Data-driven methods for dynamical systems: Quantifying predictability and extracting spatiotemporal patterns

Dimitrios Giannakis* and Andrew J. Majda

Center for Atmosphere Ocean Science
Courant Institute of Mathematical Sciences
New York University

Abstract

Large-scale datasets generated by dynamical systems arise in many applications in science and engineering. Two research topics of current interest in this area involve using data collected through observational networks or output by numerical models to quantify the uncertainty in long-range forecasting, and improve understanding of the operating dynamics. In this research expository we discuss applied mathematics techniques to address these topics blending ideas from machine learning, delay-coordinate embeddings of dynamical systems, and information theory. We illustrate these methods with applications to climate atmosphere ocean science.

Contents

*Corresponding author. Email: dimitris@cims.nyu.edu.

1 Introduction

Large-scale datasets generated by dynamical systems arise in a diverse range of disciplines in science and engineering, including fluid dynamics [1, 2], materials science [3, 4], molecular dynamics [5, 6], and geophysics [7, 8]. A major challenge in these domains is to utilize the vast amount of data that is being collected by observational networks or output by large-scale numerical models to advance scientific understanding of the operating physical processes, and reveal their predictability. For instance, in climate atmosphere ocean science (CAOS) the dynamics takes place in an infinite-dimensional phase space where the coupled nonlinear partial differential equations for fluid flow and thermodynamics are defined, and the observed data correspond to functions of that phase space, such as temperature or circulation measured over a set of spatial points. There exists a strong need for data analysis algorithms to extract and create reduced representations of the large-scale coherent patterns which are an outcome of these dynamics, including the El Niño Southern Oscillation (ENSO) in the ocean [9] and the Madden-Julian Oscillation (MJO) in the atmosphere [10]. Advances in the scientific understanding and forecasting capability of these phenomena have potentially high socioeconomic impact.

In this paper, we review work of the authors and their collaborators on data-driven methods for dynamical systems to address these objectives. In particular, in Sections 2 and 3, respectively, we present (i) methods based on data clustering and information theory to reveal predictability in high-dimensional dynamical systems, and quantify the fidelity of forecasts made with imperfect models [11–13]; and (ii) nonlinear Laplacian spectral analysis (NLSA) algorithms [14–19] for decomposition of spatiotemporal data. The common theme in these topics is that aspects of the coarse-grained geometry of the data in phase space play a role. Specifically, in (i) we discuss how the affiliation of the system state to a discrete partition of phase space can be used as a surrogate variable replacing a high-dimensional vector of initial data in relative entropy functionals measuring predictability and model error. In (ii) the coarse-grained geometry of the data will enter through discrete diffusion operators constructed using dynamics-adapted kernels to provide basis functions for temporal modes of variability analogous to linear-projection coordinates in principal components analysis (PCA, e.g., [1]).

Throughout, we illustrate these techniques with applications to CAOS. In particular, in (i) we study long-range predictability in a simple model [20] of ocean circulation in an idealized basin featuring a current analogous to the Gulf Stream and the Kuroshio Current in the Atlantic and Pacific Oceans, respectively. Such currents are known to undergo changes in configuration affecting continental-scale climate patterns on timescales spanning several months to years. Revealing the predictability of these circulation regimes is important for making skillful initial-value decadal forecasts [21]. In (ii) we present an application of NLSA to a complex spatiotemporal signal of infrared brightness temperature (T_b) acquired through satellites (the CLAUS archive [22]). Because T_b is a good proxy variable for atmospheric convection (deep-penetrating clouds are cold, and therefore produce a strong T_b signal against the emission background from the Earth's surface), an objective decomposition of such data can provide important information about a plethora of climatic processes, including ENSO, the MJO, as well as diurnal-scale processes. This application of NLSA to two-dimensional CLAUS data has not been previously published.

We conclude in Section 4 with a synthesis discussion of open problems and possible connections between the two topics.

2 Quantifying long-range predictability and model error through data clustering and information theory

2.1 Background

Since the classical work of Lorenz [23] and Epstein [24], predictability within dynamical systems has been the focus of extensive study. In the applications outlined in Section 1, the dynamics span multiple spatial and temporal scales, take place in phase spaces of large dimension, and are strongly mixing. Yet, despite the complex underlying dynamics, several phenomena of interest are organized around a relatively small number of persistent states (so-called regimes), which are predictable over timescales significantly longer than suggested by decorrelation times or Lyapunov exponents. Such phenomena often occur in these applications in variables with nearly Gaussian equilibrium statistics [25, 26] and with dynamics that is very different [27] from the more familiar gradient flows, (arising, e.g., in molecular dynamics), where long-range predictability

also often occurs [5, 28]. In certain cases, such as CAOS [29, 30] and econometrics [31], seasonal effects play an important role, resulting in time-periodic statistics. In either case, revealing predictability in these systems is important from both a practical and a theoretical standpoint.

Another issue of key importance is to quantify the fidelity of predictions made with imperfect models when (as is usually the case) the true dynamics of nature cannot be feasibly integrated, or are simply not known [32, 33]. Prominent techniques for building imperfect predictive models of regime behavior include finite-state methods, such as hidden hidden Markov models [26, 34] and cluster-weighted models [35], as well as continuous models based on approximate equations of motion, e.g., linear inverse models [36, 37] and stochastic mode elimination [38]. Other methods blend aspects of finite-state and continuous models, employing clustering algorithms to derive a continuous local model for each regime, together with a finite-state process describing the transitions between regimes [30, 39–41].

The fundamental perspective adopted here is that predictions in dynamical systems correspond to transfer of information; specifically, transfer of information between the initial data (which in general do not suffice to completely determine the state of the system) and a target variable to be forecast. This opens up the possibility of using the mathematical framework of information theory to characterize both predictability and model error [32, 33, 37, 42–49].

The prototypical problem we wish to address here is illustrated in Figure 2.1. There, our prior knowledge about an observable a_t to be predicted at time t is represented by a distribution $p(a_t)$, which in general may be time-dependent. For instance, if a_t is the temperature measured over a geographical region, then time-dependence of $p(a_t)$ would be the due to the seasonality of the Earth's climate, or an underlying slow trend occurring in a climate change scenario. Contrasted with $p(a_t)$ is the posterior distribution $p(a_t \mid X_0)$ representing our knowledge about a_t given that we have observed initial data X_0 at time $t = 0$. In the temperature forecasting example, atmospheric variables such wind fields, pressure, and moisture, as well oceanic circulation, would all be employed to make an initial-value forecast about a_t. At short times, one would expect $p(a_t \mid X_0)$ to depend very strongly on the initial data, and have mass concentrated in a significantly narrower range of a_t values than the prior distribution. This situation, where the availability of highly-resolved initial data plays a crucial role, has been termed by Lorenz [50] as a predictability problem of the first kind. On the other hand, due to mixing dynamics, the predictive information contributed by X_0 is expected to decay at late times, and eventually $p(a_t \mid X_0)$ will converge to $p(a_t)$. In this second-kind predictability problem, knowledge of the "boundary conditions" is important. In climate science, boundary conditions would include anthropogenic and volcanic emissions, changes in solar insolation, etc. At the interface between these two types of predictability problems lie long-range initial-value forecasts (e.g., [21]), which will be the focus of the work presented here. Here, the forecast lead time is short-enough so that $p(a_t \mid X_0)$ differs significantly from $p(a_t)$, but long-enough so that fine-grained aspects of the initial data contribute little predictive information beyond forecasts with coarse-grained initial data.

In all of the cases discussed above, a common challenge is that the initial data X_0 are generally high-

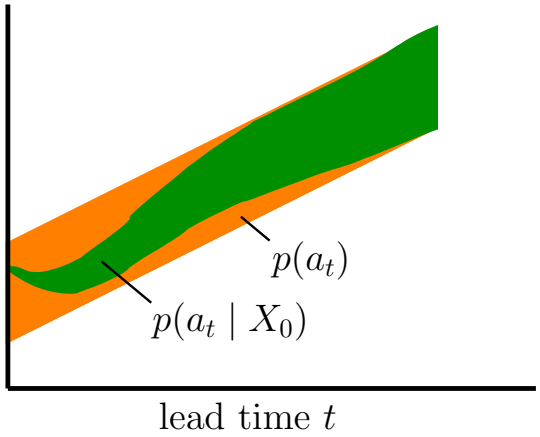

Figure 2.1. Illustration of a statistical forecasting problem. Figure adopted from [51].

dimensional even if the target observable a_t is a scalar. Indeed, in present-day numerical weather and climate forecasting one has to assimilate a very comprehensive set of initial data if only to produce a point forecast about a quantity of interest. Here, we advocate that the information-theoretic framework can be combined with data clustering algorithms to produce lower bounds to intrinsic predictability and model error, which are practically computable for high-dimensional initial data. These bounds are derived by replacing the high-dimensional space of initial data by the integer-valued affiliation function to a coarse-grained partition of that space, constructed by clustering a training dataset generated by a potentially imperfect model. The algorithm for building the partition can be general and designed according to the problem at hand—below, we describe a concrete scheme based on K-means clustering augmented with a notion of temporal persistence in cluster affiliation through running averages of initial data. We apply this scheme to study long-range predictability in an equivalent barotropic, double-gyre model of ocean circulation, and the fidelity of coarse-grained Markov models for ocean circulation regime transitions.

2.2 Information theory, predictability, and model error

2.2.1 Predictability in a perfect-model environment

We consider the general setting of a stochastic dynamical system

$$dz = F(z,t)\,dt + G(z,t)\,dW \quad \text{with} \quad z \in \mathbb{R}^m, \tag{2.1}$$

which is observed through (typically, incomplete) measurements

$$x(t) = H(z(t)), \quad x(t) \in \mathbb{R}^n, \quad n \leq m. \tag{2.2}$$

As reflected by the explicit dependence of the deterministic and stochastic coefficients in (2.1) on time and the state vector, the dynamics of $z(t)$ may be non-stationary and forced by non-additive noise. However, in the application of Section 2.5, the dynamics will be deterministic with time-independent equilibrium statistics. In particular, $z(t)$ given by the streamfunction ψ of an equivalent-barotropic ocean model, and H will be a projection operator from z to the leading 20 principal components (PCs) of ψ. We refer the reader to [13] for applications involving non-stationary stochastic dynamics.

Let $a_t = a(z(t))$ be a target variable for prediction which can be expressed as as a function of the state vector. Let also

$$X_t = (x(t), x(t-\delta t), \ldots, x(t-(q-1)\,\delta t)), \quad X_t \in \mathbb{R}^N, \quad N = qn, \quad q \geq 1, \tag{2.3}$$

with $x(t_i)$ given from (2.2), be a history of observations collected over a time window $\Delta t = (q-1)\,\delta t$. Hereafter, we refer to the observations X_0 at time $t = 0$ as initial data. Broadly speaking, the question of dynamical predictability in the setting of (2.1) and (2.2) may be posed as follows: Given the initial data, how much information have we gained about a_t at time $t > 0$ in the future? Here, uncertainty in a_t arises because of both the incomplete nature of the measurements in (2.2) and the stochastic component of the dynamical system in (2.1). Thus, it is appropriate to describe a_t via some time-dependent probability distribution $p(a_t \mid X_0)$ conditioned on the initial data. Predictability of a_t is understood in this context as the additional information contained in $p(a_t \mid X_0)$ relative to the prior distribution [14, 45, 47], $p(a_t) = \mathbb{E}_{X_0} p(a_t \mid X_0)$, which we now specify.

Throughout, we consider that our knowledge of the system before the observations become available is described by a statistical equilibrium state $p_{\text{eq}}(z(t))$, which is may be time-dependent (e.g., time-periodic [13]). An assumption made here when $p_{\text{eq}}(z(t))$ is time-independent is that $z(t)$ is ergodic, with

$$\mathbb{E}_{p_{\text{eq}}} a_t \approx \frac{1}{s} \sum_{i=0}^{s-1} a(z(t - i\,\delta t)) \tag{2.4}$$

for a large-enough number of samples s. In all of these cases, the prior distributions for a_t and X_t are the distributions $p_{\text{eq}}(a_t)$ and $p_{\text{eq}}(X_t)$ induced on these variables by $p_{\text{eq}}(z(t))$, i.e.,

$$p(a_t) = p_{\text{eq}}(a_t), \quad p(X_t) = p_{\text{eq}}(X_t). \tag{2.5}$$

As the forecast lead time grows, $p(a_t \mid X_0)$ converges to $p_{\text{eq}}(a_t)$, at which point X_0 contributes no additional information about a_t beyond equilibrium.

The natural mathematical framework to quantify predictability in this setting is information theory [52], and, in particular, the concept of relative entropy. The latter is defined as the functional

$$\mathcal{P}(p'(a_t), p(a_t)) = \mathbb{E}_{p'} \log(p'(a_t)/p(a_t)) \tag{2.6}$$

between two probability measures, $p'(a_t)$ and $p(a_t)$, and has the attractive properties that (i) it vanishes if and only if $p = p'$, and is positive if $p \neq p'$; (ii) is invariant under general invertible transformations of a_t. For our purposes, of key importance is also the so-called Bayesian-update interpretation of relative entropy. This states that if $p'(a_t) = p(a_t \mid X_0)$ is the posterior distribution of a_t conditioned on some variable X_0 and p is the corresponding prior distribution, then $\mathcal{P}(p'(a_t), p(a_t))$ measures the additional information beyond p about a_t gained by having observed X_0. This interpretation stems from the fact that

$$\mathcal{P}(p(a_t \mid X_0), p(a_t)) = \mathbb{E}_{a_t|X_0} \log(p(a_t \mid X_0)/p(a_t)) \tag{2.7}$$

is a non-negative quantity (by Jensen's inequality), measuring the expected reduction in ignorance about a_t relative to the prior distribution $p(a_t)$ when X_0 has become available [32, 52]. It is therefore crucial that $p(a_t \mid X_0)$ is inserted in the first argument of $\mathcal{P}(\cdot, \cdot)$ for a correct assessment of predictability.

The natural information-theoretic measure of predictability compatible with the prior distribution $p(a_t)$ in (2.5) is

$$\mathcal{D}(a_t \mid X_0) = \mathcal{P}(p(a_t \mid X_0), p(a_t)). \tag{2.8}$$

As one may explicitly verify, the expectation value of $\mathcal{D}(a_t \mid X_0)$ with respect to the prior distribution for X_0,

$$\mathcal{D}(a_t, X_0) = \mathbb{E}_{X_0} \mathcal{D}(a_t \mid X_0) = \mathbb{E}_{a_t, X_0} \log(p(a_t \mid X_0)/p(a_t)) = \mathcal{P}(p(a_t, X_0), p(a_t)p(X_0)) \tag{2.9}$$

is also a relative entropy; here, between the joint distribution of the target variable and the initial data and the product of their marginal distributions. This quantity is known as the mutual information between a_t and X_0, measuring the expected predictability of the target variable over the initial data [14, 43, 47].

One of the classical results in information theory is that the mutual information between the source and output of a channel measures the rate of information flow across the channel [52]. The maximum mutual information over the possible source distributions corresponds to the channel capacity. In this regard, an interesting parallel between prediction in dynamical systems and communication across channels is that the combination of dynamical system and observation apparatus [represented here by (2.1) and (2.2)] can be thought of as an abstract communication channel with the initial data X_0 as input and the target a_t as output.

2.2.2 Quantifying the error of imperfect models

The analysis in Section 2.2.1 was performed in a perfect-model environment. Frequently, however, instead of the true forecast distributions $p(a_t \mid X_0)$, one has access to distributions $p^M(a_t \mid X_0)$ generated by an imperfect model,

$$dz(t) = F^M(z, t)\, dt + G^M(z, t)\, dW. \tag{2.10}$$

Such situations arise, for instance, when one cannot afford to feasibly integrate the full dynamical system in (2.1) (e.g., simulations of biomolecules dissolved in a large number of water molecules), or the laws governing $z(t)$ are simply not known (e.g., condensation mechanisms in atmospheric clouds). In other cases, the objective is to develop reliable reduced models for $z(t)$ to be used as components of coupled models (e.g., parameterization schemes in climate models [53]). In this context, assessments of the error in the model prediction distributions are of key importance, but frequently not carried out in an objective manner that takes into account both the mean and variance [33].

Relative entropy again emerges as the natural information-theoretic functional for quantifying model error. Now, the analog between dynamical systems and coding theory is with suboptimal coding schemes. In coding theory, the expected penalty in the number of bits needed to encode a string assuming that it is drawn from a probability distribution q, when in reality the source probability distribution is p', is given by $\mathcal{P}(p', q)$

(evaluated in this case with base-2 logarithms). Similarly, $\mathcal{P}(p', q)$ with p' and q equal to the distributions of a_t conditioned on X_0 in the perfect and imperfect model, respectively, leads to the error measure

$$\mathcal{E}(a_t \mid X_0) = \mathcal{P}(p(a_t \mid X_0), p^M(a_t \mid X_0)). \tag{2.11}$$

By direct analogy with (2.7), $\mathcal{E}(a_t \mid X_0)$ is a non-negative quantity measuring the expected increase in ignorance about a_t incurred by using the imperfect model distribution $p^M(a_t \mid X_0)$ when the true state of the system is given by $p(a_t \mid X_0)$ [32, 33, 46]. As with (2.8), $p(a_t \mid X_0)$ must appear in the first argument of $\mathcal{P}(\cdot, \cdot)$ for a correct assessment of model error. Moreover, $\mathcal{E}(a_t \mid X_0)$ may be aggregated into an expected model error over the initial data,

$$\mathcal{E}(a_t, X_0) = \mathbb{E}_{X_0}\mathcal{E}(a_t \mid X_0) = \mathbb{E}_{a_t, X_0} \log(p(a_t \mid X_0)/p^M(a_t \mid X_0)). \tag{2.12}$$

However, unlike $\mathcal{D}(a_t, X_0)$ in (2.9), $\mathcal{E}(a_t, X_0)$ does not correspond to a mutual information between random variables.

Note that by writing down (2.11) and (2.12) we have tacitly assumed that the target variable can be simultaneously defined in the perfect and imperfect models, i.e., a_t can be expressed as a function of either $z(t)$ or $z^M(t)$. Even though z and z^M may lie in completely different phase spaces, in practice one is typically interested in large-scale coarse-grained target variables (e.g., the mean temperature over a geographical region of interest), which are well-defined in both the perfect and imperfect model.

2.3 Coarse-graining phase space to reveal long-range predictability

Despite their theoretical appeal, the predictability and model error measures $\mathcal{D}(a_t \mid X_0)$ and $\mathcal{E}(a_t \mid X_0)$ are frequently infeasible to evaluate in practice, the reason being that both of these measures require the evaluation of an expectation value over the initial data X_0. As stated in Section 2.1, the spaces of initial data used for making predictions in complex systems are generally high-dimensional, even if the target observable a_t is a scalar. Operationally, computing the expectation $\mathbb{E}_{X_0}$ requires evaluation of an integral over X_0 that rapidly becomes intractable as the dimension of X_0 grows. Here, we address this "curse of dimension" issue by replacing X_0 with an integer-valued surrogate variable S_0 representing the affiliation of X_0 in a partition of the initial-data space. By the data-processing inequality in information theory [52], the coarse-grained predictability and model error metrics $\mathcal{D}(a_t, S_0)$ and $\mathcal{E}(a_t, S_0)$, respectively, provide lower bounds to $\mathcal{D}(a_t, X_0)$ and $\mathcal{E}(a_t, X_0)$ which are practically computable for high-dimensional initial data.

2.3.1 Perfect-model scenario

Our method of partitioning the space of initial data, described also in [13, 13, 14], proceeds in two stages: a training stage and prediction stage. The training stage involves taking a dataset

$$\mathcal{X} = \{x(0), x(\delta t), \ldots, x((s-1)\,\delta t\}, \tag{2.13}$$

of s observation samples $x(t) \in \mathbb{R}^n$ and computing via data clustering a collection

$$\Theta = \{\theta_1, \ldots, \theta_K\}, \quad \theta_k \in \mathbb{R}^p. \tag{2.14}$$

of parameter vectors θ_k characterizing the clusters. Used in conjunction with a rule [e.g., (2.42) ahead], for determining the integer-valued affiliation $S(X_0)$ of a vector X_0 from (2.3), the cluster parameters lead to a mutually-disjoint partition of the set of initial data, viz.

$$\Xi = \{\xi_1, \ldots, \xi_K\}, \quad \xi_k \subset \mathbb{R}^N, \tag{2.15}$$

such that $S(X_0) = S_0$ indicates that the membership of X_0 is with cluster $\xi_{S_0} \in \Xi$. Thus, a dynamical regime is understood here as an element ξ_k of Ξ, and coarse-graining as a projection $X_0 \mapsto S_0$ from the (generally, high-dimensional) space of initial data to the integer-valued membership S_0 in the partition. It is important to note that $\mathcal{X}$ may consist of either observations $x(t)$ of the perfect model from (2.2), or data generated by an imperfect model [which does not have to be the same as the model in (2.10) used for prediction]. In the

latter case, the error in the training data influences the amount of information loss by coarse graining, but does not introduce biases that would lead one to overestimate predictability.

Because $S_0 = S(X_0)$ is uniquely determined from X_0, it follows that

$$p(a_t \mid X_0, S_0) = p(a_t \mid X_0). \tag{2.16}$$

The above expresses the fact no additional information about the target variable a_t is gained through knowledge of S_0 if X_0 is known. Moreover, (2.16) leads to a Markov property between the random variables a_t, X_0, and S_0, viz.

$$p(a_t, X_0, S_0) = p(a_t \mid X_0, S_0)p(X_0 \mid S_0)p(S_0) = p(a_t \mid X_0)p(X_0 \mid S_0)p(S_0). \tag{2.17}$$

The latter is a necessary condition for the predictability and model error bounds discussed below.

Equation (2.16) also implies that the forecasting scheme based on X_0 is statistically sufficient [54, 55] for the scheme based on S_0. That is, the predictive distribution $p(a_t \mid S_0)$ conditioned on the coarse-grained initial data can be expressed as an expectation value

$$p(a_t \mid S_0) = \mathbb{E}_{X_0|S_0} p(a_t \mid X_0) \tag{2.18}$$

of $p(a_t \mid X_0)$ with respect to the distribution $p(X_0 \mid S_0)$ of the fine-grained initial data X_0 given S_0. We use the shorthand notation

$$p_t^k = p(a_t \mid S_0 = k), \tag{2.19}$$

for the forecast distribution for a_t conditioned on the k-th cluster.

In the prediction stage, the p_t^k are estimated for each $k \in \{1, \ldots, K\}$ by bin-counting joint realizations of a_t and S_0, using data which are independent from the dataset $\mathcal{X}$ employed in the training stage (details about the bin-counting procedure are provided in Section 2.4). The predictive information content in the partition is then measured via coarse-grained analogs of the relative-entropy metrics in (2.8) and (2.9), viz.,

$$\mathcal{D}(a_t \mid S_0) = \mathcal{P}(p(a_t \mid S_0), p(a_t)) \quad \text{and} \quad \mathcal{D}(a_t, S_0) = \mathbb{E}_{S_0} \mathcal{D}(a_t \mid S_0). \tag{2.20}$$

By the same arguments used to derive (2.9), it follows that the expected predictability measure $\mathcal{D}(a_t, S_0)$ is equal to the mutual information between the target variable a_t at time $t \geq 0$ and the membership S_0 of the initial data in the partition at time $t = 0$. Note the formula

$$\mathcal{D}(a_t, S) = \sum_{k=1}^{K} \pi_k \mathcal{D}_t^k, \quad \text{with} \quad \mathcal{D}_t^k = \mathcal{P}(p_t^k, p_{\text{eq}}), \quad \pi_k = p(S = k). \tag{2.21}$$

Two key properties of $\mathcal{D}(a_t, S)$ are:

(i) It provides a lower bound to the predictability measure $\mathcal{D}(a_t, X_0)$ in (2.9) determined from the fine-grained initial data X_0, i.e.,

$$\mathcal{D}(a_t, X_0) \geq \mathcal{D}(a_t, S_0); \tag{2.22}$$

(ii) Unlike $\mathcal{D}(a_t, X_0)$, which requires evaluation of an integral over X_0 that rapidly becomes intractable as the dimension of X_0 grows (even if the target variable is scalar), $\mathcal{D}(a_t, S_0)$ only requires evaluation of a discrete sum over S_0.

Equation (2.22), which is known in information theory as data-processing inequality [14, 48], expresses the fact that coarse-graining, $X_0 \mapsto S(X_0)$, can only lead to conservation or loss of information. In particular, it can be shown [13] that the Markov property in (2.17) leads to the relation

$$\mathcal{D}(a_t, X_0) = \mathcal{D}(a_t, S_0) + \mathcal{I}, \tag{2.23}$$

where

$$\mathcal{I} = \mathbb{E}_{a_t, X_0, S_0} \log(p(X_0 \mid a_t, S_0)/p(X_0 \mid S_0)) \tag{2.24}$$

is a non-negative term measuring the loss of predictive information due to coarse-graining of the initial data. Because the non-negativity of $\mathcal{I}$ relies only on the existence of a coarse-graining function meeting the

condition (2.16), and not on the properties of the training data $\mathcal{X}$ used to construct that function, there is no danger of over-estimating predictability through $\mathcal{D}(a_t, S_0)$, even if an imperfect model is employed to generate $\mathcal{X}$. Thus, $\mathcal{D}(a_t, S_0)$ can be used practically as a sufficient condition for predictability, irrespective of model error in $\mathcal{X}$ and/or suboptimality of the clustering algorithm.

In general, the information loss $\mathcal{I}$ will be large at short lead times, but in many applications involving strongly-mixing dynamical systems, the predictive information in the fine-grained aspects of the initial data will rapidly decay as t grows. In such scenarios, $\mathcal{D}(a_t, S_0)$ provides a tight bound to $\mathcal{D}(a_t, X_0)$, with the crucial advantage of being feasibly computable with high-dimensional initial data. Of course, failure to establish predictability on the basis of $\mathcal{D}(a_t, S_0)$ does not imply absence of intrinsic predictability, for it could be that $\mathcal{D}(a_t, S_0)$ is small because $\mathcal{I}$ is comparable to $\mathcal{D}(a_t, X_0)$.

Since relative entropy is unbounded from above, it is useful to convert $\mathcal{D}(a_t, S_0)$ into a predictability score lying in the unit interval,

$$\delta_t = 1 - \exp(-2\mathcal{D}(a_t, S_0)). \tag{2.25}$$

Joe [56] shows that the above definition for δ_t is equivalent to a squared correlation measure, at least in problems involving Gaussian random variables.

2.3.2 Quantifying the model error in long-range forecasts

Consider now an imperfect model that, as described in Section 2.2.2, produces prediction distributions

$$p_t^{Mk} = p^M(a_t \mid S_0 = k) \tag{2.26}$$

which may be systematically biased away from p_t^k in (2.19). Similarly to Section 2.3.1, we consider that the random variables a_t, X_0, and S_0 in the imperfect model have a Markov property,

$$p^M(a_t, X_0, S) = p^M(a_t \mid X_0, S_0)p(X_0 \mid S_0)p(S_0) = p^M(a_t \mid X_0)p(X_0 \mid S_0)p(S_0), \tag{2.27}$$

where we have also assumed that the same initial data and cluster affiliation function are employed to compare the perfect and imperfect models [i.e., $p^M(X_0 \mid S_0) = p(X_0 \mid S_0)$ and $p^M(S_0) = p(S_0)$]. As a result, the coarse-grained forecast distributions in (2.26) can be determined via [cf. (2.18)]

$$p^M(a_t \mid S_0) = \mathbb{E}_{X_0|S_0} p^M(a_t \mid X_0). \tag{2.28}$$

In this setup, an obvious candidate measure for predictability follows by writing down (2.20) with p_t^k replaced by p_t^{Mk}, i.e.,

$$\mathcal{D}^M(a_t, S_0) = \mathbb{E}_{S_0}\mathcal{D}^M(a_t \mid S_0) = \sum_{k=1}^{K} \pi_k \mathcal{D}_t^{Mk}, \quad \text{with} \quad \mathcal{D}_t^{Mk} = \mathcal{P}(p_t^{Mk}, p_{\text{eq}}^M). \tag{2.29}$$

By direct analogy with (2.22), $\mathcal{D}^M(a_t, S_0)$ is a non-negative lower-bound of $\mathcal{D}^M(a_t, X_0)$. Clearly, an important deficiency of this measure is that by being based solely on forecast distributions internal to the model it fails to take into account model error, or "ignorance" of the imperfect model in (2.10) relative to the perfect model in (2.1) [15, 32, 33]. Nevertheless, $\mathcal{D}^M(a_t, S_0)$ provides an additional metric to discriminate between imperfect models with similar $\mathcal{E}(a_t, X_0)$ scores from (2.12), and estimate how far a given imperfect forecast is from the model's equilibrium distribution. For the latter reasons, we include $\mathcal{D}^M(a_t, S_0)$ as part of our model assessment framework. Following (2.25), we introduce for convenience a unit-interval normalized score,

$$\delta_t^M = 1 - \exp(-2\mathcal{D}^M(a_t, S_0)). \tag{2.30}$$

Next, note the distinguished role that the imperfect-model equilibrium distribution plays in (2.29): If $p_{\text{eq}}^M(a_t)$ differs systematically from the equilibrium distribution $p_{\text{eq}}(a_t)$ in the perfect model, then $\mathcal{D}^M(a_t, S_0)$ conveys false predictability at *all* times (including $t = 0$), irrespective of the fidelity of $p^M(a_t \mid S_0)$ at finite times. This observation leads naturally to the requirement that long-range forecasting models must reproduce the equilibrium statistics of the perfect model with high fidelity. In the information-theoretic framework of Section 2.2.2, this is expressed as

$$\varepsilon_{\text{eq}} \ll 1, \quad \text{with} \quad \varepsilon_{\text{eq}} = 1 - \exp(-2\mathcal{E}_{\text{eq}}(a_t)) \tag{2.31}$$

and

$$\mathcal{E}_{\rm eq}(a_t) = \mathcal{P}(p_{\rm eq}(a_t), p^M_{\rm eq}(a_t)). \tag{2.32}$$

Here, we refer to the criterion in (2.31) as equilibrium consistency; an equivalent condition is called fidelity [57], or climate consistency [15] in CAOS work.

Even though equilibrium consistency is a necessary condition for skillful long-range forecasts, it is not a sufficient condition. In particular, the model error $\mathcal{E}(a_t, X_0)$ at finite lead time t may be large, despite eventually decaying to a small value at asymptotic times. The expected error in the coarse-grained forecast distributions is expressed in direct analogy with (2.12) as

$$\mathcal{E}(a_t, S_0) = \mathbb{E}_{S_0}\mathcal{E}(a_t \mid S_0) = \sum_{k=1}^{K} \pi_k \mathcal{E}^k_t, \quad \text{with} \quad \mathcal{E}^k_t = \mathcal{P}(p^k_t, p^{Mk}_t), \tag{2.33}$$

and corresponding error score

$$\varepsilon_t = 1 - \exp(-2\mathcal{E}^K_t), \quad \varepsilon_t \in [0, 1). \tag{2.34}$$

Similar arguments to those used to derive (2.23) lead to a decomposition [13]

$$\mathcal{E}(a_t, X_0) = \mathcal{E}(a_t, S_0) + \mathcal{I} - \mathcal{J} \tag{2.35}$$

of the model error $\mathcal{E}(a_t, X_0)$ into the coarse-grained measure $\mathcal{E}(a_t, S_0)$, the information loss term $\mathcal{I}$ due to coarse graining in (2.24), and a term

$$\mathcal{J} = \mathbb{E}_{a_t, X_0, S_0} \log(p^M(a_t \mid X_0)/p^M(a_t \mid S_0)) \tag{2.36}$$

reflecting the relative ignorance of the fine-grained and coarse-grained forecast distributions in the imperfect model. The important point about $\mathcal{J}$ is that it obeys the bound [13]

$$\mathcal{J} \leq \mathcal{I}. \tag{2.37}$$

As a result, $\mathcal{E}(a_t, S_0)$ is a lower bound of the fine-grained error measure $\mathcal{E}(a_t, X_0)$ in (2.12), i.e.,

$$\mathcal{E}(a_t, X_0) \geq \mathcal{E}(a_t, S_0). \tag{2.38}$$

Because of (2.38), a detection of a significant $\mathcal{E}(a_t, S_0)$ is sufficient to reject a forecasting scheme based on the fined-grained distributions $p^M(a_t \mid X_0)$. The reverse statement, however, is generally not true. In particular, the error measure $\mathcal{E}(a_t, X_0)$ may be significantly larger than $\mathcal{E}(a_t, S_0)$, even if the information loss $\mathcal{I}$ due to coarse-graining is small. Indeed, unlike $\mathcal{I}$, the $\mathcal{J}$ term in (2.35) is not bounded from below, and can take arbitrarily large negative values. This is because the coarse-grained forecast distributions $p^M(a_t \mid S_0)$ are determined through (2.28) by averaging the fine-grained distributions $p^M(a_t \mid X_0)$, and averaging can lead to cancellation of model error. Such a situation with negative $\mathcal{J}$ cannot arise with the forecast distributions of the perfect model, where, as manifested by the non-negativity of $\mathcal{I}$, coarse-graining can at most preserve information.

In summary, our framework for assessing long-range coarse-grained forecasts with imperfect models takes into consideration all of $\varepsilon_{\rm eq}$, ε_t, and δ^M_t as follows:

- $\varepsilon_{\rm eq}$ must be small, i.e., the imperfect model should be able to reproduce with high fidelity the distribution of the target variable a_t at asymptotic times (the prior distribution, relative to which long-range predictability is measured).
- The imperfect model must have correct statistical behavior at finite times, i.e., ε_t must be small at the forecast lead time of interest.
- At the forecast lead time of interest, the additional information beyond equilibrium δ^M_t must be large, otherwise the model has no utility compared with a trivial forecast drawn for the equilibrium distribution.

In order to evaluate these metrics in practice, the following two ingredients are needed. (i) The training dataset $\mathcal{X}$ in (2.13), to compute the cluster parameters Θ from (2.14). (ii) Simultaneous realizations of a_t (in both the perfect and imperfect models) and $x(t)$ [which must be statistically independent from the data in (i)], to evaluate the cluster-conditional distributions p_t^k and p_t^{Mk}. Note that neither access to the full state vectors $z(t)$ and $z^M(t)$ of the perfect and imperfect models, nor knowledge of the equations of motions is required to evaluate the predictability and model error scores proposed here. Moreover, the training dataset $\mathcal{X}$ can be generated by an imperfect model. The resulting partition in that case will generally be less informative in the sense of the $\mathcal{D}(a_t, S_0)$ and $\mathcal{E}(a_t, S_0)$ metrics, but, so long as (ii) can be carried out with small sampling error, $\mathcal{D}(a_t, S_0)$ and $\mathcal{E}(a_t, S_0)$ will still be lower bounds of $\mathcal{D}(a_t, X_0)$ and $\mathcal{E}(a_t, X_0)$, respectively. See [13] for an example where $\mathcal{D}(a_t, S_0)$ and $\mathcal{E}(a_t, S_0)$ reveal long-range predictability and model error despite substantial model error in the training data.

2.4 K-means clustering with persistence

We now describe a method based on K-means clustering and running-average smoothing of training and initial data that is able to reveal predictability beyond decorrelation time in the ocean model in Section 2.5, as well as in stochastic models with nonlinearities [14]. Besides the number of clusters (regimes) K, our algorithm has two additional free parameters. These are temporal windows, $\Delta t'$ and Δt, used to take running averages of $x(t)$ in the training and prediction stages, respectively. This procedure, which is reminiscent of kernel density estimation methods [58], leads to a two-parameter family of partitions as follows.

First, set an integer $q' \geq 1$, and replace $x(t)$ in (2.13) with the averages over a time window $\Delta t' = (q'-1)\,\delta t$, i.e.,

$$x^{\Delta t'}(t) = \sum_{i=1}^{q'} x(t-(i-1)\,\delta t)/q'. \tag{2.39}$$

Next, apply K-means clustering [59] to the above coarse-grained training data. This leads to a set of parameters Θ from (2.14) that minimize the sum-of-squares error functional,

$$L(\Theta) = \sum_{k=1}^{K} \sum_{i=q'-1}^{s-1} \gamma_k(i\,\delta t) \|x^{\Delta t'}(i\,\delta t) - \theta_k^{\Delta t'}\|_2^2, \tag{2.40}$$

where

$$\gamma_k(t) = \begin{cases} 1, & k = \Gamma(t), \\ 0, & \text{otherwise}, \end{cases} \qquad \Gamma(t) = \underset{j}{\operatorname{argmin}} \|x^{\Delta t'}(t) - \theta_j^{\Delta t'}\|_2, \tag{2.41}$$

is the weight of the k-th cluster at time $t = i\,\delta t$, and $\|v\|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$ denotes the Euclidean norm. Note that temporal persistence of $\Gamma(t)$ is an outcome of running-average smoothing of the training data.

In the second (prediction) stage of the procedure, data $X_0 = (x(0), x(-\delta t), \ldots, x(-(q-1)\,\delta t))$ of the form (2.3) are collected over an interval $[-\Delta t, 0]$ with $\Delta t = (q-1)\,\delta t$, and their average $x^{\Delta t}(0)$ is computed via an analogous formula to (2.39). It is important to note that the initial data X_0 used in the prediction stage are independent of the training dataset. The affiliation function S is then given by

$$S(X_0) = \underset{k}{\operatorname{argmin}}(\|x^{\Delta t}(0) - \theta_k^{\Delta t'}\|_2); \tag{2.42}$$

i.e., $S(X_0)$ depends on both Δt and $\Delta t'$. Because $x^{\Delta t}$ can be uniquely determined from the initial-data vector X_0, (2.42) provides a mapping from X_0 to $\{1, \ldots, K\}$, defining the elements of the partition in (2.15) through

$$\xi_k = \{X_t : S(X_t) = k\}. \tag{2.43}$$

Physically, the width of Δt controls the influence of the past history of the system relative to its current state in assigning cluster affiliation. If the target variable exhibits significant memory effects, taking the running average over a window comparable to the memory time scale should lead to gains of predictive information $\mathcal{D}(a_t, S_0)$, at least for lead times of order Δt or less. We provide an example of this behavior in Section 2.5.

For ergodic dynamical systems satisfying (2.4), the cluster-conditional densities p_t^k in (2.19) may be estimated as follows. First, obtain a sequence of observations $x(t')$ [independent of the training dataset $\mathcal{X}$ in (2.13)] and the corresponding time series $a_{t'}$ of the target variable. Second, using (2.42), compute the membership sequence $S_{t'} = S(X_{t'})$ for every time t'. For given lead time t, and for each $k \in \{1, \ldots, K\}$, collect the values

$$\mathcal{A}_t^k = \{a_{t+t'} : S_{t'} = k\}. \tag{2.44}$$

Then, set distribution bin boundaries $A_0 < A_1 < \ldots$, and compute the occurrence frequencies

$$\hat{p}_t^k(i) = N_i/N, \tag{2.45}$$

where N_i is the number of elements of $\mathcal{A}_t^k$ lying in $[a_{i-1}, a_i]$, and $N = \sum_i N_i$. Note that the A_i are vector-valued if a_t is multi-variate. By ergodicity, in the limit of an infinite number of bins and samples, the estimators $\hat{p}_t^k(i)$ converge to the continuous densities p_t^k in (2.19). The equilibrium distribution $p_{\rm eq}(a_t)$ and the cluster affiliation probabilities π_k in (2.21) may be evaluated in a similar manner. Together, the estimates for p_t^k, $p_{\rm eq}$, and π_k are sufficient to determine the predictability metrics $\mathcal{D}_t^k$ from (2.20). In particular, if a_t is a scalar variable (as will be the case below), the relative-entropy integrals in (2.20) can be carried out by standard one-dimensional quadrature, e.g., the trapezoidal rule. This simple procedure is sufficient to estimate the cluster-conditional densities with little sampling error for the univariate target variables in Section 2.5. For non-ergodic systems and/or lack of availability of long realizations, more elaborate methods (e.g., [60]) may be required to produce reliable estimates of $\mathcal{D}(a_t, S_0)$.

We close this section with an important point about the forecast distributions from (2.19): Because p_t^k are evaluated independently for each pair $\Delta\mathcal{T} = (\Delta t, \Delta t')$ of running-average intervals, there is no reason why one should use the same $p_t^k|_{\Delta\mathcal{T}}$ for all lead times. In particular, given a collection $\{\Delta\mathcal{T}_1, \Delta\mathcal{T}_2, \ldots\}$ of coarse-graining parameters, the natural forecast distribution to use are the ones that maximize the expected predictability (2.21), viz.

$$p_t^{*k} = p_t^k|_{\Delta\mathcal{T}_i}, \quad i = \operatorname*{argmax}_j \mathcal{D}(a_t, S_0)|_{\Delta\mathcal{T}_j}, \tag{2.46}$$

with corresponding predictability score

$$\mathcal{D}^*(a_t, S_0) = \mathcal{D}(a_t, S_0)|_{\Delta\mathcal{T}_i}, \quad \delta_t^* = 1 - \exp(-2\mathcal{D}^*(a_t, S_0)). \tag{2.47}$$

We will see in Section 2.5 that the p_t^{*k} can contain significantly more information than the individual forecast distributions p_t^k.

2.5 Demonstration in a double-gyre ocean model

The so-called 1.5-layer model [20] describes the dynamics of wind-driven ocean circulation as the motion of two immiscible, vertically-averaged layers of fluid of different density under the influence of wind-induced shear, Coriolis force (in the β-plane approximation), and subgrid-scale diffusion. The lower layer is assumed to be infinitely deep and at rest, whereas the upper layer is governed by a quasigeostrophic equation for the streamfunction $\psi(\mathbf{r}, t)$ (which, in this case is equal to the interface displacement) at position $\mathbf{r} = (x, y)$ and time t, giving the velocity vector $\mathbf{v} = (\partial_y\psi, -\partial_x\psi)$. The kinetic and potential energies, respectively given by $E_{\rm kin} = C_{\rm kin} \int d\mathbf{r}\, \|\mathbf{v}(\mathbf{r}, t)\|^2$ and $E_{\rm pot} = C_{\rm pot} \int d\mathbf{r}\, \psi^2(\mathbf{r}, t)$ with $C_{\rm kin}$, $C_{\rm pot}$ constants, make up the total energy, $E = E_{\rm kin} + E_{\rm pot}$. The latter will be one of our main prediction observables.

We adopt throughout the model parameter values in Section 2 of McCalpin and Haidvogel [20], as well as their canonical asymmetric double-gyre wind forcing. With this forcing, the 1.5-layer model develops an eastward-flowing separating jet configuration analogous to the Gulf Stream in the North Atlantic. Moreover, the model features the essential dynamical mechanisms of equivalent barotropic Rossby waves, lateral shear instability, and damping. The model was integrated by R. Abramov using a pseudospectral code on a 180×140 uniform grid of size $\Delta r = 20$ km, and 4th-order Runge-Kutta timestepping of size $t^+ = 3$ hours. The resulting time-averaged streamfunction and its standard deviation, $\Psi(\mathbf{r}) = \langle\psi(\mathbf{r}, t)\rangle$ and $\sigma(\mathbf{r}) = \langle\psi^2(\mathbf{r}, t) - \Psi^2(\mathbf{r})\rangle^{1/2}$, where $\langle f(t)\rangle = \int_0^T dt\, f(t)/T$ denotes empirical temporal averaging, are shown in Figure 2.2. In that figure, the eastward jet is seen to separate from the western boundary $x = 0$ approximately at meridional coordinate $y = 0$, and to follow a characteristic sinusoid path as it penetrates into the basin. The meridional asymmetry

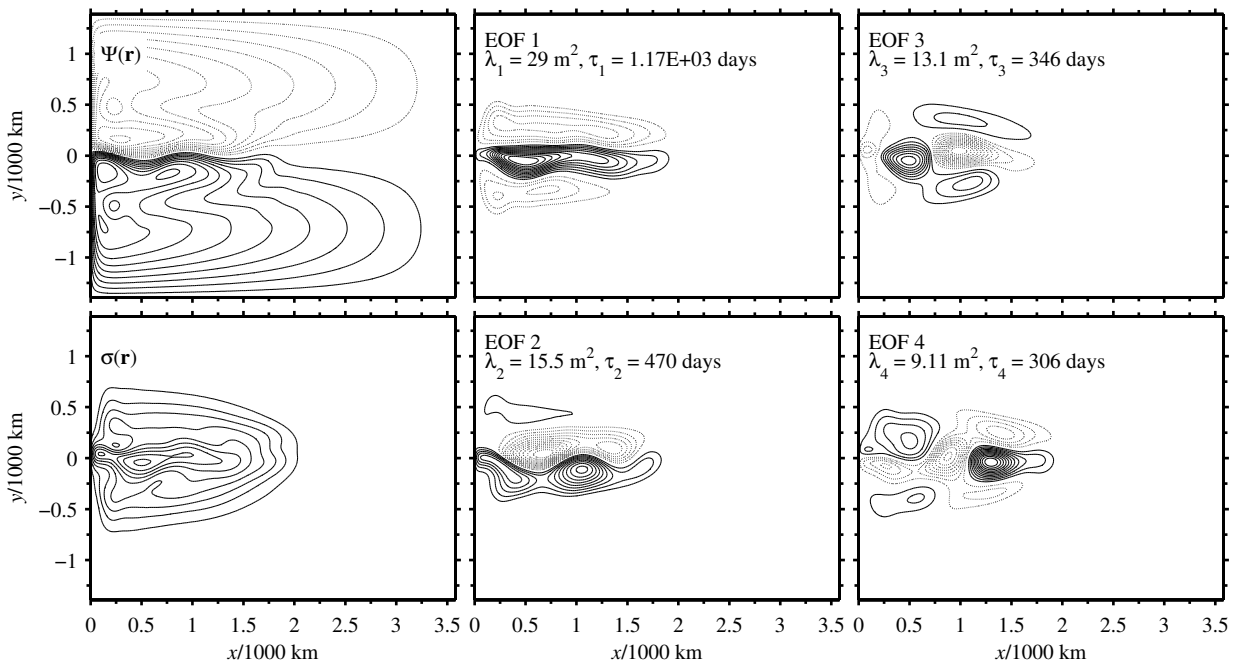

Figure 2.2. The time-averaged state, $\Psi(\mathbf{r})$, its standard deviation, $\sigma(\mathbf{r})$, and the leading four streamfunction-metric EOFs, evaluated using an equilibrated realization of the 1.5-layer model of length $T = 10{,}000$ years sampled every $\delta t = 20$ days. The contour levels in the panels for $\Psi(\mathbf{r})$ and $\sigma(\mathbf{r})$ are spaced by 12.5 m, spanning the interval $[-150, 150]$ m. Contours are drawn every 12.5 arbitrary units in the panels for $\mathrm{EOF}_i(\mathbf{r})$, which also indicate the corresponding eigenvalues and correlation times, respectively λ_i and τ_i. Solid and dotted lines correspond to positive and negative contour levels, respectively. The separation point of the eastward jet is located near the coordinate origin, $\mathbf{r} = (x, y) = (0, 0)$. The eigenvalues and EOFs are the solutions of the eigenproblem $\int d\mathbf{r}'\, C(\mathbf{r}, \mathbf{r}')\mathrm{EOF}_i(\mathbf{r}') = \lambda_i \mathrm{EOF}_i(\mathbf{r})$ associated with the covariance matrix $C(\mathbf{r}, \mathbf{r}') = \int_0^T dt\, \psi'(\mathbf{r}, t)\psi'(\mathbf{r}', t)/T$, where $\psi'(\mathbf{r}, t)$ is the streamfunction anomaly. With this definition, the physical dimension of the λ_i is (length)^2. The correlation times are given by $\tau_i = \int_0^T dt\, |\rho_i(t)|$, where ρ_i is the autocorrelation function of the corresponding PC (see Figure 2.3).

of the wind forcing is manifested in the somewhat stronger anti-cyclonic gyre in the southern portion of the domain.

The phenomenological study of McCalpin and Haidvogel [20] has determined that in this parameter regime the time of viscous decay of westward-propagating eddies can be either small, comparable, or large relative to the drift time taken for the eddies to reach the western boundary current (the drift time increases with the eastward position in the domain where an eddy forms). The eddies that survive long-enough to reach the western meridional boundary perturb the eastward current, resulting in a meander-like pattern. Otherwise, in the absence of eddy interaction, the current penetrates deeply into the basin. As shown in Figure 2.2, most of the variance of the time-averaged state is concentrated in the portion of the domain occupied by the jet.

Because of the intermittent nature of the current-eddy interaction, the model exhibits interesting low-frequency variability, characterized by infrequent transitions between a small number of metastable states. These metastable states may be differentiated by their distinct ranges of energy content (e.g., Figure 2.7). Empirically, three metastable states have been identified, consisting of high, middle, and low-energy configurations [20, 61]. As illustrated in Figure 2.6, the high-energy state is dominated by a strong elongated jet, which penetrates deep into the basin. On the other hand, the jet is significantly weakened in the low-energy state, where the most prominent features are meandering flow structures. The middle-energy state is characterized by a moderately-penetrating jet that correlates strongly with the spatial configuration of the mean state. Yet, in spite of the prominent regime behavior, the equilibrium distributions of the leading PCs and the energy are unimodal (Figure 2.3). Note that regime behavior accompanied by unimodality in the equilibrium

statistics arises more generally in geophysical flows [26].

In what follows, we view the solution of the 1.5-layer model as the true signal (2.1) from nature, i.e., we set $z(t) = \psi(\mathbf{r}, t)$. Moreover, we consider that $z(t)$ is observed through the leading 20 PCs of the streamfunction, $\mathrm{PC}_i(t) = \int d\mathbf{r}\, \mathrm{EOF}_i(\mathbf{r})\psi'(\mathbf{r}, t)$, where $\mathrm{EOF}_i(\mathbf{r})$ is the i-th empirical orthogonal function in the streamfunction metric (see the caption to Figure 2.3 for a definition), and $\psi'(\mathbf{r}, t) = \psi(\mathbf{r}, t) - \Psi(\mathbf{r})$ is the streamfunction anomaly. Thus, the observation vector from (2.2), $x(t) = H(z(t)) = (\mathrm{PC}_1(t), \ldots, \mathrm{PC}_{20}(t))$, is 20-dimensional.

2.5.1 Predictability bounds for coarse-grained observables

For our clustering and forecast distribution calculations we took a time series $x(t)$ consisting of a total of $s = 1.6 \times 10^5$ samples taken uniformly every $\delta t = 20$ days. That is, the total observation time span is $s\,\delta t = 3.2 \times 10^6$ days ≈ 8767 years. Our training dataset $\mathcal{X}$ (2.13) is the first half of that time series, i.e., $t \in [0, T]$, with $T = 1.6 \times 10^6$ days.

The prediction observables a_t considered in this study are the energy E and the leading-two streamfunction PCs. In light of the conventional low-, middle-, and high-energy phenomenology of the 1.5-layer model [20, 61], energy is a natural observable to consider for long-range forecasting. Moreover, the time-averaged spatial features of the circulation regimes are well captured by the leading PCs. We used the portion of the time series with $t > T$ to compute the cluster-conditional time-dependent distributions p_t^k (2.19) for these observables via the procedure described in Section 2.4. Thus, the data used to estimate p_t^k are independent of the input data to the algorithm for evaluating the cluster coordinates θ_k in (2.14).

All forecast densities were estimated by binning the $s/2$ prediction samples in $n_B = 100$ bins of uniform width. The entropy integrals in the predictability metric $\mathcal{D}(a_t, S_0)$ in (2.21) were evaluated via the standard trapezoidal rule. We verified the robustness of our results against sampling errors by halving the length of the prediction time series, and repeating the calculation of p_t^k for each half. Quadrature errors were assessed by halving the number n_B of distribution bins, and re-evaluating $\mathcal{D}(a_t, S_0)$. In all cases, the predictability scores in Figure 2.4 did not change significantly. Moreover, we tested for robustness of the computed cluster-coordinates θ_k in (2.14) by using either of the two halves of our training data. This did not impart significant changes in the spatial structure of the regimes in Figure 2.6.

Following the strategy laid out in Section 2.4, we vary the running-average time intervals $\Delta t'$ and Δt, used respectively to coarse-grain $\mathcal{X}$ and the time series (2.3) of initial data, seeking to maximize (for the given choice of observable and forecast lead time t) the information content $\mathcal{D}(a_t, S_0)$ from (2.21) beyond equilibrium [or, equivalently, the predictability score δ_t in (2.25)] in the resulting partition from (2.15). In Figure 2.4 we display a sample of the δ_t results for fixed $\Delta t' = 1000$ days (i.e., a value comparable to the decorrelation time, $t_1 = 1165$ days, of PC_1), and representative values of short and long initial-data windows, respectively $\Delta t = 0$ and $\Delta t = \Delta t' = 1000$ days. For the time being, we consider models with either $K = 3$ or $K = 7$ clusters, and subsequently (in Section 2.5.2) study in more detail the relevance of these choices from a physical standpoint.

There are a number of important points to be made about Figure 2.4. First, for the chosen observables, the predictability score δ_t^* (2.47) of the optimal partitions is significant for prediction horizons that exceed the longest decorrelation time in the X_t components used for clustering by a large margin. The fact that decorrelation times are poor indicators of intrinsic long-range predictability has been noted in other CAOS applications [37]. Here, the decay in the δ_t^* score for energy over one e-folding time corresponding to t_1 is $\delta_{t_1}^*/\delta_0^* \simeq 0.7$, or a factor of five weaker decay than $e^{-2} \simeq 0.14$ expected for a purely exponential decay (the comparison is with e^{-2} rather than e^{-1} because δ_t^* is associated with squared correlations). Predictability of energy remains significant up to $t \simeq 3000$ days ($\delta_{3000}^*/\delta_0^* \simeq 0.07$), or three times the decorrelation time of PC_1. This means that predictions approaching the decadal scale are possible for E, given knowledge at time $t = 0$ of the system's affiliation to the regimes associated with partition Ξ in (2.15). Note that no fine-grained information about the initial conditions is needed to make these forecasts. Uncertainty in initial conditions is a well-known obstacle in long-range forecasts [21, 62–64].

Second, as illustrated by the discrepancy between the δ_t scores evaluated for $\Delta t = 0$ and 1000 days, the time window Δt that maximizes the information beyond equilibrium in the partition depends on both the observable and the forecast lead time. More specifically, in the calculations used to produce the δ_t^* versus t lines in Figure 2.4, the optimal Δt for mid-term prediction ($t \lesssim 500$ days) of the energy is around 500–1000 days, but that value rapidly decreases to essentially no coarse-graining ($\Delta t = 0$) when t extends

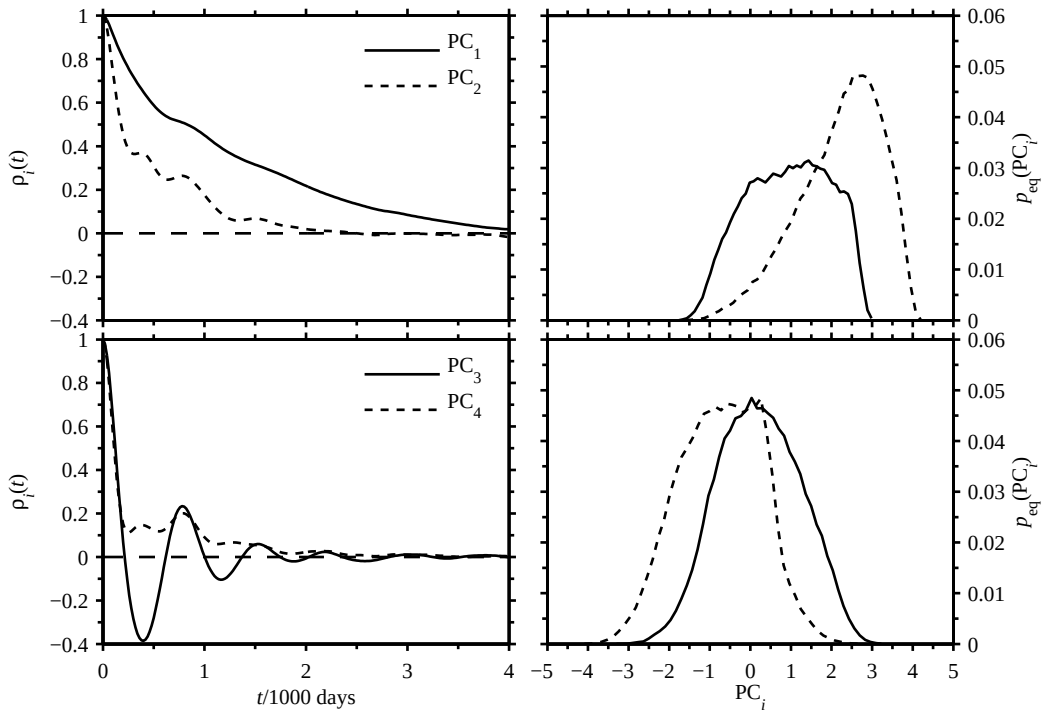

Figure 2.3. Empirical autocorrelation functions, $\rho_i(t) = \int_0^T dt' \, \mathrm{PC}_i(t')\mathrm{PC}_i(t'+t)/T$, and equilibrium densities, $p_{\mathrm{eq}}(\mathrm{PC}_i)$, of the leading four streamfunction PCs. Among these PCs, only PC_3 has significantly negative values in $\rho_i(t)$. All the autocorrelation functions of PC_i with $i \in [5, 20]$ (not shown here) take negative values. Note that $p_{\mathrm{eq}}(\mathrm{PC}_i)$ are all unimodal, yet the system exhibits long-lived affiliations to regimes (see Figure 2.7).

beyond the two-year horizon. On the other hand, $\Delta t = 0$ is optimal for all values of the prediction lead time t in the case of the PCs. The fact that the optimal Δt for long-range forecasting is small is beneficial from a practical standpoint, since it alleviates the need of collecting initial data over long periods.

Third, as alluded to in the beginning of this section, the $K = 7$ partitions carry significantly higher predictive information than the $K = 3$ ones for mid-range forecasts (up to three years), but that additional information is lost in the large lead-time regime. In particular, the δ_t^* scores of the $K = 3$ and $K = 7$ models meet at approximately $t = 2000$ days for E, 500 days for PC_1, and 1000 days for PC_2.

A final point about Figure 2.4 pertains to the non-monotonicity of δ_t [equivalently, $\mathcal{D}(a_t, S_0)$] for E. It is a general result, sometimes referred to as the generalized second law of thermodynamics, that if the dynamics of an observable are Markovian, then the corresponding relative entropy $\mathcal{D}(a_t, S_0)$ decreases monotonically with t [45, 49, 52]. Thus, the increasing portion of the $\delta_t(E)$ versus t curve for $\Delta t = 0$ and $t \lesssim 500$ days is a direct evidence of non-Markovianity of the energy observable. As discussed in Section 2.5.3, this has important implications for model error when the corresponding cluster affiliation sequence is approximated by a Markov process.

Non-Markovianity of energy is consistent with the fact that longer running-average windows are favored for optimal predictions of this observable for moderate lead times. Physically, as follows from (2.42), the width of Δt controls the influence of the past history of the system relative to its current state in assigning cluster affiliation. If a prediction observable exhibits significant memory effects, taking the running average over a window comparable to the memory time scale should lead to gains of predictive information, at least for lead times of order Δt or less. This is reflected in the δ_t results for energy in Figure 2.4, where forecasts made using a 1000-day averaging window are more skillful than the corresponding forecasts with $\Delta t = 0$, provided that the lead time does not exceed 500 days or so.

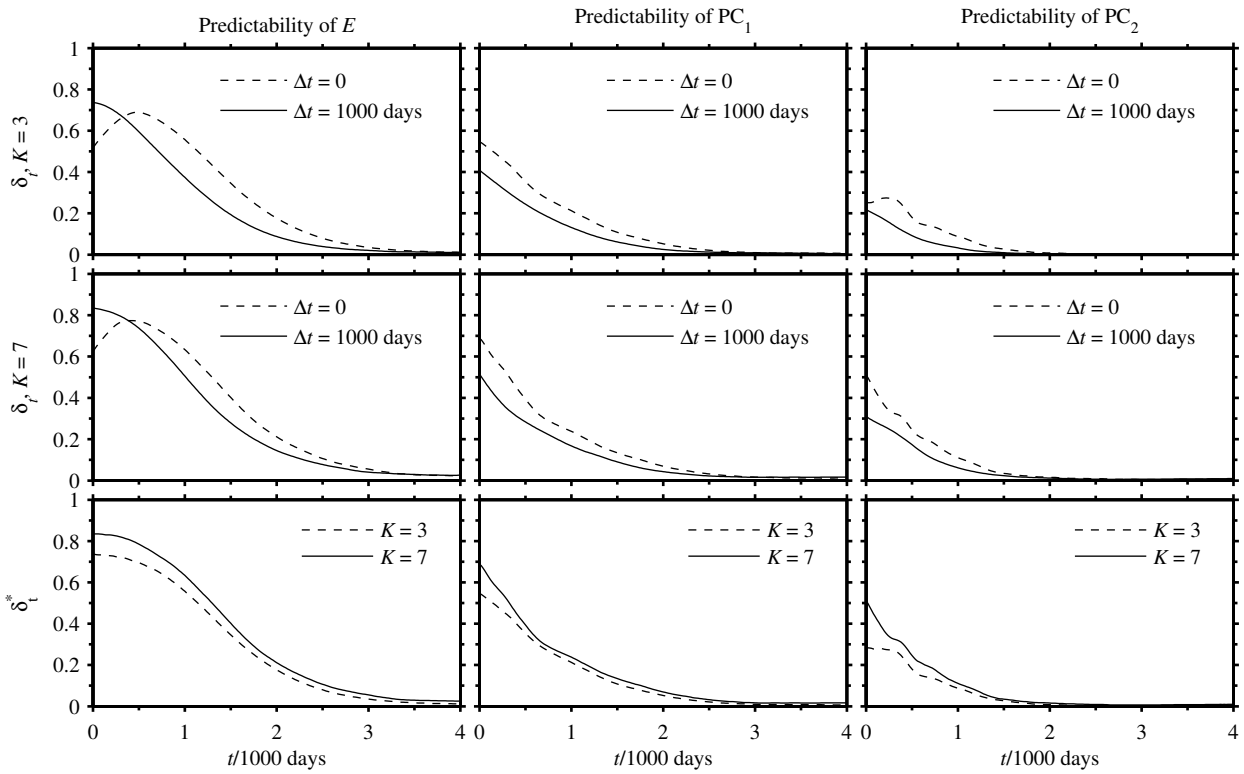

Figure 2.4. The information content (predictability score) δ_t (2.25) in $K = 3$ and $K = 7$ partitions (2.15) as a function of prediction horizon t for the energy E and the leading two PCs of the streamfunction. Two values for the running-average interval for initial cluster affiliation are displayed ($\Delta t = 0$ and 1000 days), as well as the optimal score δ_t^* for various values of Δt in the interval $[0, 1000]$ days. In all cases, the running-average interval for coarse-graining the training dataset is $\Delta t' = 1000$ days. The δ_t lines for energy with $\Delta t = 0$ illustrate that the decay of relative entropy to equilibrium may be non-monotonic; a behavior that cannot be replicated by Markov models (see Figure 2.8). The $K = 7$ partitions have higher information content than the $K = 3$ ones in the leading PCs (i.e., the large-scale structures in the flow) for $t \lesssim 600$ days, or about half the decorrelation time of the leading PC (see Figure 2.3). However, $K = 7$ contributes essentially no additional predictive information beyond $K = 3$ for decadal forecasts.

2.5.2 The physical properties of the regimes

We now study the spatial and temporal properties of the regimes associated with the coarse-grained partitions of Section 2.5.1. For concreteness, we focus on a $K = 3$ partition with running-average windows $(\Delta t, \Delta t') = (1000, 1000)$ days; see [14] for results with $K = 7$ partitions and other running-average windows. The $K = 3$ partition was motivated by the analyses in [20, 61], which associate the meandering, mean-flow resembling, and extensional circulation regimes of the 1.5-layer model with bands of low, moderate, and high values of the energy observable. More specifically, the chosen Δt value is a reasonable compromise for simultaneously maximizing the predictability metrics in Figure 2.5 for energy and the leading two PCs.

The key objects facilitating our study of the physical properties of the regimes are the cluster-conditional mean and standard deviation of the streamfunction anomaly, $\psi'_k(\mathbf{r}) = \langle \psi'(\mathbf{r}, t) \rangle_k$ and $\sigma_k(\mathbf{r}) = \langle (\psi'(\mathbf{r}, t) - \psi'_k(\mathbf{r}))^2 \rangle_k^{1/2}$, which are shown in Figure 2.6. Here, $\langle \cdot \rangle_k$ denotes expectation value with respect to p_t^k from (2.19) at $t = 0$, which, by ergodicity (2.4), can be evaluated by taking temporal averages conditioned on $S(X_t) = k$. First, it is clear from Figure 2.6 that the circulation regimes identified by the K-means clustering algorithm with $K = 3$ and running averaging are in good agreement with the semi-empirical phenomenology established for 1.5-layer double-gyre ocean models [20, 61]. Specifically, state 1, which has a low expected value of energy, $E_1 = \langle E(t) \rangle_1 = 3.5 \times 10^{17}$ J, features a meandering jet pattern; state 2, with $E_2 = 3.9 \times 10^{17}$ J resembles

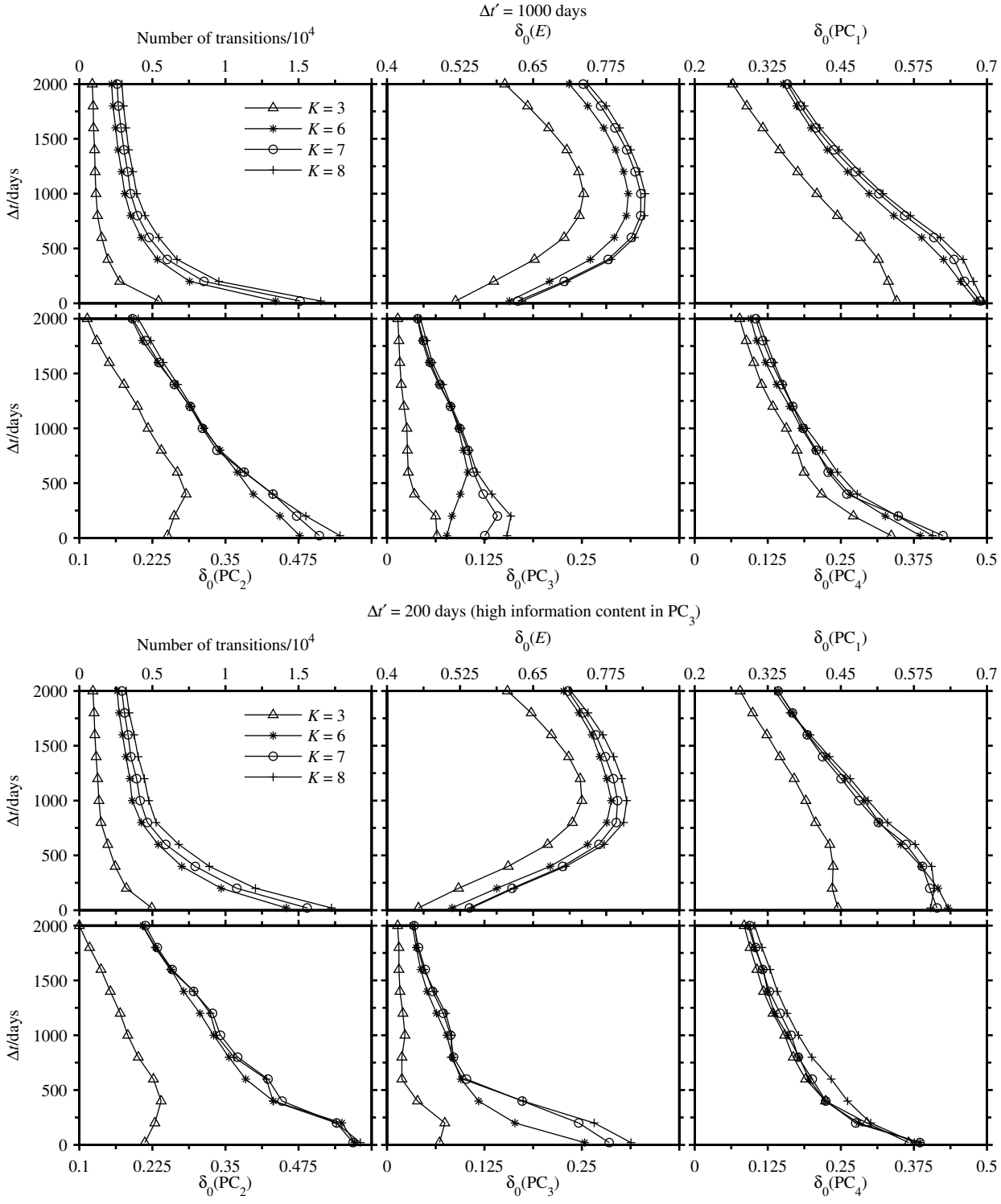

Figure 2.5. The dependence of the number of transitions and the relative-entropy predictability score δ_t (2.25) on the running-average interval Δt (initial data for prediction), evaluated at time $t = 0$ for the energy E and leading four PCs for partitions with $K \in \{3, 6, 7, 8\}$ clusters. The running-average interval for coarse-graining the training data is $\Delta t' = 1000$ days and 200 days, respectively in the top and bottom set of panels.

the time-averaged state $\Psi(\mathbf{r})$; and state 3 is dominated by a strong, deeply-penetrating jet, and has large mean energy $E_3 = 4.2 \times 10^{17}$ J. As one might expect from the corresponding relative increase in information content (see Figure 2.5), the basic spatial features of the $K = 3$ regimes are captured with significantly higher fidelity by $K = 7$ partitions (see [11]).

Turning to the temporal aspects of the switching process between the regimes, Figure 2.7 illustrates that the cluster affiliation sequence $S_t = S(X_t)$ from (2.42) of the $K = 3$ partition of observation space with $\Delta t = 1000$ days leads to a natural splitting of the energy time series into persistent regimes (with decadal mean lifetimes), as expected from the high information content of that partition about energy. As remarked in Section 2.4, imposing temporal regularity in S_t is frequently a challenge in CAOS applications (e.g., standard K-means analysis of this dataset results in unphysical, high-frequency transitions between the regimes), but it emerges here automatically by virtue of coarse-graining the training dataset and the interval Δt for initial cluster affiliation. It is important to emphasize, however, that persistence is not synonymous with skill. For instance, the δ_0 score for PC_1 in Figure 2.5 is a decreasing function of Δt, even though the persistence of the regimes exhibits a corresponding increase (as indicated by the drop in the number of transitions with Δt). Information theory allows one to tell when a persistent cluster affiliation sequence actually carries information for prediction (or classification, as is the case for the $t = 0$ examples considered here), or is too crude of a description of the intrinsic low-frequency dynamics.

2.5.3 Markov models of regime behavior in the 1.5-layer ocean model

We now apply the tools developed in Section 2.2.2 to assess the model error in Markov models of regime behavior in the 1.5-layer model. Throughout, we treat the output the 1.5-layer model as the perfect model, and Markov models of the switching process between the regimes identified in Section 2.5.1 and 2.5.2 as imperfect reduced models with dynamical model error. In particular, we introduce model error by evaluating the forecast distributions p_t^{Mk} in (2.26) under the assumption that the affiliation sequence $\Gamma(t)$ in (2.41) is a Markov process. The Markov assumption for $\Gamma(t)$ is made frequently in cluster analyses of time series in atmosphere-ocean science [26, 30, 34, 39, 40, 65, 66], but as we demonstrate in Section 2.5.4, can lead to false

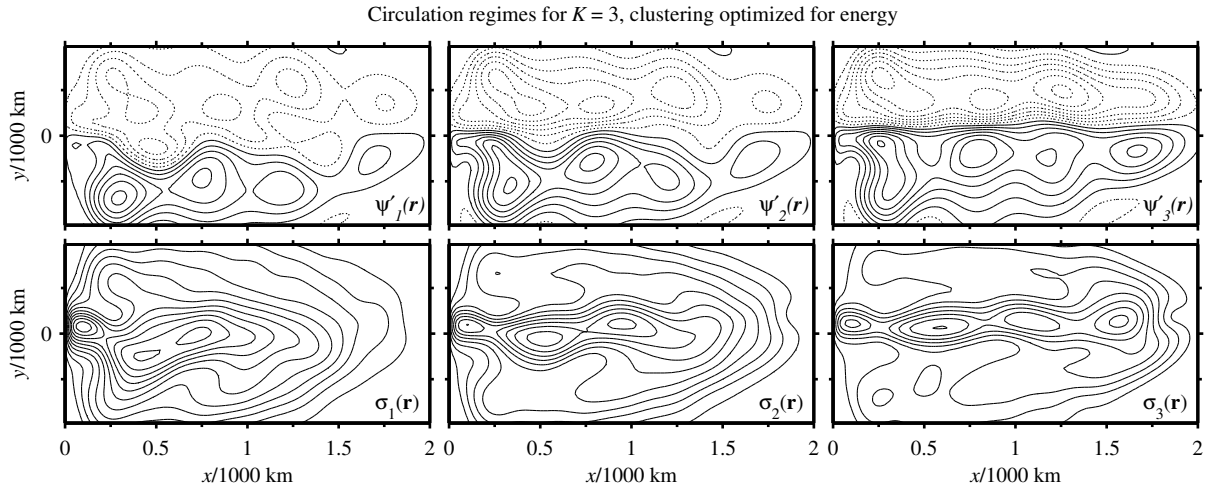

Figure 2.6. Mean streamfunction anomaly, $\psi'_k(\mathbf{r})$, and its standard deviation, $\sigma_k(\mathbf{r})$, conditioned on the clusters of a $K = 3$ partition. The contour-level spacing for $\psi'_k(\mathbf{r})$ and $\sigma_k(\mathbf{r})$ is 25 m and 10 m, respectively. Solid and dotted lines respectively represent positive and negative contour levels. This partition of observation space has been evaluated using running-average windows of duration $\Delta t = \Delta t' = 1000$ days, and is optimized for maximal information content beyond equilibrium about energy (see Figure 2.5). The spatial features of the circulation regimes identified here via running-average K-means clustering are in good agreement with the meandering (ψ'_1), mean-flow resembling (ψ'_2), and extensional (ψ'_3) phases of the jet in the McCalpin and Haidvogel [20] phenomenology, with correspondingly low, moderate, and high values of mean energy (see Figure 2.7).

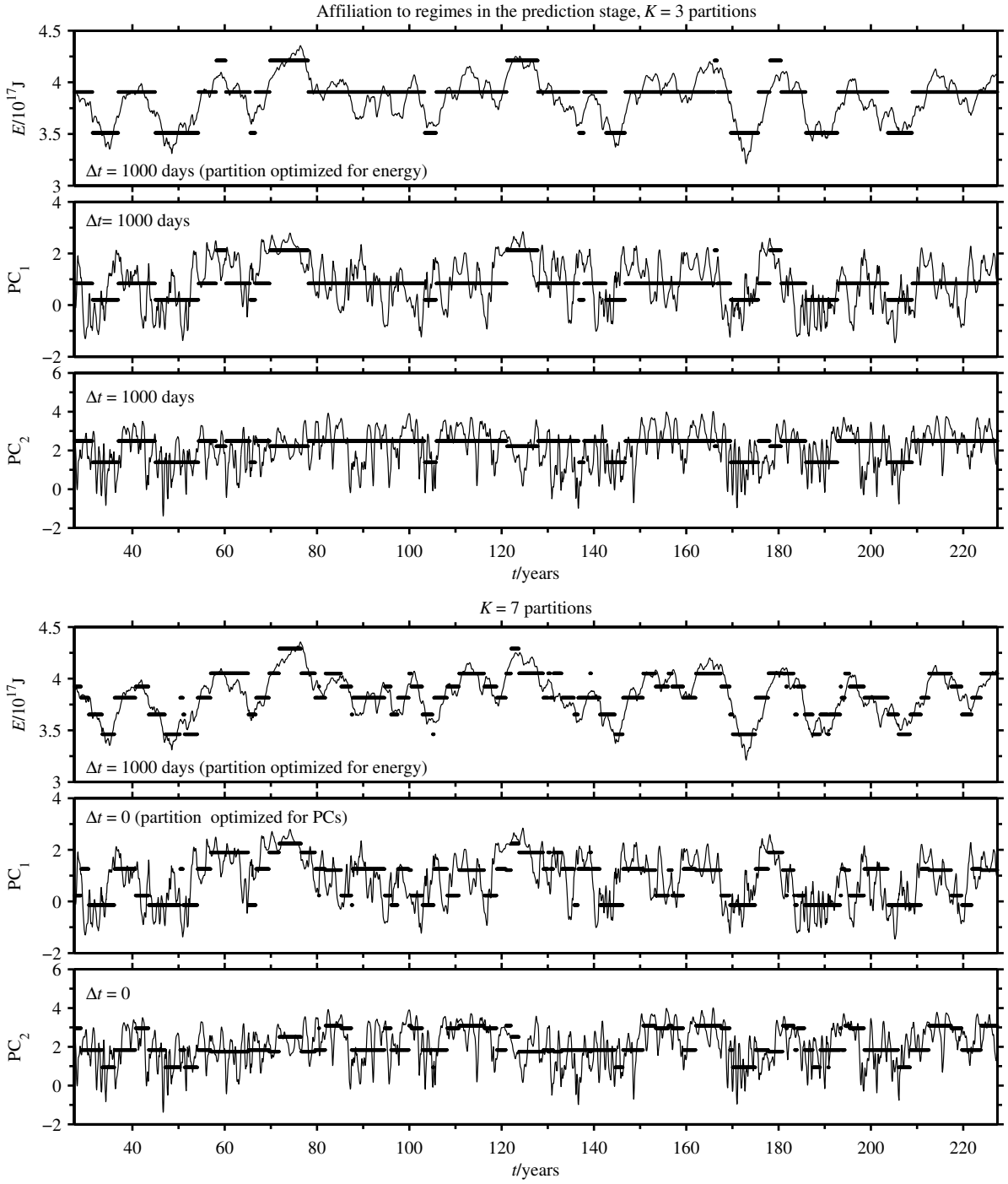

Figure 2.7. Time series of the energy E and the leading two PCs spanning a 200-year interval. The thick horizontal lines show the discrete energy and PC affiliation sequences, $\langle E\rangle_{S(t)}$ and $\langle \mathrm{PC}_i\rangle_{S(t)}$, where $\langle\cdot\rangle_k$ denotes cluster-conditional expectation value, and $S(t)$ is the cluster affiliation sequence in (2.42). Throughout, the running-average window in the training stage is $\Delta t' = 1000$ days, and regimes are ordered in order of increasing $\langle E\rangle_k$. In the $K = 7$ panels, the running-average window Δt is chosen so that the corresponding partitions of observation space contain high information about energy ($\Delta t = 1000$ days), or the leading PCs ($\Delta t = 0$; see Figure 2.5).

predictability, as measured by the $\mathcal{D}^M(a_t, S_0)$ metric in (2.29). The benefit of the scheme presented here is that false predictability can be detected directly through the measures of model error $\mathcal{E}_{\rm eq}$ and $\mathcal{E}(a_t, S_0)$.

The fundamental assumption in the Markov models studied here is that there exists a $K \times K$ generator matrix L, such that

$$\mathsf{P}(t)_{ij} = p(\Gamma(t) = j \mid \Gamma(0) = i) = \exp(t\mathsf{L})_{ij}, \tag{2.48}$$

where $\Gamma(t)$ is defined in (2.41), and t is an integer multiple of the sampling interval δt. In general, the existence of L is not guaranteed, even if $\Gamma(t)$ is indeed Markovian. Nevertheless, one may always try to estimate a Markov generator that is consistent with the given realization $\Gamma(t)$ using one of the available algorithms in the literature [67, 68], and verify a posteriori its consistency by computing $\mathcal{E}(a_t, S_0)$ from (2.33) for prediction observables a_t of interest. Here, the cluster-conditional probabilities (2.26) in the Markov model are given (via Bayes' theorem) by

$$p_t^{Mk} = \sum_{j=1}^{K} \exp(t\mathsf{L})_{kj} \phi^j, \tag{2.49}$$

where

$$\phi^k = p(a_t | \Gamma(t) = k) \tag{2.50}$$

are the distributions for a_t conditioned on the value of $\Gamma(t)$ for the training data. These distributions can be estimated by cluster-conditional bin counting of simultaneous realizations of a_t and $\Gamma(t)$, as described in Section 2.4. As with Section 2.5.1, our primary observables of interest are the total energy in the flow, E, and the leading two PCs of the streamfunction.

Since for sufficiently long training time series the ϕ^k are equivalent to the p_t^k distributions in (2.19) evaluated at $t = 0$ with equal running-average windows in the training and prediction stages (i.e., $\Delta t = \Delta t'$), the model probabilities in (2.49) have no error at time $t = 0$; i.e., $\mathcal{E}_0^k$ in (2.33) is zero by construction. Moreover, $\mathcal{E}_t^k$ will vanish as $t \to \infty$ for models that meet the equilibrium consistency condition in (2.31) for

$$p_{\rm eq}^M = \sum_{k=1}^{K} \pi_k^M \phi^k, \tag{2.51}$$

where $\pi^M = (\pi_1^M, \ldots, \pi_K^M)$ is the equilibrium distribution of the Markov model, defined by the requirement for all t,

$$\sum_{i=1}^{K} \pi_i^M \mathsf{P}(t)_{ij} = \pi_j^M. \tag{2.52}$$

However, due to dynamical model error, $\mathcal{E}(a_t, S_0)$ will generally be nonzero for finite and nonzero t.

2.5.4 The model error in long-range predictions with coarse-grained Markov models

To construct our Markov models, we took the same training data used in Section 2.5.1 consisting of the leading 20 PCs of the streamfunction in the 1.5-layer model, $x(t) = (\mathrm{PC}_1(t), \ldots, \mathrm{PC}_{20}(t))$, and computed affiliation sequences $\Gamma(t)$ from (2.41), applying the procedure described in Section 2.4 for various choices of K and running-average windows $\Delta t'$. In each case, we determined P by fitting the generator matrix L in (2.48) to $\Gamma(t)$ using the Bayesian algorithm of Metzner et al. [68]. We checked for robustness against sampling errors by repeating our calculations using either of the two halves of the training time series. This resulted to relative changes of order 10^{-3} in L, as measured by the ratio of Frobenius norms $\|\delta\mathsf{L}\|/\|\mathsf{L}\|$. Likewise, the changes in the results in Figure 2.8 were not significant.

The two main dangers with the assumption that $\Gamma(t)$ has the Markov property are that (i) the Markov model fails to meet the equilibrium consistency condition in (2.31), i.e., the Markov equilibrium distribution deviates systematically from the truth; and (ii) the discrepancy $\mathcal{D}^M(a_t, S_0)$ in (2.29) from equilibrium of the Markov model measures false predictability, e.g., for a Markov model that relaxes to equilibrium unjustifiably slowly. The latter two pitfalls may arise independently of one another, since the discrepancy $\mathcal{E}(a_t, S_0)$ in (2.12) of a model from the truth as it relaxes to equilibrium can be large for some forecast lead time t, even if the model error $\mathcal{E}_{\rm eq}$ in equilibrium is small. Nevertheless, the $\mathcal{E}(a_t, S_0)$ and $\mathcal{E}_{\rm eq}(a_t)$ metrics [and the corresponding

normalized error scores ϵ_t and $\epsilon_{\rm eq}$ in (2.34) and (2.31), respectively] allow one to detect these types of error a posteriori, given the Markov matrix P fitted to the data.

In Figure 2.8 we illustrate these issues through a comparison between a $K = 7$ and a $K = 3$ Markov model. The seven-state model was constructed using the $\Gamma(t)$ affiliation sequence of the $K = 7$ partitions of Section 2.5.1; i.e., the training time series $x(t)$ was coarse-grained using a running-average window of $\Delta t' = 1000$ days. The same training time series was used to evaluate the generator of the $K = 3$ Markov model, but in this case $\Delta t'$ was increased to 2000 days. First, it is clear from the graphs for ϵ_t that the seven-state Markov model asymptotes to an incorrect equilibrium distribution. For this reason, the relative-entropy score δ_t^M measures false predictability for all forecast horizons, including $t = 0$. On the other hand, the $K = 3$ model of Figure 2.8 does meet climate equilibrium consistency (with $\epsilon_{\rm eq} \sim 10^{-5}$), which means that for this model δ_0^M is a true measure of classification skill beyond equilibrium. That model, however, experiences a gradual ramp-up of ϵ_t, peaking at around $t = 2000$ days, and as a result, its predictions cannot be deemed accurate beyond, say, a horizon $t \gtrsim 1000$ days.

Note now how the second pitfall might lead one to believe that the seven-state Markov model is more skillful than the three-state one: The smallest non-trivial eigenvalue, $\mu_1 = (\log \lambda_1)/\delta t \simeq -1/(4000 \text{ days})$, of the generator matrix of the $K = 7$ model has smaller absolute value than the corresponding eigenvalue, $\mu_1 \simeq -1/(3000 \text{ days})$, of the $K = 3$ model. That is, for long-enough prediction horizons, the seven-state model relaxes more slowly to equilibrium than the three-state model, i.e., it is more persistent. By monitoring $\epsilon_{\rm eq}$ and ϵ_t it is possible to identify models with false persistence as illustrated above.

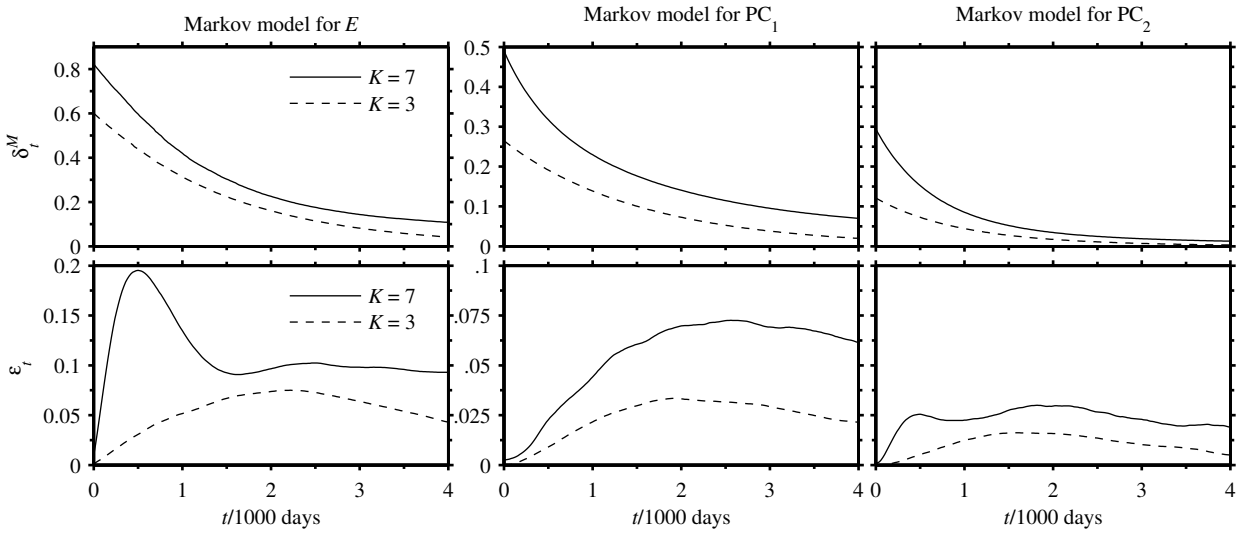

Figure 2.8. Internal predictability score, δ_t^M, and model error, ϵ_t, of $K = 3$ and $K = 7$ Markov models as a function of the forecast lead time t. The observables under consideration are the energy E and the leading two PCs. The coarse-graining interval $\Delta t'$ for creating the partition in the training stage is $\Delta t' = 2000$ and 1000 days, respectively for $K = 3$ and $K = 7$, with corresponding model error in the energy equilibrium $\epsilon_{\rm eq} \sim 10^{-5}$ and 0.068; i.e., the three-state Markov model meets equilibrium consistency (2.31), but the seven-state model does not. At finite t, small values of ϵ_t mean that the relative-entropy distance δ_t^M is an appropriate surrogate for the true predictive skill of the Markov models. On the other hand, if ϵ_t and δ_t^M are both large, then δ_t^M is biased, and measures false predictive skill. The equilibration time of the Markov models (given by $-1/\mu_1$, where μ_1 is the first non-trivial eigenvalue of the generator of the Markov process) is 3000 days and 4260 days, respectively for $K = 3$ and $K = 7$. Thus, in this example the most erroneous Markov model has the largest false skill and is also the most persistent.

3 Nonlinear Laplacian spectral analysis (NLSA) algorithms for decomposition of spatiotemporal data

3.1 Background

In a wide range of disciplines in science and engineering, including those outlined in Section 1, there exists a strong interest in extracting physically-meaningful information about the spatial and temporal variability of data from models, experiments, or observations with the goal of enhancing the understanding of the underlying dynamics. Frequently, observations of the system under study are incomplete; i.e., only a subset of the full phase space is accessible.

A classical way of attacking this problem is through singular spectrum analysis (SSA), or one of its variants [69–74]. 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 $\mathbb{R}^N$ (here referred to as lagged-embedding space) using the method of delays [75–78], and then performing a truncated singular value decomposition (SVD) of the matrix X containing the embedded data. In this manner, information about the dynamical process is extracted from the left and right singular vectors of X with the l largest singular values. The left (spatiotemporal) singular vectors form a set of so-called extended empirical orthogonal functions (EEOFs) in $\mathbb{R}^N$, which, at each instance of time, are weighted by the corresponding principal components (PCs) determined from the right (temporal) singular vectors to yield a rank-l approximation of X.

A potential drawback of this approach is that it is based on minimizing an operator norm which may be unsuitable for signals generated by nonlinear dynamical systems. Specifically, the PCs are computed by projecting onto the principal axes of the l-dimensional ellipsoid that best fits the covariance of the data in lagged-embedding space in the least-squares sense. This construction is optimal when the data lies on a linear subspace of $\mathbb{R}^N$, but nonlinear processes and/or observation functions will in general produce data lying on a nonlinear submanifold $\mathcal{M} \subset \mathbb{R}^N$ with non-Gaussian distributions departing significantly from the ellipsoid defined by the covariance operator of the data. Physically, a prominent manifestation 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 [79, 80].

Despite their inherently nonlinear character, such datasets possess natural linear structures, namely Hilbert spaces $\Lambda^0\mathcal{M}$ of square-integrable functions on $\mathcal{M}$ with inner product inherited from the volume form of a Riemannian metric induced on the data by lagged embedding. Moreover, intrinsically discrete analogs $\Lambda^0 M$ of $\Lambda^0\mathcal{M}$ can be constructed empirically for the set $M \subset \mathcal{M}$ of observed data using techniques from discrete exterior calculus (DEC, e.g., [81–83]). These spaces may be thought of as the collection of all possible weights that can be assigned to the data samples when making a low-rank reconstruction, i.e., they are analogous to the temporal spaces of SSA. Based on these observations, it is reasonable to develop algorithms for data decomposition which are based on suitably-defined maps from $\Lambda^0 M$ to lagged-embedding space $\mathbb{R}^N$. Such maps, denoted here by A, have the advantage of being simultaneously linear and compatible with the nonlinear geometry of the data.

Here, we advocate that this approach, implemented via algorithms developed in machine learning [84, 85], can reveal important aspects of complex, high-dimensional signals, which are not accessible to classical SSA. In this framework, which we refer to as nonlinear Laplacian spectral analysis (NLSA), an orthonormal basis for $\Lambda^0 M$ is constructed through eigenfunctions of a diffusion operator associated with a kernel in lagged-embedding space with explicit dependence on the dynamical vector field on $\mathcal{M}$ generating the data. Projecting the data from embedding space onto these eigenfunctions then gives a matrix representation of A leading, via SVD, to a decomposition of the dataset into a biorthonormal set of spatiotemporal patterns.

3.2 Mathematical framework

We consider a scenario where the dynamics is described by a flow $F_t : \mathcal{F} \mapsto \mathcal{F}$ operating in a phase space $\mathcal{F}$, and evolving on an attractor $\mathcal{M} \subseteq \mathcal{F}$. Moreover, observations are taken uniformly in time with a timestep $\delta t > 0$ on the attractor via a smooth vector-valued function $G : \mathcal{F} \mapsto \mathbb{R}^n$, forming a dataset $\mathsf{x} = (x_1, \ldots, x_s)$ with

$$x_i = G(z_i), \quad z_i = F_{t_i} z_0, \quad t_i = i\,\delta t, \quad z_0 \in \mathcal{M}. \tag{3.1}$$

In general, we are interested in cases where the observation function is incomplete, i.e., the x_i alone are not sufficient to uniquely determine the state of the system in $\mathcal{M}$. Geometrically, this means that the image

manifold $G(\mathcal{M}) \subset \mathbb{R}^n$ is not diffeomorphic to $\mathcal{M}$. Our objective is to produce a decomposition of x_i into a set of l spatiotemporal patterns $\hat{x}_i^k$,

$$x_i \approx \tilde{x}_i = \sum_{k=1}^{l} \hat{x}_i^k, \tag{3.2}$$

taking into account the the underlying dynamical system operating on $\mathcal{M}$. That is, compared to generic point clouds of data, an additional structure that we have at our disposal here is the time ordering of the observations, which carries meaningful information about F_t. Therefore, we seek that the decomposition (3.2) depends on that time ordering.

The methodology employed here to address this objective consists of four basic steps: (i) Embed the observed data in a vector space $\mathbb{R}^N$ of dimension greater than n via the method of delays; (ii) construct a linear map A_l taking a Hilbert space of scalar functions on $\mathcal{M}$ representing temporal patterns to the spatiotemporal patterns in $\mathbb{R}^N$; (iii) perform a singular value decomposition (SVD) in a basis of orthonormal diffusion eigenfunctions to extract the spatial and temporal modes associated with A_l; (iv) project the modes from $\mathbb{R}^N$ to data space $\mathbb{R}^n$ to obtain the spatiotemporal patterns $\hat{x}_t^k$ in (3.2). Below, we provide a description of each step, as well as an outline of SSA algorithms to draw connections between the two approaches. Further details of the NLSA framework, as well as pseudocode, are presented in [14, 16]. A Matlab implementation is available upon request from the corresponding author.

Hereafter, we shall consider that $\mathcal{M}$ has integer dimension m and is compact and smooth, so that a well-defined continuous spectral theory exists [86]. Moreover, we assume that the dynamical vector field $\dot{F}$ induced on $\mathcal{M}$, given by

$$\dot{F}(f) = \lim_{t \to 0} (f(F_t z) - f(z))/t, \quad \text{with} \quad z \in \mathcal{M}, \quad f \in C^\infty \mathcal{M}, \tag{3.3}$$

is also smooth. We emphasize, however, that the smoothness assumptions for $\mathcal{M}$ and $\dot{F}$ are to be viewed as "Platonic ideals" serving as a guidelines for algorithm design and analysis, but seldom encountered in practice (e.g., due to finite number of samples and/or non-smoothness of the attractor). Operationally, one works in the intrinsically discrete framework of spectral graph theory [87] and DEC [81, 83], which exist independently of the continuous theory, even if the latter was used as a means of gaining insight.

3.2.1 Time-lagged embedding

This step is familiar from the qualitative theory of dynamical systems [75–78]. Under generic conditions, the image of $z_i \in \mathcal{M}$ in embedding space, $\mathbb{R}^N$, under the delay-coordinate mapping,

$$H(z_i) = X_i = (G(x_i), G(x_{i-1}), \dots G(x_{i-(q-1)})), \quad X_i \in \mathbb{R}^N, \quad N = qn, \tag{3.4}$$

lies on a manifold $H(\mathcal{M})$ which is diffeomorphic to $\mathcal{M}$, provided that the dimension N is sufficiently large. Thus, given a sufficiently-long embedding window $\Delta t = (q-1)\,\delta t$, we obtain a representation of the attractor underlying our incomplete observations.

Broadly speaking, preprocessing the data by time lagged-embedding produces both topological and geometrical changes. In particular, the topology of the embedded dataset $H(\mathcal{M})$ will be different from that of the original $G(\mathcal{M})$ data if the observation map is incomplete, recovering the manifold structure of the attractor lost through partial observations. An implication of this is that the time series X_i in (3.4) becomes Markovian, or, equivalently, the dynamical vector field $\dot{F}$ on $\mathcal{M}$ from (3.3), is carried along by means of the derivative map DH to a smooth vector field $\dot{F}_* = DH\,\dot{F}$ on $H(\mathcal{M})$.

Besides topological properties, time-lagged embedding also influences the geometry of the data, in the sense that the Riemannian metric h induced on $\mathcal{M}$ by pulling back the canonical inner product of $\mathbb{R}^N$ depends explicitly on the dynamical flow generating the data. To see this, let $(u^1, \dots, u^m)$ be a coordinate system in a neighborhood of $z_i \in \mathcal{M}$. In this coordinate system, the induced metric h at z_i has components

$$h_{\mu\nu}|_i = \sum_{\alpha=1}^{N} \frac{\partial X_i^\alpha}{\partial u^\mu} \frac{\partial X_i^\alpha}{\partial u^\nu} = \sum_{\alpha=1}^{n} \sum_{j=0}^{q-1} \frac{\partial x_{i-j}^\alpha}{\partial u^\mu} \frac{\partial x_{i-j}^\alpha}{\partial u^\nu} = \sum_{j=0}^{q-1} g_{\mu\nu}|_{i-j}, \tag{3.5}$$

where g is the induced metric on the original dataset $G(\mathcal{M})$, and X_i^α (x_i^α) the components of X_i (x_i) in an orthonormal basis of $\mathbb{R}^N$ ($\mathbb{R}^n$). It therefore follows that the induced metric on $\mathcal{M}$ is a "running-averaged"

version of the induced metric on $G(\mathcal{M})$ along orbits of the dynamical system in (3.1) of temporal extent Δt. In other words, the results of a data analysis algorithm operating in $\mathbb{R}^N$ that processes the data based on distance-based affinity metrics will depend on the dynamical flow F_t.

Note that this effect takes place even if the observation map is complete [i.e., $G(\mathcal{M})$ is a diffeomorphic copy of $\mathcal{M}$], which suggests that time-lagged embedding may be used as a tool to control the dataset geometry even in fully observed dynamical systems. This question has been studied in detail in recent work by Berry and collaborators [88], who have established a correspondence between the Lyapunov metric of a dynamical system acting along the most stable Osedelets subspace and the induced metric in a suitable lagged-embedding space.

3.2.2 Overview of singular spectrum analysis

The classical linear framework for creating low-rank approximations of a dataset in lagged-embedding space is essentially identical to principal components analysis (PCA) or proper orthogonal decomposition (POD) algorithms [1, 72], apart from the fact that one obtains a biorthonormal basis of temporal–spatiotemporal patterns rather than the usual temporal–spatial basis. Algorithms of this type are interchangeably called SSA [71, 74]), singular system analysis [70], and EEOFs [69]. We use the term SSA to refer to this family of algorithms.

Let $\mathsf{X} = (X_1, \ldots, X_S)$ with $S = s - q + 1$ be the data matrix in lagged embedding space, dimensioned $N \times S$. In SSA, the biorthonormal basis that optimally fits the data is constructed through the SVD of X,

$$\begin{gathered} \mathsf{X} = \mathsf{U}\mathsf{\Sigma}\mathsf{V}^T, \\ \mathsf{U} = (U_1, \ldots, U_N), \quad \mathsf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{N,S\}}), \quad \mathsf{V} = (V_1, \ldots, V_S), \\ U_i \in \mathbb{R}^N, \quad \sigma_i \geq 0, \quad V_i \in \mathbb{R}^S, \\ U_i^T U_j = V_i^T V_j = \delta_{ij}. \end{gathered} \tag{3.6}$$

Here, U and V are orthogonal matrices of dimension $N \times N$ and $S \times S$, respectively, and $\mathsf{\Sigma}$ a diagonal matrix with nonnegative diagonal entries ordered in decreasing order. In (3.6), the j-th column of V gives rise to a function of time (a PC)

$$\tilde{v}_j(t_i) = V_{ij}. \tag{3.7}$$

Moreover, the corresponding column U_j of U represents a spatiotemporal process $u_j(\tau_i)$ in of time duration Δt, viz.

$$u_j(\tau_i) = (U_{n(i-1)+1,j}, \ldots, U_{ni,j}) \in \mathbb{R}^n, \quad \tau_i = (i-1)\,\delta t. \tag{3.8}$$

A rank-l approximation X_l of the dataset in lagged embedding space is then constructed through the leading l singular vectors from (3.6),

$$\mathsf{X}_l = \mathsf{U}_l\mathsf{\Sigma}_l\mathsf{V}_l = \mathsf{X}\mathsf{V}_l^T\mathsf{V}_l, \quad \mathsf{U}_l = (U_1, \ldots, U_l), \quad \mathsf{\Sigma}_l = \mathrm{diag}(\sigma_1, \ldots, \sigma_l), \quad \mathsf{V}_l = (V_1, \ldots, V_l). \tag{3.9}$$

It is a standard result from linear algebra that X_l is the optimal rank-l approximation of X with respect to the Frobenius norm of linear operators.

More abstractly, implicit to (3.6) is the notion that the dataset induces a linear map $\mathsf{X} : \mathbb{R}^S \mapsto \mathbb{R}^N$ taking the so-called "chronos" space of temporal patterns, $\mathbb{R}^S$, to the "topos" space of spatiotemporal patterns, $\mathbb{R}^N$, via matrix multiplication [72], i.e., $f \mapsto \mathsf{X}f$ with $f \in \mathbb{R}^S$. This picture will be useful in the development of NLSA algorithms ahead.

3.2.3 Spaces of temporal patterns

Another useful way of interpreting the chronos modes is to view them as scalar functions on the data manifold. In particular, we think of the components of each $(f_1, \ldots, f_S) \in \mathbb{R}^S$, as the values of a function $f : \mathcal{M} \mapsto \mathbb{R}$ sampled at the states z_i in (3.1), i.e., $f(z_i) = f_i$. In particular, to each V_j from (3.6) we associate a scalar function $v_j(z_i) = V_{ij}$. The main tenet in NLSA algorithms is that the extracted temporal modes should belong in low-dimensional families of "well-behaved" functions on $\mathcal{M}$. The function space in question is spanned by the leading eigenfunctions of diffusion operators on $\mathcal{M}$, as we now discuss.

Let $\Lambda^p\mathcal{M}$ denote the vector space of smooth p-form fields on $\mathcal{M}$. For our purposes, a diffusion operator $\mathcal{L}$ will be an elliptic, second-order differential operator acting on scalar functions in $\Lambda^0\mathcal{M}$, which annihilates constant functions, i.e.,

$$f = \text{const.} \quad \Longrightarrow \quad \mathcal{L}f = 0. \tag{3.10}$$

An important theoretical result (e.g., [89, 90]) is that every diffusion operator $\mathcal{L}$ induces a Riemannian geometry on $\mathcal{M}$, in the sense that one can associate to $\mathcal{L}$ a unique metric tensor k. More specifically, to every $\mathcal{L}$ there corresponds a unique codifferential operator $\delta : \Lambda^1\mathcal{M} \mapsto \Lambda^0\mathcal{M}$ which produces the factorization

$$\mathcal{L} = \delta d, \tag{3.11}$$

and gives the metric implicitly through the relation

$$\langle \delta\omega, f\rangle_k = \int_{\mathcal{M}} \delta\omega\, f\, d\mu = \langle \omega, df\rangle_k = \int_{\mathcal{M}} \sum_{\alpha,\beta=1}^{m} k^{\alpha\beta} w_\alpha\, d_\beta f\, d\mu. \tag{3.12}$$

Here, f and ω are arbitrary smooth scalar functions and one-forms, $\langle\cdot,\cdot\rangle_k$ the Hodge inner product between p-forms, $d : \Lambda^p\mathcal{M} \mapsto \Lambda^{p+1}\mathcal{M}$ the exterior derivative, $d\mu = \sqrt{\det k}\, du^1 \wedge \cdots \wedge du^m$ the volume form of k, and $k^{\alpha\beta}$ the components of the "inverse metric" associated with k in the u^μ coordinates. A local expression for the codifferential in terms of the metric is

$$\delta(\omega) = -\frac{1}{\sqrt{\det k}} \sum_{\alpha,\beta=1}^{m} \frac{\partial}{\partial u^\alpha}\left(k^{\alpha\beta}\sqrt{\det k}\omega_\beta\right). \tag{3.13}$$

It follows from (3.13) that the Riemannian metric k associated with $\mathcal{L}$ has the property that the corresponding codifferential δ is the adjoint of d with respect to the Hodge inner product. This construction leads naturally to a normalized Dirichlet form

$$E_k(f) = \frac{\langle f, \mathcal{L}f\rangle_k}{\|f\|_k^2} = \frac{\langle df, df\rangle_k}{\|f\|_k^2} \geq 0, \quad \|f\|_k = \langle f, f\rangle_k^{1/2}, \tag{3.14}$$

which characterizes how strongly oscillatory a scalar function f is. Note that $E_k(f)$ depends significantly on k.

Let $\phi_0, \phi_1, \ldots$ be normalized eigenfunctions of $\mathcal{L}$ with corresponding eigenvalues $\lambda_0, \lambda_1, \ldots$,

$$\mathcal{L}\phi_i = \lambda_i\phi_i, \quad \langle\phi_i, \phi_j\rangle_k = \delta_{ij}, \quad 0 = \lambda_0 < \lambda_1 \leq \lambda_2 \cdots. \tag{3.15}$$

The basic requirement in NLSA is that the recovered patterns $v_j(z_i)$ of temporal variability should have bounded Dirichlet form with respect to a Riemannian metric [see (3.29) ahead] constructed in lagged embedding space with an explicit dependence on the dynamical vector field $\dot{F}$. Specifically, for a function $f = \sum_i c_i\phi_i$ we require that $c_i = 0$ for $i > l$, or, equivalently, $E_k(f) \leq \lambda_l$. Operationally, this criterion is enforced by introducing the l-dimensional space of functions spanned by the leading l eigenfunctions of $\mathcal{L}$,

$$\Phi_l = \text{span}\{\phi_0, \ldots, \phi_{l-1}\}, \quad \dim\Phi_l = l, \tag{3.16}$$

and replacing the linear map X in (3.6) by a linear map A_l whose domain is Φ_l. We will return to this point in Section 3.2.6.

We remark that this viewpoint is fundamentally different from SSA and related variance-optimizing algorithms. In those algorithms, the unimportant features of the data are spanned by vectors in embedding space onto which the dataset projects weakly. On the other hand, in NLSA, the unimportant features are those which require large-λ_i basis functions on the data manifold to be described. In particular, there may be temporal modes of variability that carry a small portion of the variance of the total signal, but are "large-scale" on $\mathcal{M}$ in the sense of small E_k. Such modes will generally not be accessible to SSA algorithms.

3.2.4 Discrete formulation

In practical applications, one has seldom access to a densely sampled smooth manifold $\mathcal{M}$. However, using the machinery of DEC [81, 83] and spectral graph theory [84, 85], it is possible to design an algorithm which has the same fundamental properties as the continuous formulation in Section (3.2.3), but is intrinsically discrete (i.e., not a discretization of a continuous theory). In this regard, the main role of the continuous picture is to provide a guideline for building a discrete algorithm.

Let $M = \{z_q, z_{q+1}, \dots, z_{S+q-1}\} \subset \mathcal{M}$ be the discrete set of states on the attractor which are available for data analysis after time-lagged embedding [the initial $q-1$ states, $z_1, \dots, z_{q-1}$, must be discarded in order to apply (3.4)]. The first step of the classical procedure for building a discrete diffusion operator L analogous to $\mathcal{L}$ in (3.11) is to identify the spaces $\Lambda^0 M$ and $\Lambda^1 M$ of scalar-valued functions and 1-forms on the discrete dataset M. These spaces consist of functions defined on the vertices M and edges $M \times M$, respectively, of an undirected graph, whose nodes $1, \dots, S$ correspond to the states $z_q, \dots, z_{S+q-1}$ of the dynamical system at which observations are taken. That graph is further equipped with an ergodic, reversible[1] Markov chain (typically constructed through a kernel, as described in Section 3.2.5) whose state space is M. That is, we have

$$\sum_{i=1}^{S} \pi_i p_{ij} = \pi_j, \quad \pi_i p_{ij} = \pi_j p_{ji}, \tag{3.17}$$

where p and π are the transition probability matrix and invariant distribution of the Markov chain, respectively. The latter are used to construct the inner products $\langle \cdot, \cdot \rangle_p$ of $\Lambda^0 M$ and $\Lambda^1 M$ via the formulas

$$\langle f, f' \rangle_p = \sum_{i=1}^{S} \pi_i f(i) f'(i), \quad \langle \omega, \omega' \rangle_p = \sum_{i,j=1}^{S} \pi_i p_{ij} \omega([ij]) \omega'([ij]), \tag{3.18}$$

where $f, f' \in \Lambda^0 M$, $\omega, \omega' \in \Lambda^1 M$, and $[ij]$ is the edge connecting vertices i and j. Introducing the discrete exterior derivative $d : \Lambda^0 M \mapsto \Lambda^1 M$ with $df([ij]) = f(j) - f(i)$, the codifferential $\delta : \Lambda^1 M \mapsto \Lambda^0 M$ is defined as the adjoint of d with respect to the $\langle \cdot, \cdot \rangle_p$ inner product. That is, for any $f \in \Lambda^0 M$ and $\omega \in \Lambda^1 M$ we have

$$\langle \omega, df \rangle_p = \langle \delta \omega, f \rangle_p. \tag{3.19}$$

An explicit formula for δ (which must be modified if p is not reversible) is

$$\delta \omega(i) = \sum_{j=1}^{S} p_{ij} (\omega([ji]) - \omega([ij])). \tag{3.20}$$

Equations (3.18) and (3.20) are the discrete counterparts of (3.12) and (3.13), respectively.

With these definitions, the discrete diffusion operator is constructed in direct analogy to (3.11), viz.

$$L = \delta d, \quad Lf(i) = 2 \sum_{i,j=1}^{S} p_{ij} (f(i) - f(j)), \quad f \in \Lambda^0 M. \tag{3.21}$$

This operator provides a tool for computing orthonormal basis functions of $\Lambda^0 M$ through its associated eigenfunctions,

$$L\phi_i = \lambda_i \phi_i, \quad \langle \phi_i, \phi_j \rangle_p = \delta_{ij}, \quad 0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots. \tag{3.22}$$

Moreover, it provides a measure for the oscillatory character of a function $f \in \Lambda^0 M$ through the associated Dirichlet form [cf. (3.14)], $E_p(f) = \langle f, Lf \rangle_p / \|f\|_p^2$. The spaces of admissible temporal patterns in NLSA,

$$\Phi_l = \mathrm{span}\{\phi_0, \dots, \phi_{l-1}\}, \tag{3.23}$$

are modeled after (3.16), and have the property

$$E_p(f) \leq \lambda_l \quad \text{for every } f \in \Phi_l. \tag{3.24}$$

[1] Reversibility of the Markov chain is not strictly necessary, but it simplifies the expression for the codifferential in (3.20). Moreover, Markov chains derived from kernels are reversible by construction. See [82] for more general expressions applicable to non-reversible Markov chains, as well as higher-order forms.

3.2.5 Dynamics-adapted kernels

In order to turn the framework of Section 3.2.4 into a complete algorithm, we must specify the Markov transition probability matrix p associated with the discrete diffusion operator L in (3.21). Here, we follow the widely-adopted approach in the literature [e.g., 84, 85, 88, 91–94], whereby p is constructed through a suitable local kernel whose asymptotic properties provide a connection between L and a diffusion operator $\mathcal{L}$ in the continuum limit. In other words, the design of the Markov matrix employed in the discrete algorithm is informed by the asymptotic properties of the kernel and the Riemannian geometry associated with $\mathcal{L}$ via (3.12). Because kernels can be computed using only observed quantities in data space without having to know a priori the structure of the underlying phase space $\mathcal{F}$ and flow F_t, this approach opens the possibility of model-free analysis of dynamical system data [93].

Recall that a kernel is a function which maps pairs of states in $\mathcal{M}$ to a positive number, and decays exponentially fast away from the basepoint at which it is evaluated. A standard choice in this context is the isotropic Gaussian kernel [84, 85, 91, 92],

$$\bar{K}_\epsilon(z_i, z_j) = \exp(-\|H(z_i) - H(z_j)\|^2/\epsilon), \tag{3.25}$$

where ϵ is a positive parameter and $\|\cdot\|$ the canonical Euclidean norm. In writing down (3.25) using the delay-coordinate map H from (3.4) we have made explicit the fact that our focus is kernels defined lagged-embedding space. In NLSA, we work with a "locally scaled" kernel

$$K_{\delta t}(z_i, z_j) = \exp(-\|H(z_j) - H(z_i)\|^2/(\|\xi_i\|\|\xi_j\|)), \quad \xi_i = X_i - X_{i-1}, \tag{3.26}$$

where ξ_i is the displacement vector between temporal nearest neighbors. One may explicitly verify that the quantities ξ_i and ξ_j are finite-difference approximations to the dynamical vector field carried along to lagged-embedding space [94], i.e.,

$$\dot{F}_*|_{z_i} = \xi_i/\delta t + O(\delta t). \tag{3.27}$$

Thus, the NLSA kernel depends on the dynamics implicitly through time-lagged embedding, and explicitly through ξ.

Given a choice of kernel such as the examples above, it is possible to construct a Markov transition probability p by performing suitable normalizations to convert $K_{ij} = K(z_i, z_j)$ to a row-stochastic matrix. Here, we adopt the normalization procedure developed by Coifman and Lafon [85] in the diffusion map (DM) family of algorithms. In DM, the Markov matrix p is constructed from K by introducing a scalar parameter α and performing the sequence of operations

$$\begin{aligned} Q(z_i) &= \sum_{z_j \in M} K(z_i, z_j), \quad \tilde{K}(z_i, z_j) = K(z_i, z_j)/(Q_i Q_j)^\alpha, \\ \tilde{Q}(z_i) &= \sum_{z_j \in M} \tilde{K}(z_i, z_j), \quad p_{ij} = \tilde{K}(z_i, z_j)/\tilde{Q}(z_i). \end{aligned} \tag{3.28}$$

In [85] it was shown that with this definition of p and $\alpha = 1$ the discrete diffusion operator L associated with an isotropic, exponentially decaying kernel converges as $\epsilon \to 0$ (and the sample number S grows at least as fast as $\epsilon^{-m/2-1}$ [95]) to $\mathcal{L} = \Delta$, where Δ is the Laplace-Beltrami operator associated with the Riemannian metric induced on $\mathcal{M}$ through the embedding $\mathcal{M} \mapsto H(\mathcal{M})$. Importantly, the $L \to \Delta$ convergence holds even if the sampling density on $\mathcal{M}$ is non-uniform relative to the Riemannian volume form. The metric associated with the kernel in (3.25) is the induced metric h in (3.5) determined by the delay-coordinate mapping.

This result was extended to the case of anisotropic kernels by Berry [93], who showed that under relatively weak assumptions [which are met by both of the kernels in (3.25) and (3.26)] the asymptotic diffusion operator and the induced metric k is determined by the Hessian of K evaluated at $z_j = z_i$. In particular, in the limit $\delta t \to 0$ the locally-scaled kernel (3.26) leads to the induced metric [94]

$$k = h/\|\dot{F}\|_h^2, \quad \|\dot{F}\|_h^2 = h(\dot{F}, \dot{F}). \tag{3.29}$$

Motivated by this asymptotic result, we work throughout with the $\alpha = 1$ DM normalization in conjunction with the locally scaled kernel of (3.26).

It follows from (3.29) that the outcome of the $\|\xi_i\|$ scaling factors in (3.26) is a conformal transformation of the metric h in lagged-embedding space with a conformal factor $\|\dot{F}\|_h^2$ given by the squared magnitude of the "phase-space velocity" $\dot{F}$. In [14] this feature was found to be crucial for successful dimensional reduction of a dynamical system with chaotic metastability. Additional properties of ξ_i, in particular, its angle relative to the relative displacement vector $H(z_i) - H(z_j)$, can be incorporated in so-called "cone" kernels with stronger invariance properties [94].

A further desirable outcome of the local scaling by the $\|\xi_i\|$ factors is that the diffusion operator $\mathcal{L}$ and the associated Φ_l spaces of temporal patterns from (3.16) become conformally invariant. In particular, equivalence classes of datasets related by conformal transformations of the metric in lagged embedding space,

$$\tilde{h}|_z = r(z)h|_z, \quad r(z) > 0, \tag{3.30}$$

lead asymptotically to the same $\mathcal{L}$ operator as $\delta t \to 0$.

Scale invariance is also beneficial in situations where H is a composite map $H : \mathcal{M} \mapsto \mathbb{R}^{N_1} \oplus \mathbb{R}^{N_2}$ such that $H(z) = (H_1(z), H_2(z))$ where both H_1 and H_2 are embeddings of $\mathcal{M}$. This scenario arises in practice when one has access to multivariate observations with distinct physical units, but there is no natural way of choosing a norm for the product space $\mathbb{R}^{N_1} \oplus \mathbb{R}^{N_2}$. Because the ratios $\|H_\beta(z_i) - H_\beta(z_j)\|^2 / \|\xi_i^{(\beta)}\| \|\xi_j^{(\beta)}\|$, $\beta \in \{1, 2\}$, are invariant under scaling of the data by a constant (including change of units), the kernels (3.26) computed individually for H_1 and H_2 can be combined into a single product kernel without having to introduce additional scaling parameters, namely

$$K_{\delta t}(z_i, z_j) = \exp\left(-\frac{\|H_1(z_i) - H_1(z_j)\|^2}{\|\xi_i^{(1)}\| \|\xi_j^{(1)}\|} - \frac{\|H_2(z_i) - H_2(z_j)\|^2}{\|\xi_i^{(2)}\| \|\xi_j^{(2)}\|} \right). \tag{3.31}$$

A climate science application of this technique can be found in [18].

3.2.6 Singular value decomposition

Having established the procedure to obtain the temporal spaces Φ_l in (3.23), the next step in NLSA is to form a family of linear maps $\mathsf{A}_l : \Phi_l \mapsto \mathbb{R}^N$, which are represented by $N \times l$ matrices with elements

$$A^\alpha{}_j = \langle X^\alpha, \phi_j \rangle_p = \sum_{i=1}^{S} \pi_i X^\alpha(i) \phi_j(i), \quad X^\alpha(i) = \langle e_\alpha, X_i \rangle_{\mathbb{R}^N}. \tag{3.32}$$

Here, X^α is the scalar-valued function in $\Lambda^0 M$ giving the α-th component of the observed data in an orthonormal basis $e_1, \ldots, e_N$ of $\mathbb{R}^N$. That is, a function $f = \sum_{k=1}^{l} c_k \phi_{k-1}$ in Φ_l , is mapped to $y = \mathsf{A}_l(f)$ with $y = (y^1, \ldots, y^N)$ and $y^\alpha = \sum_{j=1}^{l} A^\alpha{}_j c_j$. These linear maps replace the corresponding map for SSA in (2.1), enforcing the condition (3.24) on the discrete Dirichlet form. The spatial and temporal patterns associated with A_l follow in analogy with (3.6) by performing the SVD

$$\mathsf{A}_l = \mathsf{U}_l \mathsf{\Sigma}_l \mathsf{V}_l^T, \tag{3.33}$$

where $\mathsf{U}_l = (U_1, \ldots, U_N)$ and V_l are $N \times N$ and $l \times l$ orthogonal matrices, and $\mathsf{\Sigma} = \mathrm{diag}(\sigma_1^{(l)}, \ldots, \sigma_{\min\{N,l\}}^{(l)})$ a diagonal matrix of nonnegative singular values. Here, the matrix elements of V_l are expansion coefficients in the ϕ_i basis of Φ_l. In particular, the j-th column of V_l corresponds to a function $v_j \in \Lambda^0 M$ and a function of time $\tilde{v}_j$ given by

$$v_j(i) = \tilde{v}_j(t_i) = \sum_{k=1}^{l} v_{kj} \phi_{k-1}(i), \tag{3.34}$$

The above are the NLSA analogs of the chronos modes (3.7) in classical SSA. By the orthogonality properties of the ϕ_i basis functions, the v_j are orthogonal with respect to the inner product in (3.18). Note that unlike the rank-l truncated U_l matrix from SSA in (3.9), the first l columns of U_l from NLSA are not equal to the first l columns of U_{l+1} (the same is true for $\mathsf{\Sigma}_l$ and V_l). Moreover, the temporal patterns in (3.34) are not linear projections of the data onto the corresponding spatiotemporal patterns U_i.

Consider now the rank-l approximation X_l of the signal X in lagged-embedding space obtained by using the SVD of A_l,

$$\mathsf{X}_l = \mathsf{U}_l \mathsf{\Sigma}_l \mathsf{V}_l^T \mathsf{\Phi}_l^T = \mathsf{X} \mathsf{\Pi} \mathsf{\Phi}_l \mathsf{\Phi}_l^T. \tag{3.35}$$

Here $\mathsf{\Pi} = \mathrm{diag}(\pi_1, \ldots, \pi_S)$ is a diagonal matrix containing the invariant distribution (Riemannian measure) in (3.17), and $\mathit{\Phi}_l = (\phi_0, \ldots, \phi_{l-1})$ an $S \times l$ matrix of eigenfunction values. It follows by comparing (3.35) with (3.9) that the rank-l approximations of the signal in NLSA and SSA differ in their filtering kernel. NLSA filters the data by the diffusion kernel, $\mathsf{\Pi \Phi}_l \mathsf{\Phi}_l$, whereas SSA by the covariance kernel, $\mathsf{V}_l \mathsf{V}_l^T$. Note that besides differences in X_l, the spatiotemporal patterns corresponding to individual singular-vector pairs [i.e., the $\hat{\mathsf{X}}_j$ terms in (3.37)] may differ substantially between the two methods.

As discussed in Sections 3.2.3 and 3.2.4, the parameter l controls the "wavenumber" on the data manifold resolved by the diffusion eigenfunctions spanning $\mathit{\Phi}_l$. On the one hand, working at a tight truncation level (small l) is desirable in order to produce a parsimonious description of the data with minimal risk of overfitting (the variance of the discrete eigenfunction ϕ_l increases with l for a fixed number of samples S [95]). At the same time, a too drastic truncation will inevitably lead to important features of the data being unexplained. A useful heuristic criterion for selecting l is to monitor a relative spectral entropy D_l, measuring changes in the energy distribution among the modes of A_l as l grows [15]. This measure is given by the formula

$$D_l = \sum_{i=1}^{l} p_i^{(l+1)} \log(p_i^{(l+1)} / \hat{p}_i^{(l+1)}), \tag{3.36}$$

with $p_i^{(l)} = (\sigma_i^{(l)})^2 / (\sum_i^l (\sigma_i^{(l)})^2)$, $\hat{p}_i^{(l)} = (\hat{\sigma}_i^{(l)})^2 / (\sum_i^l (\hat{\sigma}_i^{(l)})^2)$, and $(\hat{\sigma}_1^{(l)}, \ldots, \hat{\sigma}_{l-1}^{(l)}, \hat{\sigma}_{(l)}^l) = (\sigma_1^{(l-1)}, \ldots, \sigma_{l-1}^{(l-1)}, \sigma_{l-1}^{(l-1)})$. The appearance of qualitatively new features in the spectrum of A_l is accompanied by spikes in D_l (e.g., Figure 3.2a), suggesting that a reasonable truncation level is the minimum l beyond which D_l settles to small values. Note that the compressed representation of the data in the $N \times l$-sized A_l results in substantial gains in computational efficiency compared to the SVD of the full data matrix X in large-scale applications where the ambient space dimension N and the sample number S are both large (e.g., Section 3.3). Of course, in NLSA one has to perform the pairwise kernel evaluations to form the diffusion operator L, but this computation can be straightforwardly parallelized. Moreover, by virtue of the exponential decay of the kernel, the eigenvalue problem (3.23) can be carried out efficiently using sparse iterative solvers.

Our experience from applications ranging from low-dimensional models [14], to comprehensive numerical models [14, 15, 18], and real-world observations [17, 19], has been that the locally-scaled kernel in (3.26) in conjunction with the $\alpha = 1$ DM normalization in (3.28) and the $\mathit{\Phi}_l$-restricted SVD in (3.33), leads to superior timescale separation and ability to detect physically-meaningful low-variance patterns which are not accessible to classical linear-projection techniques such as PCA and SSA. However, a complete theoretical understanding of the SVD procedure, as well as its potential limitations, is still lacking.

3.2.7 Projection to data space

The final step in the NLSA pipeline is to construct the spatiotemporal patterns $\hat{x}_i^j$ in n-dimensional data space associated with the corresponding singular vectors and values, $\{U_j, V_j, \sigma_j^{(l)}\}$, of the A_l map in (3.32). Because A_l is a linear map, this procedure is significantly more straightforward and unambiguous than in methods based on nonlinear mapping functions (e.g., [96, 97]), and consists of two steps: (i) Compute the $N \times S$ matrix $\hat{\mathsf{X}}_j$ containing the j-th spatiotemporal pattern in lagged embedding space, $\hat{\mathsf{X}}_j = U_j \sigma_j^l V_j^T \mathsf{\Phi}_l^T$; (ii) decompose each column of $\hat{\mathsf{X}}_j$ into q blocks $\hat{x}_{ij}$ of size n,

$$\hat{\mathsf{X}}_j = \begin{pmatrix} \uparrow & & \uparrow \\ \hat{X}_1^j & \cdots & X_S^j \\ \downarrow & & \downarrow \end{pmatrix} = \begin{pmatrix} \hat{x}_{11} & \cdots & \hat{x}_{1s'} \\ \vdots & \ddots & \vdots \\ \hat{x}_{q1} & \cdots & \hat{x}_{qs'} \end{pmatrix}, \tag{3.37}$$

and take the average over the lagged embedding window,

$$\mathsf{x}_j = (\hat{x}_1^j, \ldots, \hat{x}_s^j), \quad \hat{x}_i^j = \sum_{k=1}^{\min\{q,i\}} \hat{x}_{j,i-k+1} / \min\{q, i\}. \tag{3.38}$$

This leads to s samples in n-dimensional data space, completing the decomposition in (3.2).

3.3 Analysis of infrared brightness temperature satellite data for tropical dynamics

Satellite imagery has been used to study convection-coupled tropical disturbances since the 1970s. Substantial advances in the understanding of tropical waves have been made through linear theories and diagnostics guided by these theories (e.g., [98]). However, convection-coupled tropical motions are highly nonlinear and multiscale. Among the most notable examples is the Madden-Julian oscillation (MJO, e.g., [10]); an eastward-propagating, planetary-scale envelope of organized tropical convection. Originating in the Indian Ocean and propagating eastward over the Indonesian Maritime Continent until its decay in the Western Pacific, the MJO has gross scales in the 30–90-day intraseasonal time range and zonal wavenumber of order 1–4. It dominates the tropical predictability in subseasonal time scales, exerting global influences through tropical–extratropical interactions, affecting weather and climate variability, and fundamentally interfacing the short-term weather prediction and long-term climate projections [99].

Conventional methods for extracting MJO signals from observations and models are linear, including linear bandpass filtering, regression, and EOFs [100]. On the other hand, theory development has suggested that the MJO is a nonlinear oscillator [101, 102]. With a nonlinear temporal filter, the observed MJO appears to be a stochastically driven chaotic oscillator [103].

In this section, we apply NLSA to extract the spatiotemporal patterns of the MJO and other convective processes from satellite infrared brightness temperature over the tropical belt. This analysis is an extension of the work in [17, 19], which considered one-dimensional (1D) latitudinally-averaged data instead of the two-dimensional (2D) infrared brightness temperature field studied here.

3.3.1 Dataset description

The Cloud Archive User Service (CLAUS) Version 4.7 multi-satellite infrared brightness temperature (denoted T_b) [22] is used for this study. Brightness temperature is a measure of the Earth's infrared emission in terms of the temperature of a hypothesized blackbody emitting the same amount of radiation at the same wavelength [$\sim$ 10–11 μm in the CLAUS data]. It is a highly correlated variable with the total longwave emission of the Earth. In the tropics, positive (negative) T_b anomalies are associated with reduced (increased) cloudiness. The global CLAUS T_b data are on a 0.5° longitude by 0.5° latitude fixed grid, with three-hour time resolution from 00 UTC to 21 UTC, spanning July 1, 1983 to June 30, 2006. T_b values range from 170 K to 340 K with approximately 0.67 K resolution.

The subset of the data in the global tropical belt between 15°S and 15°N was taken to create a spatiotemporal dataset sampled at a uniform longitude-latitude grid of $n_x \times n_y = 720 \times 60$ gridpoints (with data space dimension $n = n_x n_y = 43{,}200$) and $s = 67{,}208$ temporal snapshots. Prior to analysis, the missing gridpoint values (less than 1% of ns) were filled via linear interpolation in time. A portion of the data for the period of the observational campaign Tropical Ocean Global Atmosphere Coupled Ocean Atmosphere Response Experiment (TOGA COARE, November 1992 to February 1993) [104], which is studied in Section 3.3.3, is shown in Figure 3.1 in a time-longitude plot of T_b averaged about the equator.

3.3.2 Modes recovered by NLSA

We have applied the NLSA algorithm described in Section 3.2 using an embedding window spanning $\Delta t = 64$ days (d). This amounts to an embedding space dimension $N = qn \approx 2.2 \times 10^7$ for the $\delta t = 3$ hour sampling interval ($q = \Delta t/\delta t = 512$) and 0.5° resolution of our dataset. This choice of embedding window was motivated from our objective to resolve propagating structures such as the MJO with intraseasonal (30–90-d) characteristic timescales. In comparison, Kikuchi et al. [105] used a 10 d window with 5 d increments in their analysis of outgoing longwave radiation (OLR) data using EEOFs. Unlike conventional approaches [105–107], neither bandpass filtering nor seasonal detrending was applied prior to processing the data via NLSA.

For the calculation of the diffusion eigenfunctions in (3.22), we computed the pairwise kernel values from (3.26) in embedding space using brute force, and evaluated the Markov matrix retaining nonzero entries for 5000 nearest neighbors per datapoint.The resulting spectral entropy D_l, computed via (3.36) and normalized to the unit interval by applying the transformation [56]

$$\delta_l = (1 - \exp(-2D_l))^{1/2}, \tag{3.39}$$

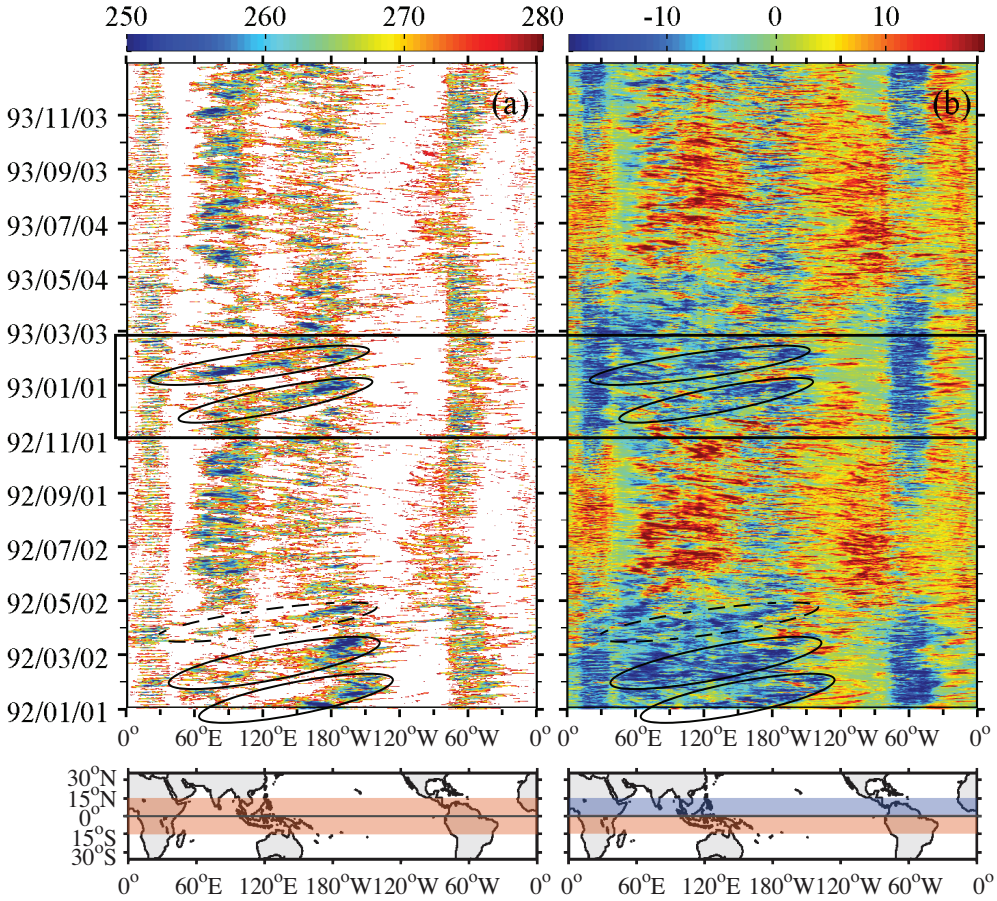

Figure 3.1. Time-longitude section of (a) symmetric and (b) antisymmetric brightness temperature T_b data (in K) from CLAUS for the period 1992–1993. In (a) only thresholded values < 280 K are shown to emphasize convective activity. The bottom map in (a) indicates that the symmetric component was obtained via averaging over 15°S to 15°N. The antisymmetric component in (b) was obtained by subtracting the values at the northern latitudes from the corresponding southern latitudes. The boxed interval corresponds to the TOGA-CORE period. Ovals mark significant MJO events.

is shown in Figure 3.2a. As described in Section 3.2.6, D_l exhibits a series of spikes as l increases from small to moderate values ($l \sim 20$), which correspond to qualitatively new spatiotemporal patterns entering in the spectrum of A_l. Eventually, D_l settles to small values for $l \gtrsim 25$. On the basis of the results in Figure 3.2a, hereafter we set the temporal space dimension in (3.23) to $l = 27$. The singular values $\sigma_i^{(l)}$ of the associated A_l linear map from (3.32) are displayed in Figure 3.2b.

With these NLSA parameter values, the recovered spatiotemporal modes describe several convective processes of interest operating in a wide range of spatial and temporal scales. Representative temporal patterns $\tilde{v}_j(t_i)$ from (3.34) are shown in Figure 3.3. Snapshots of the corresponding spatiotemporal patterns $\hat{x}_i^j$ from (3.38) are displayed in Figures 3.4 and 3.5. The properties of these modes are as follows.

- *Modes (1, 2) and (6,7).* As manifested by the prominent lines at the once- and twice-per year frequencies in their temporal Fourier spectra, these modes respectively describe annual (Figures 3.3a and 3.3b) and semiannual (Figures 3.3f and 3.3g) periodic processes, which are expected to be prominent in tropical T_b signals. In the spatial domain, Modes $(1, 2)$ are characterized by T_b anomalies of opposite sign in the North and South Hemispheres. The December 25 snapshot of Figure 3.4a corresponds to the dry season in the tropical North Hemisphere and wet season in the tropical South Hemisphere. The semiannual

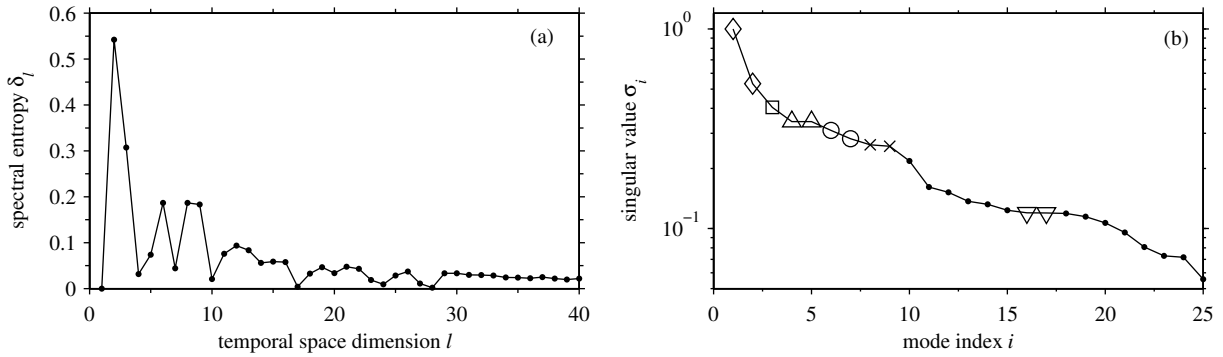

Figure 3.2. (a) Spectral entropy δ_l from (3.39) and (b) singular values $\sigma_i^{(l)}$ of the A_l linear map (3.32) with $l = 27$. The highlighted modes in (b) are the ($\Diamond$) annual; ($\square$) interannual; ($\triangle$) symmetric diurnal; ($\bigcirc$) MJO; ($\times$) semiannual; ($\triangledown$) antisymmetric diurnal modes. See Figure 3.3 for the corresponding temporal patterns.

pair captures the march of intensified deep convection in the intratropical convergence zone (ITCZ), South Pacific convergence zone (SPCZ), Indo-Pacific warm pool, monsoons, and tropical storm tracks across the Southern and Northern Hemispheres. This pair of modes explains the apparent migration and amplitude modulation of convection signals due to the north-south asymmetry in land mass and bathymetry within the tropical belt.

- *Mode 3.* Due to the characteristic dipole pattern over the equatorial Pacific (Figure 3.4c) and presence of spectral power on interannual timescales (Figure 3.3c) this mode is interpreted as the El Niño Southern Oscillation (ENSO) [9] mode. The ENSO can enhance or inhibit MJO propagation by preconditioning the environment of the Western Pacific. For instance, the snapshot of Figure 3.4c coincides with an amplifying El Niño phase with warm sea surface temperature and enhanced convection (i.e., negative T_b anomaly), which is conducive to MJO propagation in the Western Pacific. On the other hand, MJO propagation is inhibited during La Niña years (not shown here).

- *Modes (8, 9).* This two-mode family corresponds the manifestation of the MJO in T_b data. Characterized by broad intraseasonal peaks (20–90 days) in their frequency spectra and phase-locked in quadrature (Figures 3.3h and 3.3i), these modes describe a 5000 km-scale eastward-propagating envelope of organized convection (Figures 3.4f and 3.5). This structure originates over the Indian Ocean, and propagates eastward until it decays upon reaching the cold waters of the Central Pacific. The presence of semiannual lines in the frequency spectra of Modes $(8, 9)$ is consistent with the fact that MJO events occur preferentially in boreal winter (November–March).

- *Modes* $(4, 5)$ *and* $(16, 17)$. Characterized by dominant peaks over the once-per-day frequency in their Fourier spectra (Figures 3.3d, 3.3e, 3.3j, and 3.3k), these modes describe diurnal convective variability. The corresponding spatiotemporal patterns (Figures 3.4b and 3.4g) are most prominent over land, where the diurnal cycle of convection is most active. The major difference between these modes is that $(4, 5)$ are predominantly symmetric about the equator, whereas $(16, 17)$ are predominantly antisymmetric. The symmetric pair is active year-round, but the antisymmetric pair is strongly modulated by the seasonal cycle.

The availability of this family of modes, determined by an objective algorithm requiring no preprocessing of the data, enables one to study interdependencies between modes of convection across multiple temporal and spatial scales. For instance, the interdependencies between ENSO, MJO, and the diurnal cycle are topics of significant current interest [19, and references therein]. Such analyses are outside the scope of this paper, but we refer the reader to [19] for a study involving NLSA modes from 1D CLAUS T_b data averaged about the equator, where a comparison between the NLSA and SSA modes is also made.

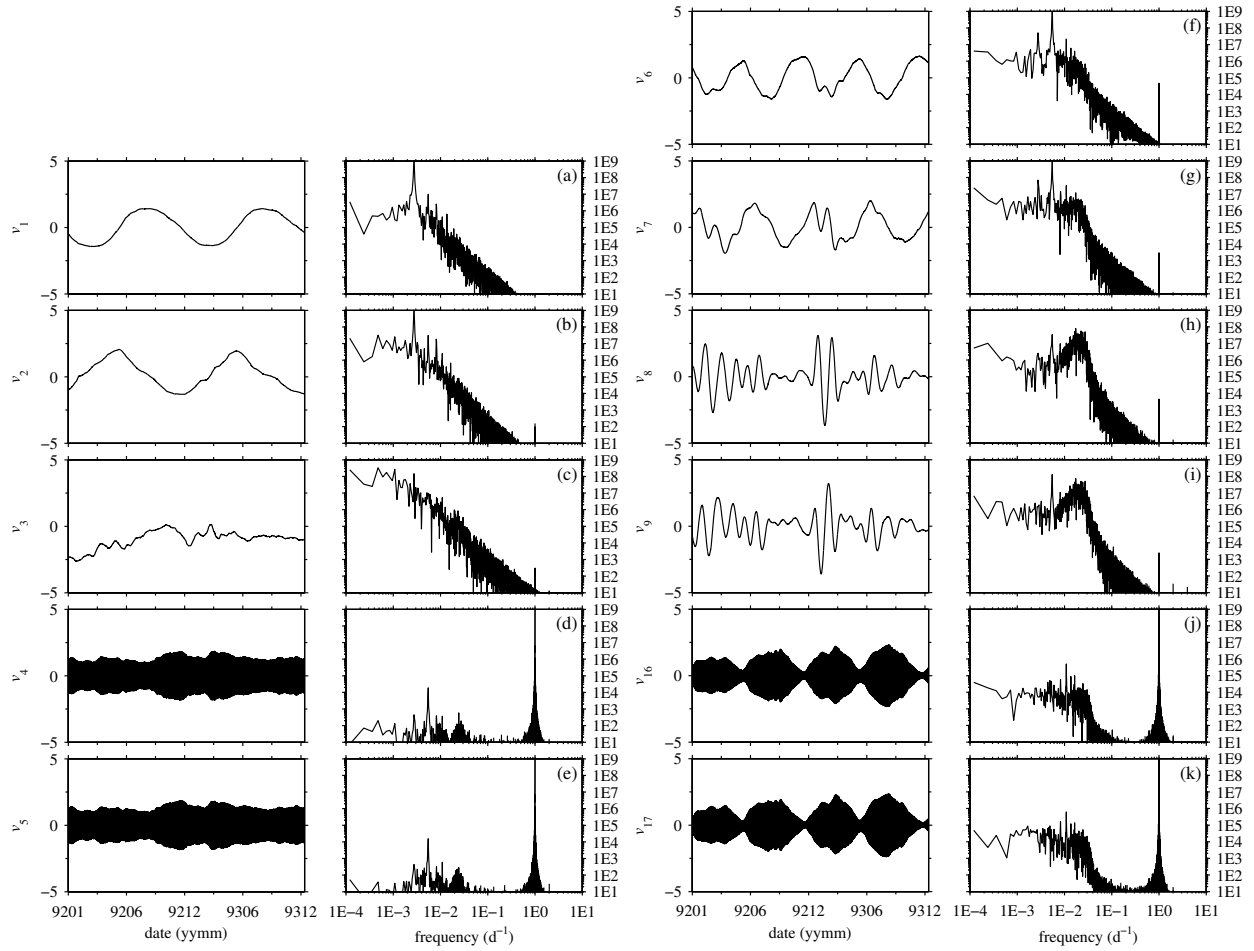

Figure 3.3. Representative NLSA temporal patterns of CLAUS T_b data for the period of January 1992–December 1993 and their frequency power spectra. (a, b) Annual modes; (c) interannual (ENSO) mode; (d, e) symmetric diurnal pair; (f, g) semiannual modes; (h, i) MJO pair; (j, k) antisymmetric diurnal pair.

3.3.3 Reconstruction of the TOGA COARE MJOs

Two complete MJO events were observed during the TOGA COARE period (e.g., [108]). Figure 3.5 shows reconstructions of these events based on the NLSA modes. This reconstruction captures the salient features of the propagating envelope of deep convection associated with the MJO, including the initiation of enhanced deep convection (hence cold anomalies) over the Indian Ocean, the passage over the Maritime Continent, and the arrival and demise near the date line. The first event started near 75° E in late November, subsequently crossed the Maritime Continent around 100°–150°E, then disappeared near 170°W around January 10. The second event, being slightly faster than the first, started around January 5, and reached the central Pacific in early February. A third event started in March, after the end of the TOGA COARE period.

The TOGA COARE period was coincident with the amplifying phase of an El Niño event; therefore, the convective MJO superclusters propagated further east beyond the date line, where during normal years the cold sea surface temperature is not conducive to deep convection. The eastward-propagation speed of the reconstructed MJO events is consistent with the observed $\sim$ 4–5 m s^{-1} value.

4 Synthesis

In this paper, we have reviewed two examples of applied mathematics techniques for data analysis in dynamical systems: (i) Methods for quantifying predictability and model error based on data clustering and

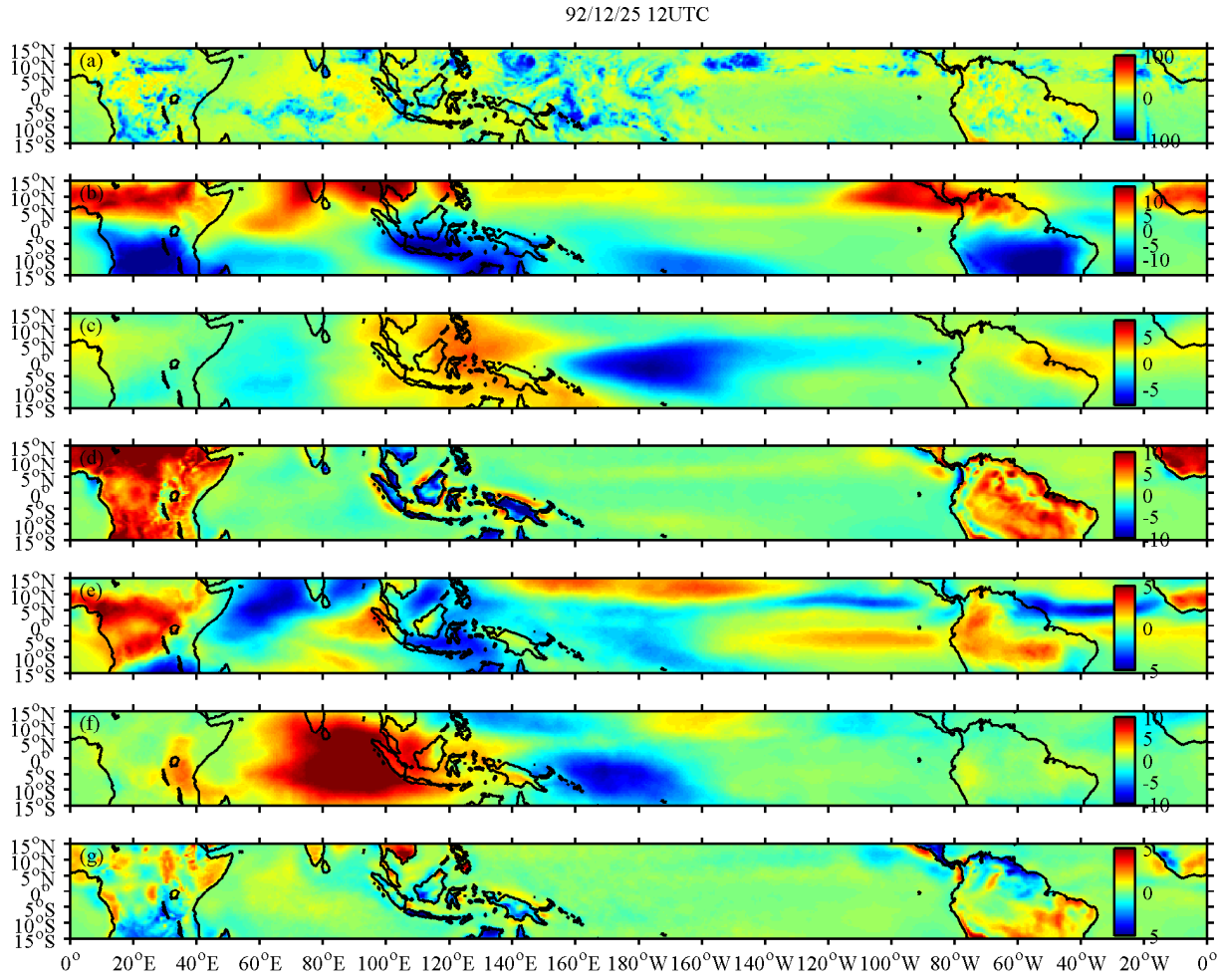

Figure 3.4. A snapshot of reconstructed CLAUS T_b (in K) data for December 25, 1992, 12UTC using the NLSA modes highlighted in Figures 3.2 and 3.3. Negative T_b anomalies (blue colors) indicate enhanced convection. Positive T_b anomalies (red colors) indicate decreased cloudiness. (a) Raw data; (b) annual modes, $x_1 + x_2$; (c) interannual (ENSO) mode, x_3; (d) latitudinally symmetric diurnal pair, $x_4 + x_5$; (e) semiannual modes, $x_6 + x_7$; (f) MJO pair, $x_8 + x_9$; (g) latitudinally antisymmetric diurnal pair, $x_{16} + x_{17}$. The prominent MJO event in (f) occurring over the Western Pacific was observed by the TOGA COARE field campaign [108].

information theory (Section 2.2); (ii) nonlinear Laplacian spectral analysis (NLSA) algorithms for extracting spatiotemporal patterns from high-dimensional data (Section 3). We have highlighted these techniques with applications to climate atmosphere ocean science; in particular, predictability assessment and Markov modeling of circulation regimes in a simple ocean model (Section 2.5), and extraction of modes of organized convection in the tropics from infrared brightness temperature (T_b) satellite data (Section 3.3).

A common theme in these methods has been the coarse-grained geometry of the data. In Section 2 we saw how a discrete partition of the space of initial data (constructed empirically through data clustering) can be used in conjunction with information theory to place practically computable lower bounds to the predictability of observables in dynamical systems and the error of forecasting these observables with imperfect models. In Section 3, the machinery of discrete exterior calculus and spectral graph theory was combined with delay-coordinate mappings of dynamical systems to extract spatiotemporal modes of variability which are describable in terms of low-dimensional sets of diffusion eigenfunctions, selected according to a "low-wavenumber" criterion on the data manifold formulated in an intrinsically discrete setting.

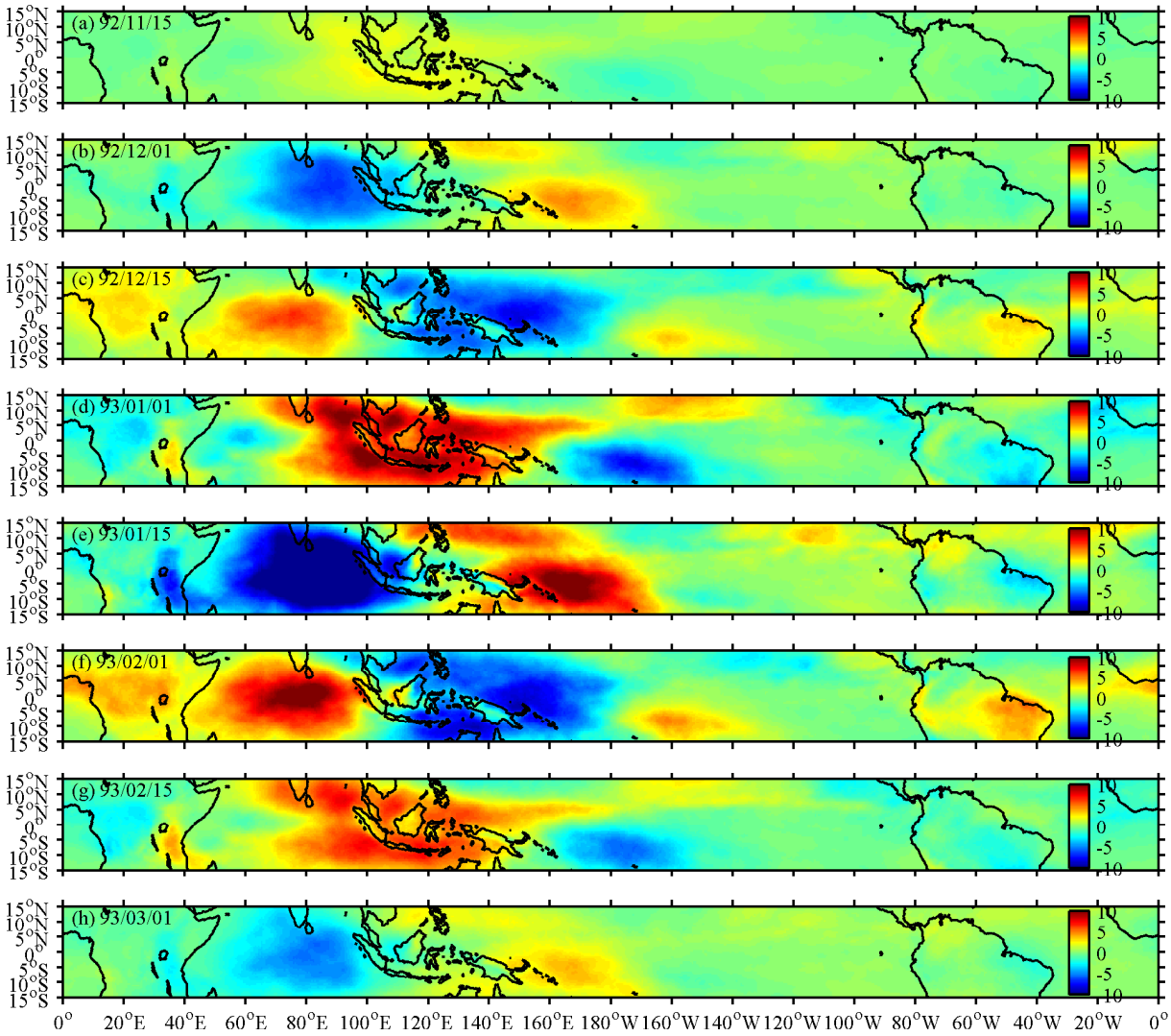

Figure 3.5. Reconstruction of the MJO waivetrain observed during the TOGA COARE intensive observing period (IOP) November 1992–March 1993. The color maps show temperature T_b anomalies (in K) obtained from the NLSA MJO modes of Figures 3.3h and 3.3i projected to data space via (3.38); i.e., x_8+x_9. Blue (red) colors correspond to increased convection (decreased cloudiness). (a) No MJO activity is present; (b, c, d) the first MJO initiates over the Indian Ocean, propagates eastward over the Indonesian Maritime Continent, and decays after reaching the 180° dateline; (e, f, g) a second, stronger, MJO event with an initiation signal over East Africa; (h) a weak third event starting at the end of the TOGA COARE IOP. See Figure 3.1 for the manifestation of these events in time-longitude section of the raw data.

The techniques in (i) and (ii) can be naturally combined. In particular, recall that in Section 2.5 data space was spanned by the leading 20 principal components (PCs) of the oceanic streamfunction. One can consider replacing the PCs with the temporal patterns recovered by NLSA, and seek predictable patterns in that space. A natural, but challenging application is to use the predictability framework of Section 2 to study MJO predictability (a problem of wide practical impact [99]) in the space of NLSA modes recovered from brightness temperature data and other relevant fields. In this case, it is likely that the complexity of the data in modal space compared to the 1.5-layer ocean model will require a modification of the K-means algorithm used in Section 2.5 in order to identify states with high MJO predictability. Likewise, we believe that it would be fruitful to explore alternative formulations to the locally-scaled NLSA kernel in (3.26) (which has been

designed having a deterministic dynamical system in mind) to deal with stochastic dynamics. The recent work in [109] should be relevant in this context. Finally, as mentioned in Section 3.2.6, an open theoretical problem is the justification (and potential improvement) of the truncated SVD step in NLSA. We plan to pursue these topics in future work.

Acknowledgments

The authors would like to thank Rafail Abramov, Tyrus Berry, Grant Branstator, Mitch Bushuk, John Harlim, Illia Horenko, and Wen-wen Tung for their explicit and implicit contributions to this work. This work was supported by the Office of Naval Research, including ONR DRI grants N25-74200-F6607 and N00014-10-1-0554, and ONR MURI grant 25-74200-F7112. The results of Section 3.3 were obtained using the CLAUS archive held at the British Atmospheric Data Centre, produced using ISCCP source data distributed by the NASA Langley Data Center.

References

[1] P. Holmes, J. L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, Cambridge, 1996.

[2] K. R. Sreenivasan and R. A. Antonia. The phenomenology of small-scale turbulence. *Annu. Rev. Fluid Mech.*, 29(1):435–472, 1997. doi:10.1146/annurev.fluid.29.1.435.

[3] M. A. Katsoulakis and D. G. Vlachos. Coarse-grained stochastic processes and kinetic Monte Carlo simulators for the diffusion of interacting particles. *J. Chem. Phys.*, 119(18):9412–9427, 2003. doi:10.1063/1.1616513.

[4] M. A. Katsoulakis, A. J. Majda, and A. Sopasakis. Intermittency, metastability and coarse-graining for coupled deterministic–stochastic lattice systems. *Nonlinearity*, 19:1021–1047, 2006. doi:10.1088/0951-7715/19/5/002.

[5] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. Computation of essential molecular dynamics by subdivision techniques I: basic concept. *Lect. Notes Comp. Sci. Eng.*, 4:98, 1999.

[6] B. Nadler, S. Lafon, R. R. Coifman, and I. Kevrikedes. Diffusion maps, spectral clustering, and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.*, 21:113–127, 2006. doi:10.1016/j.acha.2005.07.004.

[7] A. J. Majda and X. Wang. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge University Press, Cambridge, 2006.

[8] V. P. Dymnikov and A. N. Filatov. *Mathematics of Climate Modeling*. Birkhäuser, Boston, 1997.

[9] K. E. Trenberth. The definition of El Niño. *Bull. Amer. Meteor. Soc.*, 78(12):2771–2777, 1997. doi:10.1175/1520-0477(1997)078<2771:tdoeno>2.0.co;2.

[10] R. A. Madden and P. R. Julian. Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, 29(6):1109–1123, 1972. doi:10.1175/1520-0469(1972)029<1109:dogscc>2.0.CO;2.

[11] 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*, 25:1793–1813, 2012. doi:10.1175/2011jcli4143.1.

[12] 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*, 25: 1814–1826, 2012. doi:10.1175/jcli-d-11-00110.1.

[13] D. Giannakis, A. J. Majda, and I. Horenko. Information theory, model error, and predictive skill of stochastic models for complex nonlinear systems. *Phys. D.*, 241:1735–1752, 2012. doi:10.1016/j.physd.2012.07.005.

[14] D. Giannakis and A. J. Majda. Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proc. Natl. Acad. Sci.*, 109(7):2222–2227, 2012. doi:10.1073/pnas.1118984109.

[15] D. Giannakis and A. J. Majda. Comparing low-frequency and intermittent variability in comprehensive climate models through nonlinear Laplacian spectral analysis. *Geophys. Res. Lett.*, 39:L10710, 2012. doi:10.1029/2012GL051575.

[16] D. Giannakis and A. J. Majda. Nonlinear Laplacian spectral analysis: Capturing intermittent and low-frequency spatiotemporal patterns in high-dimensional data. *Stat. Anal. Data Min.*, 6(3):180–194, 2013. doi:10.1002/sam.11171.

[17] D. Giannakis, W.-w. Tung, and A. J. Majda. Hierarchical structure of the Madden-Julian oscillation in infrared brightness temperature revealed through nonlinear Laplacian spectral analysis. In *2012 Conference on Intelligent Data Understanding (CIDU)*, pages 55–62, Boulder, Colorado, 2012. doi:10.1109/cidu.2012.6382201.

[18] M. Bushuk, D. Giannakis, and A. J. Majda. Reemergence mechanisms for North Pacific sea ice revealed through nonlinear Laplacian spectral analysis. *J. Climate*, 2013. submitted.

[19] W.-w. Tung, D. Giannakis, and A. J. Majda. Symmetric and antisymmetric Madden-Julian oscillation signals in tropical deep convective systems. *J. Atmos. Sci.*, 2013. submitted.

[20] J. D. McCalpin and D. B. Haidvogel. Phenomenology of the low-frequency variability in a reduced-gravity quasigeostrophic double-gyre model. *J. Phys. Oceanogr.*, 26(5):739–752, 1996.

[21] G. A. Meehl et al. Decadal prediction. can it be skillful? *Bull. Amer. Meteor. Soc.*, 90(10):1467–1485, 2009. doi:10.1175/2009bams2778.1.

[22] K. Hodges, D.W. Chappell, G.J. Robinson, and G. Yang. An improved algorithm for generating global window brightness temperatures from multiple satellite infra-red imagery. *J. Atmos. Oceanic Technol.*, 17:1296–1312, 2000. doi:10.1175/1520-0426(2000)017<1296:aiafgg>2.0.co;2.

[23] E. N. Lorenz. The predictability of a flow which possesses many scales of moition. *Tellus*, 21(3):289–307, 1969. doi:10.1111/j.2153-3490.1969.tb00444.x.

[24] E. S. Epstein. Stochastic dynamic prediction. *Tellus*, 21:739–759, 1969. doi:10.1111/j.2153-3490.1969.tb00483.x.

[25] J. Berner and G. Branstator. Linear and nonlinear signatures in planetary wave dynamics of an AGCM: Probability density functions. *J. Atmos. Sci.*, 64:117–136, 2007. doi:10.1175/jas3822.1.

[26] A. J. Majda, C. Franzke, A. Fischer, and D. T. Crommelin. Distinct metastable atmospheric regimes despite nearly Gaussian statistics: A paradigm model. *Proc. Natl. Acad. Sci.*, 103(22):8309–8314, 2006. doi:10.1073/pnas.0602641103.

[27] C. Franzke, A. J. Majda, and G. Branstator. The origin of nonlinear signatures of planetary wave dynamics: Mean phase space tendencies and contributions from non-Gaussianity. *J. Atmos. Sci.*, 64: 3988, 2007. doi:10.1175/2006jas2221.1.

[28] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. *Linear Alg. Appl.*, 315:39, 2000. doi:10.1016/s0024-3795(00)00095-1.

[29] A. J. Majda and X. Wang. Linear response theory for statistical ensembles in complex systems with time-periodic forcing. *Comm. Math. Sci.*, 8(1):145–172, 2010.

[30] I. Horenko. On the identification of nonstationary factor models and their application to atmospheric data analysis. *J. Atmos. Sci.*, 67(5):1559–1574, 2010. doi:10.1175/2010jas3271.1.

[31] R. S. Tsay. *Analysis of Financial Time Series*. Wiley, Hoboken, 2010.

[32] M.S. Roulston and L.A. Smith. Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.*, 130(6):1653–1660, 2002. doi:10.1175/1520-0493(2002)130<1653:epfuit>2.0.co;2.

[33] A. J. Majda and B. Gershgorin. Quantifying uncertainty in climage change science through empirical information theory. *Proc. Natl. Acad. Sci.*, 107(34):14958–14963, 2010. doi:10.1073/pnas.1007009107.

[34] C. Franzke, D. Crommelin, A. Fischer, and A. J. Majda. A hidden Markov model perspective on regimes and metastability in atmospheric flows. *J. Climate*, 21(8):1740–1757, 2008. doi:10.1175/2007jcli1751.1.

[35] J. Bröcker, D. Engster, and U. Parlitz. Probabilistic evaluation of time series models: A comparison of several approaches. *Chaos*, 19(4):04130, 2009. doi:10.1063/1.3271343.

[36] C. Penland. Random forcing and forecasting using principal oscillation pattern analysis. *Mon. Weather Rev.*, 117(10):2165–2185, 1989. doi:10.1175/1520-0493(1989)117<2165:rfafup>2.0.co;2.

[37] H. Teng and G. Branstator. Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM. *Climate Dyn.*, 36:1813–1834, 2010. doi:10.1007/s00382-010-0749-7.

[38] A. J. Majda, I. I. Timofeyev, and E. Vanden Eijnden. Systematic strategies for stochastic mode reduction in climate. *J. Atmos. Sci.*, 60:1705, 2003. doi:10.1175/1520-0469(2003)060<1705:ssfsmr>2.0.co;2.

[39] C. Franzke, I. Horenko, A. J. Majda, and R. Klein. Systematic metastable regime identification in an AGCM. *J. Atmos. Sci.*, 66(9):1997–2012, 2009. doi:10.1175/2009jas2939.1.

[40] I. Horenko. On robust estimation of low-frequency variability trends in discrete Markovian sequences of atmospheric circulation patterns. *J. Atmos. Sci.*, 66(7):2059–2072, 2009. doi:10.1175/2008jas2959.1.

[41] I. Horenko. On clustering of non-stationary meteorological time series. *Dyn. Atmos. Oceans*, 49:164–187, 2010. doi:10.1016/j.dynatmoce.2009.04.003.

[42] L.-Y. Leung and G. R. North. Information theory and climate prediction. *J. Climate*, 3(1):5–14, 1990. doi:10.1175/1520-0442(1990)003<0005:itacp>2.0.co;2.

[43] T. Schneider and S. M. Griffies. A conceptual framework for predictability studies. *J. Climate*, 12(10): 3133–3155, 1999. doi:10.1175/1520-0442(1999)012<3133:acffps>2.0.co;2.

[44] K. Sobczyk. Information dynamics: Premises, challenges and results. *Mech. Syst. Signal Pr.*, 15(3): 475–498, 2001. doi:10.1006/mssp.2000.1378.

[45] R. Kleeman. Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, 59(13): 2057–2072, 2002. doi:10.1175/1520-0469(2002)059<2057:mdpuur>2.0.co;2.

[46] D. Cai, R. Kleeman, and A. J. Majda. A mathematical framework for quantifying predictability through relative entropy. *Methods Appl. Anal.*, 9(3):425–444, 2002.

[47] T. DelSole. Predictability and information theory. Part I: Measures of predictability. *J. Atmos. Sci.*, 61 (20):2425–2440, 2004. doi:10.1175/1520-0469(2004)061<2425:paitpi>2.0.co;2.

[48] T. DelSole. Predictability and information theory. Part II: Imperfect models. *J. Atmos. Sci.*, 62(9): 3368–3381, 2005. doi:10.1175/jas3522.1.

[49] R. Kleeman. Information theory and dynamical system predictability. *Entropy*, 13(3):612–649, 2011. doi:10.3390/e13030612.

[50] E. Lorenz. *The Physical Basis of Climate and Climate Modelling*, volume 16 of *GARP Publications Series*, chapter Climate Predictability, pages 132–136. World Meteorological Organization, 1975.

[51] G. Branstator and H. Teng. Two limits of initial-value decadal predictability in a CGCM. *J. Climate*, 23:6292–6311, 2010. doi:10.1175/2010jcli3678.1.

[52] T. A. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, 2 edition, 2006.

[53] I. Horenko. Nonstationarity in multifactor models of discrete jump processes, memory and application to cloud modeling. *J. Atmos. Sci.*, 68(7):1493–1506, 2011. doi:10.1175/2011jas3692.1.

[54] M. H. DeGroot and S. E. Fienberg. Assessing probability assessors: Calibration and refinements. In S. S. Gupta and J. O. Berger, editors, *Statistical Decision Theory and Related Topics III*, volume 1, pages 291–314. Academic Press, New York, 1982.

[55] J. Bröcker. Reliability, sufficiency and the decomposition of proper scores. *Q. J. R. Meteorol. Soc.*, 135: 1512–1519, 2009. doi:10.1002/qj.456.

[56] H. Joe. Relative entropy measures of multivariate dependence. *J. Amer. Stat. Assoc.*, 84(405):157–164, 1989.

[57] T. DelSole and J. Shukla. Model fidelity versus skill in seasonal forecasting. *J. Climate*, 23(18): 4794–4806, 2010. doi:10.1175/2010jcli3164.1.

[58] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, 1986.

[59] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2 edition, 2000.

[60] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov. Relative performance of mutual information estimation nethods for quantifying the dependence among short and noisy data. *Phys. Rev. E*, 76:026209, 2007. doi:10.1103/PhysRevE.76.026209.

[61] P. S. Berloff and J. C. McWilliams. Large-scale, low-frequency variability in wind-driven ocean gyres. *J. Phys. Oceanogr.*, 29:1925–1949, 1999. doi:10.1175/1520-0485(1999)029<1925:lslfvi>2.0.co;2.

[62] N. S. Keenlyside, M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner. Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, 453:84–88, 2008. doi:10.1038/nature06921.

[63] J. W. Hurrell et al. Decadal climate prediction: Opportunities and challenges. In *Proceedings of the OceanObs09 Conference: Sustained Ocean Observations and Information for Society*, pages 21–25, Venice, Italy, 2009.

[64] A. Solomon et al. Distinguishing the roles of natural and anthropogenically forced decadal climate variability: Implications for prediction. *Bull. Amer. Meteor. Soc.*, 92(2):141–156, 2011. doi:10.1175/2010bams2962.1. Submitted.

[65] M. Ghil and A. W. Robertson. "Waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proc. Natl. Acad. Sci*, 99(suppl. 1):2493–2500, 2002. doi:10.1073/pnas.012580899.

[66] I. Horenko. On simultaneous data-based dimension reduction and hidden phase identification. *J. Atmos. Sci.*, 65:1941–1954, 2008. doi:10.1175/2007jas2587.1.

[67] D. T. Crommelin and E. Vanden-Eijnden. Fitting timeseries by continuous-time Markov chains: A quadratic programming approach. *J. Comput. Phys.*, 217(2):782–805, 2006. doi:10.1016/j.jcp.2006.01.045.

[68] P. Metzner, E. Dittmer, T. Jahnke, and C. Schütte. Generator estimation of Markov jump processes based on incomplete observations equidistant in time. *J. Comput. Phys.*, 227(1):353–375, 2007. doi:10.1016/j.jcp.2007.07.032.

[69] B. C. Weare and J. S. Nasstrom. Examples of extended empirical orthogonal function analyses. *Mon. Wea. Rev.*, 110:481–485, 1982. doi:10.1175/1520-0493(1982)110<0481:eoeeof>2.0.co;2.

[70] D. S. Broomhead and G. P. King. Extracting qualitative dynamics from experimental data. *Phys. D*, 20(2–3):217–236, 1986. doi:10.1016/0167-2789(86)90031-x.

[71] R. Vautard and M. Ghil. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Phys. D*, 35:395–424, 1989. doi:10.1016/0167-2789(89)90077-8.

[72] N. Aubry, R. Guyonnet, and R. Lima. Spatiotemporal analysis of complex signals: Theory and applications. *J. Stat. Phys.*, 64:683–739, 1991. doi:10.1007/bf01048312.

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

[74] M. Ghil et al. Advanced spectral methods for climatic time series. *Rev. Geophys.*, 40:1003, 2002. doi:10.1029/2000rg000092.

[75] N. H. Packard et al. Geometry from a time series. *Phys. Rev. Lett.*, 45:712–716, 1980. doi:10.1103/physrevlett.45.712.

[76] 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. doi:10.1007/bfb0091924.

[77] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65(3–4):579–616, 1991. doi:10.1007/bf01053745.

[78] E. R. Deyle and G. Sugihara. Generalized theorems for nonlinear state space reconstruction. *PLoS ONE*, 6(3):e18295, 2011. doi:10.1371/journal.pone.0018295.

[79] N. Aubry, W.-Y. Lian, and E. S. Titi. Preserving symmetries in the proper orthogonal decomposition. *SIAM J. Sci. Comput.*, 14:483–505, 1993. doi:10.1137/0914030.

[80] D. T. Crommelin and A. J. Majda. Strategies for model reduction: Comparing different optimal bases. *J. Atmos. Sci.*, 61:2206–2217, 2004. doi:10.1175/1520-0469(2004)061<2206:sfmrcd>2.0.co;2.

[81] M. Desbrun, E. Kanso, and Y. Tong. Discrete differential forms for computational modeling. In *ACM SIGGRAPH 2005 Courses*, SIGGRAPH '05, Los Angeles, 2005. doi:10.1145/1198555.1198666.

[82] D. Zhou and C. Burges. High-order regularization on graphs. In *6th International Workshop on Mining and Learning with Graphs*, Helsinki, 2008.

[83] L. G. Grady and J. R. Polimeni. *Discrete Calculus*. Springer-Verlag, London, 2010.

[84] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15:1373–1396, 2003. doi:10.1162/089976603321780317.

[85] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21:5–30, 2006. doi:10.1016/j.acha.2006.04.006.

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

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

[88] T. Berry, R. Cressman, Z. Greguric Ferencek, and T. Sauer. Time-scale separation from diffusion-mapped delay coordinates. *SIAM J. Appl. Dyn. Sys.*, 12:618–649, 2013.

[89] S. Rosenberg. *The Laplacian on a Riemannian Manifold*, volume 31 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 1997.

[90] K. D. Elworthy, Y. Le Jan, and X.-M. Xi. *The Geometry of Filtering*. Frontiers in Mathematics. Birkhäuser, 2010.

[91] R R. Coifman et al. Geometric diffusions as a tool for harmonic analysis and structure definition on data. *Proc. Natl. Acad. Sci.*, 102(21):7426–7431, 2005. doi:10.1073/pnas.0500334102.

[92] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *J. Comput. Syst. Sci.*, 74(8):1289–1308, 2008. doi:10.1016/j.jcss.2007.08.006.

[93] T. Berry. *Model Free Techniques for Reduction of High-Dimensional Dynamics*. PhD thesis, George Mason University, 2013.

[94] D. Giannakis. Dynamics-adapted cone kernels. in preparation, 2013.

[95] A. Singer. From graph to manifold Laplacian: The convergence rate. *J. Appl. Comput. Harmon. Anal.*, 21:128–134, 2006. doi:10.1016/j.acha.2006.03.004.

[96] B. Christiansen. The shortcomings of nonlinear component analysis in identifying circulation regimes. *J. Climate*, 18:4814–4823, 2005. doi:10.1175/jcli3569.1.

[97] C. H. R. Lima, U. Lall, T. Jebara, and A. G. Barnston. Statistical prediction of ENSO from subsurface sea temperature using a nonlinear dimensionality reduction. *J. Climate*, 22:4501–4519, 2009. doi:10.1175/2009jcli2524.1.

[98] G. N. Kiladis, M. C. Wheeler, P. T. Haertel, K. H. Straub, and P. E. Roundy. Convectively coupled equatorial waves. *Rev. Geophys.*, 47(2):RG2003, 2009. doi:10.1029/2008rg000266.

[99] D. Waliser. Predictability and forecasting. In W. K. M. Lau and D. E. Waliser, editors, *Intraseasonal Variability in the Atmosphere-Ocean Climate System*, pages 389–423. Springer, Berlin, 2005.

[100] D. Waliser et al. MJO simulation diagnostics. *J. Climate*, 22:3006–3030, 2009. doi:10.1175/2008jcli2731.1.

[101] A. J. Majda and S. N. Stechmann. The skeleton of tropical intraseasonal oscillations. *Proc. Natl. Acad. Sci.*, 106:8417–8422, 2009. doi:10.1073/pnas.0903367106.

[102] A. J. Majda and S. N. Stechmann. Nonlinear dynamics and regional variations in the MJO skeleton. *J. Atmos. Sci.*, 68:3053–3071, 2011. doi:10.1175/jas-d-11-053.1.

[103] W. W. Tung, J. Gao, J. Hu, and L. Yang. Detecting chaos in heavy noise environments. *Phys. Rev. E*, 83:046210, 2011. doi:10.1103/physreve.83.046210.

[104] P. J. Webster and R. Lukas. TOGA COARE: The Coupled Ocean-Atmosphere Response Experiment. *Bull. Amer. Meteor. Soc.*, 73(9):1377–1416, 1992. doi:10.1175/1520-0477(1992)073<1377:tctcor>2.0.co;2.

[105] K. Kikuchi, B. Wang, and Y. Kajikawa. Bimodal representation of the tropical intraseasonal oscillation. *Climate Dyn.*, 38:1989–2000, 2012. doi:10.1007/s00382-011-1159-1.

[106] E. D. Maloney and D. L. Hartmann. Frictional moisture convergence in a composite life cycle of the Madden-Julian oscillation. *J. Climate*, 11:2387–2403, 1998. doi:10.1175/1520-0442(1998)011<2387:fmciac>2.0.co;2.

[107] M. C. Wheeler and H. H. Hendon. An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Weather Rev.*, 132(8):1917–1932, 2004. doi:10.1175/1520-0493(2004)132<1917:aarmmi>2.0.co;2.

[108] M. Yanai, B. Chen, and W.-w. Tung. The Madden-Julian oscillation observed during the TOGA COARE IOP: Global view. *J. Atmos. Sci.*, 57(15):2374–2396, 2000. doi:10.1175/1520-0469(2000)057<2374:tmjood>2.0.co;2.

[109] R. Talmon and R. R. Coifman. Empirical intrinsic geometry for nonlinear modeling and time series filtering. *Proc. Natl. Acad. Sci.*, 110(31):12535–12540, 2013. doi:10.1073/pnas.1307298110.