. S. Broomhead and G. P. King. Extracting qualitative dynamics from experimental data. *Phys. D*, 20(2–3):217–236, 1986.

[9] F. R. K. Chung. *Spectral Graph Theory*, volume 97 of *CBMS Regional Conference Series in Mathematics*. Americal Mathematical Society, Providence, 1997.

[10] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[11] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition on data. *Proc. Natl. Acad. Sci.*, 102(21):7426–7431, 2004.

[12] W. D. Collins et al. The community climate system model version 3 (CCSM3). *J. Climate*, 19:2122–2143, 2006.

[13] D. T. Crommelin and A. J. Majda. Strategies for model reduction: Comparing different optimal bases. *J. Atmos. Sci.*, 61:2206, 2004.

[14] E. R. Deyle and G. Sugihara. Generalized theorems for nonlinear state space reconstruction. *PLoS ONE*, 6(3):e18295, 2011.

[15] M. Ghil et al. Advanced spectral methods for climatic time series. *Rev. Geophys.*, 40(1), 2002.

[16] D. Giannakis and A. J. Majda. Quantifying the predictive skill in long-range forecasting. Part I: Coarse-grained predictions in a simple ocean model. *J. Climate*, 2011. submitted.

[17] D. Giannakis and A. J. Majda. Quantifying the predictive skill in long-range forecasting. Part II: Model error in coarse-grained Markov models with application to ocean-circulation regimes. *J. Climate*, 2011. submitted.

[18] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Techniques*. CRC Press, Boca Raton, 2001.

[19] M. Latif and T. P. Barnett. Causes of decadal climate variability over the North Pacific and North America. *Science*, 266(5185):634–637, 1994.

[20] M. Latif and T. P. Barnett. Decadal climate variability over the North Pacific and North America: Dynamics and predictability. *J. Climate*, 9:2407–2423, 1996.

[21] N. J. Mantua and S. R. Hare. The pacific decadal oscillation. *J. Oceanogr.*, 58(1):35–44, 2002.

[22] R. Mendelssohn, F. B. Schwing, and S. J. Bograd. Nonstationary seasonality of upper ocean temperature in the California Current. *J. Geophys. Res.*, 109, 2004.

[23] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Lett.*, 45:712–716, 1980.

[24] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65(3–4):579–616, 1991.

[25] F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, Berlin, 1981.

[26] H. Teng and G. Branstator. Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM. *Climate Dyn.*, 36(9–10):1813–1834, 2010.

[27] K. E. Trenberth and J. W. Hurrell. Decadal atmosphere-ocean variations in the Pacific. *Climate Dyn.*, 9(6):303–319, 1994.

[28] R. Vautard and M. Ghil. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Phys. D*, 35:395–424, 1989.