Predicting Monsoon Intraseasonal Precipitation using a Low-Order Nonlinear Stochastic Model

Nan Chen*

Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, USA.

Andrew J. Majda

Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, USA,

Center for Prototype Climate Modeling, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, UAE

C. T. Sabeerali

Center for Prototype Climate Modeling, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, UAE

Ajayamohan S. Ravindran

Center for Prototype Climate Modeling, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, UAE

*Corresponding author address: Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA.

E-mail: chennan@cims.nyu.edu

ABSTRACT

We assess the predictability of the Monsoon Intraseasonal Oscillations (MISOs) as measured by precipitation. A recent advanced nonlinear time series technique, Nonlinear Laplacian Spectral Analysis, is applied directly to the daily rainfall data without any preliminary detrending or spatiotemporal filtering to define two spatial modes associated with the MISO. The time series of these two modes are highly intermittent with large variation in amplitude from year to year in the boreal summer season. Then a recent systematic strategy for data driven physics constrained low-order stochastic modeling is applied to these time series. The result is a four dimensional nonlinear stochastic model for the two observed MISO variables as well as two hidden variables involving correlated multiplicative noise defined through energy conserving nonlinear interaction. Systematic calibration and prediction experiments with the nonlinear stochastic model show that the precipitation MISO indices can be skillfully predicted 20 to 50 days in advance and the ensemble spread in the forecast model is an accurate indicator of forecast uncertainty at long lead times. Then an effective and practical spatiotemporal reconstruction algorithm is developed, which shows the predicted spatiotemporal patterns have comparable skill as the MISO indices. It is also found that a 3-year short training period is sufficient for the model to describe the essential characteristics of the MISO and retain skillful predictions. In addition, outgoing longwave radiation is shown to be a good proxy for monsoon intraseasonal precipitation and the lagged embedding window size is crucial to reaching unbiased MISO indices.

1. Introduction

Monsoon Intraseasonal Oscillation (MISO) (Kikuchi et al. 2012; Lee et al. 2013; Sikka and Gadgil 1980; Goswami and Mohan 2001; Lau and Waliser 2011; Webster et al. 1998) is one of the prominent modes of tropical intraseasonal variability. As a slow moving planetary scale envelope of convection propagating northeastward, it is strongly associated with the boreal summer monsoon rainfall over south Asia. Due to the interaction with the mean monsoon circulation and other modes of tropical variability, the propagating characteristics of the MISO are more complex compared with the eastward-propagating Madden-Julian Oscillation (MJO) (Zhang 2005). The MISO plays an important role in determining the onset and demise of the Indian summer monsoon as well as affecting the rainfall over the Indian subcontinent (Murakami et al. 1986; Goswami and Mohan 2001; Goswami et al. 2003; Gadgil 2003). Therefore, the extended range prediction of MISO phases and real-time monitoring of the MISO have large societal impacts (Sahai et al. 2013; Abhilash et al. 2014a).

Several indices have been proposed for real-time monitoring and forecast of the MISO. The Indian Institute of Tropical Meteorology (IITM) relies on an index based on extended empirical orthogonal function (EEOF) analysis on longitudinal averaged daily rainfall anomalies for extended range prediction of MISO (Suhas et al. 2013; Sahai et al. 2013; Abhilash et al. 2014a). Another well-known MISO index (Lee et al. 2013) mimics that for the real-time multivariate MJO (RMM) index (Wheeler and Hendon 2004), and is based on the multivariate EOF analysis of daily anomalies of the zonal wind at 850h Pa and outgoing long-wave radiation (OLR). Other MISO indices (Kikuchi et al. 2012; Goswami et al. 1999) are based on similar EOF and EEOF techniques. These covariance-based approaches in general capture the spatiotemporal MISO patterns reasonably well and isolate the northeastward-propagating intraseasonal periodicity band from high-frequency band (Suhas et al. 2013; Abhilash et al. 2014a,b). Yet, the seasonal extraction and longitudinal averaging in computing these indices are sometimes ad hoc and can potentially lead to loss of predictive information or mixing with other modes. In addition, these covariance-based techniques have potential inadequacy in capturing the rare/extreme events in complex nonlinear dynamics (Crommelin and Majda 2004) which have significant societal and economic impacts.

Recently (Sabeerali et al. 2017), a new MISO index based on the Nonlinear Laplacian Spectral Analysis (NLSA) (Giannakis and Majda 2012b,a) technique was developed. NLSA is a nonlinear data analysis technique that combines ideas from lagged embedding (Packard et al. 1980; Sauer et al. 1991), machine learning (Coifman and Lafon 2006; Belkin and Niyogi 2003), adaptive weights and spectral entropy criteria to extract spatiotemporal modes of variability from high-dimensional time series. These modes are computed utilizing the eigenfunctions of a discrete Laplace-Beltrami operator, which can be thought of as a local analog of the temporal covariance matrix employed in EOF and EEOF techniques, but adapted to the nonlinear geometry of data generated by complex dynamical systems. A key advantage of NLSA over classical covariance-based techniques is that NLSA by design requires no ad hoc detrending or spatiotemporal filtering of the full data set and captures both intermittency and low frequency variability (Giannakis and Majda 2012a,b, 2013; Giannakis et al. 2012). Therefore, the NLSA-based MISO index provides an objective identification of the MISO patterns in noisy precipitation data. In addition, as reported in Sabeerali et al. (2017), the NLSA MISO modes have higher memory and predictability, stronger amplitude and higher fractional explained variance over the western Pacific, Western Ghats, and adjoining Arabian Sea regions, and more realistic representation of the regional heat sources over the Indian and Pacific Oceans compared with those extracted via EEOF analysis. Other applications of NLSA beyond the capability of EOF and EEOF in capturing both the intermittent and low-frequent modes in climate, atmosphere and ocean can be found in Székely et al. (2016a,b); Slawinska and Giannakis (2016); Giannakis and Majda (2012a, 2011); Brenowitz et al. (2016).

In this article, we assess the predictability of the MISO as measured through precipitation. A recent systematic strategy for data driven physics constrained low-order stochastic modeling of time series (Majda and Harlim 2013; Harlim et al. 2014) is applied to the two-dimensional MISO indices from NLSA (Sabeerali et al. 2017). The result is a four dimensional nonlinear stochastic model for the two MISO variables and two hidden variables. This low-order model involves correlated multiplicative noise defined through energy conserving nonlinear interactions between the observed and hidden variables as well as additive stochastic noise. The special structure of this nonlinear stochastic model allows effective data assimilation algorithm for determining the initial ensemble of the hidden variables that facilitates the ensemble prediction scheme. This nonlinear low-order stochastic model has been shown to have significant skill for determining the predictability limits of the large-scale cloud patterns of both the boreal winter MJO and boreal summer intraseasonal oscillations (Chen et al. 2014; Chen and Majda 2015a) as well as improving the prediction skill of the RMM indices (Chen and Majda 2015b). Then with the predicted MISO indices in hand, an effective and practical spatiotemporal reconstruction algorithm is developed, which overcomes the fundamental difficulty in most data decomposition techniques with lagged embedding that require extra information beyond the predicted time series in the future.

The remainder of this article is organized as follows. Section 2 describes the precipitation dataset and the MISO indices obtained from the NLSA technique. Section 3 presents the physics constrained low-order nonlinear stochastic model as well as the calibration and the effective prediction algorithm. The results of predicting the MISO indices are reported in Section 4 and the prediction of the spatiotemporal reconstructed patterns are shown in Section 5. Section 6 discusses the possibility of shortening the training period to only 3 years and then illustrates the discrepancy in forming the MISO indices with different lagged embedding sizes. Summary conclusions are included in Section 7.

2. The Precipitation MISO Indices from NLSA

The dataset utilized here is the daily Global Precipitation Climatology Project (GPCP) rainfall data (Huffman et al. 2001) over the Asian summer monsoon region (20^oS-30^oN, 30^oE-140^oE) for period 1997-2014. The spatial resolution of this dataset is $1^o \times 1^o$, amounting to $d = 5500$ grid points for the Asian summer monsoon region.

NLSA is applied to the daily GPCP dataset with a lagged embedding window of $q = 64$ days, an ideal choice for the intraseasonal time scale. A variety of extended spatial precipitation patterns emerge from the analysis but the focus here is on the two spatial patterns associated with MISO with time series depicted in Figure 1. The details of applying NLSA to daily GPCP dataset have already been described in Sabeerali et al. (2017) and are thus omitted here. It is evident from Figure 1 that these patterns are active in boreal summer and quiescent in boreal winter. It was shown in Sabeerali et al. (2017) that the NLSA MISO modes display the characteristic pattern of northeastward propagating anomalies associated with the MISO. A case study there also revealed three consecutive MISO events in the NLSA MISO modes in the boreal summer of 2004, the onset and demise phases of which are highly consistent with observations. These facts indicate that the time series depicted in Figure 1 give a reasonable representation of the full life cycle of the northward propagating boreal summer convection band and can be utilized to determine the phase and amplitude of the poleward-propagating rainfall anomalies associated with the MISO. Below, we utilize the terminology, MISO indices, for the two time series in Figure 1.

3. The Low-Order Nonlinear Stochastic Model

Denote by u_1 and u_2 the two components, MISO 1 and MSIO 2, depicted in Figure 1. The probability distribution functions (PDFs) for u_1 and u_2 are highly non-Gaussian with fat tails that indicate the temporal intermittency in the large scale precipitation patterns associated with the MSIO. The following family of low-order stochastic models are proposed to describe the intermittent variability of the time series u_1 and u_2:

$$\frac{du_1}{dt} = (-d_u u_1 + \gamma(v + v_f(t))u_1 - (a + \omega_u)u_2) + \sigma_u \dot{W}_{u_1}, \tag{1a}$$

$$\frac{du_2}{dt} = (-d_u u_2 + \gamma(v + v_f(t))u_2 + (a + \omega_u)u_1) + \sigma_u \dot{W}_{u_2}, \tag{1b}$$

$$\frac{dv}{dt} = (-d_v v - \gamma(u_1^2 + u_2^2)) + \sigma_v \dot{W}_v, \tag{1c}$$

$$\frac{d\omega_u}{dt} = (-d_\omega \omega_u) + \sigma_\omega \dot{W}_\omega, \tag{1d}$$

where

$$v_f(t) = f_0 + f_t \sin(\omega_f t + \phi). \tag{2}$$

In (1), u_1 and u_2 are the two observed MISO variables while v and ω_u are hidden unobserved variables which represent the stochastic damping and stochastic phase, respectively. In (1), $\dot{W}_{u_1}, \dot{W}_{u_2}, \dot{W}_v$ and $\dot{W}_\omega$ are independent white noise. The time periodic damping in the equations in (1a) and (1b) is utilized to crudely model the active summer season and the quiescent winter season in the seasonal cycle. The hidden variables v, ω_u interact with the observed MISO variables u_1, u_2 through energy conserving nonlinear interactions following the systematic physics constrained nonlinear regression strategies for time series developed recently (Majda and Harlim 2013; Harlim et al. 2014). The energy-conserving nonlinearity is easily seen by multiplying (1a)–(1d) by u_1, u_2, v and ω_u, respectively, and then these equations sum up. The energy change in the quadratic nonlinear terms cancels with each other and thus the energy due to the nonlinear interaction is conserved. The low-order stochastic nonlinear models in (1) are fundamentally different from those utilized earlier (Kondrashov et al. 2013; Kravtsov et al. 2005) which allow for nonlinear interactions only between the observed variables u_1, u_2 and only special linear interactions with layers of hidden variables. The physics constrained nonlinear low-order stochastic model (1)–(2) has been shown to have significant skill for determining the predictability limits of the large-scale cloud patterns of both the boreal winter MJO and boreal summer intraseasonal oscillations (Chen et al. 2014; Chen and Majda 2015a) as well as improving the prediction skill of the RMM indices by incorporating a new information-theoretic strategy in the training phase (Chen and Majda 2015b).

a. Calibration of the Nonlinear Stochastic Model

The parameters of the stochastic model in (1)–(2) are calibrated by fitting the highly non-Gaussian PDFs and autocorrelations of the two MISO variables u_1, u_2 in the training period from 1998 to 2007 as shown in Figure 1. Table 1 records the optimal parameter values while Figure 2 displays the skill of the stochastic model with these parameters in recovering the statistics of the two MISO indices. Panels (a) and (b) show that the stochastic model succeeds in capturing the autocorrelations almost perfectly for a three-month duration and even the wiggles that appear with lags around one year. Panel (c) shows that the stochastic model captures the highly non-Gaussian fat-tailed PDF of the two MISO indices due to intermittency. Panel (d) shows that the power spectrum of the two MISO indices from the data and those from the stochastic model match very well. The optimal parameters in the stochastic model from Table 1 have been determined by systematically minimizing the information distance of the equilibrium PDF of the stochastic model compared with that of the actual data (Majda and Gershgorin 2010, 2011). Details are presented in the Appendix A. Importantly, the model statistics are robust with respect to the parameter variations around these optimal values (See Appendix A).

b. Prediction Algorithm and Data Assimilation of the Hidden Variables

The ensemble prediction algorithm is applied to the nonlinear low-order stochastic model (1) for predicting the MISO time series. The algorithm involves running the forecast model (1) forward in time given the initial values. The initial data of the two state variables $\mathbf{U} = (u_1, u_2)$ are obtained directly from the observations, i.e., MISO 1 and MISO 2 indices. The more important and challenging issue is to determine the initial ensemble of the two hidden variables $\Gamma = (v, \omega_u)$. To this end, an active data assimilation algorithm is incorporated into the ensemble forecasting scheme.

The estimates of the hidden parameters $\Gamma = (v, \omega_u)$ during the training period and initialization of these parameters during the prediction phase exploit the special structure of the nonlinear low-order stochastic model (1). The equations in (1) are a conditional Gaussian system with respect to the observations $\mathbf{U} = (u_1, u_2)$, meaning that once u_1 and u_2 are given the time evolution of the distributions of $\Gamma = (v, \omega_u)$ is Gaussian. Such special feature of (1) allows the closed analytic equations for the conditional Gaussian distributions of the hidden parameters $\Gamma = (v, \omega_u)$ obtained from the posterior estimations in the Bayesian framework (Liptser and Shiryaev 2001). Appendix B contains the details and explicit equations. We utilize this fact to construct an initial ensemble for forecasting at each time instant in both the training and prediction phases for $t \in [t_0, t_1, \ldots, t_s]$ in the following way.

1. Starting from a "burn in" time t_- earlier than t_0 with arbitrary initial conditions for Γ, solve the associated analytic formula (B2) until time t_0 to obtain the conditional Gaussian distribution $p_0(\Gamma|\mathbf{U}(t_0))$. The initial ensemble of the hidden variables $\Gamma = (v, \omega_u)$ for prediction starting from t_0 is drawn from this distribution.

2. The initial ensemble for prediction starting from the next time t_1 is drawn from $p_1(\Gamma|\mathbf{U}(t_1))$, where $p_1(\Gamma|\mathbf{U}(t_1))$ is solved by running the analytic formula (B2) forward from time t_0 to t_1 with initial value $p_0(\Gamma|\mathbf{U}(t_0))$.

3. Following the same procedure, the initial distributions of the hidden variables $\Gamma = (v, \omega_u)$ for prediction starting from each time t_i are obtained "on the fly" by running the analytic formula (B2) forward from time t_{i-1} to t_i with initial value $p_{i-1}(\Gamma|\mathbf{U}(t_{i-1}))$ when the new observations up to $\mathbf{U}(t_i)$ are available.

In the prediction below with (1), we use N ensemble members with $N = 50$.

4. Results of Predicting the MISO indices

With the optimal parameters from Table 1 and the ensemble initialization scheme described in Section 3b, the prediction skill of the stochastic model in (1) for the six year prediction period from year 2008 to 2013 is presented here. The skill scores of ensemble mean prediction as a function of lead time (days) in different years are shown in Figure 3 and the comparison of the ensemble mean prediction and the truth at lead times of 15 and 25 days for MISO 1 index for all six years are shown in Figure 4. Here, the skill scores adopted are the root-mean-squared (RMS) error and pattern correlation (Corr):

$$
\begin{aligned}
\text{RMS error}(\mathbf{U}_t, \mathbf{U}_t^{pred}) &= \sqrt{\frac{\sum_{t=1}^{n}\left((u_{1,t} - u_{1,t}^{pred})^2 + (u_{2,t} - u_{2,t}^{pred})^2\right)}{n}},\\
\text{Corr}(\mathbf{U}_t, \mathbf{U}_t^{pred}) &= \frac{\sum_{t=1}^{n}\left(u_{1,t}u_{1,t}^{pred} + u_{2,t}u_{2,t}^{pred}\right)}{\sqrt{\sum_{t=1}^{n}\left(u_{1,t}^2 + u_{2,t}^2\right)}\sqrt{\sum_{t=1}^{n}\left((u_{1,t}^{pred})^2 + (u_{2,t}^{pred})^2\right)}},
\end{aligned}
$$

where n is the number of the points in the time series, and $\mathbf{U}_t = (u_{1,t}, u_{2,t})$ and $\mathbf{U}_t^{pred} = (u_{1,t}^{pred}, u_{2,t}^{pred})$ are the truth and predicted time series, respectively. It is shown in Figures 3 and 4 that the 15 day predictions are very skillful and even the 25 day predictions have highly significant skill in most years. Among different years, year 2010 has useful predictions for about 20 days while year 2011 and 2013 have skillful predictions around 25-30 days. In year 2008, 2009 and 2012, there is a significant prediction skill out to more than 50 days. Here, useful predictions are defined by 1) the RMS error in the prediction is less than the standard deviation of the truth at the equilibrium and 2) the pattern correlation between the predicted signal and the truth is above 0.5. Importantly, the prediction here yields a significantly higher skill than the conventional EEOF based indices (Suhas et al. 2013).

Both the phase and amplitude of MISO activity play important roles in determining the prediction skill in different years. For example, year 2008 has an overall strong and regular MISO activity during the whole monsoon season that results in a long predictability, while the signal to noise ratio in year 2010 is smaller than other years and thus the predictability is greatly affected. Note that although year 2009 is a drought year with weak MISO activity during the late monsoon season (September), the MISO activity in other months of 2009 boreal summer remains strong and the overall prediction skill is high. From the limited sample size (12 years) of our analyis, it is hard to derive relationship between predictability of MISO and interannual variability of monsoon. However, it appears that the drought years do not necessarily have low predictability.

In addition to the ensemble mean prediction, the ensemble spread that indicates the predictive uncertainty is another important indicator of the prediction skill. Figure 5 shows the ensemble predictions including the ensemble spread for the six years, beginning at three different dates: April 1, June 1 and October 1. It is clear from Figure 1 that April 1 is a time at the transition between the quiescent phase and the active phase of the MISO indices; June 1 is a starting date in the active mature phase while October 1 is a starting date in the decaying phase of MISO activity. As shown in Figure 5, the ensemble mean predictions for the April 1 starting date do not have any long range skill but the ensemble spread automatically predicts this lack of skill and the envelope of the ensemble predictions contains the true signal for all years and forecast times including the return to skill in the winter quiescent phase. The forecasts from June 1 obviously have skill from both the mean and ensemble spread for all years for moderate to long lead times. The forecasts starting from October 1 have both an accurate mean and small ensemble spread for all six years and for very long times.

It is easy to perform twin prediction experiments with the perfect nonlinear stochastic model in (1)–(2) where 10 year training segments of the data generated from the model are utilized to make 6 year forecasts. It is significant that this internal prediction skill of the stochastic model is comparable to its skill in predicting the MISO indices from observations (not shown here). This lends support to the fact that the nonlinear stochastic model in (1)–(2) can accurately determine the predictability limits of the two MISO indices in Figure 1.

5. The Spatiotemporal Reconstruction

With the predicted MISO indices in hand, the final step is to recover the spatiotemporal MISO patterns in physical space. This requires the combination of time series and spatial bases.

a. Method

Let z_i be an d-dimensional vector of gridded precipitation values over the South Asia monsoon region at time i. Here, i is an integer ranging from 1 to n, representing the period of training phase. The first step in NLSA is to construct a higher-dimensional, time lagged embedding dataset utilizing Takens' method of delay. Denote q be the lagged embedding window size. Then the lagged embedding matrix can be written as

$$X = \begin{pmatrix} z_1 & z_2 & \cdots & z_{n-2q+1} \\ z_2 & z_3 & \cdots & z_{n-2q+2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{q-1} & z_q & \cdots & z_{n-q-1} \\ z_q & z_{q+1} & \cdots & z_{n-q} \end{pmatrix} = \begin{pmatrix} z_1 & z_2 & \cdots & z_{N-q+1} & z_{N-q+2} & \cdots & z_{N-1} & \boxed{z_N} \\ z_2 & z_3 & \cdots & z_{N-q+2} & z_{N-q+3} & \cdots & \boxed{z_N} & z_{N+1} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \cdots & \cdots & \vdots \\ z_{q-1} & z_q & \cdots & z_{N-1} & \boxed{z_N} & \cdots & z_{N+q-3} & z_{N+q-2} \\ z_q & z_{q+1} & \cdots & \boxed{z_N} & z_{N+1} & \cdots & z_{N+q-2} & z_{N+q-1} \end{pmatrix}, \quad (3)$$

where $N = n-2q+1$. Note that q in (3) is actually dq but d is omitted here for notation simplicity. Although the lagged embedding matrix X in (3) is formed by the raw observational data, the matrix associated with each eigenmode, such as the annual mode, semi-annual mode and MISO mode, after the spatiotemporal reconstruction has essentially the same structure, except that the q entries that are represented by the same z_i in (3) may have different values. Therefore, averaging over these q entries finalizes the reconstruction of z_i. From now on, X represents the lagged embedding matrix containing only the MISO mode. Note that, since $z_1,\ldots,z_{q-1}$ and $z_{N+1},\ldots,z_{N+q-1}$ appear less than q times in (3), recovering these components requires a longer observational period.

The relationship between the spatiotemporal representation X, the time series (i.e., indices) Φ and the spatial basis A is simply given by

$$X = A\Phi^T, \quad (4)$$

where $\cdot^T$ stands for the transpose. Clearly, with the predicted MISO indices in hand, the reconstruction (4) is easily achieved as long as A is able to be reached in the predicted phase. Different from the 2-Dimensional time series that is predicted by the low-order models, it is a challenge to describe and predict the exact evolution of the high-dimensional spatial basis A. Therefore, approximations are typically included in developing A in the predicted phases. Below, the spatial basis A utilized for prediction is assumed to be a constant matrix that is completely determined in the training phase. Note that the stationary assumption of A is in general not necessary and may even result in some errors in the spatiotemporal reconstruction for nonlinear data. Nevertheless, adopting a constant matrix A greatly reduces the computational cost and facilitates an effective and practical prediction algorithm as will be demonstrated soon. As will be shown at the end of this section, the approximated reconstruction with such a constant spatial basis is actually highly consistent with the truth. In Section 6d, one alternative of creating a non-stationary spatial basis will be briefly discussed and compared with the method proposed in this section.

In light of both X and Φ in the training period, the spatial pattern A according to (4) is given by

$$A = cX\Phi, \qquad \text{with } c = \frac{1}{\|\Phi\|^2}. \tag{5}$$

Then a natural way of performing the spatiotemporal reconstruction of the predicted MISO patterns is to multiply A obtained from (5) by the predicted MISO indices. Denote X^f and Φ^f the spatiotemporal pattern and indices in the prediction period. The following relation is reached:

$$A \cdot [\Phi; \Phi^f] = [X; X^f], \tag{6}$$

where we ignore the transpose in Φ and Φ^f for notation simplicity.

However, as is shown in Appendix C, the fundamental barrier for the direct method (6) to become practical is that in order to obtain the spatiotemporal patterns at s lead days, the prediction of the time series up to $s+q$ days is required (Comeau et al. 2016). For example, in the prediction of MISO, reconstructing the spatiotemporal pattern for the next day requires the prediction of time series up to the next 65 days! In fact, to reach the last z_N as shown in (3), the information up to z_{N+q-1} is required.

One remedy to overcome such difficulty is to switch the extra future information as required in the time series Φ to that in the spatial basis A by calculating a "predicted spatial basis" $\tilde{A}$ in the training phase. This $\tilde{A}$ is obtained by taking advantage of a new lagged embedding matrix $\tilde{X}$,

$$\widetilde{X} = \begin{pmatrix} z_q & z_{q+1} & \cdots & z_{n-q} \\ z_{q+1} & z_{q+2} & \cdots & z_{n-q+1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{2q-1} & z_{2q} & \cdots & z_{n-1} \end{pmatrix}, \tag{7}$$

which is just $q-1$ units shift forward in time with respect to X in (3). Similar as A, a new spatial pattern $\tilde{A}$ is formed by

$$\tilde{A} = c\tilde{X}\Phi, \qquad \text{with } c = \frac{1}{\|\Phi\|^2}. \tag{8}$$

Replacing A by $\tilde{A}$ in (6) leads to

$$\tilde{A} \cdot [\Phi; \Phi^f] = [\tilde{X}, \tilde{X}^f]. \tag{9}$$

As shown in Appendix C, with (9), reconstructing the spatiotemporal patterns at s days in the future requires only the prediction of the time series for s days. Since $\tilde{A}$ is completely determined in the training period that involves only straightforward calculations, the formula in (9) is an effective and practical method for predicting the spatiotemporal patterns.

b. Prediction of the Spatial-Temporal Reconstructed Precipitation Fields

Figure 6 includes three phase diagrams of the MISO indices, each containing a length of one-month period. The corresponding predictions, starting from the first day of each period and lasting for 30 days, is also included. Among the three periods, a significant skillful prediction is found for July 2009 while the prediction skill of June 2008 is moderate. The true signal of June 2013 has a weak amplitude and the corresponding prediction is far from the truth. Below, the prediction of spatiotemporal patterns based on the improved method (9) is demonstrated, where the ensemble mean of prediction is utilized for spatiotemporal reconstruction.

The skill scores of the predicted spatiotemporal patterns for each of the three periods are shown in Figure 7. Consistent with the MISO indices, July 2009 has the highest prediction skill and the useful prediction lasts for 40 days. On the other hand, a higher pattern correlation is found in predicting the spatiotemporal patterns of June 2008 than that of June 2013, where the useful prediction of June 2008 is up to around 22 days. Note that, different from predicting the MISO indices, the skill scores of predicting the spatiotemporal pattern do not decrease monotonically as a function of lead time. This is due to the stationary approximation of the spatial basis $\tilde{A}$ in the prediction period. In addition, this averaged spatial basis may also lead to the amplitude underestimation in the predicted spatiotemporal reconstruction. A direct remedy is to compute the ratio of $||A||$ and $||\tilde{A}||$ in the training period and multiply this constant ratio in prediction. In fact, the value of $||\tilde{A}||$ decreases with the increases in q value. This is because the correlation between the spatial basis and the time series that with a phase lag becomes weaker when the lag increases.

Figures 8 and 9 compare the truth and the predicted spatiotemporal patterns of July 2009 and June 2008, respectively. The predicted patterns for the whole July 2009 are highly consistent with the truth, especially in the regions of Indian subcontinent and Bay of Bengal. On the other hand, despite the skillful prediction up to 20 days lead time, significant errors in the spatiotemporal patterns appear for longer time predictions of June 2008 due to the failure in predicting the precipitation in regions such as the Indian Ocean.

To understand the approximated error in $\tilde{A}$, the true spatiotemporal patterns from NLSA, which is validated in Sabeerali et al. (2017), and the approximated patterns based on $\tilde{A}\Phi$ for July 2009 are compared in the first two rows of Figure 10. Note that the truth of Φ is adopted here to exclude the error in the prediction of the time series. Despite the stationary approximation in creating the spatial basis $\tilde{A}$, the time series Φ from NLSA remains highly nonlinear and intermittent. Clearly, the approximated patterns are remarkably consistent with the truth, where the non-Gaussian and intermittent features in the spatiotemporal patterns are both retained to a high extent.

6. Discussions

a. Prediction with a 3-year short training period

A typical situation in climate science is that only a short period of observational data is available. This actually leads to one of the fundamental difficulties in prediction utilizing most non-parametric methods that require a huge amount of data for training. Suitable models that are able to describe essential characteristics of the data are usually preferred since they allow a much shorter training period. Recall in previous sections, 10 years of observations (1998-2007) were adopted for model calibration and the prediction skill were assessed for the remaining 6 years (2008-2013). Although this 10-year training window is already much shorter than that required by most non-parametric methods, it is important to understand whether an even shorter training period is possible here for the nonlinear model to obtain the information in nature.

To this end, a very short training period involving only the first three years of the time series (1998-2000) is adopted here for model calibration. Figure 11 compares the statistics of the MISO time series with different lengths, including this short 3-year training period (1998-2000), the 10-year training period adopted in previous sections (1998-2007) and the full MISO period (1998-2013). The fact that the statistics of the 10-year training period and the full MISO period almost perfectly match each other indicates the sufficiency of the 10-year training period in obtaining the unbiased information. On the other hand, the 3-year training period, including one weak year (1998), one moderate year (1999) and one strong year (2000) of MISO activity, also has highly consistent statistics with those associated with the full MISO time series, including the non-Gaussian fat-tailed PDFs, the power spectrums and the autocorrelations up to 1.5 months. Therefore, the information of the full MISO indices are well reflected in this short 3-year training period. Due to the robustness of the model parameters (Appendix A), the calibrated parameters based on this 3-year short training period are nearly the same as the optimal parameters shown in Table 1. Importantly, this short training period allows the study of prediction skill for a long period back to year 2001 and the results are roughly reported here.

Figure 12 shows the skill scores and the predicted signals based on the ensemble mean prediction from year 2001 to 2007, analogous to those in Figure 3 and 4 from year 2008 to 2013. The useful prediction of these 7 years all exceeds 25 days, where in particular the skillful predictions in year 2001, 2003 and 2007 are more than 40 days. Among these 7 years, year 2002 and 2004 are recorded as drought years. A significant error is found in predicting the subdued MISO activity during August and September of year 2002, which explains its lower overall prediction skill than most of the other years. On the other hand, despite being a drought year, MISO activity during 2004 is persistently strong throughout the boreal summer. The major error in predicting MISO indices of year 2004 is in fact due to the model's failure in capturing the extremely slow oscillation frequency during August and September.

We have also check the model statistics and prediction skill by utilizing any three consecutive years between 1998-2013 as the training phase. Despite the discrepancy in the signal variance due to the strength of MISO activity in different years, the fat tails in the non-Gaussian PDFs, the peak of the power spectrums and the autorrelations up to 1.5 months all resemble those of the full MISO time series. Importantly, the ensemble prediction skill does not have significant deterioration based on different training periods.

b. MISO indices based on different lagged embedding window sizes and the corresponding prediction skill

Recall that the two MISO indices shown in Figure 1 and studied throughout this article were obtained by applying NLSA to the precipitation data with a lagged embedding window of length $q = 64$ days. Adopting $q = 64$ is natural since it is an ideal choice for representing the intraseasonal time scale and such lagged embedding window size was utilized for defining the large-scale cloud patterns of the MJO and monsoon in previous works (Chen et al. 2014; Chen and Majda 2015b,a; Tung et al. 2014). On the other hand, EEOF was also widely utilized in defining the MISO indices in literature (Suhas et al. 2013; Kikuchi et al. 2012), which involves removing the climatological mean, first a few harmonics of the seasonal cycle and then applying a much shorter embedding window with 15-20 days. Therefore, it is important to study the difference in the MISO indices by applying NLSA with different lagged embedding window sizes.

Figure 13 shows the resulting MISO indices by applying NLSA with $q = 64, 48$ and 34 as well as the corresponding statistics. Different from $q = 64$, the MISO indices with $q = 48$ and 34 have active phases in both boreal summer and winter, implying that the obtained MISO indices contain the components of the boreal winter MJO, and the associated PDFs are nearly Gaussian. In addition to being polluted by the boreal winter signal, these time series, especially with $q = 34$, also contain bi-annual, annual and semi-annual cycles, as indicated by the large bursts in the low-frequent band of the power spectrum. Another significant discrepancy with different q values lies in the causality between the two components of the MISO indices. With $q = 64$, the cross-correlation functions have significant peaks at lags around 12 days, which is nearly 1/4 of the averaged oscillation frequency and indicates the quadrature structure of MISO 1 and MISO 2. On the other hand, the cross-correlations $R_{12}(t)$ and $R_{21}(t)$ with $q = 48$ and $q = 34$ remain close to zero, and the maximum value of the lagged correlation between MISO 1 and MISO 2 indices is less than 0.3 (not shown here). These facts imply that MISO 1 and MISO 2 are nearly uncorrelated and therefore model errors appear in fitting the cross-correlations utilizing the nonlinear low-order model (1). Finally, the fast decay of autocorrelations $R_{11}(t)$ and $R_{22}(t)$ with $q = 48$ and 34 implies deterioration in the predictability of the MISO indices.

Figure 14 shows the prediction skill with different q. Here useful prediction is defined in the same way as that in Section 4: 1) the RMS error in the prediction is less than the standard deviation of the truth at the equilibrium and 2) the pattern correlation between the predicted signal and the truth is above 0.5. In addition to illustrating the prediction skill for the whole year, the prediction skill conditioned on the boreal summer time (June to September) is also emphasized. As expected, with the decrease in q, the overall prediction skill becomes worse. Nevertheless, conditioned on the boreal summer time, the prediction with $q = 48$ remains quite skillful and in particular the 15-day lead time prediction is highly consistent with the truth. This is, however, not true for the prediction with $q = 34$, where the useful prediction only lasts for 10-12 days in terms of both the whole year and only the boreal summer time.

c. Significant prediction skill of the precipitation MISO indices with parameters calibrated from OLR dataset

Most tropical rainfall is convective, which implies that OLR, a proxy for the convection, is a potential candidate to describe the precipitation in tropics. Positive (negative) OLR anomalies are associated with reduced (increased) cloudiness, hence suppressed (enhanced) deep convection. Due to the possible relationship between the OLR and the tropical precipitation anomalies, it is important to understand the skill of the low-order nonlinear stochastic model (1)–(2) in predicting the MISO indices with parameters calibrated from OLR dataset.

In Chen and Majda (2015b), the low-order nonlinear stochastic model (1)–(2) was adopted to predict the two boreal summer intraseasonal oscillation (BSISO) modes obtained by applying NLSA to the brightness temperature, a highly correlated variable with OLR, within the equatorial tropical belt from $15^o S$ to 30^oN. The dataset utilized there was Cloud Archive User Service (CLAUS) Version 4.7. To explore the strength of the correlation between OLR and precipitation, the parameters in Chen and Majda (2015b) are applied to (1)–(2) to calibrate and predict the precipitation MISO indices. For simplicity, these parameters are named as OLR-based parameters.

The OLR-based parameters are listed in the second row of Table 1 with two minor modifications. First, since the time series in Chen and Majda (2015b) were started from September instead of January, the phase parameter ϕ in Chen and Majda (2015b) is modified accordingly. Second, due to the general negative correlation between OLR and precipitation, the sign of the oscillation frequency a in Chen and Majda (2015b) is flipped. In fact, as shown in Table 1, the OLR-based parameters are quite similar to the optimal parameters utilized in the previous sections.

Panels (a)-(d) of Figure 15 show the model statistics with the OLR-based parameters. The autocorrelations, power spectrums and non-Gaussian fat-tailed PDFs are all quite consistent with the truth, implies nearly identical statistical and dynamical features in describing the precipitation MISO with the OLR-based parameters. The slight underestimation of the variance with the OLR-based parameters is due to the gap in the units of the two variables, which can easily be remedied by applying an amplitude rescaling to the OLR variable and is a secondary issue here. Panels (e) and (f) compare the 25-day lead ensemble mean predictions using the optimal parameters and the OLR-based parameters of three different years from 2008 to 2010 with strong, moderate and weak MISO activities, respectively. In fact, except a light underestimation of the amplitude due to the insufficient amplitude of the seasonal cycle damping f_t, the prediction utilizing OLR-based parameters has almost the same skill as that utilizing the optimal parameters in all the three years.

The results from both model calibration and prediction confirm a strong (negative) correlation between OLR and precipitation anomalies (Sabeerali et al. 2017).

d. Defects of creating the spatial basis based on the running average over the raw data

Recall in Section 5, despite a stationary spatial basis $\tilde{A}$ being utilized, the approximated spatiotemporal reconstructed pattern $\tilde{A}\Phi$ is highly consistent with nature (Figure 10). In addition to the strong nonlinear and intermittent time series Φ, adopting the spatiotemporal pattern $\tilde{X}$ from NLSA in determining $\tilde{A}$ is another important reason for such high consistency, which will be emphasized in this section.

Note that the true spatiotemporal patterns based on the traditional linear methods, such as EOF and EEOF, are typically reached by applying bandpass filter or running average on the raw observations over a certain prescribed window. Below, running average is applied to the raw data, the results of which in the training period are then utilized to form the spatial basis. In the prediction period, such spatial basis is multiplied by the NLSA MISO indices for the spatiotemporal reconstruction. Since the same NLSA MISO indices are utilized here and in Section 5, the discrepancy in the reconstructed spatiotemporal patterns completely lies in the spatial basis.

The details of the spatiotemporal reconstruction mentioned above is presented below. Note that a non-stationary spatial basis is adopted here based on phase decomposition in the training period. First, a q day running average is applied on raw rainfall datasets and the daily rainfall anomaly is computed at each grid point in the training period. Then the whole phase space of MISO 1 and MISO 2 is divided into S disjoint pieces, named as phases. The spatial pattern of each phase is computed by conditional averaging of this rainfall anomalies subject to the criteria that the instantaneous MISO indices have amplitude being greater than 1 standard deviation and belong to the corresponding phase. In the prediction stage, multiplying the magnitude of the predicted MISO indices by the selected spatial basis according to the location of the predicted MISO indices in the phase space. This results in the spatiotemporal reconstructed pattern. A typical situation involves dividing the whole phase space into $S = 8$ pieces with equal areas (See Figure 6) as was done in RMM and other MJO and monsoon indices (Wheeler and Hendon 2004; Székely et al. 2016a; Suhas et al. 2013; Lee et al. 2013). Since all the phases and the associated average spatial bases are determined in training period, this spatiotemporal prediction is also practical and computational efficient.

We apply this phase decomposition method with $S = 8$ to reach the spatial-temporal patterns of July 2009 and the results are shown in the third row of Figure 10. Significant differences are found compared with the truth (first row). Particularly, a reversed drought/flood pattern from July 1 to July 11 in the India subcontinent based on these two methods is observed and the amount of precipitation in Indian Ocean and Arabian Sea is significantly different within the whole month. Such errors result from the spatial basis that is determined by applying the running average over the raw data. In fact, as was pointed out in Sabeerali et al. (2017) that many important MISO features are not well represented by those linear methods, including underestimating the fractional variance over Western Ghats and the failure in capturing variability over Indo-West Pacific region which is particularly crucial in determining the propagation characteristics of MISO (Sabeerali et al. 2017). This indicates the importance of the NLSA spatial patterns in addition to the nonlinear and intermittent time series.

Replacing the non-stationary spatial basis resulting from the modes based on the running average of the raw data by the NLSA modes is a potential way to improve the spatiotemporal reconstruction in the prediction stage. Yet, the phase decomposition method has an obvious drawback that the predicted spatiotemporal patterns are discontinuous in time when the corresponding spatial basis transits from one phase to another. One remedy to the discontinuity issue is to introduce a smooth transition between different phases such as adopting a convolution with a Gaussian kernel. This remains as a future work.

7. Conclusions

A recently developed technique for nonlinear time series analysis NLSA (Giannakis and Majda 2012a,b, 2013) has been utilized to define two MISO indices from the daily GPCP rainfall data set without detrending or spatiotemporal filtering (Sabeerali et al. 2017). The observed time series have non-Gaussian fat-tailed PDFs, which is a consequence of intermittency.

Systematic strategies for physics constrained regression models (Majda and Harlim 2013; Harlim et al. 2014) suggest a four dimensional stochastic model with two hidden variables representing stochastic damping and random phasing with energy conserving nonlinear feedback interaction (Section 3). In a calibration phase, these models can successfully capture the observed non-Gaussian PDFs, power spectrums and autocorrelations. The models have a special structure that allows efficient data assimilation and ensemble initialization algorithms for the hidden variables. It is shown in Section 4 that the low-order nonlinear stochastic model has been applied to prediction of the NLSA MISO indices with forecasting skill ranging from 20 to 50 days in different years. Furthermore, the ensemble spread in the stochastic model has been shown to be an accurate predictive indicator of forecast uncertainty at long range.

An effective and practical spatiotemporal reconstruction algorithm is then proposed in Section 5, which overcomes the fundamental difficulty in most data decomposition techniques with lagged embedding that require extra information beyond the predicted time series in the future. The prediction skill of the reconstruction spatiotemporal patterns is consistent with that of the MISO indices.

A few issues are addressed in Section 6. First, the model calibration and prediction with a 3-year short training period is studied. The resulting statistics and prediction skill do not have significant deterioration compared with those based on a 10-year training period. This suggests the advantage of utilizing the low-order nonlinear model (1) over most non-parametric methods in predicting the MISO indices from a practical point of view. Second, the NLSA MISO indices based on different lagged embedding window sizes are compared. The resulting MISO indices with a shorter lagged embedding window size ($q = 48$ and $q = 34$) are polluted with other variabilities and the corresponding overall prediction skill is greatly affected. Nevertheless, with $q = 48$ days lag, the prediction conditioned on the boreal summer time remains still skillful. Thirdly, the low-order nonlinear stochastic model with OLR-based parameters remains high skill in both fitting the non-Gaussian statistics and predicting the precipitation MISO indices, implying a significant correlation between the tropical precipitation and OLR. Finally, algorithms with potential abilities to improve the prediction skill of the spatiotemporal patterns are briefly discussed, the implementation of which remains as future works.

Developing more effective and accurate spatiotemporal reconstruction algorithm remains as one of the future works. In fact, clustering method is a promising technique for recovering more detailed features of spatial basis conditioned on different phases.

Acknowledgments. The research of A.J.M is partially supported by the Office of Naval Research Grant ONR MURI N00014-16-1-2161 and the New York University Abu Dhabi Research Institute. N.C. is supported as a postdoctoral fellow following A.J.M's ONR MURI Grant. C.T.S, R.S.A and A.J.M also acknowledge the support from Monsoon Mission of the Ministry of Earth Sciences (MoES), Government of India (Grant No. MM/SERP/NYU/2014/SSC-01/002). The research of

C. T. S and R.S. A is also supported by the New York University Abu Dhabi Research Institute. The authors thank Dimitrios Giannakis for useful discussions.

APPENDIX A

Calibration of the nonlinear stochastic model with information theory

The optimal parameters in the nonlinear low-order stochastic model (1) are calibrated by systematically minimizing the information distance, i.e., the model error, in the PDF of the model π^M compared with that of the MISO index π (Majda and Gershgorin 2010, 2011; Kleeman 2002; Majda and Branicki 2012; Branicki et al. 2013),

$$\mathscr{P}(\pi, \pi^M) = \int \pi \log\left(\frac{\pi}{\pi^M}\right). \tag{A1}$$

The model error dependence on the variation of different parameters is shown in Figure A1, which indicates that the nonlinear low-order stochastic model (1) is robust with respect to the parameters around their optimal values. The huge model error with the underestimation of σ_u, f_t and γ and the overestimation of d_u is due to the failure of capturing the intermittency. Note that the model error has only a weak dependence on the background phase a since the contribution of the oscillation in the signal has been averaged out in the time-averaged PDF. However, the parameter a is crucial in describing the frequency of intraseasonal oscillation, and it is calibrated by matching the autocorrelation functions associated with the model and the truth. The other parameters d_v, σ_v, d_ω and σ_ω in describing the stochastic processes affect not only on the model error but more on the autocorrelations and power spectrums as well. A large discrepancy appears in the statistics if these parameters are outside the optimal range. The parameter f_0 is not an independent parameter given d_u and γ and therefore we fix its value. The frequency ω_f in the time-periodic damping $v_f(t)$ is prescribed to be $2\pi/12$ such that one time unit of the model corresponds to one month in reality.

The phase ϕ in $v_f(t)$ is tuned to make the strong intermittency occur in the boreal summer in accordance with the MISO indices. Note that none of the parameters is redundant in the nonlinear stochastic model (1). In fact, without the hidden variables v and ω_u, even if the time-period damping $v_f(t)$ is able to crudely describe the active phase of BSISO in the reduced linear model, a distinguished disparity is observed in the model statistics compared with the truth, indicating the intrinsic barrier (Majda and Gershgorin 2011; Majda and Branicki 2012).

Prediction with random suboptimal parameters is also studied. Here the suboptimal parameters are taken randomly between the two dotted lines in each panel of Figure 7. Comparable prediction skill is found with these random suboptimal parameters as the optimal parameters.

APPENDIX B

Mathematical details of effective data assimilation and prediction algorithm

Recall the nonlinear low-order stochastic model (1). Denote by $\mathbf{U} = (u_1, u_2)^T$ and $\mathbf{\Gamma} = (v, \omega_u)^T$. The abstract form of the low-order stochastic model (1) is given as follows:

$$d\mathbf{U}_t = [\mathbf{A}_0(t,\mathbf{U}) + \mathbf{A}_1(t,\mathbf{U})\mathbf{\Gamma}_t]dt + \mathbf{\Sigma}_U(t,\mathbf{U})d\mathbf{W}_U(t), \tag{B1a}$$

$$d\mathbf{\Gamma}_t = [\mathbf{a}_0(t,\mathbf{U}) + \mathbf{a}_1(t,\mathbf{U})\mathbf{\Gamma}_t]dt + \mathbf{\Sigma}_\Gamma(t,\mathbf{U})d\mathbf{W}_\Gamma(t), \tag{B1b}$$

where

$$\mathbf{A}_0 = \begin{pmatrix} -d_u u_1 + \gamma v_f(t) u_1 - a u_2 \\ -d_u u_2 + \gamma v_f(t) u_2 + a u_1 \end{pmatrix}, \qquad \mathbf{A}_1 = \begin{pmatrix} \gamma u_1 & -u_2 \\ \gamma u_2 & u_1 \end{pmatrix},$$

$$\mathbf{a}_0 = \begin{pmatrix} -\gamma(u_1^2 + u_2^2) \\ 0 \end{pmatrix}, \qquad \mathbf{a}_1 = \begin{pmatrix} -d_v & \\ & -d_\omega \end{pmatrix},$$

$$\mathbf{\Sigma}_U = \begin{pmatrix} \sigma_u & \\ & \sigma_u \end{pmatrix}, \qquad \mathbf{\Sigma}_\Gamma = \begin{pmatrix} \sigma_v & \\ & \sigma_\omega \end{pmatrix}.$$

The model (B1) is a conditional Gaussian system conditioned on the observations $\mathbf{U}$, meaning that once the observations $\mathbf{U}$ are given the dynamics of $\mathbf{\Gamma}$ in (B1) becomes a Gaussian system (Chen and Majda 2016). The special structure of system (B1) allows the closed analytic formulae for the evolution of the conditional Gaussian distributions of the hidden parameters v and ω_u (Liptser and Shiryaev 2001) obtained in the Bayesian framework:

$$\begin{aligned} d\mu_t =& [\mathbf{a}_0(t,\mathbf{U}) + \mathbf{a}_1(t,\mathbf{U})\mu_t]dt + (R_t\mathbf{A}_1^*(t,\mathbf{U}))(\mathbf{\Sigma}_U\mathbf{\Sigma}_U^*)^{-1}(t,\mathbf{U})\times \\ & [d\mathbf{U}_t - (\mathbf{A}_0(t,\mathbf{U}) + \mathbf{A}_1(t,\mathbf{U})\mu_t)dt], \\ dR_t =& \{\mathbf{a}_1(t,\mathbf{U})R_t + R_t\mathbf{a}_1^*(t,\mathbf{U}) + (\mathbf{\Sigma}_\Gamma\mathbf{\Sigma}_\Gamma^*)(t,\mathbf{U}) \\ & -(R_t\mathbf{A}_1^*(t,\mathbf{U}))(\mathbf{\Sigma}_U\mathbf{\Sigma}_U^*)^{-1}(t,\mathbf{U})(R_t\mathbf{A}_1^*(t,\mathbf{U}))^*\}\,dt, \end{aligned} \tag{B2}$$

where μ_t and R_t are the posterior mean and posterior covariance of the conditional distributions, respectively. The asterisk represents the complex conjugate.

As a remark, the formulae (B2) are optimal if and only if the signal is generated from system (B1). Since our observed signal, i.e., the MISO indices, are not from the nonlinear low-order s-tochastic model (B1), the evolutions of the conditional Gaussian distributions (B2) are suboptimal.

APPENDIX C

Details of the spatiotemporal reconstruction

Recall the relation between the spatial basis Φ, the time series A and the spatiotemporal patterns X given by the direct method (6)

$$A \cdot [\Phi; \Phi^f] = [X; X^f]. \tag{C1}$$

The left and right hand side of (C1) are given respectively by

$$
\begin{aligned}
A \cdot \left[\Phi; (\Phi^f)\right] &= \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_q \end{pmatrix} \cdot \begin{pmatrix} \Phi_1, \cdots, & \Phi_N, & \Phi_1^f, & \Phi_2^f, & \cdots, & \Phi_q^f \end{pmatrix} \\
&= \left(\begin{array}{ccc|cccc}
A_1\Phi_1 & \cdots & A_1\Phi_N & A_1\Phi_1^f & A_1\Phi_2^f & \cdots & \boxed{A_1\Phi_q^f} \\
A_2\Phi_1 & \cdots & A_2\Phi_N & A_2\Phi_1^f & A_2\Phi_2^f & \cdots & A_2\Phi_q^f \\
\vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
A_{q-1}\Phi_1 & \cdots & A_{q-1}\Phi_N & A_{q-1}\Phi_1^f & \boxed{A_{q-1}\Phi_2^f} & \cdots & A_{q-1}\Phi_q^f \\
A_q\Phi_1 & \cdots & A_q\Phi_N & \boxed{A_q\Phi_1^f} & A_q\Phi_2^f & \cdots & A_q\Phi_q^f
\end{array}\right)
\end{aligned}
\tag{C2}
$$

and

$$
[X; X^f] = \left(\begin{array}{cccc|ccccc}
z_1 & z_2 & \cdots & z_{n-2q+1} & z_{n-2q+2} & z_{n-2q+3} & \cdots & z_{n-q} & \boxed{z_1^f} \\
z_2 & z_3 & \cdots & z_{n-2q+2} & z_{n-2q+3} & z_{n-2q+4} & \cdots & \boxed{z_1^f} & z_2^f \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
z_{q-1} & z_q & \cdots & z_{n-q+1} & z_{n-q} & \boxed{z_1^f} & \cdots & z_{q-1}^f & z_{q-1}^f \\
z_q & z_{q+1} & \cdots & z_{n-q} & \boxed{z_1^f} & z_2^f & \cdots & z_q^f & z_q^f
\end{array}\right), \tag{C3}
$$

However, the entries with boxes in (C2) and (C3) implies that in order to obtain the spatiotemporal pattern at one lead time unit, the time series up to q lead time units are required. In other words, predicting $q = 64$ days of the time series in the future is only sufficient to achieve the spatiotemporal pattern for 1 day forward. Therefore, this method is not practical.

On the other hand, the improved method based on the new spatial basis $\tilde{A}$ in (8). Recall the relationship in (C4),

$$
\tilde{A} \cdot [\Phi; \Phi^f] = [\tilde{X}, \tilde{X}^f]. \tag{C4}
$$

The left and right hand sides can be written down explicitly,

$$
\begin{aligned}
\widetilde{A}\cdot[\Phi;\Phi^f] &= \begin{pmatrix} \widetilde{A}_1 \\ \widetilde{A}_2 \\ \vdots \\ \widetilde{A}_q \end{pmatrix} \cdot \left(\begin{array}{ccc|ccc|cccc} \Phi_1, & \cdots, & \Phi_{N-q}, & \Phi_{N-q+1}, & \cdots, & \Phi_N, & \Phi_1^f, & \Phi_2^f, & \cdots, & \Phi_q^f \end{array} \right) \\
&= \left(\begin{array}{ccc|cccc|ccc}
\widetilde{A}_1\Phi_1 & \cdots & \widetilde{A}_1\Phi_{N-q} & \widetilde{A}_1\Phi_{N-q+1} & \widetilde{A}_1\Phi_{N-q+2} & \cdots & \widetilde{A}_1\Phi_N & \boxed{\widetilde{A}_1\Phi_1^f} & \cdots & \widetilde{A}_1\Phi_q^f \\
\widetilde{A}_2\Phi_1 & \cdots & \widetilde{A}_2\Phi_{N-q} & \widetilde{A}_2\Phi_{N-q+1} & \widetilde{A}_2\Phi_{N-q+2} & \cdots & \boxed{\widetilde{A}_2\Phi_N} & \widetilde{A}_2\Phi_1^f & \cdots & \widetilde{A}_2\Phi_q^f \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\
\widetilde{A}_{q-1}\Phi_1 & \cdots & \widetilde{A}_{q-1}\Phi_{N-q} & \widetilde{A}_{q-1}\Phi_{N-q+1} & \widetilde{A}_{q-1}\Phi_{N-q+2} & \cdots & \widetilde{A}_{q-1}\Phi_N & \widetilde{A}_{q-1}\Phi_1^f & \cdots & \widetilde{A}_{q-1}\Phi_q^f \\
\widetilde{A}_q\Phi_1 & \cdots & \widetilde{A}_q\Phi_{N-q} & \widetilde{A}_q\Phi_{N-q+1} & \boxed{\widetilde{A}_q\Phi_{N-q+2}} & \cdots & \widetilde{A}_q\Phi_N & \widetilde{A}_q\Phi_1^f & \cdots & \widetilde{A}_q\Phi_q^f
\end{array} \right),
\end{aligned}
\tag{C5}
$$

and

$$
[\widetilde{X};\widetilde{\widetilde{X}}] = \left(\begin{array}{ccc|cccc|ccc}
z_q & \cdots & z_{n-2q} & z_{n-2q+1} & z_{n-2q+2} & \cdots & z_{n-q} & \boxed{z_1^f} & \cdots & z_{q-1}^f \\
z_{q+1} & \cdots & z_{n-2q+1} & z_{n-2q+2} & z_{n-2q+3} & \cdots & \boxed{z_1^f} & z_2^f & \cdots & z_q^f \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
z_{2q-2} & \cdots & z_{n-q} & z_{n-q+1} & z_{n-q} & \cdots & z_{q-1}^f & z_{q-1}^f & \cdots & z_{2q-2}^f \\
z_{2q-1} & \cdots & z_{n-q-1} & z_{n-q} & \boxed{z_1^f} & \cdots & z_q^f & z_q^f & \cdots & z_{2q-1}^f
\end{array} \right). \tag{C6}
$$

Comparing (C5) and (C6), it is clear that reconstructing the spatiotemporal patterns at s days in the future requires only the prediction of the time series for s days.

References

Abhilash, S., A. Sahai, S. Pattnaik, B. Goswami, and A. Kumar, 2014a: Extended range prediction of active-break spells of indian summer monsoon rainfall using an ensemble prediction system in ncep climate forecast system. *International Journal of Climatology*, **34 (1)**, 98–113.

Abhilash, S., and Coauthors, 2014b: Prediction and monitoring of monsoon intraseasonal oscillations over indian monsoon region in an ensemble prediction system using cfsv2. *Climate dynamics*, **42 (9-10)**, 2801–2815.

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

Branicki, M., N. Chen, and A. J. Majda, 2013: Non-Gaussian test models for prediction and state estimation with model errors. *Chinese Annals of Mathematics, Series B*, **34 (1)**, 29–64.

Brenowitz, N. D., D. Giannakis, and A. J. Majda, 2016: Nonlinear laplacian spectral analysis of rayleigh–bénard convection. *Journal of Computational Physics*, **315**, 536–553.

Chen, N., and A. J. Majda, 2015a: Predicting the cloud patterns for the boreal summer intraseasonal oscillation through a low-order stochastic model. *Mathematics of Climate and Weather Forecasting*, **1 (1)**, 1–20.

Chen, N., and A. J. Majda, 2015b: Predicting the real-time multivariate madden–julian oscillation index through a low-order nonlinear stochastic model. *Monthly Weather Review*, **143 (6)**, 2148–2169.

Chen, N., and A. J. Majda, 2016: Filtering nonlinear turbulent dynamical systems through conditional gaussian statistics. *Monthly Weather Review*, **144 (12)**, 4885–4917.

Chen, N., A. J. Majda, and D. Giannakis, 2014: Predicting the cloud patterns of the madden-julian oscillation through a low-order nonlinear stochastic model. *Geophysical Research Letters*, **41 (15)**, 5612–5619.

Coifman, R. R., and S. Lafon, 2006: Diffusion maps. *Applied and computational harmonic analysis*, **21 (1)**, 5–30.

Comeau, D., Z. Zhao, D. Giannakis, and A. J. Majda, 2016: Data-driven prediction strategies for low-frequency patterns of north pacific climate variability. *Climate Dynamics*, 1–18.

Crommelin, D., and A. Majda, 2004: Strategies for model reduction: comparing different optimal bases. *Journal of the Atmospheric Sciences*, **61 (17)**, 2206–2217.

Gadgil, S., 2003: The indian monsoon and its variability. *Annual Review of Earth and Planetary Sciences*, **31 (1)**, 429–467.

Giannakis, D., and A. J. Majda, 2011: Time series reconstruction via machine learning: Revealing decadal variability and intermittency in the north pacific sector of a coupled climate model. *CIDU*, Citeseer, 107–117.

Giannakis, D., and A. J. Majda, 2012a: Comparing low-frequency and intermittent variability in comprehensive climate models through nonlinear laplacian spectral analysis. *Geophysical Research Letters*, **39 (10)**.

Giannakis, D., and A. J. Majda, 2012b: Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proceedings of the National Academy of Sciences*, **109 (7)**, 2222–2227.

Giannakis, D., and A. J. Majda, 2013: Nonlinear laplacian spectral analysis: capturing intermittent and low-frequency spatiotemporal patterns in high-dimensional data. *Statistical Analysis and Data Mining*, **6 (3)**, 180–194.

Giannakis, D., W.-w. Tung, and A. J. Majda, 2012: Hierarchical structure of the madden-julian oscillation in infrared brightness temperature revealed through nonlinear laplacian spectral analysis. *Intelligent Data Understanding (CIDU), 2012 Conference on*, IEEE, 55–62.

Goswami, B., R. Ajayamohan, P. K. Xavier, and D. Sengupta, 2003: Clustering of synoptic activity by indian summer monsoon intraseasonal oscillations. *Geophysical Research Letters*, **30 (8)**.

Goswami, B., V. Krishnamurthy, and H. Annmalai, 1999: A broad-scale circulation index for the interannual variability of the indian summer monsoon. *Quarterly Journal of the Royal Meteorological Society*, **125 (554)**, 611–633.

Goswami, B., and R. A. Mohan, 2001: Intraseasonal oscillations and interannual variability of the indian summer monsoon. *Journal of Climate*, **14 (6)**, 1180–1198.

Harlim, J., A. Mahdi, and A. J. Majda, 2014: An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. *Journal of Computational Physics*, **257**, 782–812.

Huffman, G. J., R. F. Adler, M. M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *Journal of Hydrometeorology*, **2 (1)**, 36–50.

Kikuchi, K., B. Wang, and Y. Kajikawa, 2012: Bimodal representation of the tropical intraseasonal oscillation. *Climate dynamics*, **38 (9-10)**, 1989–2000.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy. *Journal of the atmospheric sciences*, **59 (13)**, 2057–2072.

Kondrashov, D., M. Chekroun, A. Robertson, and M. Ghil, 2013: Low-order stochastic model and past-noise forecasting of the madden-julian oscillation. *Geophysical Research Letters*, **40 (19)**, 5305–5310.

Kravtsov, S., D. Kondrashov, and M. Ghil, 2005: Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. *Journal of Climate*, **18 (21)**, 4404–4424.

Lau, W. K.-M., and D. E. Waliser, 2011: *Intraseasonal variability in the atmosphere-ocean climate system*. Springer Science & Business Media.

Lee, J.-Y., B. Wang, M. C. Wheeler, X. Fu, D. E. Waliser, and I.-S. Kang, 2013: Real-time multivariate indices for the boreal summer intraseasonal oscillation over the asian summer monsoon region. *Climate Dynamics*, **40 (1-2)**, 493–509.

Liptser, R. S., and A. N. Shiryaev, 2001: *Statistics of Random Processes II: II. Applications*, Vol. 2. Springer.

Majda, A. J., and M. Branicki, 2012: Lessons in uncertainty quantification for turbulent dynamical systems. *Discrete Cont. Dyn. Systems*, **32 (9)**, 3133–3221.

Majda, A. J., and B. Gershgorin, 2010: Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences*, **107 (34)**, 14 958–14 963.

Majda, A. J., and B. Gershgorin, 2011: Improving model fidelity and sensitivity for complex systems through empirical information theory. *Proceedings of the National Academy of Sciences*, **108 (25)**, 10 044–10 049.

Majda, A. J., and J. Harlim, 2013: Physics constrained nonlinear regression models for time series. *Nonlinearity*, **26 (1)**, 201–271.

Murakami, T., L.-X. Chen, and A. Xie, 1986: Relationship among seasonal cycles, low-frequency oscillations, and transient disturbances as revealed from outgoing longwave radiation data. *Monthly Weather Review*, **114 (8)**, 1456–1465.

Packard, N. H., J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, 1980: Geometry from a time series. *Physical review letters*, **45 (9)**, 712.

Sabeerali, C., R. Ajayamohan, D. Giannakis, and A. J. Majda, 2017: Extraction and prediction of indices for monsoon intraseasonal oscillations: an approach based on nonlinear laplacian spectral analysis. *Climate Dynamics*, 1–20.

Sahai, A., and Coauthors, 2013: Simulation and extended range prediction of monsoon intraseasonal oscillations in ncep cfs/gfs version 2 framework. *Current Science*, **104 (10)**, 1394–1408.

Sauer, T., J. A. Yorke, and M. Casdagli, 1991: Embedology. *Journal of statistical Physics*, **65 (3)**, 579–616.

Sikka, D., and S. Gadgil, 1980: On the maximum cloud zone and the itcz over indian, longitudes during the southwest monsoon. *Monthly Weather Review*, **108 (11)**, 1840–1853.

Slawinska, J., and D. Giannakis, 2016: Indo-pacific variability on seasonal to multidecadal timescales. part i: Intrinsic sst modes in models and observations. *arXiv preprint arXiv:1604.01742*.

Suhas, E., J. Neena, and B. Goswami, 2013: An indian monsoon intraseasonal oscillations (miso) index for real time monitoring and forecast verification. *Climate dynamics*, **40 (11-12)**, 2605–2616.

Székely, E., D. Giannakis, and A. J. Majda, 2016a: Extraction and predictability of coherent intraseasonal signals in infrared brightness temperature data. *Climate Dynamics*, **46 (5-6)**, 1473–1502.

Székely, E., D. Giannakis, and A. J. Majda, 2016b: Initiation and termination of intraseasonal oscillations in nonlinear laplacian spectral analysis-based indices. *Mathematics of Climate and Weather Forecasting*, **2 (1)**, 1–25.

Tung, W.-w., D. Giannakis, and A. J. Majda, 2014: Symmetric and antisymmetric convection signals in the madden–julian oscillation. part i: basic modes in infrared brightness temperature. *Journal of the Atmospheric Sciences*, **71 (9)**, 3302–3326.

Webster, P. J., V. O. Magana, T. Palmer, J. Shukla, R. Tomas, M. Yanai, and T. Yasunari, 1998: Monsoons: Processes, predictability, and the prospects for prediction. *Journal of Geophysical Research: Oceans*, **103 (C7)**, 14 451–14 510.

Wheeler, M. C., and H. H. Hendon, 2004: An all-season real-time multivariate mjo index: Development of an index for monitoring and prediction. *Monthly Weather Review*, **132 (8)**, 1917–1932.

Zhang, C., 2005: Madden-julian oscillation. *Reviews of Geophysics*, **43 (2)**.

LIST OF TABLES

Table 1. Top: Optimal parameters for the nonlinear low-order stochastic model (1). The parameters $d_u, a, \gamma, f_0, f_t, \omega_f, d_v$ and d_ω have units m^{-1}; σ_u, σ_v and σ_ω have units $\mathrm{m}^{-1/2}$; ϕ is dimensionless. Here m stands for month. Bottom: the parameters from Chen and Majda (2015b), predicting the two boreal summer intraseasonal oscillation (BSISO) modes obtained by applying NLSA to the OLR dataset, the Cloud Archive User Service (CLAUS) Version 4.7. The comparison between Section 6c. 39

	d_u	d_v	d_ω	σ_u	σ_v	σ_ω	γ	a	f_0	f_t	ω_f	ϕ
I). Optimal parameters	0.8	0.6	0.5	0.5	0.5	0.7	0.3	4.1	1.0	4.7	$2\pi/12$	-2
II). OLR-based parameters	0.9	0.9	0.5	0.3	0.8	1.0	0.3	4.25	1.0	4.0	$2\pi/12$	-1.4

TABLE 1. Top: Optimal parameters for the nonlinear low-order stochastic model (1). The parameters $d_u, a, \gamma, f_0, f_t, \omega_f, d_v$ and d_ω have units m^{-1}; σ_u, σ_v and σ_ω have units m$^{-1/2}$; ϕ is dimensionless. Here m stands for month. Bottom: the parameters from Chen and Majda (2015b), predicting the two boreal summer intraseasonal oscillation (BSISO) modes obtained by applying NLSA to the OLR dataset, the Cloud Archive User Service (CLAUS) Version 4.7. The comparison between Section 6c.

LIST OF FIGURES

Fig. 1. Left: MISO indices from NLSA ranging from 1998/01/01 to 2013/12/31. The period up to 2007/12/31 is utilized as the training period in Section 3a for model calibration and the prediction skill is tested for year 2008 to 2013 in Section 4. Right: the highly non-Gaussian probability density functions (PDFs) with fat-tails of the MISO indices in the training period, where the sub-panel shows the PDF in logarithm scale. 42

Fig. 2. Comparison of the statistics of the MISO indices and those from the nonlinear low-order stochastic model (1)–(2) with optimal parameters in Table 1. (a) Long-term autorrelation and cross-correlation function up to 2 years. (b) Short-term autocorrelation function and cross-correlation functions up to 3 months. (c) PDFs of the MISO indices and the model signals. (d) Power spectrum of MISO indices and model signals. 43

Fig. 3. Skill scores with RMS error and pattern correlation for prediction utilizing ensemble mean in different years. The dashed line in the top panel shows the standard deviation of the equilibrium distribution associated with the MISO indices and that in the bottom panel shows the threshold of Corr = 0.5. Useful predictions are defined by 1) the RMS error in the prediction is less than the standard deviation of the truth at the equilibrium and 2) the pattern correlation between the predicted signal and the truth is above 0.5. 44

Fig. 4. Prediction of MISO 1 at 15- and 25-day lead utilizing ensemble mean. 45

Fig. 5. Long-range prediction of MISO 1 for 8 months starting from different dates. April 1 is a time at the transition between the quiescent phase and the active phase of the MISO indices; June 1 is a starting date in the active mature phase while October 1 is a starting date in the decaying phase of MISO activity. 46

Fig. 6. Phase diagrams of MISO 1 and MISO 2 (blue) and the corresponding predictions [ensemble mean (red) and 50 ensemble members (green)], starting from 2009/07/01, 2008/06/01 and 2013/06/01 and lasting for 30 days. Blue dots and small red rectangles indicate the truth and prediction for every 5 days. 47

Fig. 7. Skill score with pattern correlation (left) and RMS error (right) for predicting the reconstructed spatiotemporal patterns as a function of lead time (days) for July 2009 (black), June 2008 (red) and June 2013 (green). 48

Fig. 8. Reconstruction and prediction of the spatiotemporal patterns of July 2009 starting from 1 July 2009. The prediction at day 1, 4, 8, 12, 16, 20, 24 and 30 are shown. 49

Fig. 9. Same as Figure 8 but for June 2008. 50

Fig. 10. Comparison of the truth (first row), the approximated spatiotemporal patterns from NLSA based on (9) (second row) and the patterns in which the spatial basis is obtained based on the modes by applying running average to the raw data as described in Section 5 (third row). . 51

Fig. 11. Comparison of the statistics based on the full period of MISO indices (1998-2013; black), training period utilized in Section 3a (1998-2007; red) and the 3-year short training period in Section 6a (1998-2000; green). 52

Fig. 12. Left: Skill score with RMS error and pattern correlation for predicting the MISO indices in different years from 2001 to 2007. As in Figure 3, the two dashes lines indicate the standard deviation of the MISO indices at climatology and the value with Corr = 0.5, which serve

as the threshold for the useful prediction. Right: 25-day lead prediction of MISO 1 at four different years. The model parameters are listed in the top row of Table 1. 53

Fig. 13. Comparison of the indices, obtained by applying NLSA with different lagged embedding window sizes ($q = 64, 48$ and 34 days), and the associated statistics. 54

Fig. 14. Comparison of the prediction skill of the indices obtained by applying NLSA with different lagged embedding window sizes (top: $q = 64$; middle: $q = 48$; bottom: $q = 34$). Left: useful prediction for the full year and for only the boreal summer time (June to September). Right: 15-day lead prediction for the indices obtained with different lagged embedding window sizes. 55

Fig. 15. Applying the parameters in Chen and Majda (2015b) to calibrate and predict the MISO precipitation indices. Panel (a) is the autocorrelation of u_1 up to 2 years. Panels (b) are for the autocorrelation of u_1 and crosscorrelation between u_1 and u_2 up to 3 months. Panels (c) are for the PDF in linear and logarithm scales, respectively. Panel (d) compares the power spectrum of u_1. Panels (e) and (f) compare the 25-day lead ensemble mean predictions with the optimal parameters and the parameters borrowed from OLR dataset in Chen and Majda (2015b). 56

Fig. A1. Sensitive test. Left: Model error (via information distance) as functions of different parameters. Right: RMS error in the autocorrelation functions as functions of those parameters related to the phase. The black dots indicate the optimal parameters as listed in Table 1. The two dotted lines in each panel indicate the range of randomly-picked suboptimal parameters in the test of prediction. 57

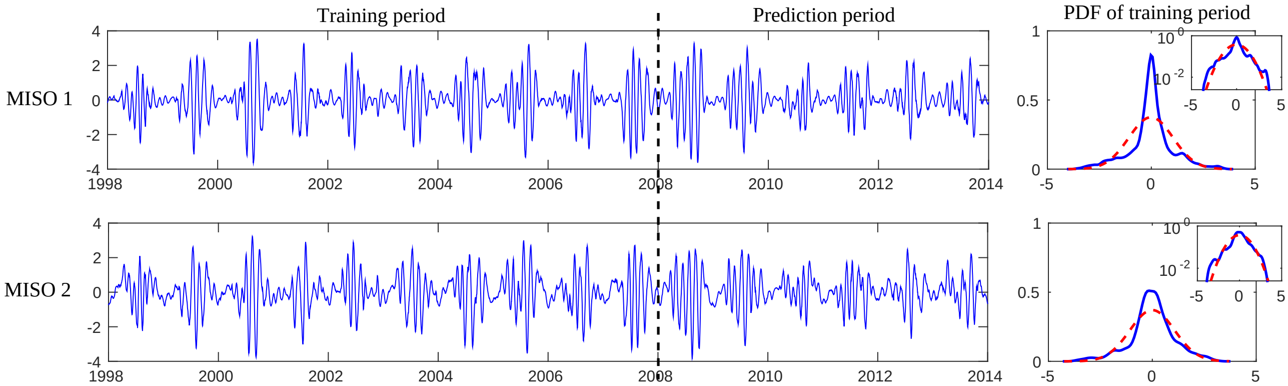


FIG. 1. Left: MISO indices from NLSA ranging from 1998/01/01 to 2013/12/31. The period up to 2007/12/31 is utilized as the training period in Section 3a for model calibration and the prediction skill is tested for year 2008 to 2013 in Section 4. Right: the highly non-Gaussian probability density functions (PDFs) with fat-tails of the MISO indices in the training period, where the sub-panel shows the PDF in logarithm scale.

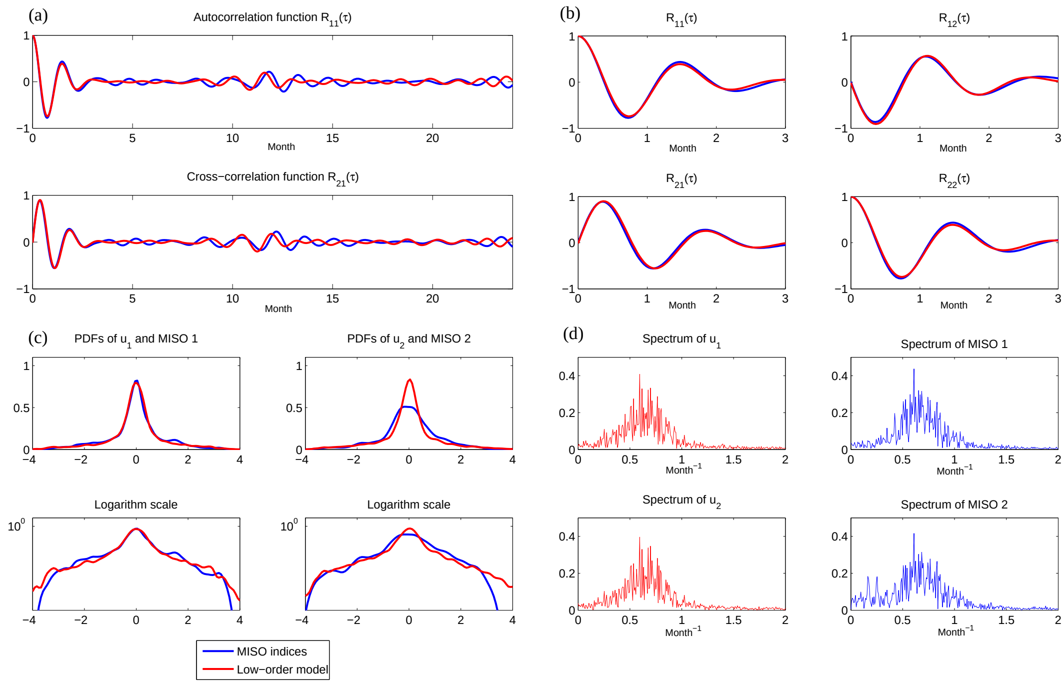
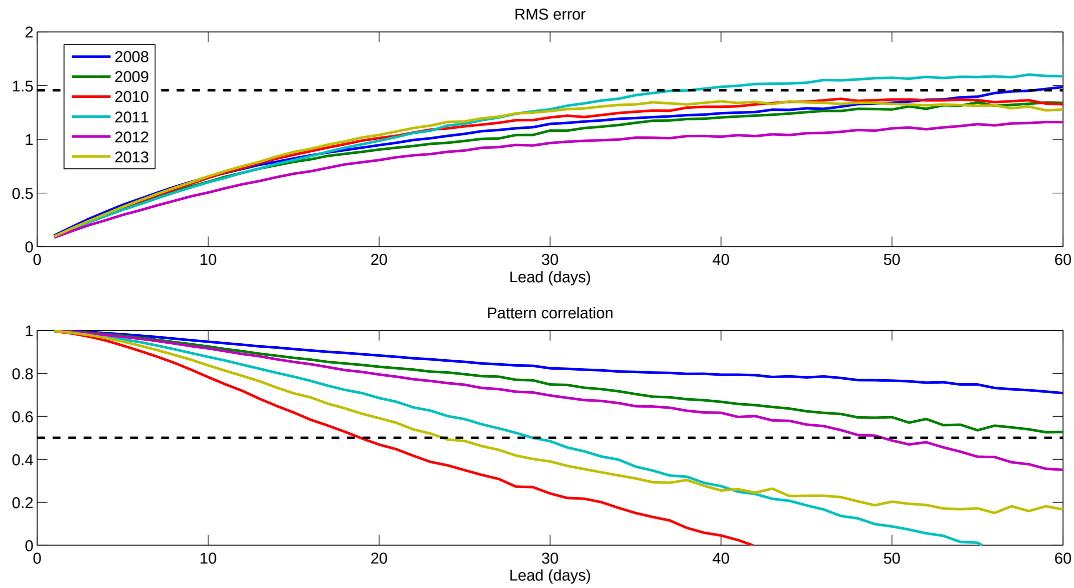


FIG. 2. Comparison of the statistics of the MISO indices and those from the nonlinear low-order stochastic model (1)–(2) with optimal parameters in Table 1. (a) Long-term autorrelation and cross-correlation function up to 2 years. (b) Short-term autocorrelation function and cross-correlation functions up to 3 months. (c) PDFs of the MISO indices and the model signals. (d) Power spectrum of MISO indices and model signals.



806 FIG. 3. Skill scores with RMS error and pattern correlation for prediction utilizing ensemble mean in different
807 years. The dashed line in the top panel shows the standard deviation of the equilibrium distribution associated
808 with the MISO indices and that in the bottom panel shows the threshold of Corr $= 0.5$. Useful predictions are
809 defined by 1) the RMS error in the prediction is less than the standard deviation of the truth at the equilibrium
810 and 2) the pattern correlation between the predicted signal and the truth is above 0.5.

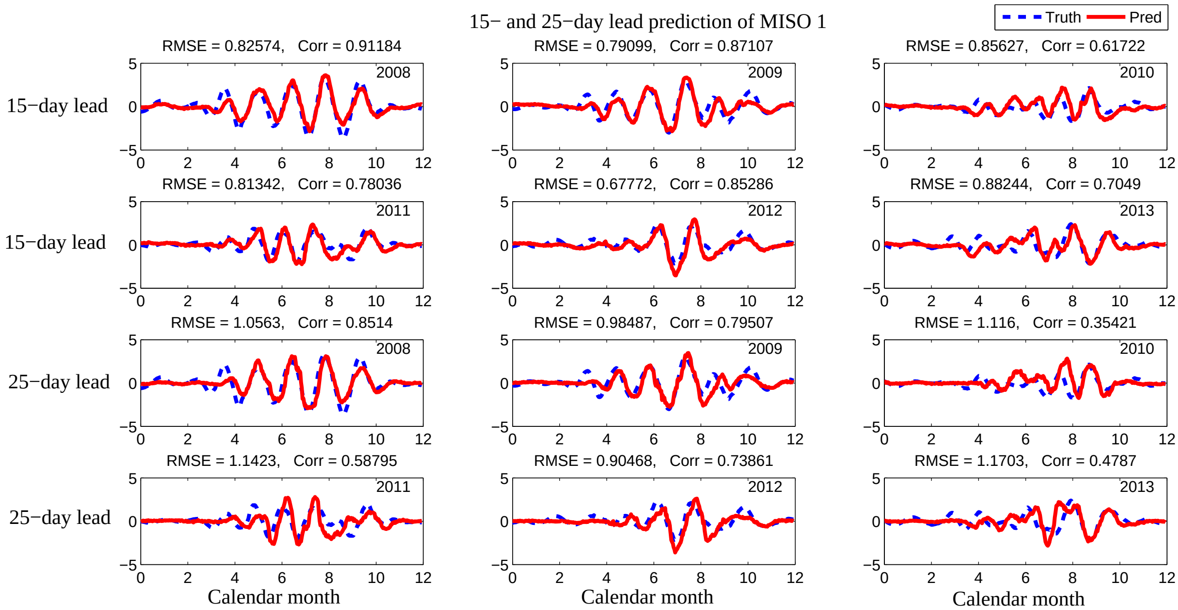


FIG. 4. Prediction of MISO 1 at 15- and 25-day lead utilizing ensemble mean.

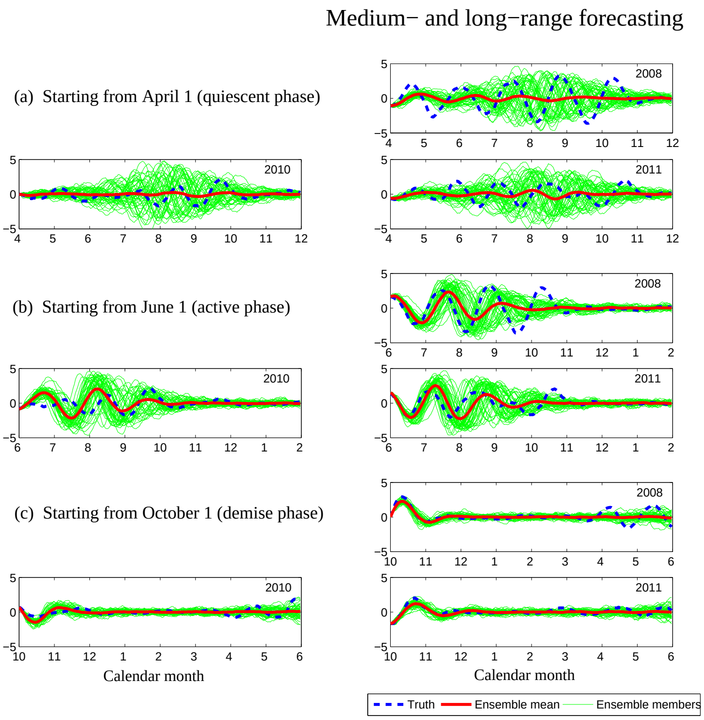


FIG. 5. Long-range prediction of MISO 1 for 8 months starting from different dates. April 1 is a time at the transition between the quiescent phase and the active phase of the MISO indices; June 1 is a starting date in the active mature phase while October 1 is a starting date in the decaying phase of MISO activity.

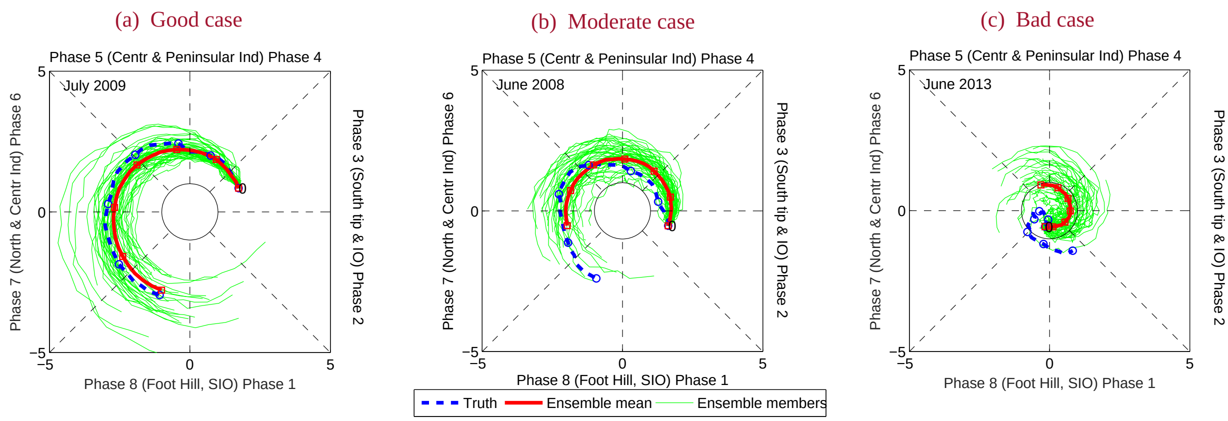
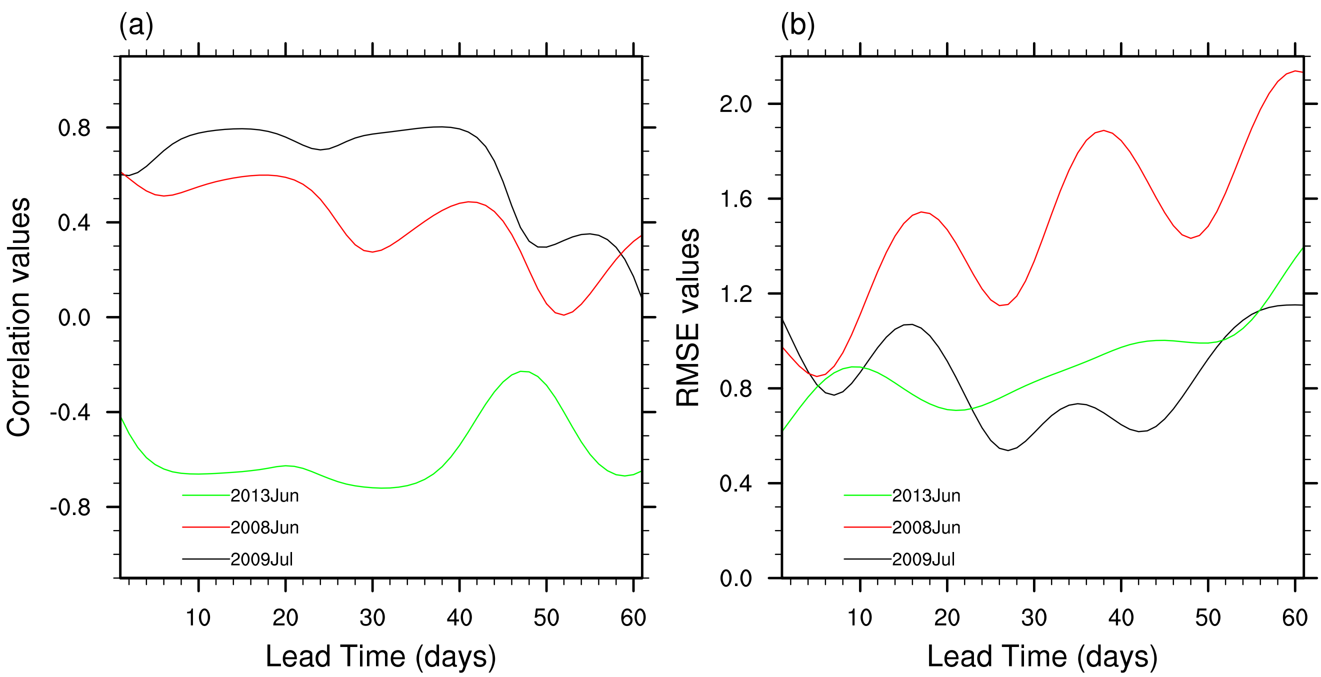


FIG. 6. Phase diagrams of MISO 1 and MISO 2 (blue) and the corresponding predictions [ensemble mean (red) and 50 ensemble members (green)], starting from 2009/07/01, 2008/06/01 and 2013/06/01 and lasting for 30 days. Blue dots and small red rectangles indicate the truth and prediction for every 5 days.



817 FIG. 7. Skill score with pattern correlation (left) and RMS error (right) for predicting the reconstructed
818 spatiotemporal patterns as a function of lead time (days) for July 2009 (black), June 2008 (red) and June 2013
819 (green).

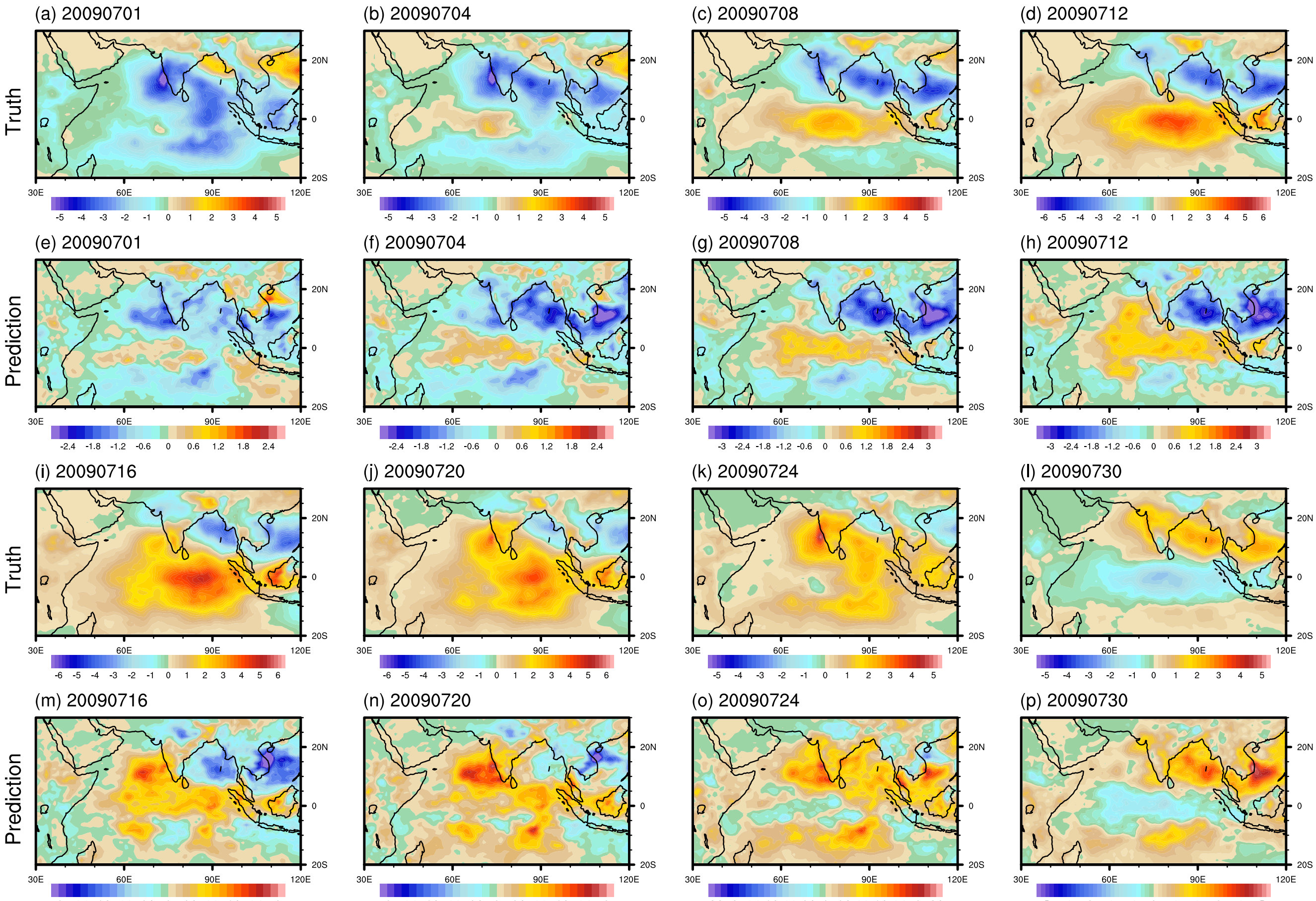


FIG. 8. Reconstruction and prediction of the spatiotemporal patterns of July 2009 starting from 1 July 2009. The prediction at day 1, 4, 8, 12, 16, 20, 24 and 30 are shown.

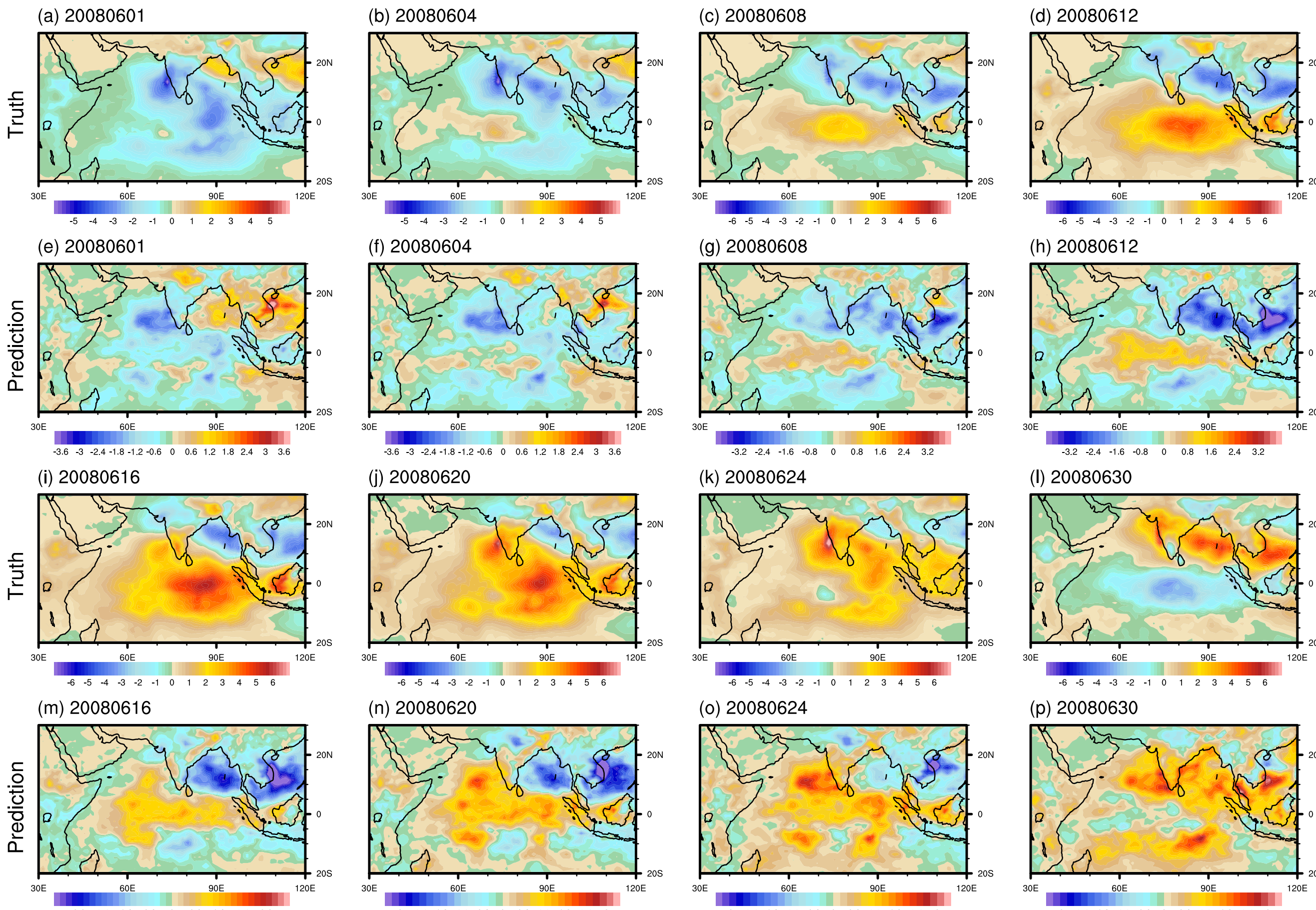
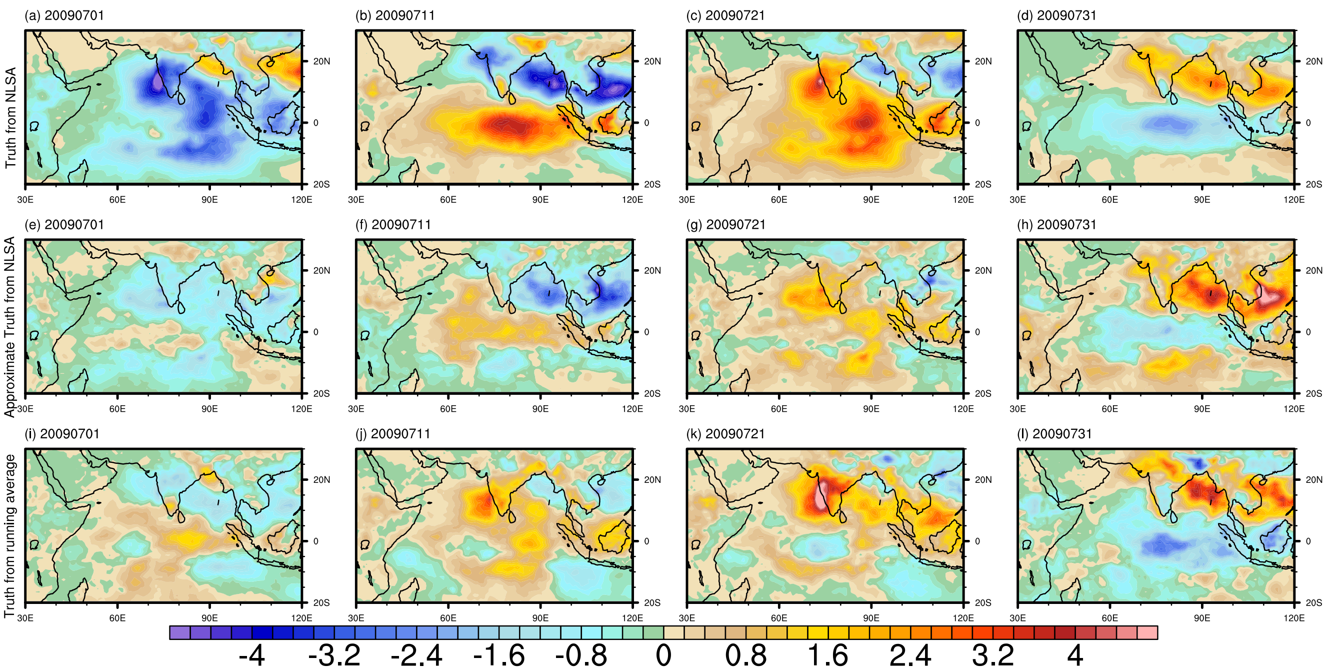


FIG. 9. Same as Figure 8 but for June 2008.



822 FIG. 10. Comparison of the truth (first row), the approximated spatiotemporal patterns from NLSA based on
823 (9) (second row) and the patterns in which the spatial basis is obtained based on the modes by applying running
824 average to the raw data as described in Section 5 (third row).

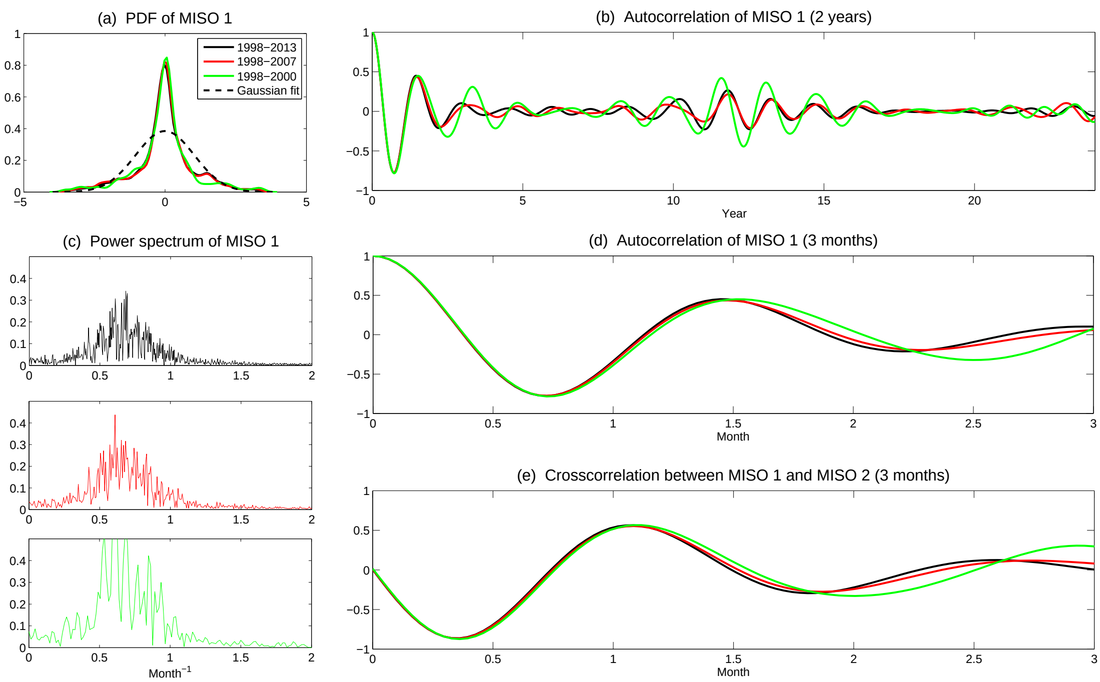


FIG. 11. Comparison of the statistics based on the full period of MISO indices (1998-2013; black), training period utilized in Section 3a (1998-2007; red) and the 3-year short training period in Section 6a (1998-2000; green).

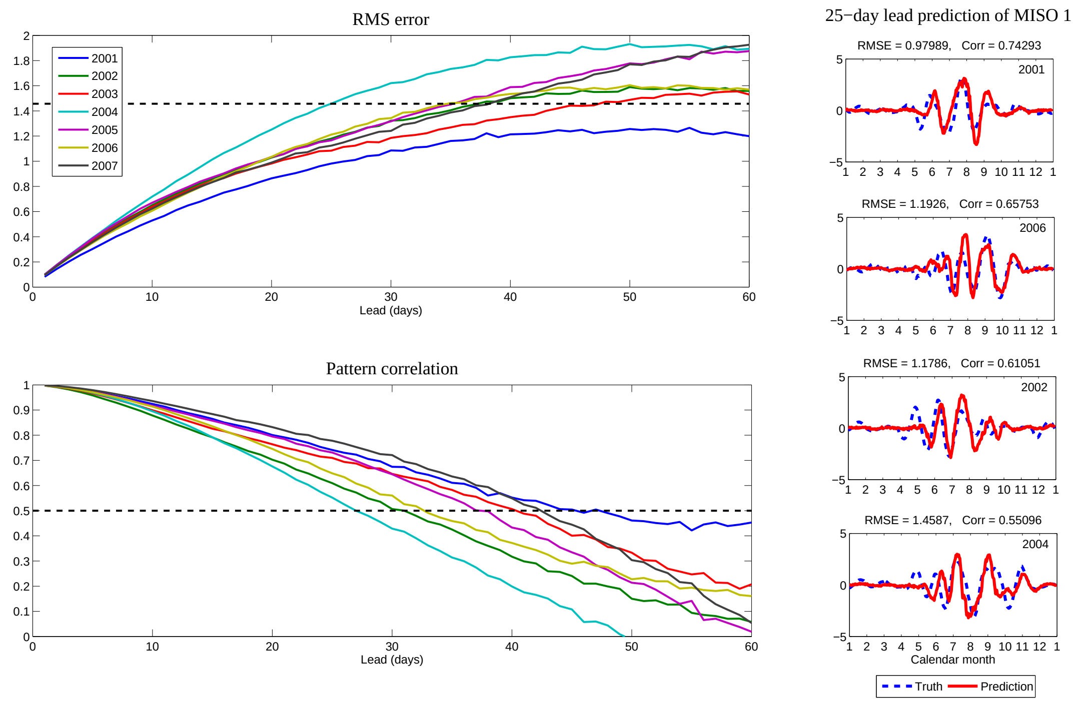
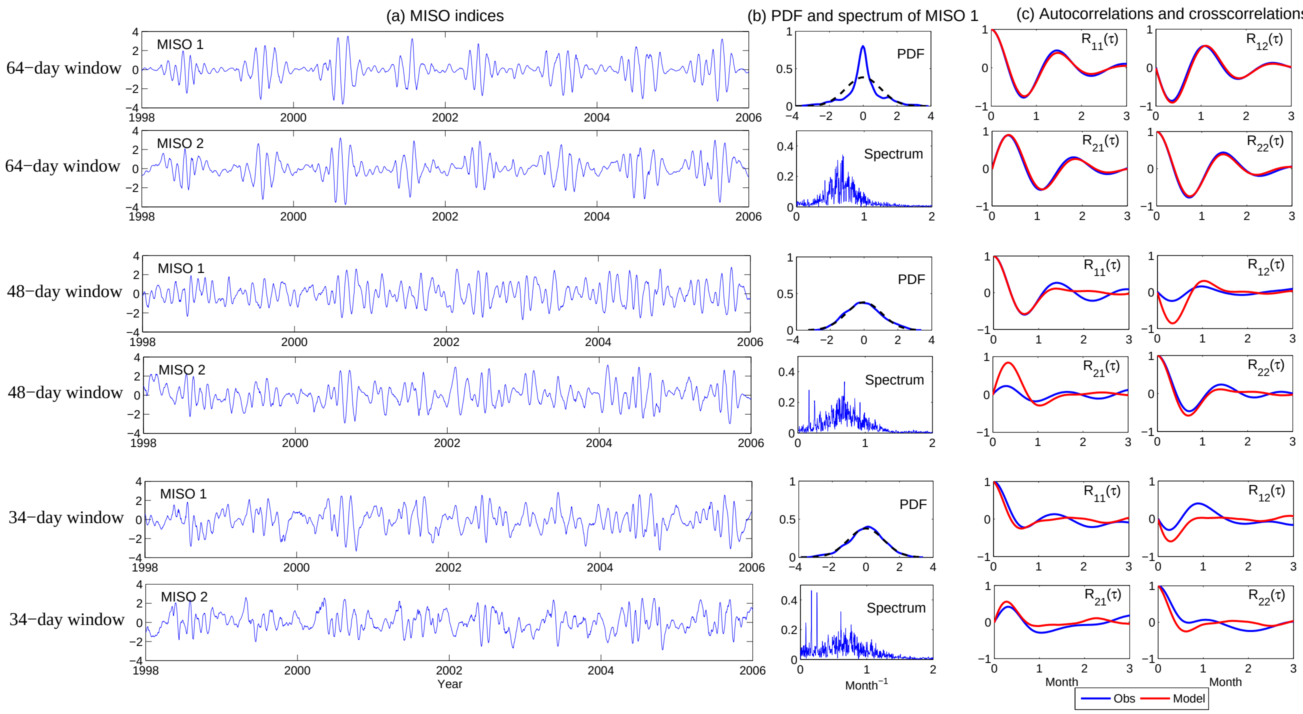


FIG. 12. Left: Skill score with RMS error and pattern correlation for predicting the MISO indices in different years from 2001 to 2007. As in Figure 3, the two dashes lines indicate the standard deviation of the MISO indices at climatology and the value with Corr = 0.5, which serve as the threshold for the useful prediction. Right: 25-day lead prediction of MISO 1 at four different years. The model parameters are listed in the top row of Table 1.



833 FIG. 13. Comparison of the indices, obtained by applying NLSA with different lagged embedding window

834 sizes ($q = 64, 48$ and 34 days), and the associated statistics.

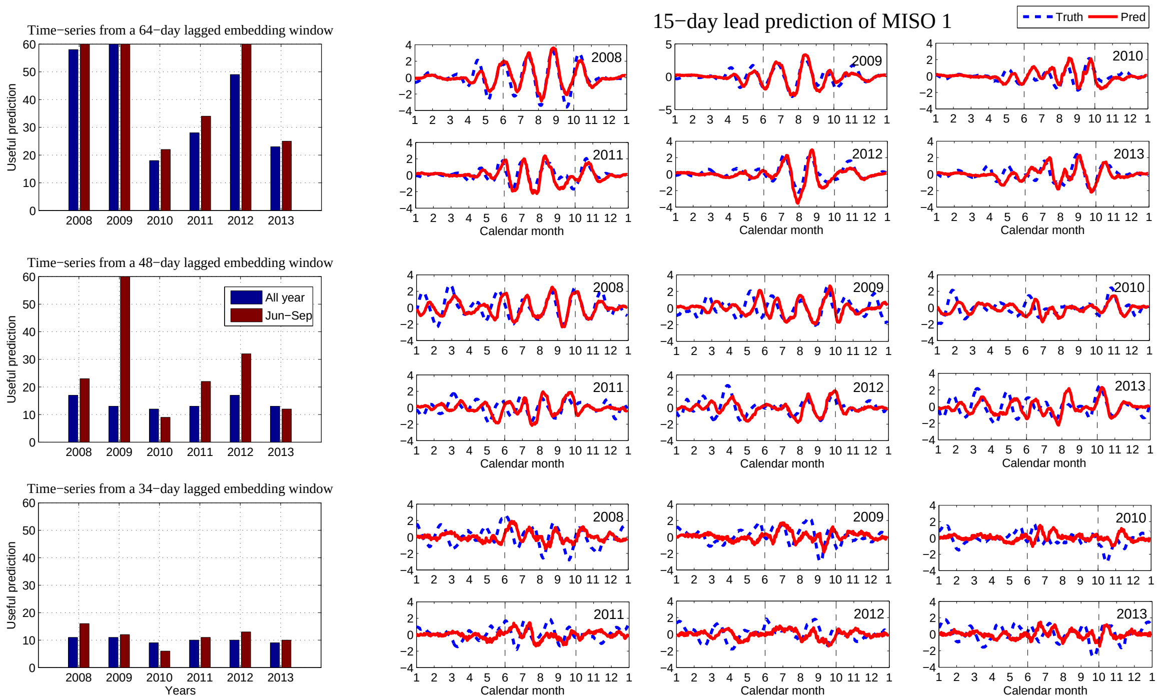


FIG. 14. Comparison of the prediction skill of the indices obtained by applying NLSA with different lagged embedding window sizes (top: $q = 64$; middle: $q = 48$; bottom: $q = 34$). Left: useful prediction for the full year and for only the boreal summer time (June to September). Right: 15-day lead prediction for the indices obtained with different lagged embedding window sizes.

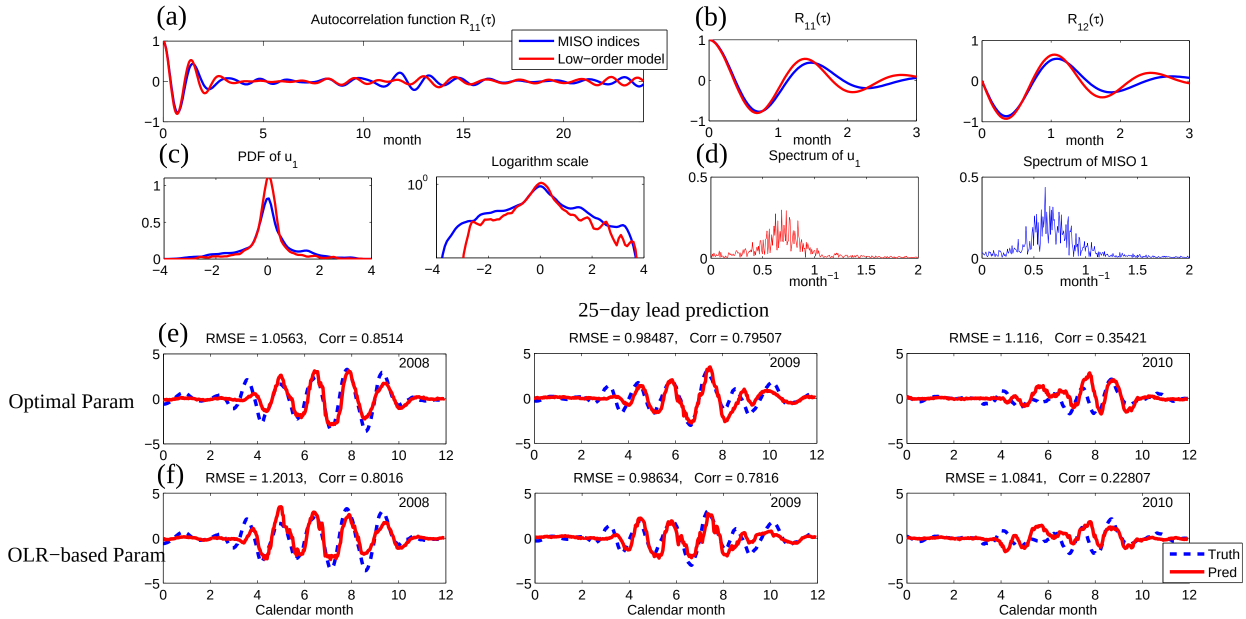
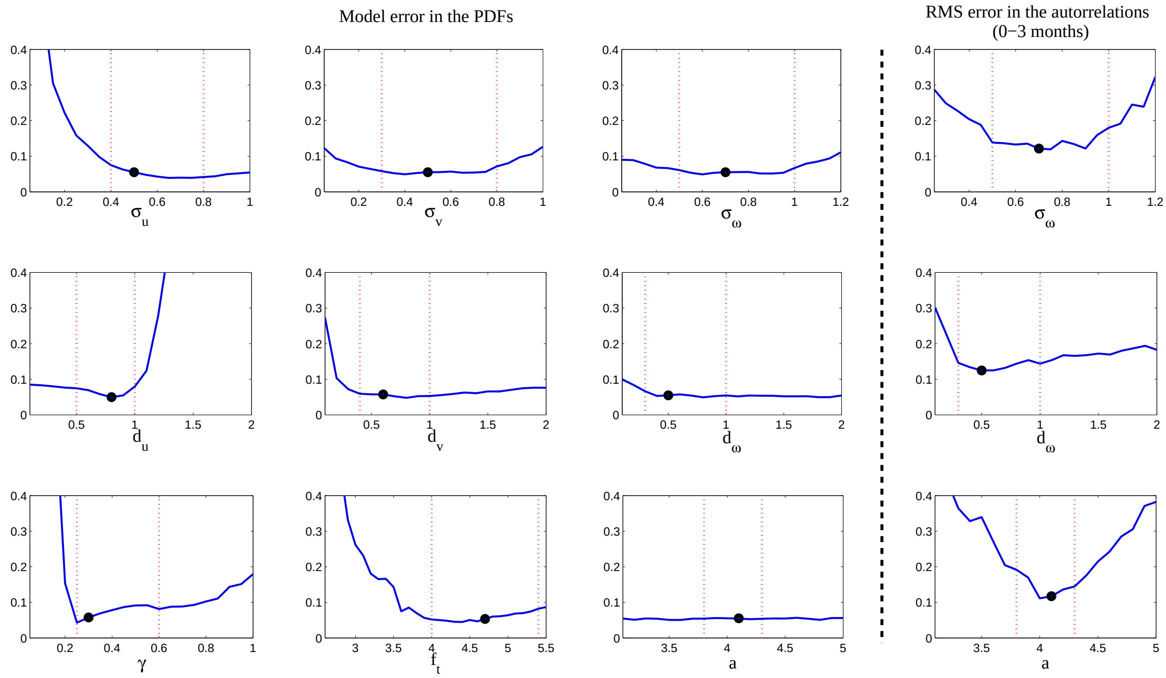


FIG. 15. Applying the parameters in Chen and Majda (2015b) to calibrate and predict the MISO precipitation indices. Panel (a) is the autocorrelation of u_1 up to 2 years. Panels (b) are for the autocorrelation of u_1 and crosscorrelation between u_1 and u_2 up to 3 months. Panels (c) are for the PDF in linear and logarithm scales, respectively. Panel (d) compares the power spectrum of u_1. Panels (e) and (f) compare the 25-day lead ensemble mean predictions with the optimal parameters and the parameters borrowed from OLR dataset in Chen and Majda (2015b).



845 Fig. A1. Sensitive test. Left: Model error (via information distance) as functions of different parameters.
846 Right: RMS error in the autocorrelation functions as functions of those parameters related to the phase. The
847 black dots indicate the optimal parameters as listed in Table 1. The two dotted lines in each panel indicate the
848 range of randomly-picked suboptimal parameters in the test of prediction.