Clim Dyn (2016) 46:1473–1502
DOI 10.1007/s00382-015-2658-2

Extraction and predictability of coherent intraseasonal signals in infrared brightness temperature data

Eniko Székely[1] · Dimitrios Giannakis[1] · Andrew J. Majda[1]

Received: 25 October 2014 / Accepted: 14 May 2015 / Published online: 29 May 2015

Abstract This work studies the spatiotemporal structure and regime predictability of large-scale intraseasonal oscillations (ISOs) of tropical convection in satellite observations of infrared brightness temperature (T_b). Using nonlinear Laplacian spectral analysis (NLSA), a data analysis technique designed to extract intrinsic timescales of dynamical systems, the T_b field over the tropical belt 15°S − 15°N and the years 1983–2006 (sampled every 3 h at 0.5° resolution) is decomposed into spatiotemporal modes spanning interannual to diurnal timescales. A key advantage of NLSA is that it requires no preprocessing such as bandpass filtering or seasonal partitioning of the input data, enabling simultaneous recovery of the dominant ISOs and other patterns influenced by or influencing ISOs. In particular, the eastward-propagating Madden–Julian oscillation (MJO) and the poleward-propagating boreal summer intraseasonal oscillation (BSISO) naturally emerge as distinct families of modes exhibiting non-Gaussian statistics and strong intermittency. A bimodal ISO index constructed via NLSA is found to have significantly higher discriminating power than what is possible via linear methods. Besides MJO and BSISO, the NLSA spectrum contains a multiscale hierarchy of modes, including the annual cycle and its harmonics, ENSO, and modulated diurnal modes. These modes are used as predictors to quantify regime predictability of the MJO amplitude in T_b data through a cluster-based framework. It is found that the most predictable MJO regimes occur before the active-MJO season (November–December), when ENSO has a strong influence on the future statistical behavior of MJO activity. In forecasts initialized during the active-MJO period (February), both ENSO and the current state of MJO are significant predictors, but the predictive information provided by the large-scale convective regimes in T_b is found to be smaller than in the early-season forecasts.

Keywords Tropical intraseasonal oscillations · MJO · Dimension reduction · Regime predictability

1 Introduction

Organized tropical convection is a key element of global climate dynamics with direct impact on both short-term weather forecasting and long-term climate projections. Among the dominant modes of tropical variability, intraseasonal oscillations (ISOs) play a key role in explaining large-scale convective organization at subseasonal timescales while also influencing the global climate through extratropical interactions (Lau and Waliser 2011). The behavior of tropical ISOs is strongly influenced by the annual cycle (Wang and Rui 1990; Zhang and Dong 2004), resulting in significant differences between the coherent propagating patterns that emerge during boreal winter and boreal summer. The dominant boreal winter ISO is the well-known Madden–Julian oscillation (MJO; Madden and Julian 1971, 1972), a 30–90-day eastward-propagating pattern with zonal wavenumber 1–4. The dominant boreal summer ISO (BSISO) has a more emphasized poleward-propagating pattern with a weakened eastward propagation (Wang and Rui 1990; Kikuchi et al. 2012). Besides being

Electronic supplementary material The online version of this article (doi:10.1007/s00382-015-2658-2) contains supplementary material, which is available to authorized users.

✉ Eniko Székely
eszekely@cims.nyu.edu

[1] Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY, USA

influenced by the annual cycle, ISOs interact with several modes of variability of the coupled atmosphere-ocean system. These modes include interannual modes, in particular the El Niño Southern Oscillation (ENSO) (Lau and Chan 1985; Kessler 2001; Hendon et al. 2007; Lau 2011), as well as the diurnal cycle (Chen and Houze 1997; Tian et al. 2006; Ichikawa and Yasunari 2008). However, despite that tropical ISOs are a major source of predictability on intraseasonal timescales (Waliser 2011), accurate simulation of the MJO and other ISOs by large-scale dynamical models remains elusive (Hung et al. 2013).

A significant challenge in understanding the behavior of ISOs and their connections to other modes of variability is that the phenomena themselves are defined subjectively through some data analysis technique (Straub 2013; Kiladis et al. 2014). In the case of MJO and BSISO, the extensive range of techniques in the literature include spacetime filtering (Wheeler and Kiladis 1999; Kiladis et al. 2005; Kikuchi and Wang 2010), empirical orthogonal functions (EOFs) (Lo and Hendon 2000; Maloney and Hartmann 1998; Kessler 2001; Wheeler and Hendon 2004; Kikuchi et al. 2012; Ventrice et al. 2013; Kiladis et al. 2014), as well as hybrid filtering–EOF approaches (Roundy and Schreck 2009). These techniques have been employed to extract ISO signals and construct indices from various data sources, with outgoing longwave radiation (OLR) and brightness temperature (T_b) typically employed as proxies for convective activity, and zonal winds, streamfunctions, and velocity potential data commonly used to represent circulation. Radiation and circulation data have also been combined to create multivariate indices taking into account both convection and circulation aspects of ISOs (Wheeler and Hendon 2004; Ventrice et al. 2013). While the coarse-grained properties of dominant ISOs such as the MJO and BSISO have been fairly consistent among these methods, significant differences exist in the details, including the identification of significant events (Straub 2013). Such differences impede the scientific understanding of ISOs as well as advances in their simulation and forecasting via numerical models.

Arguably, the discrepancies between ISO analyses in the literature are at least partly caused by the various types of ad hoc preprocessing steps to which the data is subjected prior to the extraction of ISO signals. Spacetime filtering methods require the selection of windows in the wavenumber-frequency domain containing the signal of interest, and this often requires an estimation of a background spectrum (Wheeler and Kiladis 1999; Kiladis et al. 2005). In EOF-based methods, various preprocessing techniques such as bandpass filtering (Kessler 2001; Kikuchi et al. 2012; Maloney and Hartmann 1998; Lo and Hendon 2000), seasonal partitioning (Kikuchi et al. 2012), and running averaging (Kiladis et al. 2014) are commonly applied prior to analysis to isolate the intraseasonal component of the data from other signals such as ENSO and the seasonal and diurnal cycles. EOFs and the related extended EOFs (EEOFs) are also prone to lack of physical interpretability due to imposition of the orthogonality constraint (Horel 1981; Kessler 2001; Groth and Ghil 2011). Another preprocessing approach common to both spacetime filtering and EOF approaches is to reduce the initial two-dimensional (2D) spatial data to one-dimensional (1D) representations through either symmetric or antisymmetric latitudinal averaging in space (Kikuchi and Wang 2010; Kikuchi et al. 2012; Tung et al. 2014a) or the Fourrier domain (Wheeler and Kiladis 1999; Kiladis et al. 2005). Even though averaging is justified on theoretical grounds (Matsuno 1966), it will invariably lead to loss of information compared to the full 2D data.

Recently, in an effort to extract the MJO and other signals of interest for organized tropical convection with minimal preprocessing of the data, Giannakis et al. (2012a) and Tung et al. (2014a, b) (hereafter, collectively TGM) have carried out an analysis of T_b data from the CLAUS multisatellite archive (Hodges et al. 2000) through a data analysis technique called nonlinear Laplacian spectral analysis (NLSA; Giannakis and Majda 2012b, 2013, 2014). Blending ideas from machine learning (Belkin and Niyogi 2003; Coifman and Lafon 2006) and delay-coordinate maps of dynamical systems (Takens 1981; Broomhead and King 1986; Sauer et al. 1991), NLSA seeks to extract spatiotemporal patterns from high-dimensional time series which are intrinsic to the dynamical system generating the data (in the present application, the coupled atmosphere-ocean climate system). A key ingredient of this technique is to replace the covariance operator used in singular spectrum analysis (SSA) (Ghil et al. 2002) and the equivalent EEOF analysis by a discrete Laplace–Beltrami operator constructed from the cloud of data lagged over a Takens embedding window. The eigenfunctions of this operator form a natural orthonormal basis for functions on the nonlinear data manifold sampled by the data, providing superior timescale separation (Berry et al. 2013) and the ability to capture temporally intermittent and modulated patterns (Giannakis and Majda 2012b; Bushuk et al. 2014). Such patterns may carry low variance and may fail to be captured by variance-greedy algorithms such as SSA, yet may play an important dynamical role (Aubry et al. 1993; Crommelin and Majda 2004). In the standard version of NLSA, the Laplacian eigenfunctions are combined with singular value decomposition (SVD) techniques (Aubry et al. 1991) to construct biorthonormal spatial and temporal patterns analogous to EEOFs and principal components (PCs), respectively.

Using NLSA, TGM extracted a hierarchy of spatiotemporal modes of variability from symmetrically and antisymmetrically averaged, but otherwise unprocessed, T_b data spanning interannual to diurnal timescales. In

particular, the NLSA spectra for the symmetric and antisymmetric data were both found to contain in-quadrature pairs of MJO modes with intermittent envelopes differing significantly from the corresponding SSA patterns. Besides MJO, the NLSA modes included periodic modes representing the annual and semiannual harmonics of the yearly climatology, an interannual ENSO mode, as well as interannual–intraseasonal modes active in the Indo-Pacific Ocean describing ENSO-modulated intraseasonal patterns. In addition, a variety of diurnal modes were extracted featuring clear modulation relationships with MJO and ENSO. Tung et al. (2014b) used indices derived from the symmetric and antisymmetric MJO modes from NLSA to construct phase composites of various kinematic and thermodynamic fields revealing significant differences in the energetics and propagation characteristics of predominantly symmetric and antisymmetric MJO events. Yet, despite these attractive features, the analysis of TGM was restricted to 1D latitudinal averages, obscuring certain aspects of convective variability such as the influence of the Maritime Continent on MJO propagation. Moreover, the temporal and spatiotemporal patterns studied by TGM were obtained after postprocessing the Laplacian eigenfunctions through an SVD-based rotation and thus subject to an orthogonality constraint in physical space.

In this paper, we extend the analysis of TGM to study via NLSA full 2D brightness temperature data over the equatorial belt $15°\text{S} - 15°\text{N}$. Applying no preprocessing to the data, we construct a hierarchy of spatiotemporal modes yielding additional insights on organized convective variability which were not available via the 1D analysis. In particular, a clear separation of the dominant ISOs emerges through Laplacian eigenfunctions representing the eastward-propagating boreal winter MJO and the poleward-propagating boreal summer ISO. These eigenfunctions are characterized by intermittent envelopes and project onto non-orthogonal patterns in the spatial domain. As a result, the MJO and BSISO signals become mixed if these eigenfunctions are rotated through SVD to produce orthogonal patterns in space. Here, we use the MJO and BSISO eigenfunctions from NLSA to construct a bimodal ISO index analogous to the index of Kikuchi et al. (2012), without having to perform an ad hoc seasonal partitioning of the data. We find that the NLSA-based index has significantly higher discriminating power than the corresponding bimodal index constructed via EEOF- and SSA-type algorithms. Elsewhere (Chen et al. 2014), it is shown that the NLSA MJO modes can be accurately described via stochastic nonlinear oscillator models where intermittency is an outcome of time-dependent damping and phase. Besides the MJO and BSISO, the Laplace–Beltrami eigenfunctions from NLSA describe a multiscale hierarchy of patterns of interest, including the seasonal cycle and its harmonics, interannual modes, and ISO-modulated diurnal modes.

A further objective of this work is to quantify the regime predictability of the MJO amplitude as represented by the associated NLSA-based index. Below, we address this question using the information-theoretic framework of Giannakis and Majda (2012c, d) (hereafter, GM) and Giannakis et al. (2012b), adapted to variables with cyclostationary statistics. This framework derives lower bounds for predictability through coarse-grained partitions of a space of predictor variables constructed via clustering algorithms. In this work, the space of predictors will be a 17-dimensional space spanned by the leading Laplacian eigenfunctions from NLSA representing interannual and intraseasonal variability, as well as the annual cycle and its harmonics. The partitions therefore correspond to large-scale convective regimes (e.g., El Niño events) extracted from the T_b data by the eigenfunctions.

Our results show that the interannual convective regimes embedded in T_b provide significant predictability of the MJO amplitude for forecasts initialized in November–December, i.e., before the active-MJO season. In particular, we find that the statistical predictability of the MJO amplitude as measured by the NLSA index is especially high during El Niño years, with a reemergence of predictability associated with MJO wavetrains taking place at intraseasonal-scale (60–80 days) leads. During the active-MJO season (January–February), the large-scale convective regimes identified via clustering contribute less to predictability, but nevertheless significant La Niña events, as well as the current MJO activity at initialization time, both emerge as significant predictors. Overall, our study objectively quantifies the role of planetary-scale convective regimes in the statistical predictability of the MJO amplitude.

The rest of the paper is organized as follows. Section 2 describes the two approaches used to first extract salient temporal and spatiotemporal features from the T_b data, and then assess the predictability of the MJO amplitude by coarse-graining predictor variables through data clustering. The infrared brightness temperature data used in this study is described in Sect. 3. The hierarchy of the spatiotemporal patterns extracted from the T_b data via NLSA and the associated ISO indices are presented and discussed in Sect. 4. Section 5 contains the predictability analysis for MJO conditioned on coarse-grained convective regimes. We conclude in Sect. 6 with a review of the main contributions of this work and perspectives for future developments. The interested reader can jump directly to Sects. 3, 4 and 5 for a description of the data and the results. The temporal patterns and spatial snapshots presented here are accompanied by their respective spatiotemporal reconstructions as videos (Movies 1 and 2). A brief comparison of the modes recovered by NLSA and the corresponding modes recovered by

SSA and EEOF analysis is also provided as electronic supplementary material.

2 Methods

2.1 NLSA algorithms

NLSA (Giannakis and Majda 2012a, 2013, 2014) is a dimension reduction technique used for the extraction of spatiotemporal patterns from high-dimensional data generated by dynamical systems. Such systems are highly nonlinear in nature, yet they can be described at a coarse scale by low-dimensional geometric structures embedded in the ambient data space. NLSA's aim is to recover this underlying low-dimensional structure, generally modeled as a manifold, and describe it in terms of a reduced set of basis functions. The core of the analysis consists of three steps: (1) time-lagged embedding using Takens method of delays; (2) construction of a discrete Laplace–Beltrami operator via kernel methods from machine learning and harmonic analysis applied to Takens embedding space; (3) construction of an empirical set of basis functions for feature extraction and dimension reduction through the eigenfunctions of the Laplace–Beltrami operator. The Laplace–Beltrami eigenfunctions are nonlinear analogs to PCs, and provide a low-dimensional representation of the high-dimensional input data. In addition, an SVD-based rotation may be performed if one is interested in identifying biorthonormal spatiotemporal patterns associated with these eigenfunctions that maximally explain the variance of the input data. Here, however we do not carry out this procedure as the spatiotemporal patterns associated with the pure eigenfunctions are already physically interpretable without the basis rotation taking place in the SVD step. Moreover, SVD enforces orthogonality of the recovered spatial patterns, and in Sect. 4 ahead we will find that the dominant ISO patterns are in fact non-orthogonal in space. Note that the clustering results in Sect. 5 are invariant under orthogonal transformations such as SVD.

2.1.1 Time-lagged embedding

Consider an n-dimensional time series

$$x(t_i) = (x^1(t_i), \ldots, x^n(t_i))$$

consisting of s samples taken at times $t_i = i\,\delta t$ with a uniform sampling interval δt. A standard approach in state-space reconstruction methods for dynamical systems, as well as in EEOF analysis and SSA, is to embed the n-dimensional samples into a higher-dimensional space of lagged sequences of observations (hereafter, called Takens embedding space, or delay-coordinate space). Specifically, given an integer parameter $q \geq 1$ (the number of lags), $x(t_i)$ is mapped to the sequence

$$X(t_i) = (x(t_i), x(t_i - \delta t), \ldots, x(t_i - (q-1)\,\delta t)) \quad (1)$$

of dimension $N = nq$. We use the notation $X = (X(t_q), \ldots, X(t_s))$ to represent the $N \times S$ data matrix in the N-dimensional delay-coordinate space, where $S = s - q + 1$ is the number of samples available for analysis after time-lagged embedding. The temporal extent of the embedding window is $\Delta t = q\,\delta t$.

For sufficiently large q and under weak assumptions on the underlying dynamical system and the observation function, this operation recovers the topology of the attractor of the dynamical system lost by partial observations (Takens 1981; Broomhead and King 1986; Sauer et al. 1991). In other words, the time-lagged embedded data observations $X(t_i)$ sample a manifold $\mathscr{M}$ which is in a one-to-one correspondence with the attractor of the underlying dynamical system, even if some of the dimensions of that attractor have been projected away in the snapshots $x(t_i)$.

Besides the topology of the data, however, time-lagged embedding affects its geometry (Giannakis and Majda 2012b; Berry et al. 2013; Giannakis and Majda 2014). That is, Euclidean distances $||X(t_i) - X(t_j)||$ between the time-lagged data depend not only on the states at times t_i and t_j, but also on the dynamical trajectory that the system followed to arrive at those states. This dynamical dependence carries over to the global covariance matrix $C = XX^T$ utilized in EEOF analysis and SSA to recover spatial and temporal modes of superior timescale separation than what is possible through classical PCA. However, global covariances are not intrinsic to the nonlinear manifold geometry of the data, as C is invariant only under rigid-body rotations of X. Rather, the intrinsic geometry of the data is characterized through a local notion of distance (a Riemannian metric) varying smoothly over $\mathscr{M}$, requiring only preservation of neighborhood distances for faithful low-dimensional representation (a significantly weaker requirement than global-distance preservation). In effect, time-lagged embedding leaves an "imprint" of the dynamics on the Riemannian geometry of the data, and NLSA uses operators compatible with this geometry which are constructed empirically from data to perform dynamics-adapted dimension reduction and feature extraction.

2.1.2 Discrete Laplace–Beltrami operator

Among the major recent advances in machine learning and harmonic analysis has been the development of theory and algorithms to construct geometrical operators for data analysis through local kernels, i.e., pairwise measures of similarity decaying exponentially in data space (e.g., Belkin and Niyogi 2003; Coifman and Lafon 2006; Hein et al. 2005;

Singer 2006; Berry and Sauer 2014). Different kernels will induce different geometries on the data, and NLSA is based on a kernel formulated in Takens embedding space, viz.

$$K(X(t_i), X(t_j)) = \exp\left(-\frac{||X(t_i) - X(t_j)||^2}{\varepsilon ||\zeta(t_i)|| ||\zeta(t_j)||}\right). \tag{2}$$

Here, ε is a positive parameter controlling the bandwidth of the kernel, and $\zeta(t_i) = X(t_i) - X(t_{i-1})$ measures the local phase space velocity (time tendency) of the data. The quantities $\zeta(t_i)$ can be interpreted as finite-difference approximations of the vector field in phase space driving the dynamics (Giannakis 2014).

Evaluating (2) pairwise for all the data sample leads to an $S \times S$ symmetric kernel matrix K with elements $K_{ij} = K(X(t_i), X(t_j))$. This matrix operates on temporal patterns in a similar manner as the temporal covariance matrix $X^T X$. Specifically, because the data are ordered in time, a discretely sampled function of time $v(t_i)$ can be represented by an S-dimensional column vector $v = (v_1, \ldots, v_S)^T$ with $v(t_i) = v_i$, and the kernel acts on v via standard matrix multiplication, Kv. At the same time, $v(t_i)$ can be thought of as sampling a function f on the manifold such that $v(t_i) = f(X(t_i))$. As the bandwidth parameter ε in (2) and the sampling interval δt become small (i.e., in the limit of large data), K becomes sensitive to the structure of f only at local neighborhoods on the manifold, approximating the action of a differential operator defined on $\mathscr{M}$.

Through a sequence of normalizations of the kernel matrix [see (3) ahead], the limit differential operator can be arranged to be the Laplace–Beltrami operator associated with a Riemannian metric on $\mathscr{M}$ that depends on K. The spectral properties of this operator (i.e., its eigenvalues and eigenvectors) depend strongly on the Riemannian metric, and are known to be useful for nonlinear dimension reduction and pattern extraction as discussed below. We thus think of K as inducing a Riemannian geometry to the data, and the results of data analysis through operators based on K can be studied and interpreted via the properties of that geometry. For kernels formulated in Takens embedding space, including (2), the induced Riemannian geometry enhances timescale separation capability by favoring stable Lyapunov directions of the dynamical system (Berry et al. 2013), such as quasi-periodic orbits associated with coherent oscillations. In NLSA, the $||\zeta(t_i)||$ scaling factors additionally improve the skill of the algorithm to capture temporally intermittent patterns (Giannakis and Majda 2012b).

Leaving out further theoretical details to other references (Giannakis and Majda 2012b, 2014; Giannakis 2014), the discrete Laplace–Beltrami operator is represented by an $S \times S$ matrix, L, constructed by performing the following sequence of normalizations proposed in the diffusion maps algorithm of Coifman and Lafon (2006):

$$\begin{aligned} \tilde{K}_{ij} &= \frac{K_{ij}}{Q_i Q_j}, \quad Q_i = \sum_{j=1}^{S} K_{ij}, \\ P_{ij} &= \frac{\tilde{K}_{ij}}{D_i}, \quad D_i = \sum_{j=1}^{S} \tilde{K}_{ij}, \\ L_{ij} &= I - P_{ij}. \end{aligned} \tag{3}$$

Note that by virtue of the exponential decay of the kernel, L can be made sparse by retaining only the largest $k_{\mathrm{nn}} \ll S$ entries per row, significantly reducing the computational cost to obtain eigenvectors compared to the temporal covariance matrix.

In the limit of large data, the discrete operator represented by L converges to the continuous Laplace–Beltrami operator on the manifold, even if the sampling density on $\mathscr{M}$ is non uniform. The latter is a particularly attractive feature for our purposes, since the data are sampled at a fixed time interval and we have no control of the sampling density. Note that no a priori knowledge of $\mathscr{M}$ and its geometry is required to carry out the procedure outlined above. Moreover, even though the convergence results formally apply in the limit of large data, in practice we operate far from this regime requiring dense sampling of the turbulent attractor for the atmosphere. Instead, an assumption which is far more likely to hold in practice is that the full attractor exhibits certain low-dimensional coarse geometric structures associated with phenomena such as ENSO and the MJO, and the sampling of the attractor during the 23-year period of the CLAUS archive is sufficiently dense to recover these structures.

2.1.3 *Empirical basis functions and low-dimensional representations*

In the analysis in Sect. 4, L will be employed to perform feature extraction and dimension reduction through its eigendecomposition

$$L\phi_i = \lambda_i \phi_i, \quad \text{with} \quad i \in \{0, 1, 2, \ldots\}. \tag{4}$$

As stated in Sect. 2.1.2, the eigenvectors $\phi_i = (\phi_{1i}, \ldots, \phi_{Si})^T$ can be interchangeably interpreted as discretely sampled functions on $\mathscr{M}$, or as time series $\phi_i(t_j) = \phi_{ji}$. These time series can be thought of as nonlinear analogs to PC time series arising in EOF analysis. It is a standard result (Belkin and Niyogi 2003; Coifman and Lafon 2006) that the Laplacian eigenfunctions form an orthonormal basis with respect to the weighted inner product

$$(\phi_i, \phi_j) = \sum_{k=1}^{S} D_k \phi_{ki} \phi_{kj} = \delta_{ij},$$

where the weights D_k are given by (3), and can be interpreted as the volume (Riemannian measure) occupied by the samples $X(t_k)$ on $\mathscr{M}$.

The Laplacian eigenfunctions ϕ_i form a natural basis to describe quantities of interest on the manifold (including the data vectors themselves), which is moreover adapted to the dynamics by virtue of the kernel in (2). For instance, in Sect. 4.1 time-lagged embedding will be essential to the ability of the eigenfunctions to separate the interannual, annual, intraseasonal, and diurnal timescales present in the T_b data. We note that the eigenvalues λ_i do not measure explained variance, but have a geometrical interpretation as an average gradient of ϕ_i on the manifold, i.e., $\lambda_i^{1/2}$ corresponds to a "wavenumber" on the manifold for ϕ_i (Giannakis and Majda 2014). Consequently, using the leading few ϕ_i is equivalent to selecting those features which vary slowly on the nonlinear manifold sampled by the data, thus reducing noise and parameter sensitivity, while avoiding overfitting.

Here, of particular interest is the use of ϕ_i to recover spatiotemporal patterns in the ambient data space using the eigenfunctions as convolution filters. First, the data in the delay-coordinate space is recovered through the operation

$$\tilde{X}_i = XD\phi_i\phi_i^T, \tag{5}$$

and then the columns of $\tilde{X}_i$ are decomposed into q blocks of dimension n similarly to SSA techniques (Ghil et al. 2002). The average value over the blocks in the time-lagged embedded reconstruction $\tilde{X}_i$ provides the reconstructed spatiotemporal patterns $\tilde{x}_i$ in the original data space, giving a decomposition of the form $x(t_j) \approx \sum_i \tilde{x}_i(t_j)$. While this decomposition does not maximize explained variance, it has been shown in various contexts to have high skill in recovering dynamically significant patterns, including patterns with intermittency (Giannakis and Majda 2012b; Bushuk et al. 2014; Tung et al. 2014a).

Another important property of the Laplacian eigenfunctions, which will become relevant in the predictability analysis in Sect. 5, is that that they can be used as nonlinear dimension reduction coordinates preserving the local neighborhood structure of the data. In particular, a dimension reduction map Φ taking the data in the N-dimensional delay-coordinate space to an l-dimensional Euclidean space with $l \ll N$ can be constructed by choosing an l-element set of eigenfunction indices, $\{j_1, \ldots, j_l\}$ (which are typically, but not always, consecutive), and setting

$$\Phi(X(t_i)) = (\phi_{ij_1}, \ldots, \phi_{ij_l}), \quad \text{with} \quad \Phi(X(t_i)) \in \mathbb{R}^l. \tag{6}$$

Hereafter, we will use the shorthand notation $\Phi_{t_i} = \Phi(X(t_i))$. It can be shown that for a sufficiently large l, dimension reduction maps of this form preserve the manifold structure of the data (Belkin and Niyogi 2003; Coifman and Lafon 2006; Jones et al. 2008; Portegies 2014), smoothly mapping nearby points in the high-dimensional data space to nearby points in the low-dimensional representation. Note that the required number of eigenfunctions depends only on intrinsic properties of the manifold (such as its intrinsic dimension and curvature), and can be significantly smaller than the number of PCs required in linear-projection approaches.

2.2 Cluster-based measures of regime predictability

We adopt an information-theoretic framework whereby predictability of a quantity of interest $r_{t+\tau}$ at lead time τ given data Φ_t observed at initialization time t is measured by the additional information in the forecast distribution $p(r_{t+\tau} \mid \Phi_t)$ beyond the climatology $p(r_{t+\tau})$ (Kleeman 2002; Majda et al. 2002). In the predictability analysis in Sect. 5, the predictand $r_{t+\tau}$ (also referred to as the response variable) will be the NLSA-based MJO index from (16a), and the predictors Φ_t will be the eigenfunction-based reduced coordinates from (6). An important property of both Φ_t and $r_{t+\tau}$ is cyclostationarity due to the seasonal cycle. In particular, we consider that all distributions are periodic in the initialization time t with period 1 y; e.g.,

$$p(r_{t+\tau} \mid \Phi_t) = p(r_{t'+\tau} \mid \Phi_{t'}) \tag{7}$$

whenever t and t' differ by an integer multiple of 1 y.

The natural information-theoretic functional to measure the gain of information in the forecast distribution relative to climatology is relative entropy (e.g., Cover and Thomas 2006), given in this case by

$$\mathscr{D}(r_{t+\tau} \mid \Phi_t) = \int dr_{t+\tau}\, p(r_{t+\tau} \mid \Phi_t) \log \frac{p(r_{t+\tau} \mid \Phi_t)}{p(r_{t+\tau})}. \tag{8}$$

Note that in (8) and hereafter we abuse notation using $p(\cdot)$ to represent both a probability measure and its corresponding probability density function (PDF). The quantity $\mathscr{D}(r_{t+\tau} \mid \Phi_t)$ is non-negative, and can be thought of as a (non-symmetric) "distance" between the conditional distribution $p(r_{t+\tau} \mid \Phi_t)$ and the prior $p(r_{t+\tau})$. In particular, relative entropy has the desirable properties that it is invariant under nonlinear invertible transformations of $r_{t+\tau}$, and vanishes if and only if $p(r_{t+\tau} \mid \Phi_t) = p(r_{t+\tau})$. The latter is expected to occur at long lead times when the system has lost its predictability. Moreover, evaluation of (8) does not require knowledge of the dynamical system generating Φ_t and r_t (in the present context, the Earth's climate system). Rather, $\mathscr{D}(r_{t+\tau} \mid \Phi_t)$ only depends on the statistics of r_t and Φ_t through the corresponding time-shifted PDFs. In what follows, we will take advantage of this property to estimate predictability via an empirically computable relative entropy [defined in (12) ahead] which is a lower bound of $\mathscr{D}(r_{t+\tau} \mid \Phi_t)$. Because these estimates are computed from observations of nature without invoking an explicit model

for the time-evolution of the data, we refer to the relative entropies in (8) and (12) as "measures of predictability." In particular, we do not use the term "measures of prediction skill" as this term is more appropriate for methods based on explicit parametric models with model error, such as regression models (Lo and Hendon 2000; Waliser 2011).

Equation (8) measures predictability of $r_{t+\tau}$ for a single realization of the initial data. The expected predictability over all initial data is given by taking the expectation value of (8) with respect to the distribution $p(\Phi_t)$ of the initial data, i.e.,

$$\mathscr{I}(r_{t+\tau}, \Phi_t) = \int d\Phi_t \, p(\Phi_t) \mathscr{D}(r_{t+\tau} \mid \Phi_t). \tag{9}$$

In information theory, $\mathscr{I}(r_{t+\tau}, \Phi_t)$ is called mutual information between the random variables $r_{t+\tau}$ and Φ_t. As one can explicitly verify, mutual information is symmetric in its arguments, and it is equal to the relative entropy between the joint distribution $p(r_{\tau+t}, \Phi_t)$ and the product $p(r_{\tau+t})p(\Phi_t)$ of the marginals. Thus, $\mathscr{I}(r_{t+\tau}, \Phi_t)$ inherits the properties of relative entropy outlined above, including the fact that it vanishes if and only if $p(r_{\tau+t}, \Phi_t) = p(r_{\tau+t})p(\Phi_t)$; this occurs at late times when the predictor and response variables are statistically independent. Because of these and other desirable properties, mutual information has been proposed as a more fundamental notion of predictability than variance-based measures (Leung and North 1990; DelSole 2004).

Despite its attractive properties, mutual information is often challenging to estimate in practice. This is because, in a typical forecasting scenario, the initial data are multivariate even if the predictand is scalar, rendering the estimation of the PDFs and expectation value with respect to Φ_t prone to sampling errors. One way of addressing this issue is to assume that the joint and prior distributions are all Gaussian, and evaluate the integrals in (9) invoking the analytical expression for the relative entropy of Gaussian random variables (e.g., Kleeman 2002). However, while Gaussianity holds in linear dynamical models with Gaussian initial conditions, both the predictor and predictand variables of interest here are highly non-Gaussian (see, e.g., Figs. 8, 11).

GM have developed a method to address these issues which involves replacing the multivariate predictors Φ_t with an integer-valued variable k_t indicating the affiliation of Φ_t to a discrete partition of the space of initial data. This partition can be represented by a family $\Xi = \{\xi_1, \ldots, \xi_K\}$ of K mutually disjoint subsets such that every Φ_t lies in one and only one element ξ_{k_t} of Ξ. Setting aside for now the issue of how to construct Ξ, we think of the affiliation function

$$k_t = S(\Phi_t) \tag{10}$$

associated with any partition as a projection map from the multivariate initial data Φ_t to a coarse-grained representation k_t. In this framework, the expected predictability of $r_{t+\tau}$ is measured through its mutual information with respect to k_t, i.e.,

$$\mathscr{I}(r_{t+\tau}, k_t) = \sum_{k_t=1}^{K} p(k_t) \mathscr{D}(r_{t+\tau} \mid k_t), \tag{11}$$

where $p(k_t)$ is the occupation probability of subset ξ_{k_t}, and

$$\mathscr{D}(r_{t+\tau} \mid k_t) = \int dr_{t+\tau} \, p(r_{t+\tau} \mid k_t) \log \frac{p(r_{t+\tau} \mid k_t)}{p(r_{t+\tau})} \tag{12}$$

is the predictability of $r_{t+\tau}$ with respect to the coarse-grained forecast distribution $p(r_{t+\tau} \mid k_t)$. Note that while the relationships

$$p(k_t) = \int_{\Phi_t \in \xi_{k_t}} d\Phi_t \, p(\Phi_t)$$

$$p(r_{t+\tau} \mid k_t) = \int_{\Phi_t \in \xi_{k_t}} d\Phi_t \, p(r_{t+\tau} \mid \Phi_t) p(\Phi_t \mid k_t),$$

hold, in practice one does not have to evaluate neither $p(r_{t+\tau} \mid \Phi_t)$ nor $p(\Phi_t)$. Instead, a key feature of the scheme is that the coarse-grained PDFs $p(r_{t+\tau} \mid k_t)$ and $p(k_t)$ [and hence (11) and (12)] can be stably estimated using significantly fewer samples than (8) and (9), respectively.

Another fundamental property of $I(r_{t+\tau}, k_t)$, which follows from the so-called data-processing inequality in information theory (Cover and Thomas 2006), is the bound

$$\mathscr{I}(r_{t+\tau}, k_t) \leq \mathscr{I}(r_{t+\tau}, \Phi_t). \tag{13}$$

This relationship shows that $\mathscr{I}(r_{t+\tau}, k_t)$ provides a lower bound to the fined-grained predictability score $\mathscr{I}(r_{t+\tau}, \Phi_t)$, and this bound is practically computable from multivariate non-Gaussian predictors. Of course, the extent to which $\mathscr{I}(r_{t+\tau}, k_t)$ approaches $\mathscr{I}(r_{t+\tau}, \Phi_t)$ depends significantly on the partition Ξ and the lead time τ, but detection of significant predictability at the lead time of interest with respect to any partition is sufficient to deduce that the full system is predictable at that lead time. In particular, if several partitions are available, one can evaluate (11) for every partition and choose the maximum mutual information at each τ. Another advantage of this framework is that the organization of the initial data into the partition allows one to study how dynamical regimes, i.e., coarse-grained features of Φ_t, affect predictability of the quantity of interest.

2.2.1 *Constructing the partition for cyclostationary data*

While the predictability measures in (11) and (12) can be evaluated for any partition, GM have developed a method which builds Ξ empirically by applying the K-means clustering algorithm (MacQueen 1967) to a training dataset $\{\Phi_{t_1}, \ldots, \Phi_{t_{S'}}\}$, optionally using running averages to induce

temporal persistence in the affiliation function in (10). This approach was adequate to reveal long-range regime predictability in a simple ocean model with time-independent equilibrium statistics, and should generally perform well in systems where $p(\Phi_t)$ and $p(r_t)$ are both independent of t. In the present application, however, a single partition Ξ is likely to perform poorly in terms of the predictability bound in (13) due to the statistical cyclostationarity of the predictors and the predictand.

To address this issue, we modify the framework of GM replacing the global partition Ξ by a family of partitions $\Xi_{t'}$ labeled by a time stamp t' in the interval $[0, 1]$ y. In the new framework, the affiliation function $S(\Phi_t)$ in (10) is computed with the respect to the partition $\Xi_{t'}$ with t' equal to t modulo 1 y. Moreover, to account for non-convex geometrical structures in the predictor variables, we use kernel K-means (Schölkopf et al. 1998; Dhillon et al. 2004) instead of the traditional K-means algorithm.

Both K-means and kernel K-means minimize an objective cost function that is the sum of the distances from each data point to their respective centers. However, while K-means operates in the coordinate space of the points Φ_{t_i}, kernel K-means first transforms the data into a higher-dimensional inner-product space (called feature space) through an implicit transformation function ψ, and rewrites the cost function in the new space in terms of the transformation ψ. Let $\langle \cdot, \cdot \rangle$ be the inner product of feature space, and $||u||_\psi = \sqrt{\langle u, u \rangle}$ the corresponding norm (i.e., the distance in feature space). Formally, the partition $\Xi_t = \{\xi_1, \ldots, \xi_K\}$ in kernel K-means is given by the solution of the minimization problem

$$\min_{\xi_k} \sum_{k=1}^{K} \sum_{i:\Phi_{t_i} \in \xi_k} \|\psi(\Phi_{t_i}) - \mu_k\|_\psi^2, \tag{14}$$

where $\mu_k = \frac{1}{m_k} \sum_{j:\Phi_{t_j} \in \xi_k} \psi(\Phi_{t_j})$ are the cluster centers, and m_k the number of points in cluster k. In practice, the transformation ψ has no explicit expression, but because (14) depends only on pairwise inner products in feature space, it is sufficient to specify these inner products through a kernel function $\mathscr{K}$ such that $\mathscr{K}(\Phi_{t_i}, \Phi_{t_j}) = \langle \psi(\Phi_{t_i}), \psi(\Phi_{t_j}) \rangle$. Here, we use the Gaussian kernel,

$$\mathscr{K}(\Phi_{t_i}, \Phi_{t_j}) = e^{-||\Phi_{t_i} - \Phi_{t_j}||^2/(2\sigma^2)}, \tag{15}$$

where $||\cdot||$ is the canonical Euclidean norm in the space of predictors, and σ a positive bandwidth parameter. We use this kernel due to its ability to detect non-convex clusters, but any kernel leading to a symmetric positive semidefinite matrix (i.e., a Gramian matrix) is valid. In particular, the K-means algorithm is a special case of kernel K-means with the covariance kernel, $\mathscr{K}(\Phi_{t_i}, \Phi_{t_j}) = \Phi_{t_i}^T \Phi_{t_j}$. Note that (15) is not related to the NLSA kernel in (2) employed to extract spatiotemporal modes.

3 Dataset

The Cloud Archive User Service (CLAUS) satellite infrared brightness temperature (T_b) data (Hodges et al. 2000) recorded over 23 years from July 1, 1983 to June 30, 2006 is used for this study. In the tropics, positive (negative) T_b anomalies are associated with reduced (increased) cloudiness, thus providing a surrogate for tropical convection. The data is sampled over the tropical belt from 15°S to 15°N with a resolution of 0.5° (in both longitude and latitude) generating 2D samples with $n_{\text{long}} = 720$ longitude and $n_{\text{lat}} = 61$ latitude gridpoints. We use the full 2D gridpoint T_b values arranged prior to analysis into vectors $x(t)$ of dimension $n = n_{\text{long}} \times n_{\text{lat}} = 43{,}920$. Observations are collected at an interval of $\delta t = 3$ h, producing a dataset with $s = 67{,}208$ samples over the 23 years of the CLAUS record.

The data contains the intensive observing period (IOP) of the Tropical Ocean Global Atmosphere Coupled Ocean Atmosphere Response Experiment (TOGA COARE) which took place from November 1, 1992 to February 28, 1993. Two complete MJO events were observed in that period, and have subsequently been studied extensively in the literature (Lin and Johnson 1996a, b; Tung et al. 1999; Yanai et al. 2000). For this reason, in Sect. 4 ahead we will employ a two-year period from January 1, 1992 to December 31, 1993 encompassing the TOGA COARE IOP to discuss the spatiotemporal modes recovered via NLSA. Movie 1(a) shows the raw data for this two-year reference period, which was also used in the 1D analysis of TGM.

4 Spatiotemporal modes of infrared brightness temperature

Following TGM, the lag-embedded data $X(t)$ are constructed via (1) using an intraseasonal embedding window $\Delta t = 64$ days. The number of lags corresponding to the $\delta t = 3$ h sampling interval is $q = \Delta t/\delta t = 64 \times 8 = 512$, meaning that the embedded data vectors populate a space of dimension $N = nq \approx 2.3 \times 10^7$. Embedded in this high-dimensional data space are intrinsically low-dimensional nonlinear subsets corresponding to the salient modes of tropical variability, such as the annual cycle and its harmonics, and interannual, intraseasonal, and diurnal modes. The purpose of NLSA is to extract from the high-dimensional ambient data space a reduced representation of these modes through the Laplacian eigenfunctions ϕ_i in (4) and the corresponding spatiotemporal patterns $\tilde{X}_i$ obtained via the convolution filters in (5). Later in this section, we discuss how the eastward-propagating MJO and the poleward-propagating BSISO naturally emerge as in-quadrature pairs

of Laplacian eigenfunctions with temporal features such as intermittency differing significantly from conventional EOF-based indices. Subsequently, in Sect. 5, we will use the leading-few eigenfunctions as coordinates for a low-dimensional space of predictor variables to quantify MJO regime predictability. We verified the robustness of our results against the choice of embedding window by computing eigenfunctions for $\Delta t = 30$ and $\Delta t = 90$ days. The $\Delta t = 90$ days eigenfunctions are in good agreement with our $\Delta t = 64$ days nominal choice. The $\Delta t = 30$ days eigenfunctions exhibit increased mixing between the different timescales, but the qualitative features of the modes are generally preserved.

The input parameters required to build the discrete Laplace–Beltrami operator in (3) are: (1) the kernel bandwidth parameter ε in (2); (2) the number $k_{\rm nn}$ of nearest neighbors used to build the discrete Laplacian matrix in (3). Here, we work with $\varepsilon = 2$ and $k_{\rm nn} = 5000$ ($\sim$10 % of the dataset), which are also the parameter values used by TGM. In separate calculations, we verified that qualitatively similar results can be obtained for ε and $k_{\rm nn}$ in the intervals 1.5–5 and 3000–10,000, respectively. Note that because of the exponential decay of the kernel, the performance of the algorithm is not limited by the ambient space dimension N, but rather by the intrinsic dimension of the nonlinear data manifold. The latter is significantly smaller than N, and likely also smaller than the dimension of the linear subspaces constructed via EOF- and EEOF-based approaches. Note also that we have applied no preprocessing such as intraseasonal bandpass filtering and seasonal detrending prior to the analysis, thus reducing the risk of introducing subjective features in the recovered temporal and spatiotemporal patterns.

Throughout this study, we restrict attention to the leading 25 Laplace–Beltrami eigenfunctions obtained with the parameter values stated above, as we find that this set is sufficient to capture the salient features of large-scale convective organization on interannual to diurnal timescales. Representative eigenfunctions from this group are shown in Fig. 1 for a two-year portion of the time series covering the TOGA COARE IOP, along with their frequency spectra. The dynamic evolution of the corresponding spatiotemporal patterns is displayed in Movie 1. Figure 2 shows the spectrum of the corresponding eigenvalues. As remarked in Sect. 2.1, the eigenvalues can be interpreted as "wavenumbers" on the nonlinear data manifold $\mathscr{M}$, and do not measure explained variance (though in practice the leading eigenfunctions tend to coincide with high-variance patterns such as the seasonal cycle and ENSO). By restricting attention to the leading Laplacian eigenfunctions we are selecting the features of the data which have large scale in the intrinsic nonlinear geometry in lagged embedding space, and are therefore qualitatively robust with respect to sampling and parameter selection.

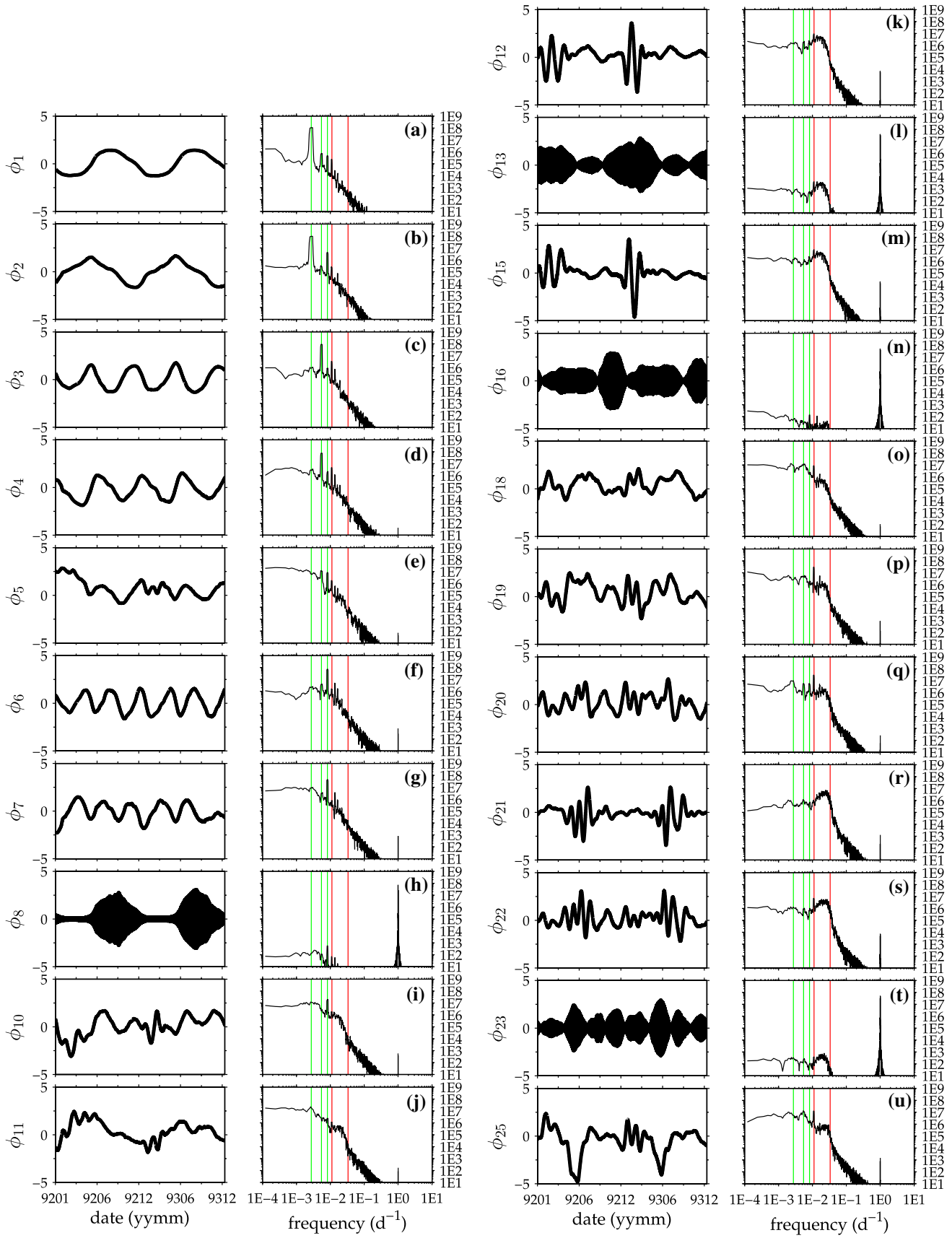
Note that the fact that the leading ϕ_i have large scale as functions on $\mathscr{M}$ does not imply that the corresponding temporal patterns all have low frequency. This is because the temporal features of the $\phi_i(X(t_j))$ time series depend on both the geometrical structure of ϕ_i on $\mathscr{M}$ as well as the sampling trajectory on $\mathscr{M}$ traced out by dynamical evolution; e.g., the $\phi_i(X(t_j))$ time series may exhibit rapid oscillations if the sampling trajectory frequently traverses level sets of ϕ_i. Indeed, the time series in Fig. 1 feature a broad range of timescales, including annual, interannual, intraseasonal and diurnal timescales. In what follows, we discuss the properties of these modes with reference to Fig. 1 and the dynamical evolution in Movie 1. Also, see Fig. 2 for a summary of the physical identification of the eigenfunctions depicted in Fig. 1 and Movie 1. For completeness, we have included a brief description of the modes recovered through SSA or EEOF analysis applied to the CLAUS T_b data using the same, $\Delta t = 64$ d, embedding window in the supplementary material.

4.1 Hierarchy of spatiotemporal modes

4.1.1 Annual and semiannual periodic modes

The leading four eigenfunctions describe two annual and two semiannual periodic patterns ($\{\phi_1, \phi_2\}$ and $\{\phi_3, \phi_4\}$, respectively) as indicated by the frequency spectra in Fig. 1a–d. Mode ϕ_1 is, to a good approximation, a pure sinusoidal wave with a frequency of 1 year^{-1}. Its spatiotemporal reconstruction in Movie 1(b) shows a characteristic winter to summer pattern in the two hemispheres with strongest amplitudes over land (Africa, the Maritime Continent, and South America). The strongest anomalies in this mode appear in winter and summer months (December–March and June–August), and are considerably weakened in spring and autumn (April–May and September–November). Mode ϕ_2 features strong variability over the Pacific and Atlantic Oceans associated with the annual movement of the Intertropical Convergence Zone (ITCZ) between the two hemispheres. This mode is in quadrature with mode ϕ_1, and is therefore active mainly in spring and autumn, but also bears a discernible semiannual peak in its spectrum.

Modes $\{\phi_3, \phi_4\}$ are characterized by dominant semiannual frequency peaks, and exhibit significant variability over both land and sea. Together, these modes describe semiannual variability in deep convection consistent with the twice-a-year equatorial crossing of the ITCZ (which is most prominent over Africa, the Maritime Continent, and the central Pacific) and cross-equatorial monsoon circulation. Mode ϕ_3 [see Movie 1(c)] features significant T_b anomalies over the Amazon, which is likely associated with the South American Monsoon (enhanced convection in boreal autumn). Mode ϕ_4 (and to a lesser extent ϕ_3) exhibits


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)
(q)
(r)
(s)
(t)
(u)
ϕ_1
ϕ_2
ϕ_3
ϕ_4
ϕ_5
ϕ_6
ϕ_7
ϕ_8
ϕ_{10}
ϕ_{11}
ϕ_{12}
ϕ_{13}
ϕ_{15}
ϕ_{16}
ϕ_{18}
ϕ_{19}
ϕ_{20}
ϕ_{21}
ϕ_{22}
ϕ_{23}
ϕ_{25}
5
0
−5
9201 9206 9212 9306 9312
date (yymm)
1E−4 1E−3 1E−2 1E−1 1E0 1E1
frequency (d^{-1})
1E9 1E8 1E7 1E6 1E5 1E4 1E3 1E2 1E1

◀ **Fig. 1** Leading Laplace–Beltrami eigenfunctions ϕ_i for the time interval January 1, 1992 to December 31, 1993, together with their associated power spectral densities (PSDs). The *vertical green lines* indicate the 1/year, 2/year and 3/year frequencies, while the *vertical red lines* indicate the 1/(90 days) and 1/(30 days) frequencies. The PSDs were estimated via the multitaper method (Thomson 1982; Ghil et al. 2002). The eigenfunctions discussed in the main text are as follows. **a**, **b** Annual modes; **c**, **d** semiannnual modes; **e** ENSO mode; **f, g** four-month quasi-periodic modes; **h** four-month modulated diurnal mode; **k, m** MJO modes; **l** MJO-modulated diurnal modes; **r, s** BSISO mode; **t** BSISO-modulated diurnal modes. *Note* that the diurnal eigenfunctions always appear as twofold-degenerate pairs, so for this reason the second eigenfunctions in each pair (i.e., ϕ_9, ϕ_{14}, ϕ_{17}, ϕ_{24}) are not shown here

a poleward-propagating center of convection in the Indian Ocean moving towards India. Among the leading four periodic modes, the amplitude of ϕ_1 is almost twice as high as the amplitudes of the other three modes.

4.1.2 Interannual modes

The dominant interannual mode in the NLSA spectrum, ϕ_5 (see Fig. 1e), describes the signature of ENSO events in deep convection. As shown in Movie 1(d), this mode is dominated by a strong east–west dipole with anomaly centers located at the dateline and at ~ 120°E. Physically, this dipole represents T_b patterns of anomalous Walker circulation with subsidence (decreased convection) over the Indian Ocean and ascent (increased convection) over the Western Pacific during El Niño events, and the opposite-sign pattern occurring during La Niña events (Chelliah and Arkin 1992; Chiodi and Harrison 2010). In the frequency domain, ϕ_5 exhibits a red-noise spectrum with significant power on interannual timescales, but also features discernible semiannual and four-month spectral lines.

In Fig. 3, the ENSO temporal pattern ϕ_5 is plotted together with the Multivariate ENSO Index (MEI; Wolter and Timlin 1998, 2011). MEI is computed by combining the PCs of six different observational sources (sea level pressure, zonal and meridional winds, sea surface temperature, surface air temperature and cloudiness) for each of the 12 sliding bi-monthly seasons. The strongest El Niño events with respect to eigenfunction ϕ_5 (corresponding to large positive eigenfunction values) took place during the winters of 1986–1988, the winter–spring of 1992, the winter of 1994–1995, year 1998, and the winter of 2002–2003. The strongest La Niña events (large negative ϕ_5) occurred

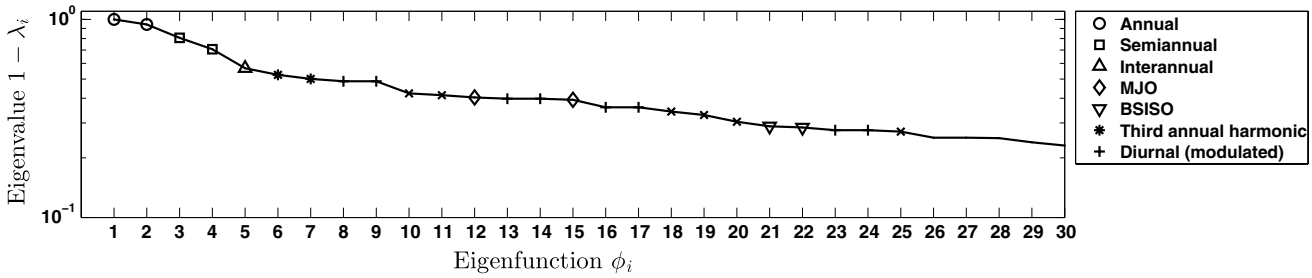


Fig. 2 Eigenvalues λ_i corresponding to the Laplace–Beltrami eigenfunctions ϕ_i. (*opencircle*) annual modes, $i \in \{1, 2\}$; (*squarebox*) semiannual modes, $i \in \{3, 4\}$; (*triangle*) interannual ENSO mode, $i = 5$; (*diamond*) boreal winter MJO modes, $i \in \{12, 15\}$; (*invertedtriangle*) boreal summer ISO modes, $i \in \{21, 22\}$; (*asterisk*) third harmonic of the annual cycle, i.e. four-month quasi-periodic modes, $i \in \{6, 7\}$; (*plus*) diurnal (modulated) modes, $i \in \{8, 9, 13, 14, 16, 17, 23, 24\}$. The remaining modes in this diagram with $i \leq 25$ (*times*) are not discussed extensively in the main text, but are included in the space of predictor variables to quantify regime predictability of the MJO amplitude

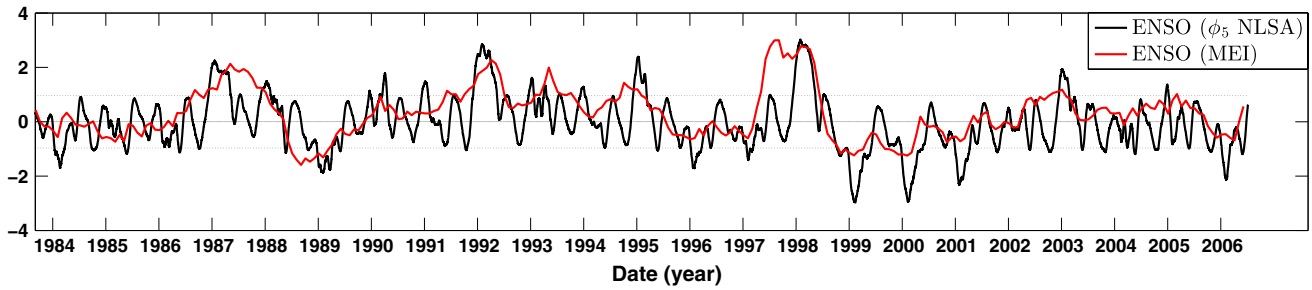


Fig. 3 The ENSO temporal pattern from NLSA (ϕ_5) and the Multivariate ENSO Index (MEI) standardized to zero mean and unit variance. The threshold for significant events (indicated by *horizontal lines*) is one standard deviation away from the mean. The two indices are consistent for the major El Niño and La Niña events

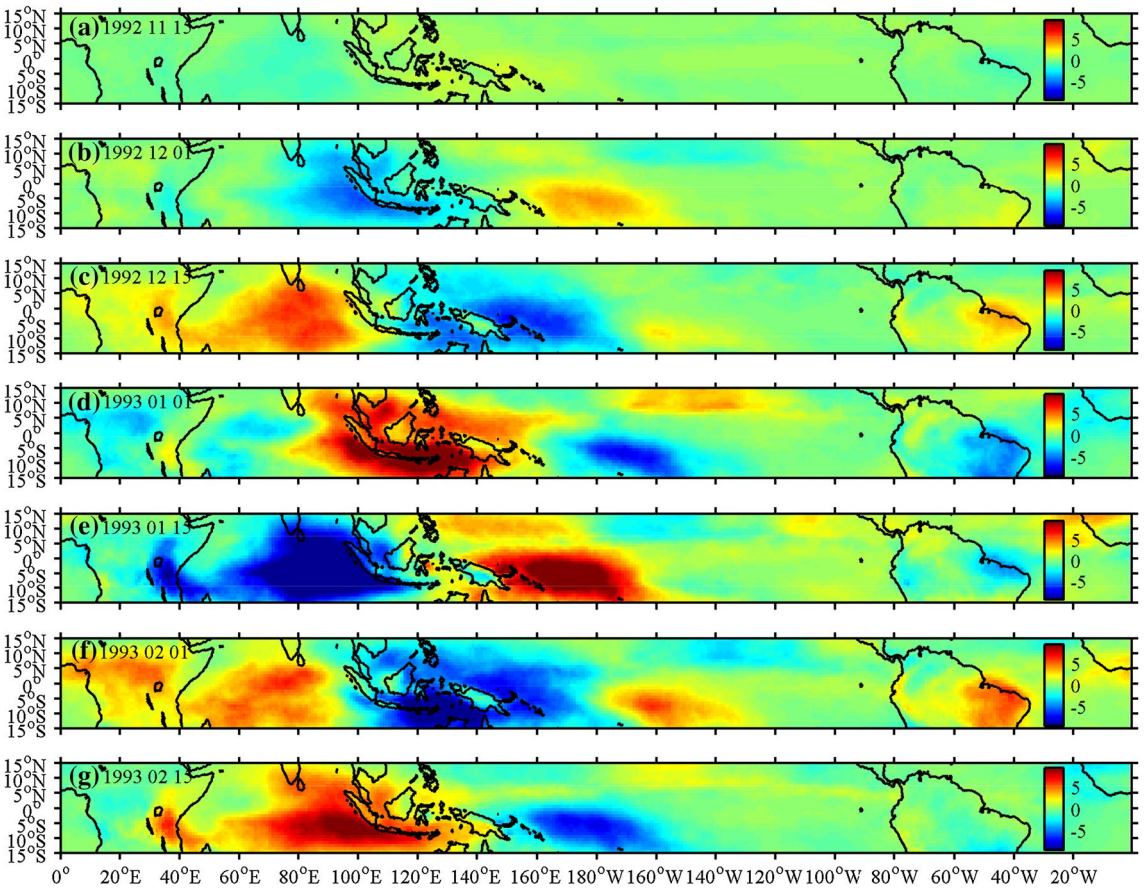


Fig. 4 Reconstruction of the MJO wavetrain observed during the TOGA COARE IOP of November 1992–March 1993. The *color maps* show T_b anomalies (in K) obtained from the NLSA MJO modes of Fig. 1k, m recovered in data space via (5). *Blue* (*red*) colors correspond to increased convection (decreased cloudiness). **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 dateline; **e, f, g** a second, stronger, MJO event with an initiation signal over East Africa. See Fig. 1 in Tung et al. (2014a) for the manifestation of these events in time-longitude sections of the raw data

during the winters of 1983–1984, 1988–1989, 1996, 1998–2001 and 2006. The results are consistent with the MEI index with a correlation of 0.57 over the 23 years of observations. Note that the frequency spectrum of our ENSO eigenfunction contains weak spectral lines at the second and third harmonics of the seasonal cycle. This periodic variability can be removed by performing a basis rotation (e.g., through SVD as described in TGM), which increases the correlation with the MEI index to 0.66.

The reconstruction in Movie 1(d) includes the strong El Niño event that took place in January–May 1992. The reconstruction also exhibits a weak reemergence of convection in the western Pacific in June 1992, followed by a weak phase of suppressed convection (La Niña) in September 1992. Both of the ends of 1992 and 1993 show a reemergence of El Niño, but this is significantly weaker than the January–May 1992 episode. Also notable are two stationary centers of enhanced convection in the Western Indian Ocean which are arranged symmetrically about the equator at ±5° latitudes. These patterns are active during the 1992 El Niño event.

4.1.3 Eastward-propagating intraseasonal modes

The leading eigenfunctions with a dominant intraseasonal component are ϕ_{12} and ϕ_{15} (Fig. 1k, m). Despite being non-consecutive in the eigenvalue ordering, these modes are phase-locked in quadrature and feature broad spectral peaks centered at $\sim 1/60\,\text{days}^{-1}$. Spatially, the reconstructions for the two boreal winters of 1992 and 1993 in Movie 1(f) show clear eastward-propagating wavetrains of enhanced and suppressed convection exhibiting the key features of the MJO, namely initiation over the Indian Ocean, propagation via the Maritime Continent to the Western Pacific warm pool, and eventual demise in the central Pacific near the dateline.

In the reconstruction of the boreal winter 1992 in Movie 1, the MJO is already active by January 1, 1992 and

remains active until the end of March developing a wavetrain of two events. The following boreal fall and winter, which encompass the TOGA COARE IOP, is reconstructed in Fig. 4. There, a moderate MJO initiates at the end of November 1992 over the Indian Ocean, and reaches the dateline by the end of December. Following this event, a stronger MJO initiates over the Indian Ocean in early January 1993, and reaches the dateline by mid-February. Notable features of this stronger event are enhanced convection over eastern tropical Brazil developing prior to initiation over the Indian Ocean, as well as an arc-like pattern of enhanced convection emanating from eastern Africa and merging with the main MJO envelope over the Indian Ocean. Tung et al. (2014b) observed a similar arc-like structure in T_b composites constructed through NLSA-based MJO indices for 1D averaged data. Overall, the reconstructions are qualitatively consistent with the two well-studied MJOs identified during the TOGA COARE IOP (Lin and Johnson 1996a, b; Tung et al. 1999; Yanai et al. 2000).

A key feature of the time series for eigenfunctions ϕ_{12} and ϕ_{15} is strong intermittency and seasonality, with most of the activity occurring during ∼4-month periods starting in late boreal winter and ending in late boreal spring. The MJO signals extracted from these eigenfunctions differ significantly from those extracted via conventional linear approaches which tend to identify MJO-like signals with persistent activity throughout the year (e.g., Wheeler and Hendon 2004). We will return to this point in Sect. 4.2. Compared to the analysis of TGM, the MJO eigenfunctions derived from the 2D data most closely resemble their counterparts derived from *antisymmetric* 1D data. The development of equatorial asymmetry within the MJO lifecycle can also be seen in Fig. 4 and Movie 1(f). There, the recovered ISO signals are approximately symmetric about the equator at initiation over the Indian Ocean, but develop a significant equatorial asymmetry as the MJO propagates across the Maritime Continent. In TGM, the MJO modes extracted from NLSA applied to 1D symmetric data have significantly weaker intermittency, and resemble more closely the corresponding modes derived via SSA.

4.1.4 Poleward-propagating intraseasonal modes

The second set of Laplacian eigenfunctions with dominant intraseasonal variability, ϕ_{21} and ϕ_{22} (Fig. 1r, s), are mainly active during boreal summer and describe the poleward-propagating BSISO. As shown in Movie 1(h), these modes initiate in boreal spring with a cluster of positive T_b (dry) anomalies forming over the northeastern Indian Ocean (cf. the MJO modes in Sect. 4.1.3, which initiate over the Indian Ocean at a wet phase). That cluster propagates northeastward towards India and the Western Pacific (eventually exiting the analysis domain), and is followed by a similar pattern of the opposite sign, creating a characteristic BSISO wavetrain. The frequency of the BSISO modes is somewhat shorter than that of the MJO ($\sim 1/40\,\text{days}^{-1}$), and a weak BSISO signal is present all year-round. In terms of initiation, BSISO appears to initiate slightly more to the north in the Indian Ocean than MJO. During the peak activity of BSISO [e.g., June 1992 and June 1993 in Movie 1(h)] appreciable anomalies also develop in the eastern Pacific to the north of the equator. This is consistent with the analysis of Zhang and Dong (2004), who find ISO-like signals in that region during boreal summer using combined zonal wind and precipitation data.

It is important to note that while the BSISO and MJO eigenfunctions are orthogonal on the data manifold, the corresponding spatiotemporal patterns are not constrained to be orthogonal in space. Visually, the non-orthogonality in the patterns is alluded by their similar spatial structure in the north of the Maritime Continent. There, the BSISO convective envelope correlates to some extent with the northern branch of the MJO envelope identified in Sect. 4.1.3. More quantitatively, the lack of orthogonality can be measured by the angle between the subspaces spanned by the projections of the data in delay-coordinate space onto the BSISO and MJO eigenfunctions, respectively. We find that the MJO and BSISO spatiotemporal subspaces form an angle of 66°. In contrast, the MJO mode identified through SSA (or the equivalent EEOFs) applied to the CLAUS T_b data contains both eastward- and poleward-propagating anomalies at the north of the Maritime Continent, and is generally active year-round (see Movie 2 and Fig. 1 in the supplementary material). It is possible that the SSA ISO modes are a mixture of the more intrinsic eigenfunctions ϕ_i identified here.

4.1.5 Higher harmonics of the annual cycle

Appearing after the annual and semiannual periodic modes and ENSO is a pair of quasi-periodic eigenfunctions, $\{\phi_6, \phi_7\}$ (Fig. 1f, g), whose dominant frequency is at the third harmonic of the annual cycle, 3/y. In the spatial reconstruction in Movie 1(e), eigenfunction ϕ_6 features a southeastward-propagating wave in the Indian Ocean, as well as appreciable variability in the North Atlantic ITCZ that propagates westward through South America into the Pacific Ocean. Eigenfunction ϕ_7 contains more power at interannual timescales than ϕ_6, and features an ENSO-like dipole in the western Pacific (not shown here) in addition to the southeastward-propagating pattern in the Indian Ocean.

It is interesting to note that the southeastward-propagating pattern of anomalously high convection in ϕ_6 crosses the Indian Ocean around the initiation phase of the MJO (see, e.g., the reconstruction in Movie 1 around January 15,

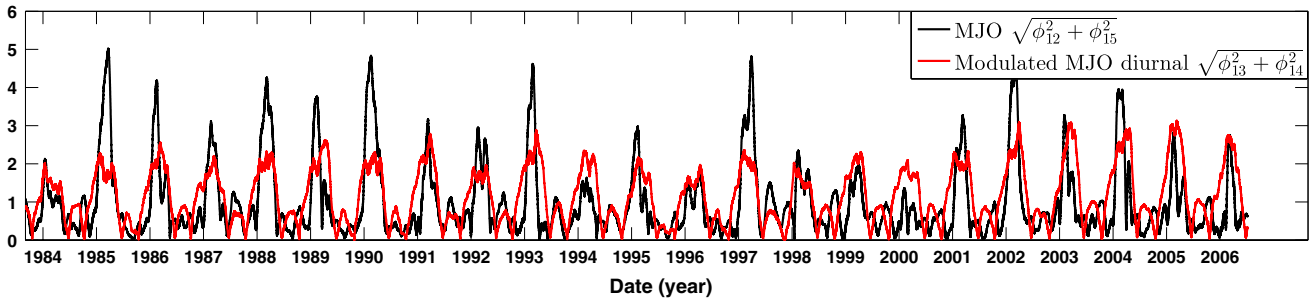


Fig. 5 The MJO index r_t^{MJO} from (16a) and the corresponding index $\sqrt{\phi_{13}^2(t)+\phi_{14}^2(t)}$ associated with the diurnal modes in Fig. 1l. The correlation of the two time series is 0.60

1993), suggesting that this mode may play a preconditioning role in the convective environment experienced by the MJO. In fact, ϕ_6 appears to be related through its phase to the timing of both the initiation and termination of the significant MJO events extracted by eigenfunctions $\{\phi_{12},\phi_{15}\}$, as shown in Fig. 10 ahead. We will discuss further this connection (as well as a similar connection between ϕ_7 and BSISO) in Sect. 4.2.1.

Besides the pair $\{\phi_6,\phi_7\}$, other eigenfunctions in the NLSA spectrum with appreciable power at the harmonics of the annual cycle are $\{\phi_{19},\phi_{20}\}$ (Fig. 1p, q) and, to a lesser extent, ϕ_{10} (Fig. 1i), ϕ_{11} (Fig. 1j), and ϕ_{18} (Fig. 1o). The dominant periodic component in $\{\phi_{19},\phi_{20}\}$ is the fourth harmonic of the annual cycle. Eigenfunctions ϕ_{10}, ϕ_{11}, and ϕ_{18} variously have power in the annual-cycle harmonics 1–4. We refrain from making further physical interpretation of these modes except to note that they might play a role in ISO preconditioning analogous to $\{\phi_6,\phi_7\}$.

4.1.6 Modulated diurnal modes

The modulated diurnal patterns always appear as pairs of twofold-degenerate eigenfunctions, namely $\{\phi_8,\phi_9\}$, $\{\phi_{13},\phi_{14}\}$, $\{\phi_{16},\phi_{17}\}$, and $\{\phi_{23},\phi_{24}\}$. Among these families, the pair $\{\phi_{13},\phi_{14}\}$ (Fig. 1l) is modulated by the amplitude of the MJO. This pair exhibits a discernible intraseasonal peak in its frequency spectrum, and has approximately equal eigenvalues to the MJO eigenfunctions (see Fig. 2). Moreover, the amplitude of this diurnal pair, depicted in Fig. 5, is correlated with the amplitude of the MJO pair to a moderately high extent (temporal correlation coefficient 0.60). As shown in Fig. 1k–m, both the MJO and the diurnal pair $\{\phi_{13},\phi_{14}\}$ have significant power on intraseasonal scales, indicating that the amplitude of MJO acts as a modulating envelope for the amplitude of these diurnal patterns. The correlation in amplitude can also be observed in the physical domain in the reconstructions in Movies 1(f) and 1(g).

In Movie 1(g), the diurnal family $\{\phi_{13},\phi_{14}\}$ exhibits strong variability over the African and South American land masses and islands of the Maritime Continent. Over South America, the reconstructed T_b anomaly field exhibits a southwestward-propagating pattern originating over eastern tropical Brazil, as well as a more pulsating pattern over the central Amazon region. Over Africa, the apparent migration of T_b anomalies takes place predominantly in a zonal, westerly direction. The spatial patterns of these diurnal modes are more complicated over the Maritime Continent, where significant variations in T_b anomalies can be seen among the islands, e.g., Borneo and New Guinea.

Besides the MJO-modulated eigenfunctions, the diurnal pair $\{\phi_{23},\phi_{24}\}$ (Fig. 1t) also exhibits an intraseasonal amplitude modulation, and in this case the modulating envelope is associated with BSISO. Similarly to the pair $\{\phi_{13},\phi_{14}\}$, the reconstructed T_b field corresponding to $\{\phi_{23},\phi_{24}\}$ displays activity over tropical Africa and South America and the Maritime Continent. However, in this case significant T_b anomalies with an apparent north-northwestward propagation also take place over the southern tips of India and Indo-China [see Movie 1(i)]. In addition, the reconstructed field from $\{\phi_{23},\phi_{24}\}$ generally exhibits smaller-scale features over Africa and South America than the MJO-modulated eigenfunctions, and tends to be confined to the north of the equator.

The other two diurnal eigenfunction families in Fig. 1, $\{\phi_8,\phi_9\}$ and $\{\phi_{16},\phi_{17}\}$, are characterized by modulating envelopes at interannual timescales and the higher harmonics of the annual cycle (in particular the third harmonic). Among these families, the pair $\{\phi_8,\phi_9\}$ is especially strongly modulated, being active during July–November and virtually quiescent during the rest of the year. We defer a more detailed study of these modes, as well as the ISO-modulated diurnal modes, to future work.

4.2 Indices for intraseasonal variability

4.2.1 NLSA-based indices

On the basis of the results in Sects. 4.1.3 and 4.1.4, we adopt eigenfunctions $\{\phi_{12},\phi_{15}\}$ and $\{\phi_{21},\phi_{22}\}$ and the

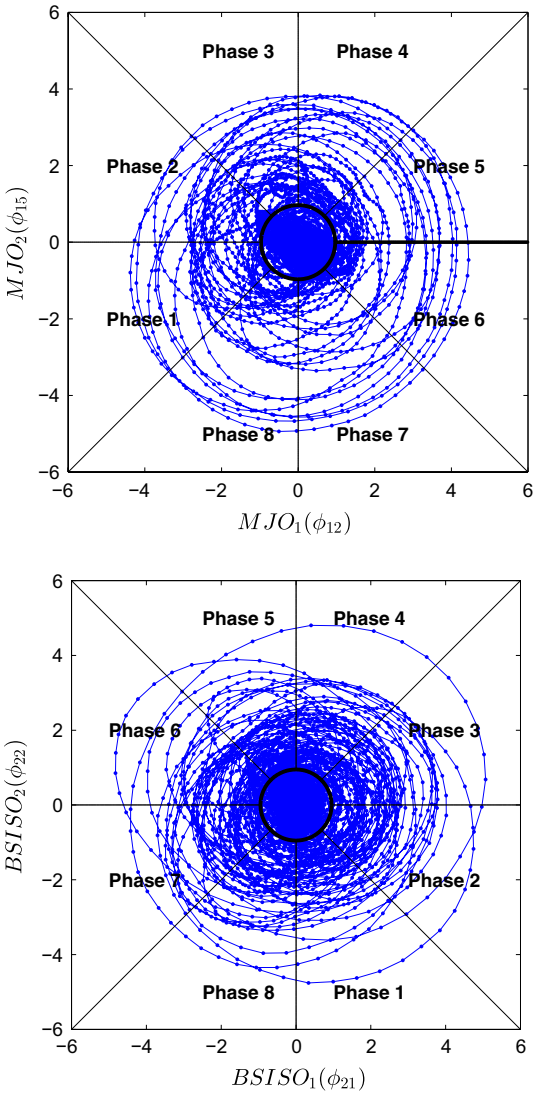


Fig. 6 Two-dimensional phase space representation of the MJO and BSISO based on the Laplace–Beltrami eigenfunctions $\{\phi_{12}, \phi_{15}\}$ and $\{\phi_{21}, \phi_{22}\}$, respectively, showing the years 1983–2006 with one sample per day plotted. The 2D phase spaces are split into 8 phases with associated composites reconstructed in Fig. 9

corresponding spatiotemporal patterns as our definition of the MJO and BSISO, respectively. These eigenfunction families, which are plotted in 2D phase diagrams in Fig. 6, are analogous to the PC families of Kikuchi et al. (2012) derived from EEOFs of seasonally-partitioned and band-pass-filtered OLR and wind data, but the seasonality of the extracted MJO and BSISO signals emerges here naturally from the Laplacian eigenfunctions applied to the full 2D T_b record. The fact that the input data were not subjected to bandpass filtering opens up the possibility to explore directly the relationships of these modes to other important modes of tropical variability, such as ENSO and the diurnal cycle. In Sect. 5, we will focus on quantifying the statistical predictability of the MJO amplitude conditioned on a reduced representation of tropical organized convection given by the Laplacian eigenfunctions discussed in Sect. 2.1. Elsewhere (Chen et al. 2014), it is shown that the NLSA-based MJO time series are well-described by a nonlinear stochastic oscillator model with predictive skill extending to 40 days. The skill of these models supports the dynamical significance of the MJO signals extracted via NLSA. Other applications of the extracted modes include phase compositing (Tung et al. 2014b) and real-time monitoring.

Following Kikuchi et al. (2012), we construct MJO and BSISO indices from the amplitudes of the corresponding Laplace–Beltrami eigenfunction pairs, i.e.,

$$r_t^{\mathrm{MJO}} = \sqrt{\phi_{12}^2(t) + \phi_{15}^2(t)}, \tag{16a}$$

$$r_t^{\mathrm{BSISO}} = \sqrt{\phi_{21}^2(t) + \phi_{22}^2(t)}. \tag{16b}$$

We use these indices, standardized to unit variance, to construct a bimodal index, $(r_t^{\mathrm{MJO}}, r_t^{\mathrm{BSISO}})$, representing the activity of the dominant ISO modes recovered by NLSA in the boreal winter and boreal summer. This bimodal index is displayed in Fig. 7a for the 1983–2006 period spanned by the data.

Significant pure MJO events with respect to the indices in (16) occur whenever $r_t^{\mathrm{MJO}} \geq 1$ and $r_t^{\mathrm{BSISO}} < 1$ after standardization. Such events can be easily identified in Region I of Fig. 7a. Similarly, significant pure BSISO events have standardized $r_t^{\mathrm{MJO}} < 1$ and $r_t^{\mathrm{BSISO}} \geq 1$. These events correspond to Region III in Fig. 7a, and can also be easily identified. Significant events with respect to both r_t^{MJO} and r_t^{BSISO} can be classified as either MJO or BSISO (Region II), but the fraction of the significant ISO events with respect to the NLSA-based indices falling in this category is relatively small. Moreover, most of the events that do belong in Region II can still be identified as either predominantly MJO or BSISO, for they are relatively far away from the $r_t^{\mathrm{MJO}} = r_t^{\mathrm{BSISO}}$ line where classification is ambiguous. For comparison we also show the bimodal ISO index from SSA in Fig. 7b and discuss it in more detail in the supplementary material along with the temporal patterns.

The strong seasonality of the ISO indices leads to significant changes in the corresponding time-dependent climatological PDFs, $p(r_t^{\mathrm{MJO}})$ and $p(r_t^{\mathrm{BSISO}})$. As shown in Fig. 8, at times t preceding the active ISO period in each case (e.g., November 1 and April 1 for MJO and BSISO, respectively), these distributions have strong peaks and large mass at $r_t \lesssim 1$. Subsequently, they evolve towards broad distributions at the peak of the active phase carrying appreciable density for significant events in the interval $1 \lesssim r_t \lesssim 4.5$ [e.g., March 1 (MJO) and August 1 (BSISO)].

The seasonality of $p(r_t^{\mathrm{MJO}})$ will play an important role in the regime predictability results in Sect. 5.

Next, we create phase composites for the MJO and BSISO by dividing the portion of the 2D phase spaces in Fig. 6 with $r_t^{\mathrm{MJO}} \geq 1$ and $r_t^{\mathrm{BSISO}} \geq 1$ (after standardization) into 8 phases, and averaging the reconstructed T_b fields in

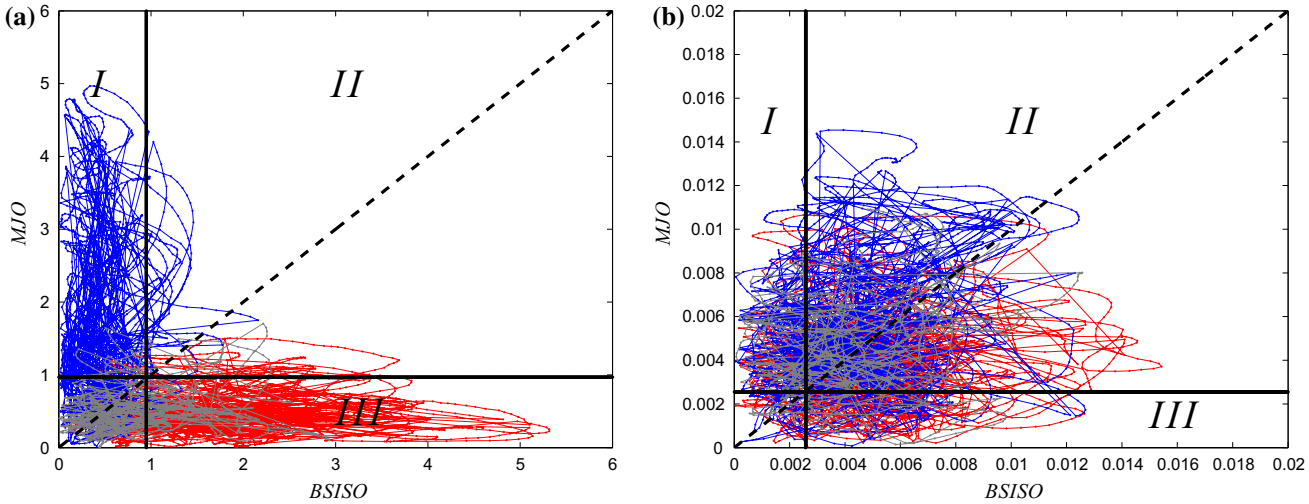


Fig. 7 Bimodal indices, $(r_t^{\mathrm{MJO}}, r_t^{\mathrm{BSISO}})$, for the MJO and BSISO from **a** NLSA and **b** SSA. Following Kikuchi et al. (2012), three periods are plotted in different colors: June–October (*red*), December–April (*blue*), and otherwise (*gray*). The *solid lines* indicate the threshold for significant ISO events corresponding to one standard deviation of the MJO and BSISO indices. Observations are classified as follows: *I* MJO events, *III* BSISO events, *II* MJO or BSISO events. The majority of significant ISO events according to the NLSA indices in (**a**) are either MJO events (*I*) occurring in December–April, or BSISO events (*II*) occurring in June–October. The classification accuracy with respect to the SSA indices in (**b**) is significantly poorer. **a** NLSA-based bimodal ISO index. **b** SSA-based bimodal ISO index

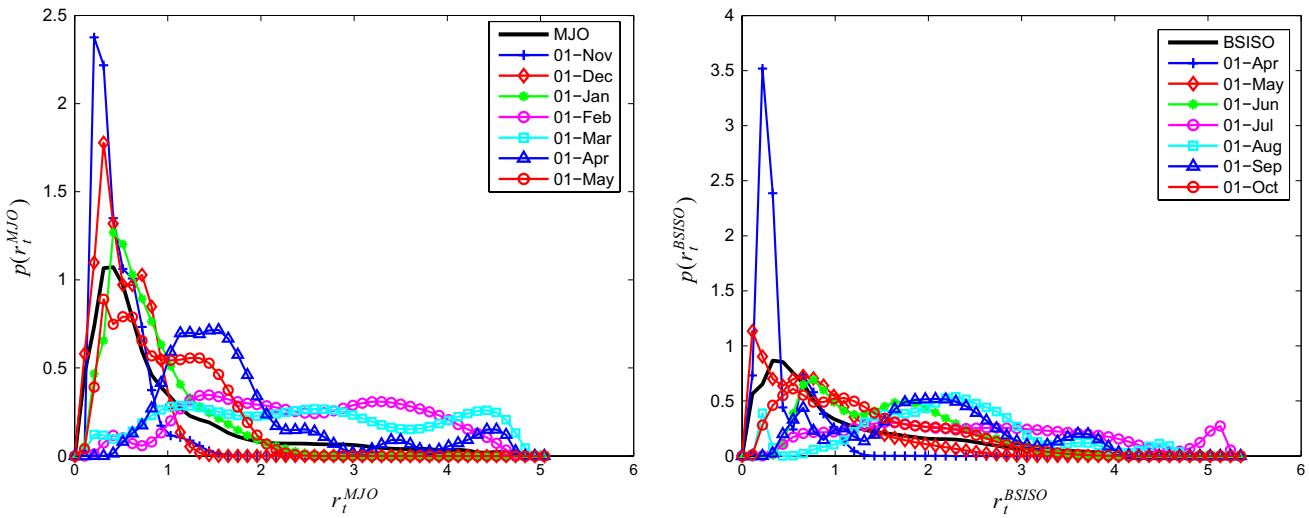
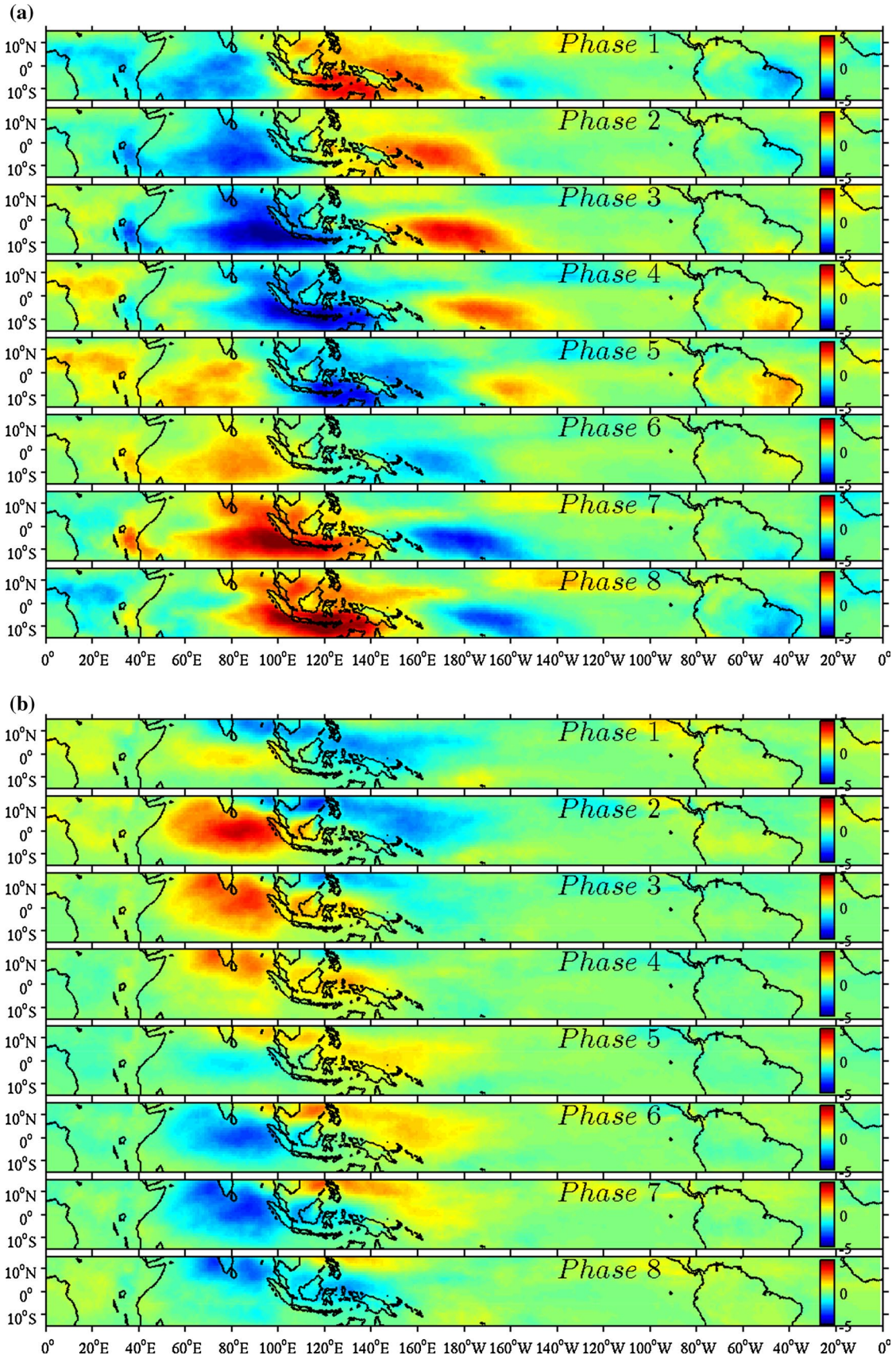


Fig. 8 Time-dependent climatological PDFs, $p(r_t)$, for the MJO and BSISO indices. For MJO, representative PDFs are shown for the autumn–spring months (November–May), and for BSISO for the spring–autumn months (April–October). The yearly (time-independent) climatological distribution is plotted in a black line in each case for reference

Fig. 9 The composite life cycles of **a** MJO and **b** BSISO reconstructed from the Laplace–Beltrami eigenfunctions in Fig. 1. The composites correspond to the 8 phases identified in Fig. 6, and exhibit eastward- and poleward-propagating patterns characteristic of MJO and BSISO, respectively. *Note* the enhanced convection signals over East Africa and eastern tropical Brazil in Phase 1 of the MJO. **a** MJO composites. **b** BSISO composites


(a)
Phase 1
Phase 2
Phase 3
Phase 4
Phase 5
Phase 6
Phase 7
Phase 8
10°N
0°
10°S
0
-5
0° 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E 180°W 160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0°
(b)
Phase 1
Phase 2
Phase 3
Phase 4
Phase 5
Phase 6
Phase 7
Phase 8
10°N
0°
10°S
0
-5
0° 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E 180°W 160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0°

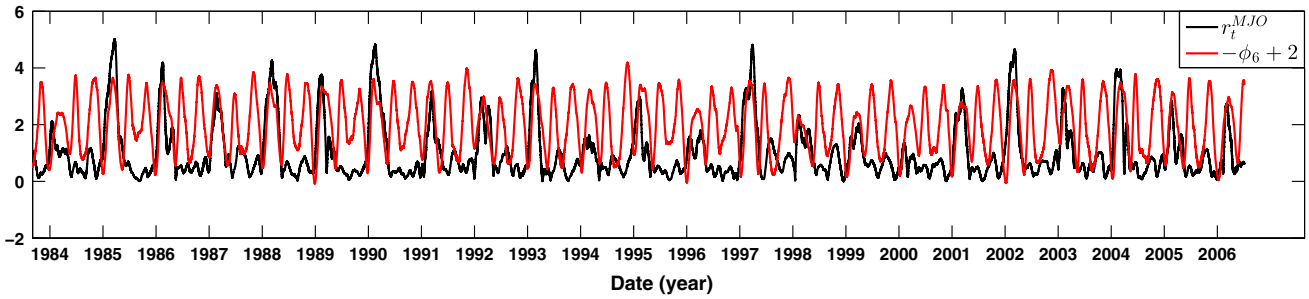


Fig. 10 The MJO index r_t^{MJO} and the quasi-periodic eigenfunction ϕ_6 associated with the third harmonic of the annual cycle. For clarity of visualization, ϕ_6 is shown here sign-inverted and shifted by a constant value. *Note* the phase correlation between the two time series, especially in relation to the initiation, duration, and termination of MJO events. The latter are seen to take place between the first two positive peaks of ϕ_6 at the beginning of each year

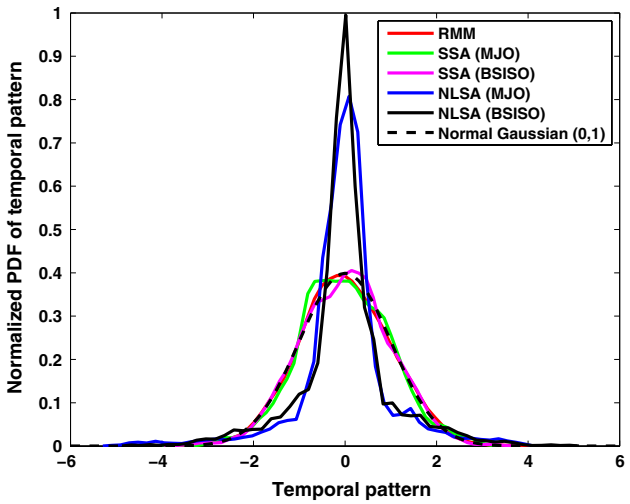


Fig. 11 PDFs of the NLSA, RMM, and SSA eigenmodes standardized to zero mean and unit variance. A normal distribution with zero mean and unit variance is plotted in a *dashed line* for reference. The temporal patterns obtained through NLSA for both MJO and BSISO have significantly heavier tails, and differ the most from the normal distribution

each phase. The resulting phase composites, displayed in Fig. 9, provide an aggregate representation of the MJO and BSISO lifecycles as extracted through the corresponding Laplace–Beltrami eigenfunctions. These lifecycles were discussed in Sects. 4.1.3 and 4.1.4 with reference to the 1992–1993 reconstruction.

In the case of MJO, initiation over the Indian ocean, accompanied by the development of anomalously high convection over eastern Africa and tropical eastern Brazil, takes place in Phases 1 and 2, and is followed by propagation over the Maritime Continent and the western Pacific warm pool in Phases 3–5 and 6–7, respectively. Termination at the dateline takes place in Phase 8. As remarked in Sect. 4.1.3, the MJO convective envelope is displaced to the south of the equator as it passes through the Maritime Continent. Moreover, the MJO envelope exhibits a strong land-sea contrast, especially over Borneo and New Guinea where the reconstructed T_b anomalies are significantly weakened.

For BSISO, we chose Phase 1 to correspond to a cluster of positive T_b anomalies developing in the central Indian Ocean, as we found that in our reconstructions the first significant BSISO event in a given year initiates at a dry phase [see Movie 1(h)]. In Phases 2–4, that cluster moves northeastward towards the Bay of Bengal and India and branches off towards the western Pacific and the Monsoon Trough, bypassing the Maritime Continent from the north. Following the dry phase of BSISO, a cluster of anomalously high convection develops in Phase 5 in the central Indian Ocean, and propagates towards India and the western Pacific in Phases 6–8, completing the BSISO cycle.

Consider now an intriguing phase relationship mentioned in Sect. 4.1.5 between the MJO index and the third harmonic of the annual cycle, ϕ_6 (Fig. 1f). As can be seen in Fig. 10, the initiation and termination of the active-MJO period in a given year (as measured by the NLSA-based r_t^{MJO} index) correlates strongly with two consecutive peaks in the $\phi_6(t)$ waveform. That waveform is quasi-periodic, so the duration of the active-MJO period will vary from year to year around the four-month period of the third harmonic of the annual cycle. A similar (though somewhat weaker) phase relationship holds between BSISO and the other third-harmonic quasi-periodic eigenfunction, ϕ_7 (Fig. 1g). These observations motivate future study of the role of the spatiotemporal patterns associated with ϕ_6 and ϕ_7, as well as the first two harmonics of the annual cycle (see Movie 1), in setting up the background convective environment underlying the initiation and termination of the MJO and BSISO. Note that while ϕ_6 appears to be related to the timing of the active-MJO period, the amplitude of the MJO within the active period is influenced by other factors, and

in particular interannual convective regimes (see Sect. 5 ahead). The latter may also be responsible for the deviations of $\phi_6(t)$ from a purely periodic signal. Note also that removal of the first three harmonics of the annual cycle is widespread practice in the development of MJO indices (e.g., Wheeler and Hendon 2004).

4.2.2 *Comparison with linear MJO and BSISO indices*

EOFs have been extensively employed to define ISO indices for prediction and real-time monitoring based on single or combined atmospheric data sources. Wheeler and Hendon (2004) proposed a widely-used MJO index based on the first two EOFs, denoted RMM1 and RMM2, of the combined fields of OLR and zonal winds at 200 and 850 hPa averaged over the equator. These modes are obtained using all-round year data after removal of the first three harmonics of the annual cycle and interannual variability from each of the three atmospheric fields. The removal of the interannual variability from these fields is necessary because time-averaged anomalies associated with the mature phase of ENSO resemble phases of MJO. Once annual and interannual variability is removed, the three fields are normalized by their global variance to balance their influence in the final output. Following this preprocessing, the first two EOFs vary mostly on the intraseasonal timescale of MJO. Together, the first two EOFs explain 25% of the variance of the signal with RMM1 lagging RMM2 by 10–15 days. In Fig. 11, we plot for comparison the PDFs of the standardized NLSA and RMM modes (for each method we chose to display only one of the two degenerate eigenmodes as they have similar distributions). The distributions for RMM are close to a normal distribution. On the other hand, the NLSA probability distributions depart significantly from the normal distribution, featuring heavy tails due to the intermittency and strong seasonality of the recovered MJO and BSISO signals. Most of the significant ISO events identified via NLSA lie in the tails of the distributions in Fig. 11.

5 Quantifying regime predictability of the MJO amplitude

5.1 Predictor and response variables

The spatiotemporal modes extracted through NLSA from infrared brightness temperature data were shown in Sect. 4 to be associated with meaningful features of organized tropical convection, including the annual cycle and its higher harmonics, interannual and intraseasonal oscillations, and the diurnal cycle. In this section, we apply the information-theoretic framework described in Sect. 2.2 to assess the regime predictability of the MJO in T_b data, using the NLSA eigenfunctions to construct a low-dimensional space of predictor variables. The response variable representing the MJO amplitude at lead time τ after forecast initialization time t will be the NLSA-based MJO index $r^{\mathrm{MJO}}_{\tau+t}$ from (16a), and in what follows we use the term "MJO predictability" to refer to predictability of the MJO amplitude. As discussed in Sect. 4.2.2, our definition of the MJO amplitude exhibits strong seasonality, and excludes ISO-like signals taking place year-round or during the boreal summer. However, this exclusion takes place objectively without subjecting the data to preprocessing. Moreover, the space of predictor variables for clustering will be the 17-dimensional space spanned by the leading 25 Laplace–Beltrami eigenfunctions (see Fig. 1), excluding the diurnals. Thus, we are creating a low-dimensional description of the system state $X(t_i)$ via a nonlinear projection map $X(t_i) \mapsto \Phi_{t_i}$ from (6) with $l = 17$. To quantify regime predictability, we further coarse-grain $\Phi_{t_i} \mapsto k_{t_i}$ in accordance with (10), where k_{t_i} is the integer-valued affiliation to a partition constructed through the kernel K-means clustering algorithm applied to the Φ_{t_i} time series. The specific eigenfunction indices used in this study are $\{j_1, \ldots, j_{17}\} = \{1, \ldots, 7, 10, 11, 12, 15, 18, \ldots, 22, 25\}$, and as stated above these eigenfunctions capture the first four harmonics of the seasonal cycle, dominant modes of interannual variability (e.g. ENSO), and the dominant modes of intraseasonal variability (MJO and BSISO).

Note that our choice to exclude the diurnal modes was not due to their presumed lack of relevance to MJO predictability, but was rather dictated by numerical considerations as their fast oscillation timescale adversely affects the performance of clustering. Indeed, the envelope correlation results in Fig. 5, as well as other studies (Chen and Houze 1997; Tian et al. 2006; Ichikawa and Yasunari 2008), suggest that the diurnal modes may contain predictive information for the MJO, but the present numerical framework is not adapted to access that information from the raw eigenfunction time series. A possible modification of the scheme, which is beyond the scope of this paper, would be to extract the modulating envelopes of these modes (e.g., via the Hilbert transform, or the modal amplitude depicted in Fig. 5), and then include the envelopes in the space of predictor variables. Similarly, our truncation of the space of predictors to eigenfunction 25 was a compromise between predictive information content (which increases with the number of eigenfunctions l) on the one hand, and performance of the clustering algorithm. In particular, note that for a fixed number of samples the predictive information content $\mathscr{I}(r^{\mathrm{MJO}}_{t+\tau}, k_t)$ in the partition may actually decrease with l, or become strongly sensitive to the input data. We find that the choice of predictors indicated above is sufficient to reveal MJO amplitude predictability on intraseasonal timescales, and its dependence on physically-meaningful large-scale

convective regimes, such as the state of ENSO and MJO at initialization time. In particular, a major challenge in clustering algorithms applied to regime identification in atmosphere ocean science is to ensure temporal persistence of clusters (Christiansen 2007; Franzke et al. 2009; Horenko 2010). In our work, temporal persistence of regimes emerges naturally by means of the timescale separation achieved by the eigenfunctions and the nonlinear kernel K-means algorithm that avoids abrupt transitions by allowing nonconvex clusters to exist in the predictor space. As a future alternative, partitioning the eigenfunction space with clustering methods enforcing temporal persistence through constrained optimization (Metzner et al. 2012) could bring new information in the regimes, potentially improving our predictability results.

5.2 Cluster analysis for MJO predictability

As discussed in Sect. 2.2, the statistics of both the predictor and response variables exhibit strong periodic time dependence with period $T = 1$ y due to the seasonal cycle [see (7)]. Ideally, one would like to independently estimate the prior and conditional PDFs of Φ_t and r_t^{MJO} for each of the $T \times \delta t = 365 \times 8 = 2920$ timestamps per calendar year, but the number of available samples from different calendar years (23 over the entire CLAUS dataset) is prohibitively small for statistically robust PDF and relative entropy estimation. To tackle this issue, two solutions were employed here: (1) a window ΔT was used around each timestamp t to increase the sample size; (2) the PDFs were estimated using a kernel density estimation (KDE) method (Bowman and Azzalini 1997). Using a moderately-small window ΔT accounts for years of uncollected data where MJO might have occurred earlier or later in the season compared to the available samples. This significantly decreases the variance in the PDF and relative entropy estimates, at the expense of introducing a small bias. Here, $\Delta T = 15$ d so that the PDFs at timestamp $t \in [0, 1]$ y are estimated from all samples in the interval $[t - \Delta T, t + \Delta T]$ plus integer multiples of 1 year, amounting to approximately $30 \times 8 \times 23 = 5520$ samples per estimate. KDE further decreases the variance of the estimators by inducing smoothness in the PDFs compared to raw histograms. In KDE, the PDF $p(u)$ of a scalar random variable to take the value u is estimated by means of a counting sum over the observed values u_i smoothed by a kernel function K_σ centered at u, i.e.,

$$p(u) = \frac{1}{C_\sigma} \sum_{i=1}^{S} K_\sigma(u, u_i),$$

where S is the number of samples, and C_σ a normalization constant computed from the requirement that $\sum_{i=1}^{S} p(u_i) = 1$. The kernel K_σ is required to be a non-negative function symmetric in its two arguments, and also depends on a smoothing parameter σ controlling the influence of the local neighborhood in the final density estimate $p(u)$. Here, we use a Gaussian kernel, $K_\sigma(u, u_i) = e^{-(u-u_i)^2/2\sigma^2}$, where the smoothing parameter is proportional to the bandwidth (standard deviation) of the kernel. For the remainder of the paper, all PDFs are estimated at 50 equispaced bins using KDE with kernel bandwidth estimated as in Bowman and Azzalini (1997) and relative entropies are computed directly via discrete sums over the bins.

Figure 8 shows the PDFs $p(r_t^{\mathrm{MJO}})$ of the NLSA MJO index estimated via this method for representative times t from November–May. Low values of r_t^{MJO} (roughly in the interval $[0, 1]$) indicate an inactive or weak MJO state. The PDFs for early November and early December contain significant probability mass in that interval. The active MJO season starts in mid to late December, with the strongest activity occurring in February and March before eventual suppression in April–May. From June–October MJO is inactive and the density estimates approximately match the November and December PDFs. Further details of the time-dependent climatological PDFs were discussed in Sect. 4.2.1.

Following the methodology described in Sect. 2.2, the kernel K-means algorithm with the Gaussian kernel in (15) is applied to subsets of the Φ_{t_i} time series to create a family of partitions Ξ_t labeled by a time stamp t in the interval $[0, 1]$ y. Again, in order to improve statistical robustness of the results, we construct each partition using the ≈ 5500 samples available for the interval $[t - \Delta T, t + \Delta T]$ plus integer multiples of 1 year with $\Delta T = 15$ days.

An important parameter in clustering algorithms, including kernel K-means, is the number of clusters K. In the scheme of GM, this parameter is set by monitoring the change in the predictive information $\mathscr{I}\left(r_{t+\tau}^{\mathrm{MJO}}, k_t\right)$ as K increases from small values, and choosing the optimal cluster number as the largest value of K beyond which there is no significant increase in $\mathscr{I}\left(r_{t+\tau}^{\mathrm{MJO}}, k_t\right)$. In this study, we also took mutual information into account, but due to the relatively small number of available samples, we additionally sought to construct partitions with cluster occupancy exceeding a threshold. This was done in order to ensure statistical robustness of the cluster-conditional forecast PDFs $p\left(r_{t+\tau}^{\mathrm{MJO}} \mid k_t\right)$ (statistical robustness was not an issue in the sample-rich datasets studied by GM).

In the kernel K-means algorithm, the geometrical structure and occupancy of the clusters are influenced by the kernel bandwidth parameter σ. Larger values of σ will tend to produce more globular, convex clusters similar to the clusters produced by standard K-means. As σ decreases, kernel K-means is able to recover nonconvex clusters, but the algorithm becomes increasingly ill-conditioned and sensitive to sampling errors. After experimenting with several combinations of

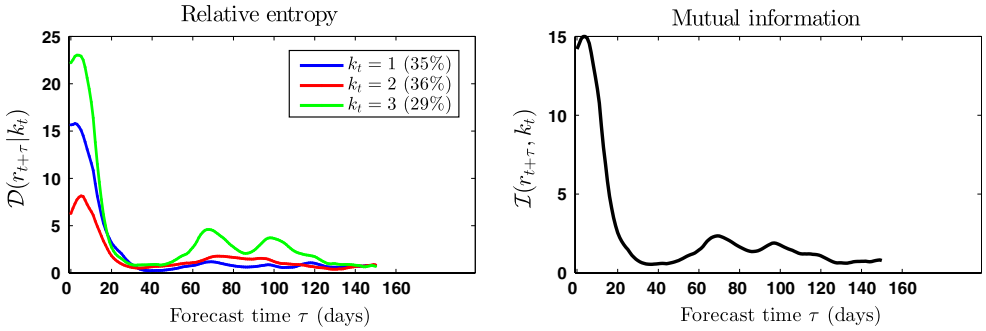


Fig. 12 Predictability scores for forecasts initialized at $t =$ December 1 as measured by the cluster-conditional relative entropy $\mathscr{D}\left(r_{t+\tau}^{\mathrm{MJO}} \mid k_t\right)$ and mutual information $\mathscr{I}\left(r_{t+\tau}^{\mathrm{MJO}}, k_t\right)$ at lead time τ associated with the La Niña ($k_t = 1$), ENSO-neutral ($k_t = 2$), and El Niño ($k_t = 3$) clusters. The El Niño cluster is the most predictable cluster for almost all leads, and also displays a reemergence of predictability for $\tau \in [65, 95]$ d. See Fig. 15 for an illustration of the difference of the cluster-conditional PDFs relative to climatology giving rise to the increased predictability of this cluster. The percentage values in the labels for k_t indicate occupation probability $p(k_t)$ of each cluster

(K, σ) with $K \leq 4$, we selected the partitions where the minimum cluster occupancy was at least 20 %. These experiments were repeated with multiple values of the bandwidth parameter σ in the interval $[0.5, 20]$ and with multiple random initializations. Overall, we found that the gain in predictive information at $K = 4$ relative to $K = 3$ for well-balanced partitions was significantly lower than the corresponding increase between $K = 3$ and $K = 2$. Moreover, several of the $K = 4$ partitions consisted of what appeared to be ad hoc subpartitions of $K = 3$ clusters. We therefore chose $K = 3$ as a reasonable compromise between predictive information gain and statistical robustness for the available number of samples.

In what follows, we present predictability results for initialization times either before the active-MJO season ($t =$ November 1 and December 1), or near the peak of the active-MJO season ($t =$ February 1). We will see that the early- and active-season partitions (and the corresponding predictability results) differ significantly, with the former being dominated by the interannual modes and the latter by both interannual and intraseasonal modes.

5.3 Early-season regime predictability

We begin by quantifying MJO regime predictability with initialization at time $t =$ December 1. As shown in Fig. 8, the prior PDF $p(r_t^{\mathrm{MJO}})$ is strongly peaked at low values of the MJO index ($\simeq 0.3$). Moreover, it features a significantly thinner right tail than the yearly climatological PDF, indicating that the MJO is mainly inactive at initialization time. We selected a well-balanced partition with $K = 3$ and kernel bandwidth $\sigma = 1.58$ using the method describe above. With this choice of parameters the occupation probabilities of the clusters are $p(k_t) = (0.35, 0.36, 0.29)$. The expected and cluster-conditional predictability corresponding to this partition, measured via the metrics $\mathscr{I}\left(r_{t+\tau}^{\mathrm{MJO}}, k_t\right)$ from (11) and $\mathscr{D}\left(r_{t+\tau}^{\mathrm{MJO}} \mid k_t\right)$ from (12), respectively, are shown in Fig. 12 for lead times τ in the interval $[0, 150]$ d. There, the expected predictability is seen to decrease rapidly after an initial period at short leads ($\tau \lesssim 7$ d), becoming essentially negligible by $\tau = 30$ d. However, predictability reemerges at later times, developing two distinct peaks at $\tau \sim$70 and 90 days, before finally decaying. As indicated by the $\mathscr{D}\left(r_{t+\tau}^{\mathrm{MJO}} \mid k_t\right)$ scores, this reemergence of predictability is mainly associated with cluster 3 in the partition.

To physically interpret these results, we first consider the cluster affiliations k_{t_i} from (10) of the data samples used to construct the partition (i.e., the samples at calendar days December 1 $\pm$ 15d), which are visualized in Fig. 13 against the $r_{t+\tau}^{\mathrm{MJO}}$ time series at representative leads. In particular, by comparing the k_{t_i} time series to historical ENSO indices it emerges that the cluster affiliations are strongly correlated with ENSO. Specifically, clusters 1, 2, and 3 are mainly occupied during La Niña (1989, 1996, 1997, 1999–2002, and 2006), ENSO-neutral (1984–1988, 1990, 1991, 1993, 2002, 2004), and El Niño years (1988, 1992, 1994, 1995, 1998, 2003, 2005), respectively.

The influence of interannual modes on the structure of the partition is also evident in the cluster-conditional marginal PDFs for individual eigenfunctions, $p(\phi_i(t) \mid k_t)$, examples of which are displayed in Fig. 14. There, the El Niño cluster ($k_t = 3$) is positive on $\{\phi_5, \phi_{10}\}$ and negative on ϕ_7. On the other hand, the La Niña cluster ($k_t = 1$) is negative on $\{\phi_5, \phi_{10}\}$ and positive on ϕ_7. The ENSO-neutral cluster ($k_t = 2$) is well distinguishable with respect to eigenfunctions $\{\phi_7, \phi_{10}, \phi_{18}\}$, where it is centered around zero. Thus, the three clusters have distinct signatures in the feature space of the eigenfunctions, and the partition of that

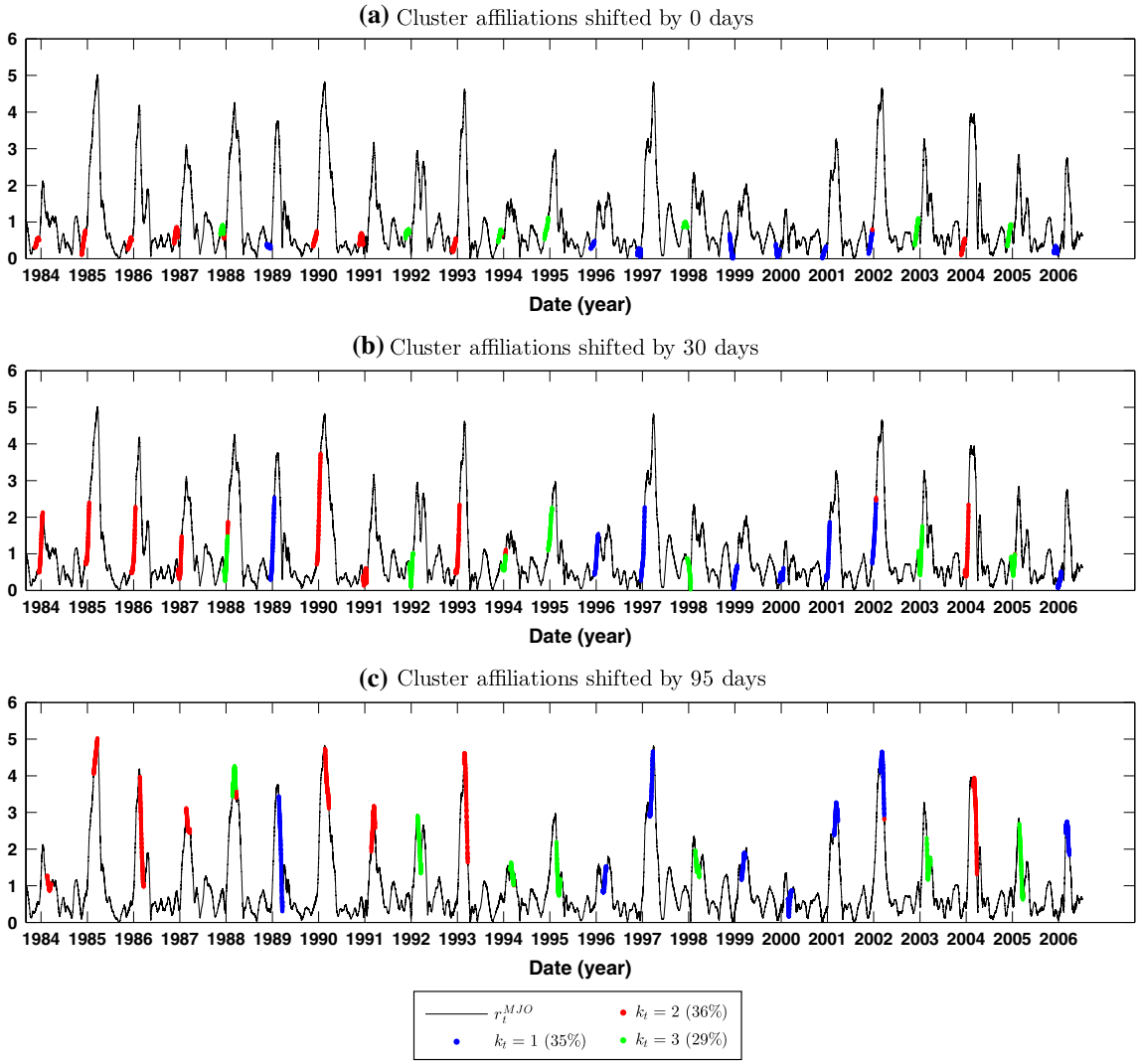


Fig. 13 The cluster affiliation sequence k_t for forecasts initialized at $t =$ December 1, visualized against the MJO index r_t^{MJO}. *Note* the double peaks in r_t^{MJO} occurring for the years occupied by the El Niño cluster ($k_t = 3$). To illustrate the time evolution of the samples in each cluster, the cluster affiliations shifted by 30 days (**b**) and 95 days (**c**) are also shown

space via kernel K-means clustering naturally corresponds to three regimes in the interannual variability of large-scale organized convection.

The three clusters also have distinct consequences on MJO regime predictability. As mentioned earlier, the El Niño cluster is the only one that exhibits a reemergence of predictability (see Fig. 12). This cluster is also the most predictable cluster for almost all lead times considered, followed by the La Niña and ENSO-neutral clusters in that sequence. Figure 15 displays the cluster-conditional PDFs $p(r_{t+\tau}^{\mathrm{MJO}} \mid k_t)$ for representative lead times in the interval $[0, 150]$ days. As is evident from the PDFs at $\tau = 65$, 85, and 95 d (which correspond to physical times $t + \tau$ during the active MJO period), the El Niño and La Niña clusters exhibit on average weaker MJO activity than the ENSO-neutral cluster. However, that activity is more predictable in the sense that the PDFs $p\left(r_{t+\tau}^{\mathrm{MJO}} \mid k_t\right)$ have larger relative entropy distance from the prior $p(r_{t+\tau})$. This result is consistent with the known tendency of anomalous ISO behavior to occur during El Niño and La Niña events (Lau 2011).

The predictive information content in the cluster-conditional PDFs is especially large for the El Niño cluster at $\tau = 65$ and 95 d. The prominent peaks of these distributions, which are not present in the prior, give rise to the observed reemergence of predictability as measured by the $\mathscr{D}\left(r_{t+\tau}^{\mathrm{MJO}} \mid k_t\right)$ and $\mathscr{I}\left(r_{t+\tau}^{\mathrm{MJO}}, k_t\right)$ metrics at those lead times. Upon inspection of the $r_{t+\tau}^{\mathrm{MJO}}$ time series in Fig. 13, one finds that during the years when the El Niño cluster is occupied $r_{t+\tau}^{\mathrm{MJO}}$ tends to form two consecutive peaks per

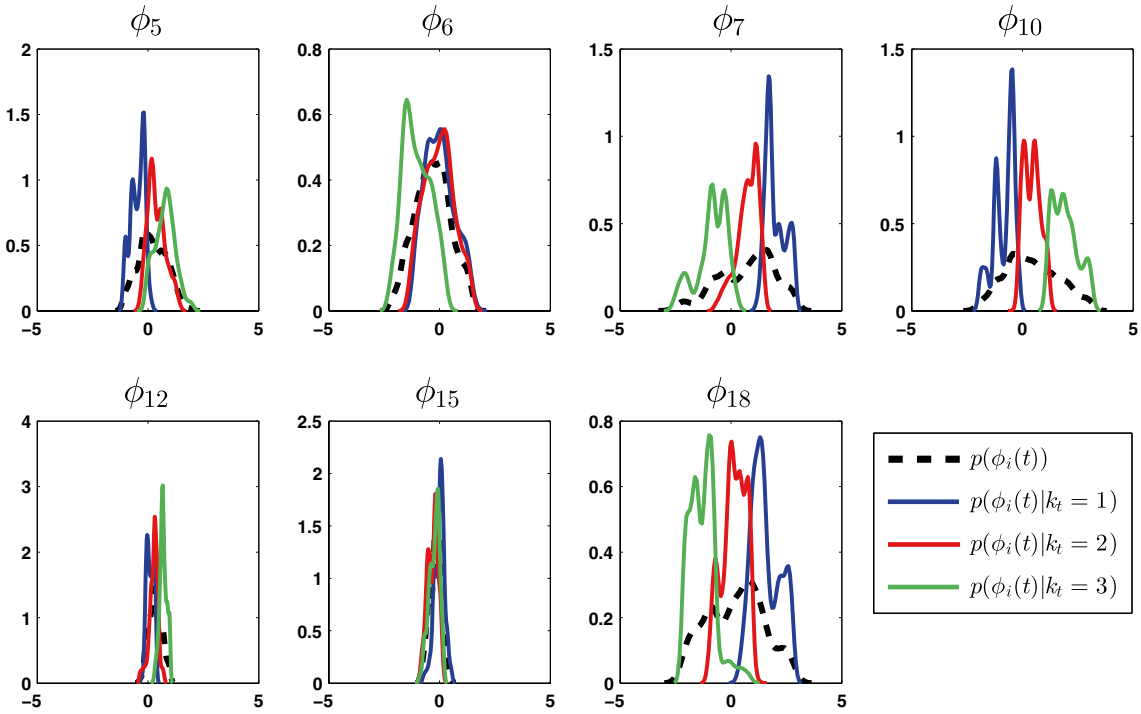


Fig. 14 Representative cluster-conditional marginal PDFs $p(\phi_i(t) \mid k_t)$ for the Laplace–Beltrami eigenfunctions used as predictors at initialization time $t =$ December 1. The ENSO eigenfunction (ϕ_5) separates into La Niña ($k_t = 1$), ENSO-neutral ($k_t = 2$) and El Niño ($k_t = 3$) clusters, corresponding to negative, $\simeq 0$, and positive ϕ_5 values, respectively. A similar separation takes place for eigenfunctions ϕ_7, ϕ_{10}, and ϕ_{18}

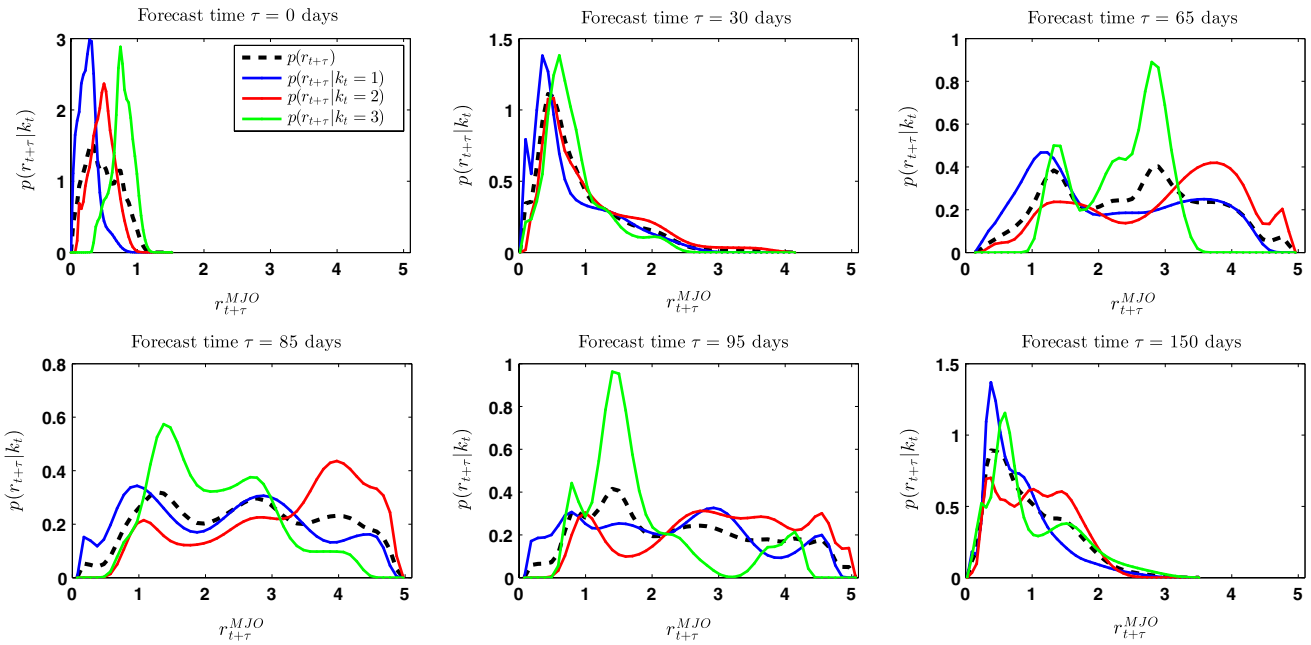


Fig. 15 Cluster-conditional PDFs, $p\left(r^{\mathrm{MJO}}_{t+\tau} \mid k_t\right)$, for the MJO index for forecasts initialized at $t =$ December 1. The prior $p(r_{t+\tau})$ at each forecast time τ is plotted with a *dashed line*. The relative entropy between the posterior cluster-conditional distributions and the prior corresponds to the predictive information of the three convective regimes represented by the clusters; i.e., La Niña ($k_t = 1$), ENSO-neutral ($k_t = 2$), and El Niño ($k_t = 3$). For leads up to 20 d, the El Niño and La Niña clusters exhibit larger differences from climatology than the ENSO-neutral cluster, giving rise to higher relative-entropy scores as shown in Fig. 12. The reemergence of predictability for the El Niño cluster is associated with the narrow peaks in the corresponding conditional PDFs at $\tau = 65$ and 95 d

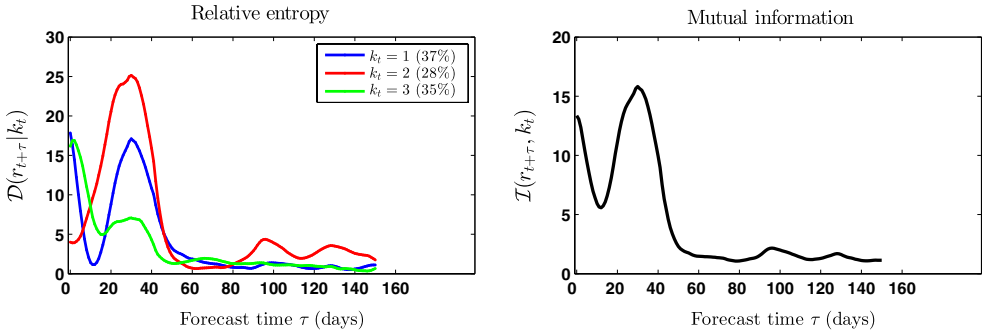


Fig. 16 Predictability scores for forecasts initialized at $t =$ November 1. The partition for November 1 has a similar structure as the December 1 partition and consists of La Niña ($k_t = 1$), El Niño ($k_t = 2$), and ENSO-neutral ($k_t = 3$) clusters. Shifted forward by 30 d, the predictability scores shown here are consistent with the $t =$ December 1 results in Fig. 12, demonstrating robustness of the partitions constructed via the kernel K-means algorithm

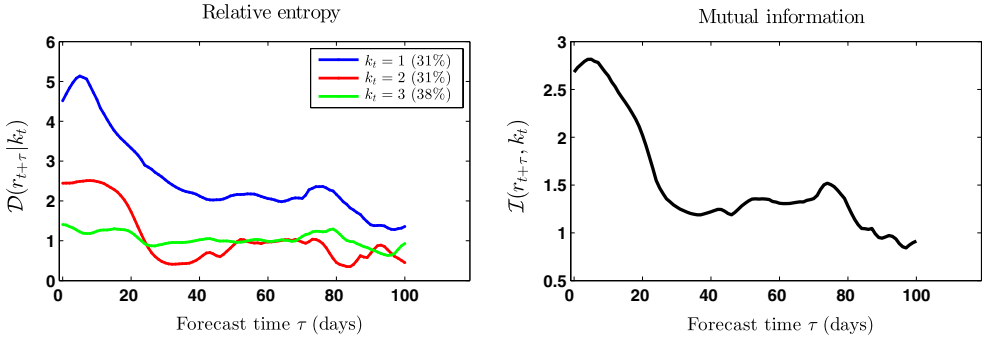


Fig. 17 Predictability scores for forecasts initialized at $t =$ February 1. In this case, the most predictive cluster at all times is a La Niña cluster ($k_t = 1$) (cf. Figs. 12 and 16). The other two clusters are both occupied during neutral and positive phases of ENSO, and they are separated with respect to the first MJO eigenfunction, ϕ_{12}. According to Fig. 6, cluster 3 (negative ϕ_{12} values in Fig. 20) mainly consists of MJO Phases 8 and 1–3, whereas cluster 2 (positive ϕ_{12}) occupies Phases 4–7

boreal winter. It is possible that during those times (which correspond to mid-February to early March) the background environment over the Warm Pool returns to a state which is more representative of ENSO-neutral years, and thus more conducive to the formation of "canonical" MJO events. It is important to note that the overall suppression of r_t^{MJO} in the La Niña and El Niño clusters does not imply complete absence of eastward-propagating MJO-like signals, for such signals may lie outside the space spanned by the eigenfunctions used to construct r_t^{MJO} (see Sect. 5.1).

As a consistency test, we have repeated this analysis for forecasts initialized even earlier in the season, at $t =$ November 1. Using the same kernel K-means parameter as in the December 1 experiments ($\sigma = 1.58$), we obtained a partition which is broadly consistent with the structure described above. That is, there is a La Niña cluster ($k_t = 1$), an El Niño cluster ($k_t = 2$), and a cluster spanning mainly ENSO-neutral years with a few samples from ENSO-active years ($k_t = 3$). The cluster-conditional and expected predictability scores, displayed in Fig. 16, are also consistent with the December 1 results (Fig. 12) advanced by $\tau = 30$ d to take into account the difference in initialization times. Specifically, following an initial decay of predictability, the highest gain of information in Fig. 16 is observed at lead time $\tau = 30$ d (at calendar day December 1), and the reemergent peaks at lead times $\tau = 95$ d and 125 d (again, a shift of 30 days with respect to the previous case).

5.4 Regime predictability during the active-MJO season

We now assess the regime predictability of MJO for forecasts initialized during the active MJO season at $t =$ February 1. Following a similar approach as Sect. 5.3, we select a partition with $K = 3$ clusters and Gaussian bandwidth $\sigma = 3.16$. The resulting predictability scores, the cluster affiliation sequence, the cluster-conditional forecast PDFs, and representative marginal PDFs of the eigenfunctions are shown in Figs. 17, 18 and 19, respectively.

Inspecting the affiliation sequences in Fig. 18, it becomes apparent that one of the clusters ($k_t = 1$) persists throughout the interval $[t - \Delta T, t + \Delta t]$ used for clustering,

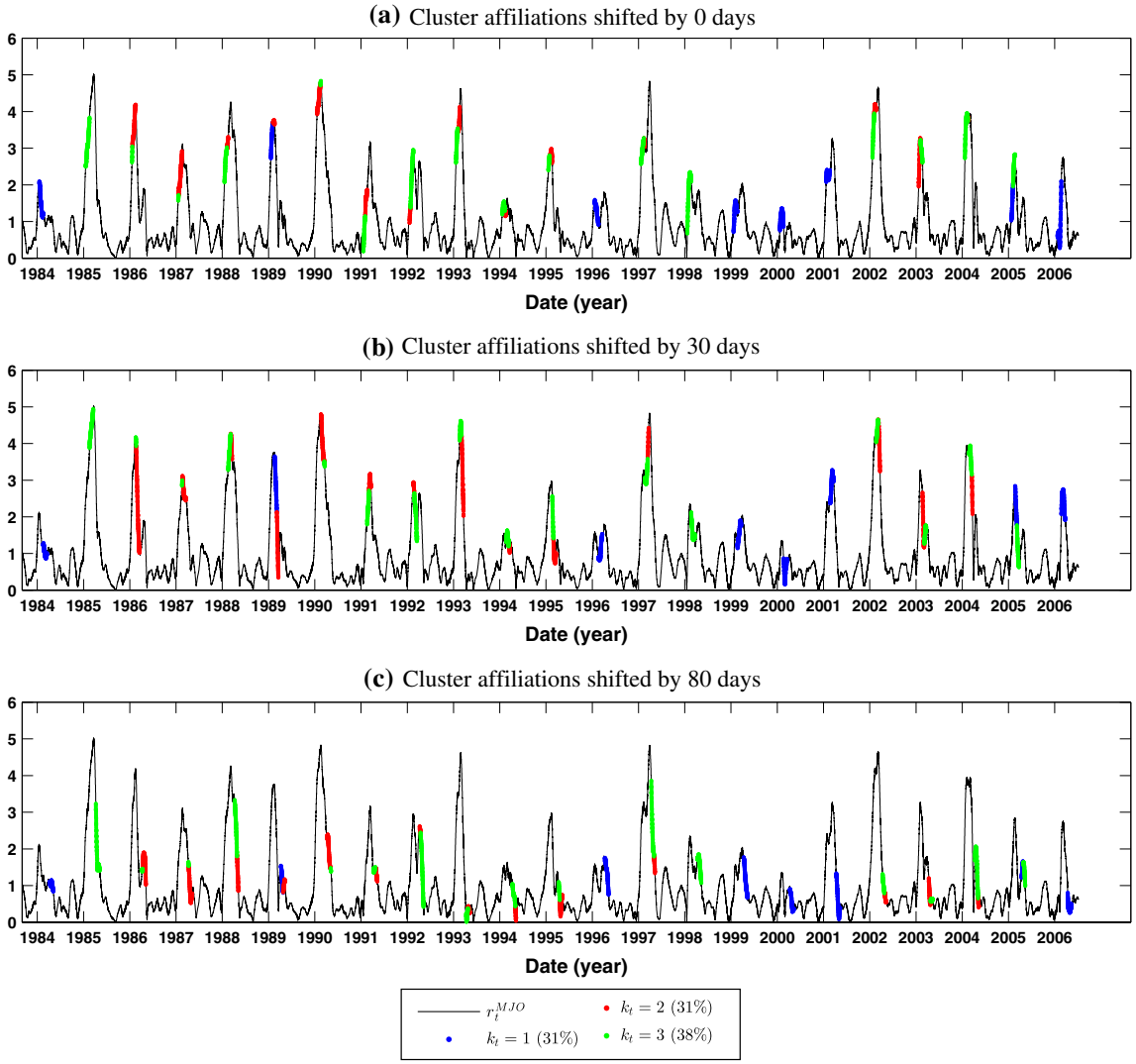


Fig. 18 The cluster affiliation sequence k_t for forecasts initialized at $t =$ February 1, visualized against the MJO index r_t^{MJO}. The La Niña cluster ($k_t = 1$) is in good agreement with the La Niña clusters found for $t =$ December 1 and November 1. The other two clusters are split on the MJO eigenfunction ϕ_{12}

whereas the affiliation to the other two clusters exhibits a switching behavior during that interval (cf. the affiliation sequences for December in Fig. 13, where all three clusters are persistent). Indeed, as is evident from the years during which cluster 1 is occupied (1984, 1989, 1996, 1999–2001, 2005–2006), as well as the cluster-conditional PDFs in Fig. 19, this cluster is mainly dominated by La Niña, and is in general agreement with the La Niña clusters identified in Sect. 5.3. According to the relative entropy results in Fig. 17, this cluster is the most predictable of the three. Intuitively, one would expect the conditional background environment during La Niña events to act detrimentally to large-scale organized convection, leading to statistically anomalous (hence predictable) MJO behavior. This picture is consistent with the conditional PDFs $p\left(r_{t+\tau}^{\text{MJO}} \mid k_t\right)$ in Fig. 19 for the La Niña cluster, which are markedly skewed towards low values of the MJO index, producing large values in relative entropy of the prior relative to these distributions.

Turning now to the remaining two clusters ($k_t = 2$ and 3), the conditional PDFs for the eigenfunctions in Fig. 20 indicate that these clusters are mainly occupied during the positive phase of the ENSO eigenfunction ϕ_5 (including El Niño events), but are generally not well distinguishable on the basis of interannual eigenfunctions. Instead, these clusters are mainly separated in terms of intraseasonal eigenfunctions, and in particular the MJO eigenfunctions themselves (see the PDF for ϕ_{12} in Fig. 20). While clusters 2 and 3 are both less predictable than cluster 1, an interesting asymmetry in predictability emerges, namely that for leads up

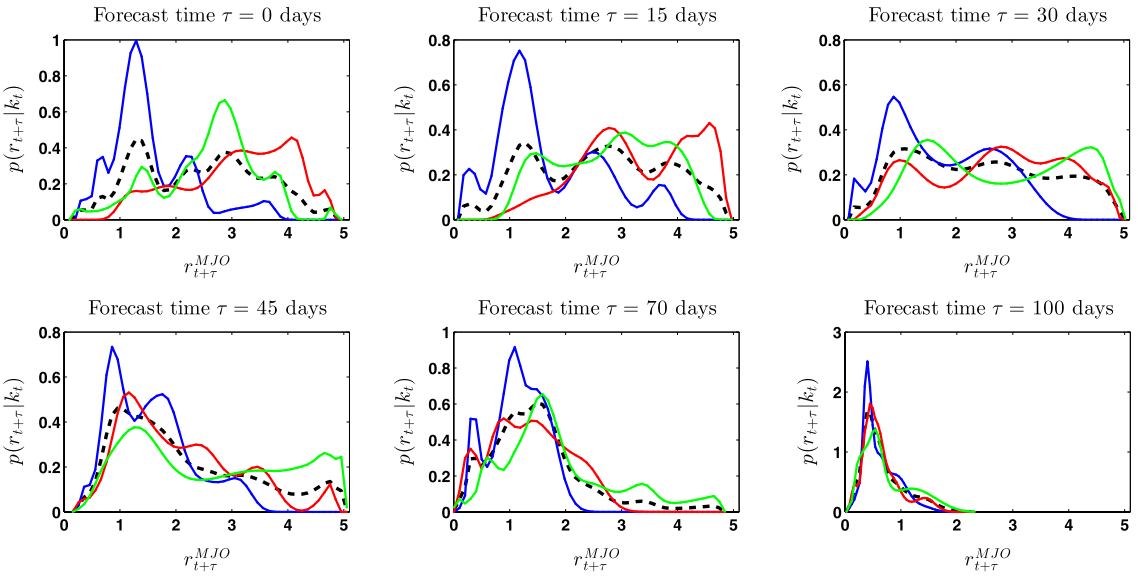


Fig. 19 Cluster-conditional PDFs $p\left(r^{\mathrm{MJO}}_{t+\tau} \mid k_t\right)$ for the MJO index for forecasts initialized at $t =$ February 1. The conditional PDFs for the La Niña cluster ($k_t = 1$) are markedly skewed towards low values of the MJO index, producing large values in the entropy of $p\left(r^{\mathrm{MJO}}_{t+\tau} \mid k_t = 1\right)$ relative to the prior $p(r^{\mathrm{MJO}}_{t+\tau})$

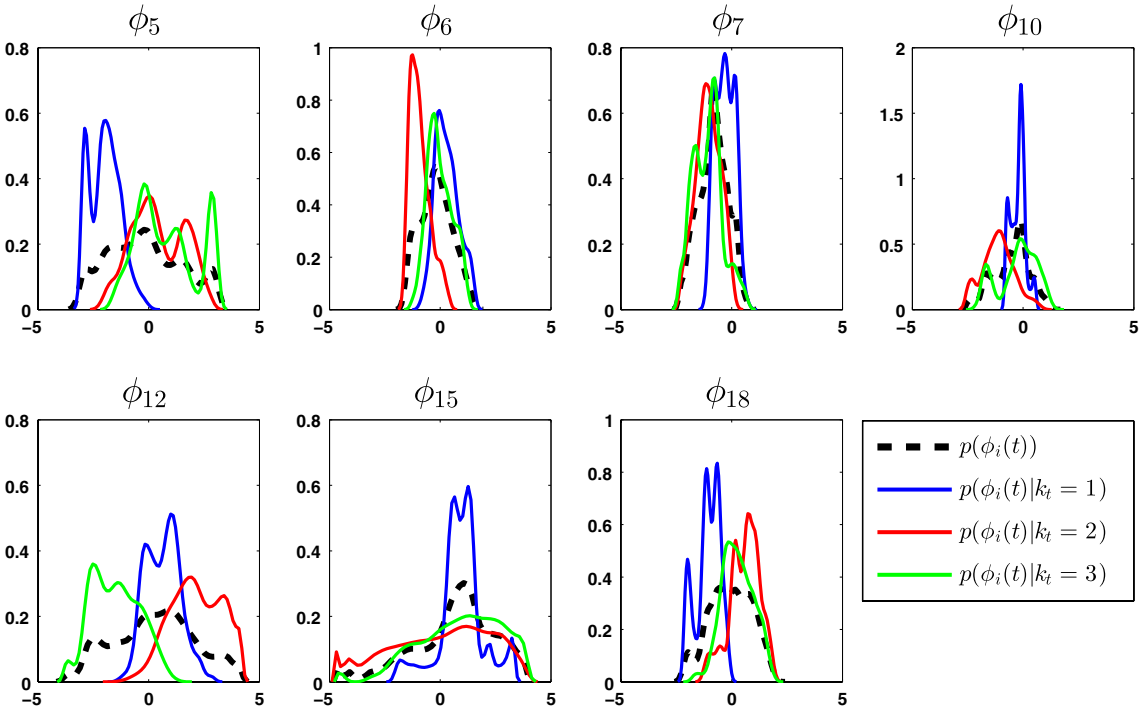


Fig. 20 Representative cluster-conditional marginal PDFs $p(\phi_i(t) \mid k_t)$ for the Laplace–Beltrami eigenfunctions used as predictors at initialization time $t =$ February 1. The La Niña cluster ($k_t = 1$) is negative on ϕ_5 and takes small values on the MJO eigenfunctions (ϕ_{12} and ϕ_{15}). Clusters 2 and 3 are mainly split with respect to the first MJO eigenfunction, ϕ_{12}

to $\tau \simeq 25$ d cluster 2 is more predictable than cluster 3. Inspecting the occupancy of these two clusters on the MJO eigenfunction space, $\{\phi_{12}, \phi_{15}\}$ (see Fig. 6), one finds that cluster 3 (negative ϕ_{12}) mainly consists of Phases 8 and 1–3, whereas cluster 2 (positive ϕ_{12}) occupies Phases 4–7. That is, according to the phase composites in Fig. 9, cluster 3

represents the initiation and development of the MJO over the Indian Ocean, and cluster 2 corresponds to its passage over the Maritime Continent and eventual decay at the dateline. When an MJO is in its mature or termination phase and forecasting is initialized in the midst of the active MJO season (in this case, February 1) there exist at least two probable scenarios for the future evolution of MJO activity, namely quiescence or initiation of a subsequent MJO in a wavetrain. On the other hand, when the MJO is in its initiation and development phase and ENSO is in a neutral or positive state, it is fairly likely that it will be able to propagate all the way to the Western Pacific. The observed asymmetry in cluster predictability is consistent with this basic intuition.

5.5 Comparison of early- and active-season predictability

A comparison of Figs. 12, 16, and 17 immediately reveals that the predictive information content in the partitions constructed for the early-season forecasts ($t =$ November 1 and December 1) differs qualitatively from the predictive information in forecasts initialized during the active season ($t =$ February 1). In particular, the early-season results exhibit an initial decay of predictability starting from large values of mutual information $\left[\mathscr{I}(r^{\mathrm{MJO}}_{t+\tau}, k_t) \simeq 15\right]$ and lasting for $\tau \sim 20$ d, which has no counterpart in the active-season result. Instead, the mutual information values for the $t =$ February 1 forecasts at $\tau \lesssim 20$ days leads are in the same range $\left[\mathscr{I}(r^{\mathrm{MJO}}_{t+\tau}, k_t) \simeq 2.5\right]$ as the predictability reemergence peaks occurring at later times ($\tau \simeq 60$ d) in the early-season experiments. In Sect. 5.3, we associated the initial large values of mutual information in the early-season forecasts to eigenfunctions with significant interannual variability, and in particular active ENSO states. Thus, our results suggest that the convective activity represented by T_b carries considerable predictive information in its interannual planetary-scale regimes. This information can be leveraged for statistical MJO forecasting initialized before the active-MJO period, but does not persist into the boreal winter months. During the active-MJO period, the dominant non-interannual large-scale predictor for MJO appears to be the state of the MJO itself. This is consistent with the stochastic modeling study of Chen et al. (2014), who found that the predictive skill of their nonlinear oscillator models is highest during strong-MJO winters.

In interpreting these observations, it is important to keep in mind certain aspects of the predictability analysis presented here. First, as already noted in Sect. 5.3, we are defining MJO activity through the NLSA-based r^{MJO}_t index. It is likely that there exist other MJO-like signals which are not represented by this index, but it appears that r^{MJO}_t is able to isolate an ISO with realistic MJO features, well-defined temporal structure (Chen et al. 2014), and non-trivial predictability conditioned on large-scale interannual convective regimes. Second, in the information-theoretic framework employed here predictability corresponds to deviations in the forecast distribution from climatology, and these deviations are not necessarily related to individual MJO initiation or termination events. In order to describe MJO initiation in the present framework, one could construct a binary random variable $\eta_\tau(t) = 1$ if $r^{\mathrm{MJO}}_t < 1$ and $r^{\mathrm{MJO}}_{t+\tau} \geq 1$ [and $\eta_\tau(t) = 0$ otherwise], and compute the mutual information $\mathscr{I}(\eta_\tau(t), k_t)$. It is possible that the predictability properties of $\eta_\tau(t)$ (and an analogous binary variable representing termination) are significantly different from the predictability of r^{MJO}_t. Third, in accordance with the data-processing inequality in (13), the lower predictability scores in the $t =$ February 1 forecasts conditioned on the coarse-grained initial data k_t does not imply absence of predictability conditioned on finer-scale initial data. Indeed, we saw in in Sect. 4.2.1 that the MJO mode recovered by NLSA exhibits anomalous convection signals over eastern Africa and eastern tropical Brazil (Phase 1 in Fig.9), so it is possible that the convective activity localized in those regions could provide higher predictive information than what is available in the large-scale regimes represented by k_t. In summary, our study firmly establishes the role of planetary-scale interannual convective regimes in early-season MJO forecasting with intraseasonal leads, but leaves open the possibility that finer-scale predictors are important during the active-MJO season.

6 Conclusions

In this paper, we have studied the dominant large-scale modes of organized tropical convection and the regime predictability of the MJO amplitude in satellite observations of infrared brightness temperature (T_b) over the tropical belt $15°\mathrm{S} - 15°\mathrm{N}$. In contrast to earlier studies, which rely on preprocessing steps such as spectral windowing, bandpass filtering, and seasonal partitioning to isolate the signals of interest, our objective has been to extract physically meaningful temporal and spatiotemporal modes of organized convection applying no preprocessing to the data. To that end, we used a nonlinear data analysis technique (NLSA; Giannakis and Majda 2012a, 2013, 2014) designed to extract intrinsic timescales from high-dimensional data generated by dynamical systems. The key difference of NLSA from classical eigendecomposition techniques such as SSA, EEOFs, and related algorithms is that the covariance matrix is replaced by a discrete Laplace–Beltrami operator on the nonlinear data manifold. This operator is computed empirically from data using kernel methods, and its eigenfunctions (which can be thought as

nonlinear analogs to PCs) provide natural basis functions for dimension reduction and feature extraction tailored to the manifold structure of the data. In particular, by using a kernel defined in Takens embedding space, the Laplace–Beltrami eigenfunctions from NLSA are predisposed to reveal intrinsic timescales of the dynamical system generating the data.

The hierarchy of modes extracted from the T_b data via NLSA spans interannual timescales, the annual cycle and its harmonics, ISOs, and diurnal modes. A major result of this analysis has been that the eastward-propagating boreal-winter MJO and the poleward-propagating BSISO emerge naturally as distinct families of modes characterized by strong intermittency and seasonality. These modes were used here to create a bimodal ISO index with significantly higher discriminating power than what is possible through conventional linear approaches. Moreover, the MJO and BSISO patterns from NLSA are non-orthogonal in space, and exhibit finer-detail structures than their SSA-based counterparts, including a signal of enhanced convection over eastern Africa and eastern tropical Brazil at the initiation phase of the MJO.

Because no preprocessing has been applied to the data, the NLSA modes provide an objective framework to study MJO-BSISO interactions, as well as interactions of ISOs with other phenomena such as ENSO and the diurnal cycle. Notably, the NLSA spectrum contains families of diurnal modes active over Africa, the Maritime Continent, and South America with envelopes modulated by MJO and BSISO. These modes should provide a useful basis to study ISO-diurnal cycle interactions. Intriguingly, we observed that the initiation and termination of the active-MJO period in a given year (as measured by the corresponding NLSA index) correlates strongly with a quasi-periodic mode at the third harmonic of the annual cycle featuring a south-eastward-propagating anomaly over the Indian Ocean. In other work (Chen et al. 2014), the well-defined temporal structure of the MJO eigenfunctions was exploited to construct nonlinear stochastic oscillator models with MJO forecast skill extending to 40-day leads, and low parametric sensitivity.

Empirical MJO forecasting was a major objective in this work too, and was approached here from the standpoint of statistical initial-value predictability conditioned on coarse-grained initial data (regimes). Specifically, we constructed a space of predictor variables representing the state of large-scale convective organization through the leading-few eigenfunctions from NLSA, and partitioned that space into a discrete set of regimes using the kernel K-means clustering algorithm. Following Giannakis and Majda (2012c, d) and Giannakis et al. (2012b), we assessed MJO predictability by measuring the information gain (relative entropy) of the empirical forecast distribution of the NLSA MJO index conditioned on the membership of the initial data to the clusters in the partition.

In our experiments, the predictive information content for MJO in large-scale convective regimes was especially high for forecasts initialized before the active-MJO season, as early as November 1. The main contributor to the increased predictability was the interannual component of T_b variability, and in particular significant El Niño and La Niña events. The occurrence of such events alters significantly the statistical properties of the NLSA-derived MJO index, and as a result its information-theoretic predictability persists over intraseasonal-scale leads (60–80 days). Experiments initialized during the active MJO period (initialization time at February 1) generally yielded weaker MJO predictability conditioned on coarse-grained T_b regimes than the early-season experiments. In this case, predictability was governed by both ENSO as well as the current state of the MJO and extended up to $\simeq 40$ d leads. This result is consistent with Chen et al. (2014), who found similar predictability limits for MJO forecasts with their stochastic model initialized in strong MJO winters.

The analysis presented here was performed using only infrared brightness temperature data. However, additional information, such as lower- and upper-level zonal winds carry important information beyond the pure T_b data (Wheeler and Hendon 2004), and will be incorporated in future work. Also, the emergence of a clear northeastward-propagating BSISO mode in the domain $15°\text{S} - 15°\text{N}$ with relatively limited northward extent motivates an extension of the domain further to the north, enabling the study of interactions with the Indian Monsoon. Finally, methods for computing the eigenfunction values (and hence our ISO indices) from newly acquired data can be used for real-time monitoring and also in nonparametric forecasting applications (Zhao and Giannakis 2015). We plan to study these topics in future work.

Acknowledgments The research of Andrew J. Majda and Dimitrios Giannakis is partially supported by ONR MURI grant 25-74200-F7112. Eniko Székely is supported as a postdoctoral fellow through this grant. The authors wish to thank Wen-wen Tung for stimulating discussions.

References

Aubry N, Guyonnet R, Lima R (1991) Spatiotemporal analysis of complex signals: theory and applications. J Stat Phys 64:683–739

Aubry N, Lian W-Y, Titi ES (1993) Preserving symmetries in the proper orthogonal decomposition. SIAM J Sci Comput 14:483–505

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

Berry T, Sauer T (2014) Local kernels and the geometric structure of data. J Appl Comput Harmon. Anal (submitted)

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

Bowman A, Azzalini A (1997) Applied smoothing techniques for data analysis. Oxford University Press, Oxford

Broomhead DS, King GP (1986) Extracting qualitative dynamics from experimental data. Phys D 20(2–3):217–236

Bushuk M, Giannakis D, Majda AJ (2014) Reemergence mechanisms for North Pacific sea ice revealed through nonlinear Laplacian spectral analysis. J Clim 27:6265–6287

Chelliah M, Arkin P (1992) Large-scale interannual variability of monthly outgoing longwave radiation anomalies over the global tropics. J Climate 5(4):371–389

Chen N, Majda AJ, Giannakis D (2014) Predicting the cloud patterns of the Madden–Julian Oscillation through a low-order nonlinear stochastic model. Geophys Res Lett 41(15):5612–5619

Chen SS, Houze RA (1997) Diurnal variation and life-cycle of deep convective systems over the tropical Pacific warm pool. Q J Roy Meteorol Soc 123(538):357–388

Chiodi AM, Harrison DE (2010) Characterizing warm-ENSO variability in the equatorial pacific: an OLR perspective. J Climate 23:2428–2439

Christiansen B (2007) Atmospheric circulation regimes: can cluster analysis provide the number? J Climate 20(10):2229–2250

Coifman RR, Lafon S (2006) Diffusion maps. Appl Comput Harmon Anal 21:5–30

Cover TA, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley-Interscience, Hoboken

Crommelin DT, Majda AJ (2004) Strategies for model reduction: comparing different optimal bases. J Atmos Sci 61:2206–2217

DelSole T (2004) Predictability and information theory. Part I: measures of predictability. J Atmos Sci 61(20):2425–2440

Dhillon IS, Guan Y, Kulis B (2004) Kernel k-means, spectral clustering and normalized cuts. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '04, ACM, New York, pp 551–556

Franzke C, Horenko I, Majda AJ, Klein R (2009) Systematic metastable regime identification in an AGCM. J Atmos Sci 66(9):1997–2012

Ghil M et al (2002) Advanced spectral methods for climatic time series. Rev Geophys 40:1-1–1-41

Giannakis D (2015) Dynamics-adapted cone kernels. SIAM J Appl Dyn Sys 14(2):556–608

Giannakis D, Majda AJ (2012a) Limits of predictability in the North Pacific sector of a comprehensive climate model. Geophys Res Lett 39:L24602

Giannakis D, Majda AJ (2012b) Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. Proc Natl Acad Sci 109(7):2222–2227

Giannakis D, Majda AJ (2012c) Quantifying the predictive skill in long-range forecasting. Part I: Coarse-grained predictions in a simple ocean model. J Climate 25:1793–1813

Giannakis D, Majda AJ (2012d) 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

Giannakis D, Majda AJ (2013) Nonlinear Laplacian spectral analysis: capturing intermittent and low-frequency spatiotemporal patterns in high-dimensional data. Stat Anal Data Min 6(3):180–194

Giannakis D, Majda AJ (2014) Data-driven methods for dynamical systems: quantifying predictability and extracting spatiotemporal patterns. In: Melnik R (ed) Mathematical and computational modeling: with applications in engineering and the natural and social sciences. Wiley, Hoboken, p 288

Giannakis D, Tung W-W, Majda AJ (2012) 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). Boulder, Colorado, pp 55–62

Giannakis D, Majda AJ, Horenko I (2012b) Information theory, model error, and predictive skill of stochastic models for complex nonlinear systems. Phys D 241:1735–1752

Groth A, Ghil M (2011) Multivariate singular spectrum analysis and the road to phase synchronization. Phys Rev E 84:036206

Hein M, Audibert J-Y, von Luxburg U (2005) From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians. In: Learning theory, volume 3559 of Lecture notes in computer science. Springer, Berlin, pp 470–485

Hendon HH, Wheeler MC, Zhang C (2007) Seasonal dependence of the MJO–ENSO relationship. J Climate 20:531–543

Hodges K, Chappell D, Robinson G, Yang G (2000) An improved algorithm for generating global window brightness temperatures from multiple satellite infra-red imagery. J Atmos Oceanic Technol 17:1296–1312

Horel JD (1981) A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500 mb height field. Mon Weather Rev 109:2080–2092

Horenko I (2010) On identification of nonstationary factor models and their application to atmospheric data analysis. J Atmos Sci 67(5):1559–1574

Hung M-P, Lin J-L, Wang W, Kim D, Shinoda D, Weaver SJ (2013) MJO and convectively coupled equatorial waves simulated by CMIP5 climate models. J Climate 26:6185–6214

Ichikawa H, Yasunari T (2008) Intraseasonal variability in diurnal rainfall over New Guinea and the surrounding oceans during austral summer. J Climate 21:2852–2868

Jones PW, Maggioni M, Schul R (2008) Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels. Proc Natl Acad Sci 105:1803

Kessler WS (2001) EOF representations of the Madden–Julian oscillation and its connection with ENSO. J Climate 14:3055–3061

Kikuchi K, Wang B (2010) Spatiotemporal wavelet transform and the multiscale behavior of the Madden–Julian oscillation. J Climate 23:3814–3834

Kikuchi K, Wang B, Kajikawa Y (2012) Bimodal representation of the tropical intraseasonal oscillation. Climate Dyn 38:1989–2000

Kiladis GN, Straub KH, T HP (2005) Zonal and vertical structure of the Madden–Julian oscillation. J Atmos Sci 62:2790–2809

Kiladis GN, Dias J, Straub KH, Wheeler MC, Tulich SN, Kikuchi K, Weickmann KM, Ventrice MJ (2014) A comparison of OLR and circulation-based indices for tracking the MJO. Mon Weather Rev 142:1697–1715

Kleeman R (2002) Measuring dynamical prediction utility using relative entropy. J Atmos Sci 59(13):2057–2072

Lau K, Chan P (1985) Aspects of the 40–50 day oscillation during the northern winter as inferred from outgoing longwave radiation. Mon Weather Rev 113:1889–1909

Lau WKM (2011) El Niño Southern oscillation connection. In: Intraseasonal variability in the atmosphere-ocean climate system. Springer, pp 297–334

Lau WKM, Waliser DE (2011) Intraseasonal variability in the atmosphere-ocean climate system. Springer, Berlin

Leung L-Y, North GR (1990) Information theory and climate prediction. J Climate 3(1):5–14

Lin X, Johnson RH (1996a) Heating, moistening, and rainfall over the western pacific warm pool during TOGA COARE. J Atmos Sci 53(22):3367–3383

Lin X, Johnson RH (1996b) Kinematic and thermodynamic characteristics of the flow over the western pacific warm pool during TOGA COARE. J Atmos Sci 53(5):695–715

Lo F, Hendon H (2000) Empirical extended-range prediction of the Madden–Julian oscillation. Mon Weather Rev 128:2528–2543

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 281–297

Madden RA, Julian PR (1971) Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. J Atmos Sci 28(5):702–708

Madden RA, Julian PR (1972) Description of global-scale circulation cells in the tropics with a 40–50 day period. J Atmos Sci 29(6):1109–1123

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

Maloney ED, Hartmann DL (1998) Frictional moisture convergence in a composite life cycle of the Madden–Julian oscillation. J Climate 11:2387–2403

Matsuno T (1966) Quasi-geostrophic motions in the equatorial area. J Meteorol Soc Jpn 44:25–43

Metzner P, Putzig L, Horenko I (2012) Analysis of persistent nonstationary time series and applications. Comm App Math Comp Sci 7:175–229

Portegies J (2014) Embeddings of Riemannian manifolds with heat kernels and eigenfunctions. arXiv:1311.7568

Roundy PE, Schreck CJI (2009) A combined wavenumberfrequency and time-extended EOF approach for tracking the progress of modes of large-scale organized tropical convection. Q J R Meteorol Soc 135:161–173

Sauer T, Yorke JA, Casdagli M (1991) Embedology. J Stat Phys 65(3–4):579–616

Schölkopf B, Smola A, Müller K (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10:1299–1319

Singer A (2006) From graph to manifold Laplacian: the convergence rate. J Appl Comput Harmon Anal 21:128–134

Straub KH (2013) MJO initiation in the real-time mutlivariate MJO index. J Climate 26:1130–1151

Takens F (1981) Detecting strange attractors in turbulence. Dynamical systems and Turbulence, Warwick 1980, volume 898 of lecture notes in mathematics. Springer, Berlin, pp 366–381

Thomson DJ (1982) Spectrum estimation and harmonic analysis. Proc IEEE 70:1055–1096

Tian B, Waliser DE, Fetzer EJ (2006) Modulation of the diurnal cycle of tropical deep convective clouds by the MJO. Geophys Res Lett 33(20):1–6

Tung W-W, Lin C, Chen B, Yanai M, Arakawa A (1999) Basic modes of cumulus heating and drying observed during TOGA-COARE IOP. Geophys Res Lett 26(20):3117–3120

Tung W-W, Giannakis D, Majda AJ (2014a) Symmetric and antisymmetric signals in MJO deep convection. Part I: basic modes in infrared brightness temperature. J Atmos Sci 71:3302–3326

Tung W-W, Giannakis D, Majda AJ (2014b) Symmetric and antisymmetric signals in MJO deep convection. Part II: kinematics and thermodynamics. J Atmos Sci (in Revision)

Ventrice M, Wheeler M, Hendon H, Schreck C, Thorncroft C, Kiladis G (2013) A modified multivariate Madden–Julian oscillation index using velocity potential. Mon Weather Rev 141(12):4197–4210

Waliser D (2011) Predictability and forecasting. In: Lau WK, Waliser DE (eds) Intraseasonal variability in the atmosphere-ocean climate system. Springer, Berlin, pp 433–468

Wang B, Rui H (1990) Synoptic climatology of transient tropical intraseasonal convection a nomalies: 1975–1985. Meteorol Atmos Phys 44:43–61

Wheeler M, Kiladis GN (1999) Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber-frequency domain. J Atmos Sci 56(3):374–399

Wheeler MC, Hendon HH (2004) An all-season real-time multivariate MJO index: development of an index for monitoring and prediction. Mon Weather Rev 132(8):1917–1932

Wolter K, Timlin M (1998) Measuring the strength of ENSO events: how does 1997/98 rank? Weather 53(9):315–324

Wolter K, Timlin M (2011) El Niño/Southern oscillation behaviour since 1871 as diagnosed in an extended multivariate enso index (MEI.ext). Int J Climatol 31(7):1074–1087

Yanai M, Chen B, Tung W-W (2000) The Madden–Julian oscillation observed during the TOGA COARE IOP: global view. J Atmos Sci 57(15):2374–2396

Zhang C, Dong M (2004) Seasonality in the Madden–Julian oscillation. J Climate 17:3169–3180

Zhao Z, Giannakis D (2015) Analog forecasting with dynamics-adapted kernels. Nonlinearity (submitted). arXiv:1412.3831