Objective Identification of Nonlinear Convectively Coupled Phases of Monsoon Intraseasonal Oscillation: Implications for Prediction

R. CHATTOPADHYAY, A. K. SAHAI, AND B. N. GOSWAMI

Indian Institute of Tropical Meteorology, Pune, India

(Manuscript received 22 March 2007, in final form 20 September 2007)

ABSTRACT

The nonlinear convectively coupled character of the summer monsoon intraseasonal oscillation (ISO) that manifests in its event-to-event variations is a major hurdle for skillful extended-range prediction of the active/break episodes. The convectively coupled character of the monsoon ISO implies that a particular nonlinear phase of the precipitation ISO is linked to a unique pattern of the large-scale variables. A methodology has been presented to capture different nonlinear phases of the precipitation ISO using a combination of a sufficiently large number of dynamical variables. This is achieved through a nonlinear pattern recognition technique known as self-organizing map (SOM) involving six daily large-scale circulation indices. It is demonstrated that the nonlinearly classified states of the large-scale circulation isolated at the SOM nodes without involving any information on rainfall are strongly linked to different phases of evolution of the rainfall ISO, including the active and break phases. While a lower SOM classification involving 9 different states identify the composite phases of the rainfall ISO, a higher SOM classification involving 81 states can identify different shades of composite phase of the rainfall ISO. The concept of isolating the nonlinear states, as well as the technique of doing so, is robust as almost identical phases of precipitation ISO are identified by the large-scale circulation indices derived from two different reanalysis datasets, namely, the 40-yr ECMWF Re-Analysis (ERA-40) and the NCEP–NCAR reanalysis.

The ability of the SOM technique to isolate spatial structure and evolutionary history of nonlinear convectively coupled states of the summer monsoon ISO opens up a new possibility of extended-range prediction of summer monsoon ISO. This knowledge is used to develop an analog technique for predicting different phases of monsoon ISO. Skillful four-pentad lead prediction of rainfall over central India is demonstrated with the model using only large-scale circulation fields. A major strength of the model is that it can easily be used for real-time extended-range prediction of monsoons.

1. Introduction

Vigorous intraseasonal oscillations (ISOs) in the form of active and break episodes are integral part of the Indian summer monsoon (see Goswami 2005 for a review). Prediction of the active and break episodes 2–3 weeks in advance is of great importance as sowing, harvesting, and water management for agriculture within the season depends crucially on the rainfall associated with these phases of the monsoon ISOs. Initially described in terms of rainfall (Ramaswamy 1962; Ramamurthy 1969) over India, the mean spatial structure of rainfall and circulation fields associated with active and break conditions (Krishnamurti and Subrahmanyam 1982; Krishnamurti et al. 1985; Webster et al. 1998; Krishnan et al. 2000; Annamalai and Slingo 2001; Annamalai and Sperber 2005; Goswami 2005) have very large spatial scale extending far beyond the Indian continent. One important character of these intraseasonal spells is the repeated northward propagation of the zonally oriented cloud band from south of equator to about 25°N in this region (Sikka and Gadgil 1980; Yasunari 1979). Further, the nonlinear relationship between the rainfall and the large-scale circulation indicates that the active–break spells are related to a convectively coupled oscillation consistent with theory (Goswami and Shukla 1984; Jiang et al. 2004; Wang 2005). This underlying large-scale spatial pattern together with relatively slow evolution has led to the optimism for extended-range prediction of these phases of the monsoon ISOs (Goswami and Xavier 2003; Web-

Corresponding author address: Dr. A. K. Sahai, Climate and Global Modeling Division, Indian Institute of Tropical Meteorology, Pune, 411 008, Maharashtra, India.
E-mail: sahai@tropmet.res.in

DOI: 10.1175/2007JAS2474.1

ster and Hoyos 2004). However, these subseasonal fluctuations have considerable event-to-event and year-to-year variability that limits the predictability of the active and break phases. The event-to-event variability (i.e., the variability of rainfall intensity and duration over a large region) of these phases is a result of the quasiperiodic nature of the monsoon. The quasiperiodic character, in turn, is a manifestation of the nonlinearity of the convectively coupled ISO. Our recognition that the monsoon ISO is nonlinear is based on the observation that the monsoon ISO is quasiperiodic. In other words, it has a broadband spectrum with two major periodicities around 15 and 40 days but appreciable power at all frequencies with periods between 10 and 80 days. This broadband nature of the frequency spectrum may be due to nonlinear interaction between the dominant periodicities [for which different physical models of scale selection exist, e.g., Chatterjee and Goswami (2004); Goswami and Shukla (1984)] and higher and lower periodicities. The linear prediction techniques, such as regression (Goswami and Xavier 2003; Webster and Hoyos 2004), use the averaged spatial structure and averaged evolution of the oscillatory component and fail in predicting the event-to-event variability of the quasiperiodic oscillation. Improvement of the skill of prediction of the active and break spells can come only if one could objectively identify and characterize different shades of each phase of the oscillation (e.g., active and break conditions) and their evolution. Such an objective method is presented in this study.

The primary manifestation of the Indian summer monsoon ISO being the rainfall fluctuations (active–break cycles), it can be described by a rainfall index [precipitation (PR) index] constructed from rainfall [India Meteorological Department (IMD) high-resolution gridded data, Rajeevan et al. (2006)] averaged over the monsoon trough region (15°–25°N, 70°–85°E). The normalized PR index for two arbitrarily selected years is shown in Fig. 1. A standardized anomaly of the PR index greater (less) than +1 (−1) is associated with active (break) situations. The state vector of the atmosphere associated with each phase of the nonlinear convectively coupled monsoon ISO would generally have large dimension. To isolate the distinct phases of the nonlinear oscillation, therefore, a number of atmospheric parameters would be required. To achieve this goal, we identify a large number of dynamical indices that are related to the monsoon precipitation ISO. First, three such indices are obtained from the fact that the dominant monsoon ISO has large spatial scale such that both the intraseasonal and interannual variability of the seasonal mean are governed by a common spatial mode of variability (Ferranti et al. 1997; Goswami and Ajaya Mohan 2001; Sperber et al. 2000; Lawrence and Webster 2001). The seasonal-mean Indian summer monsoon and its variability are often described by certain large-scale circulation indices, such as based on vertical shear of zonal wind [the WY index (Webster and Yang 1992)], vertical shear of meridional wind [the GO index (Goswami et al. 1999)], and meridional shear of zonal wind [the WF index (Wang and Fan 1999)]. As a result of similarity between the spatial structures of the seasonal mean and the dominant ISO mode, the large-scale indices of the seasonal mean could also be used as indices of the dominant ISO mode (e.g., Flatau et al. 2001). We use the daily values of these indices to represent the monsoon ISO except that the area averaging for the GO index has been extended to 10°S, keeping in mind the meridional extent of the dominant ISO mode (Goswami 2005). These three indices seem to represent the large-scale low frequency component of monsoon ISO, namely the 30–60-day mode. However, a significant part of monsoon ISO variance is contributed by the 10–20-day mode having relatively smaller spatial structure (Chatterjee and Goswami 2004). To capture this component of ISO variability, we introduce three other indices of local character. They are chosen from mean sea level pressure (MS index), specific humidity (SH index), and geopotential height (GP index). The averaging regions for the last three indices are chosen (Table 1) based on a high simultaneous correlation with the PR index during the summer monsoon period (1 June–30 September). The indices used in the study and their averaging areas are shown in Table 1. These indices constructed from daily reanalysis [the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40); Uppala et al. 2005] during the summer monsoon season (1 June–30 September) and normalized by their individual standard deviation are shown for two summers in Fig. 1. For a linear convectively coupled oscillation, the PR index and the circulation indices should have high linear correlation between them (maybe with some lead or lag) and the phase relationship between the two should remain fixed for all events. As seen from Fig. 1, while there is a linear correlation between PR and circulation indices, it is weak and the phase relationship between them changes from event to event. This is further illustrated in Fig. 2 where scatterplots between the circulation indices and the PR index are shown. The mean of scatter for each bin of PR increases linearly but tends to flatten and saturate for larger values of PR. This clearly indicates a certain degree of nonlinearity in the relationship.

Based on the evidence of nonlinearity in the relationship between precipitation and circulation, we propose

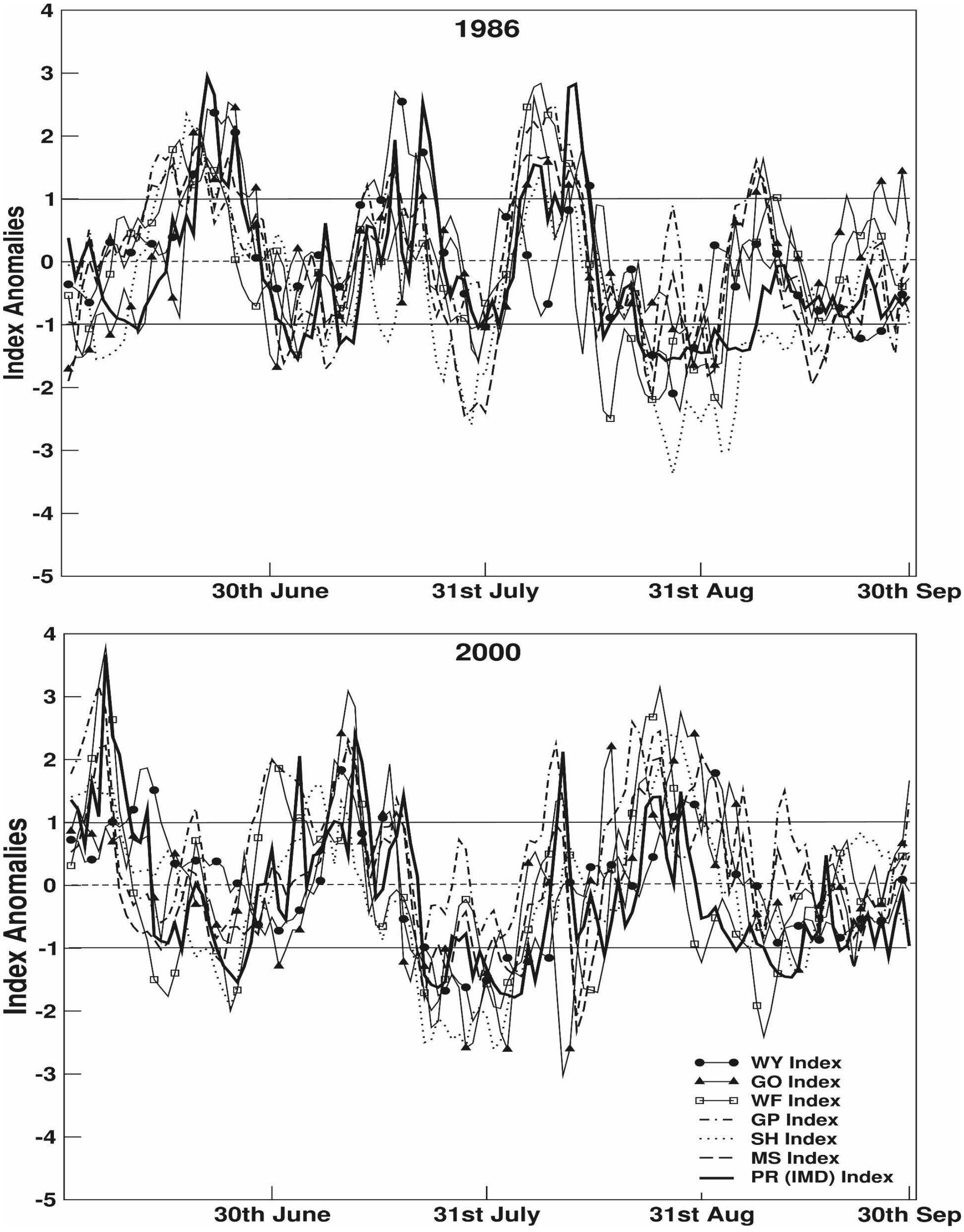

FIG. 1. Time series of all the indices of summer monsoon during the summers (1 Jun–30 Sep) of 1986 and 2000 (the years are arbitrarily chosen). The mean sea level pressure and the geopotential height are multiplied by −1.0 to make them comparable with other indices.

that the monsoon ISOs are nonlinear convectively coupled oscillations. Such a quasiperiodic or nonlinear monsoon ISO could be described by a sinusoidal oscillation, each phase of which has a spread. In other words, each phase, such as the active or break phases, has a different shade. The evolutionary history of each shade of a given oscillation may be different. As a long history of the IMD daily rainfall over India as well as large-scale daily circulation data [e.g., National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis and ERA-40] are currently available, an effective analog

TABLE 1. List of all indices and the corresponding regions used to define these area-averaged indices.

1	PR index: rainfall (15°–25°N, 70°–85°E)
2	GO index: V_{850}(10°S–30°N, 70°–110°E) − V_{200}(10°S–30°N, 70°–110°E)
3	WF index: U_{850}(5°–15°N, 40°–80°E) − U_{850}(20°–30°N, 60°–90°E)
4	WY index: U_{850}(0°–20°N, 40°–110°E) − U_{200}(0°–20°N, 40°–110°E)
5	MS index: MSL (15°–25°N, 65°–95°E)
6	SH index: SH_{850} (15°–25°N, 65°–95°E)
7	GP index: GP_{500} (10°–20°N, 65°–95°E)
8	*U*-shear index: U_{850}(15°S–5°N, 100°–140°E) − U_{200}(15°S–5°N, 100°–140°E)
9	Omega (vertical velocity at 500 mb) index: ω_{500}(0°–7.5°N, 50°–115°E) − ω_{500}(10°–20°N, 80°–150°E)
10	Mean sea level pressure shear index: MSL(10°–20°N, 110°–150°E) − MSL(15°S–5°N, 40°–60°E)

prediction for monsoon ISO could be developed, if different shades of each phase and their evolutionary history could be characterized uniquely. The convectively coupled character of the oscillation indicates that each shade of a given phase is characterized by a unique relationship between the circulation indices that, in turn, would be uniquely related to that shade of the PR index. Based on this conceptual framework, a methodology is presented in this study to identify different shades of active and break (as well as other) phases of the monsoon ISO entirely from large-scale circulation indices, and it is demonstrated that they are uniquely and strongly connected to shades of the rainfall index. This is accomplished by using a classification scheme based on unsupervised learning artificial neural network technique known as self-organizing map (SOM) (Kohenen 1990). The technique essentially brings out a series of nonlinearly coupled states described by different unique combinations of the large-scale indices. We believe that the technique and methodology presented in this study will establish a strong base for real-time prediction of active and break spells of the Indian monsoon. The data used in the study is described in section 2, and the SOM algorithm and how it is applied for this study is described in section 3. In section 4, we show that, even with a limited number of degrees of freedom in the classification scheme based *only* on the large-scale circulation fields, it is possible to identify different phases of monsoon ISO in rainfall in a quantitative way. In section 5, we show that, with increased degrees of

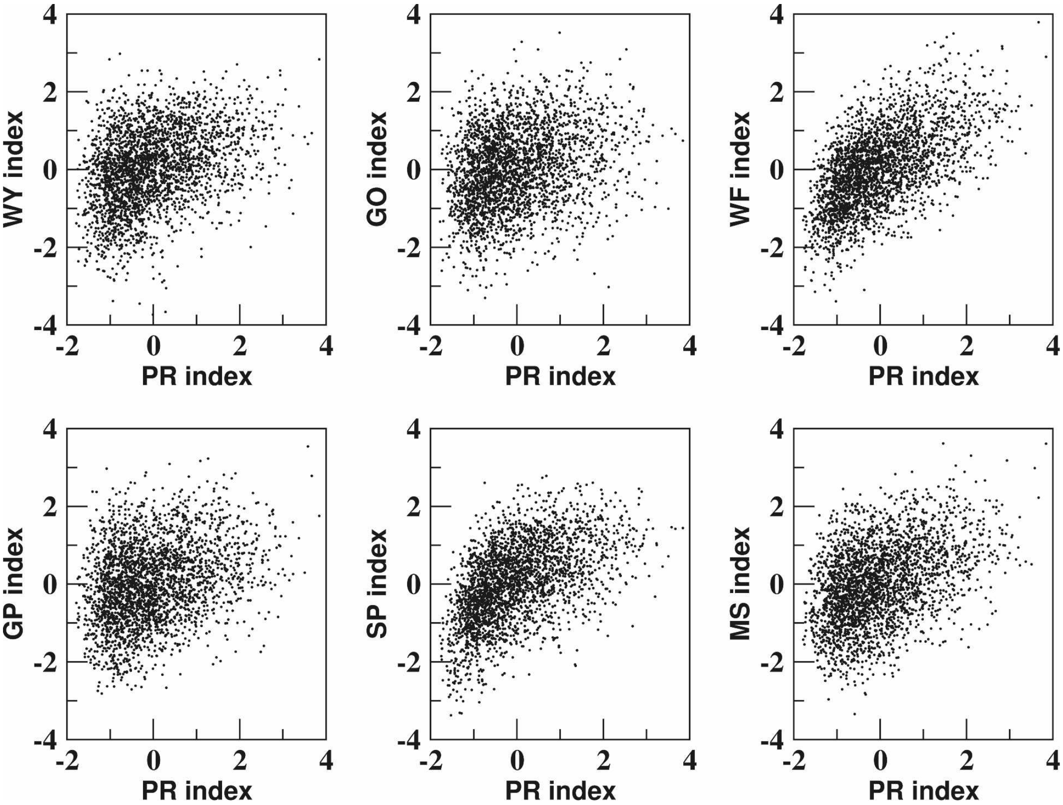

FIG. 2. Scatterplots of the large-scale circulation indices of monsoon ISO for the 22 summers of ERA-40 data (1980–2001) vs the PR index from the IMD rainfall data for the same period.

freedom in the classification scheme, different shades of each phase of ISO could be identified and the composite evolutionary history of each shade is constructed from the 54 yr of NCEP–NCAR reanalysis data (1951–2004). Using this knowledge, an analog technique for real-time prediction of rainfall over central India is described in section 6 and skillful prediction up to the fourth pentad in advance is demonstrated. The results are summarized in section 7.

2. Data

The six indices used in the SOM classification (numbers 2–7, Table 1) are constructed from the ERA-40 (Uppala et al. 2005; additional information is available online at http://www.ecmwf.int/research/era/) dataset for a period of 22 yr (1980–2001). To train the SOM nodes corresponding to different shades of the active and break phases a large sample size is desirable. For this purpose, we use the NCEP–NCAR 54 yr (1951–2004) of data (Kalnay et al. 1996). The NCEP–NCAR reanalysis data is also used because the latest ERA data is not available (after 2002) and to develop an operational prediction scheme the NCEP–NCAR reanalysis data, which is updated regularly, is more suitable. The IMD daily gridded high-resolution rainfall data based on observations from about 1803 rain gauge stations between 1951 and 2004 (Rajeevan et al. 2006) is used for comparison and validation purposes. The rainfall index (PR index) is constructed based on this data. The spatial pattern associated with different phases of the PR index is also obtained from this dataset. Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP) pentad rainfall (Xie and Arkin 1997) data without model enhancement, interpolated to daily values, is also used to see the evolution of large-scale rainfall patterns associated with the SOM nodes. Daily anomalies of the variables involved in different indices are obtained with respect to the corresponding daily smoothed climatology. The daily climatology is calculated based on the length of data used, namely 22 yr for ERA-40 data, 54 yr for NCEP–NCAR reanalysis data, and so on. The smoothed daily climatology is obtained by applying a 5-day running mean to the daily climatology. Then we constructed the daily standardized anomaly of each of the six indices based on the daily mean and standard deviation for both ERA-40 and NCEP–NCAR reanalysis data. We use the ERA-40 circulation data to develop the SOM technique and to demonstrate its strength in identifying different phases of monsoon ISO in precipitation. The robustness of the technique is further established when we show that the results are independent of the reanalysis data used and that almost identical results are obtained even with the NCEP–NCAR reanalysis dataset. For the prediction purpose, we also constructed three more indices (last three indices in Table 1) from NCEP–NCAR reanalysis data. The reason for choosing these three indices and their application will be described in section 6.

3. Methodology

The SOM algorithm in brief

The self-organizing map is basically a pattern recognition technique or cluster algorithm based on unsupervised learning neural networks (i.e., the learning process without prior knowledge of the data domain or human intervention). This method is similar to standard iterative clustering algorithms such as *k*-means clustering (see, e.g., Gutiérrez et al. 2004 for more details). In this study we will use the Kohenen model (Kohenen 1990) of SOM which belongs to the class of vector coding algorithms (Haykin 1999, chapter 9). Given an *N*-dimensional (N-D) data space consisting of cloud of data points (input variables), the SOM algorithm distributes an arbitrary number of nodes (or cluster centers) in the form of a 1D or 2D regular lattice in such a way that it is the representative of the multidimensional distribution function, thereby facilitating data compression and easy visualization. Mathematically speaking, this is a process of a topology conserving projection from an original higher dimensional data space into the lower dimensional lattice (Haykin 1999, chapter 9). Each node is uniquely defined by a reference vector (or code vector) consisting of weighing coefficients. Each weighing coefficient of the reference vector is associated with a particular input variable. The essential part of SOM is to adjust the reference vectors to the N-D data cloud (input vector) through some unsupervised learning process. This is achieved through a user-defined iterative cycle adapting the reference vector in accordance with the input vector. This adaptation is the minimization of Euclidean distance between the reference vector for any *j*th node $\mathbf{W}_j$ and the input data vector $\mathbf{X}$, that is, to find

$$\min\|\mathbf{X} - \mathbf{W}_i\|.$$

For a particular data record only one node wins and is called the "winner node." An "optimal" mapping will be such that the winner node also changes the neighbor nodes as defined by the user. This inclusion of the neighborhood makes the SOM classification nonlinear since each node has to be adjusted relative to its neighbor. This training cycle may be continued for n times and may be mathematically described as

$$\mathbf{W}_j(n+1) = \begin{cases} \mathbf{W}_j(n) + c(n)[\mathbf{x}(n) - \mathbf{W}_j(n)], & j \in R_j(n) \\ \mathbf{W}_j(n), & \text{otherwise.} \end{cases} \quad (1)$$

Here $\mathbf{W}_j(n)$ is the reference vector for the jth node for nth training cycle; $\mathbf{x}(n)$ is the input vector; $R_j(n)$ is the predefined neighborhood around the node j; and $c(n)$ is the neighborhood kernel, which defines the neighborhood. The neighborhood kernel may be a monotonic decreasing function of n ($0 < c(n) < 1$, called the "bubble") or it may be of Gaussian type:

$$\alpha(n) \exp\left[\frac{-\|r_j - r_i\|^2}{2\sigma^2(n)}\right], \quad (2)$$

where $\alpha(n)$ and $\sigma(n)$ are constants monotonically decreasing with n. Here $\alpha(n)$ is the learning rate that determines the "velocity" of the learning process and $\sigma(n)$ is the amplitude that determines the width of the neighborhood kernel. The r_j and r_i are the coordinates of the nodes j and i in which the neighborhood kernel is defined. In the present study we have used a Gaussian neighborhood. The free software for SOM (available online at http://www.cis.hut.fi/research/som-research/) has been used in this study.

The SOM reference vectors span the data space and each node represents the position approximating the mean of the nearby samples in the data space. The other important advantage is that the smaller (larger) number of SOM nodes are allocated when the data is sparse (dense), see Fig. 1, (Hewitson and Crane 2002), and also SOM arranges the distribution of nodes in such a way that similar nodes are located close together and dissimilar nodes are farther apart. In an artificial example, the above features are clearly documented by Hewitson and Crane (2002). Are these features equally apparent in real 2D atmospheric data? To test this we construct the 2D dataset comprising the spatial mean and spatial standard deviation of each of the 122 days of the summer monsoon season (1 June–30 September) from 54 yr of IMD rainfall data over central India (12°–22°N, 72°–85°E). Thus we have $122 \times 54 = 6588$ data points. We wish to map such a big sample on 10×10 or 100 nodes. The number of nodes are chosen arbitrarily (as we shall see, the choice depends on the physical requirements of the problem in question). In Fig. 3a we plotted the scatterplot of the mean versus standard deviation (smaller points) for 6588 points and then mapped SOM nodes onto the data points (black circles). It can be seen that the nodes are placed continuously and densely in the region with more data points (between 0 and 10 units along abscissa) and sparsely where there are few data points (above 15 units along abscissa) and also indicates the nonlinearity in the data preserving the topology. Thus, it is demonstrated that the advantageous features of SOM are intact for real atmospheric data. Using the SOM routine one can also find out the dates clustered at each node and plot the input variable (here the mean or the standard deviation) on those dates for any nodes. Such a plot may be used to visualize a "pattern" of any variable associated with each node.

The SOM algorithm has been used in various disciplines (e.g., Palakal et al. 1995; Chen and Gasteiger 1997). In meteorology, the SOM is used for synoptic classification of weather states (Cavazos 1999; Hewitson and Crane 2002), climate study and downscaling of seasonal forecasts (Malmgren and Winter 1999; Gutiérrez et al. 2005), cloud classification (Ambroise et al. 2000), ENSO variability and diagnostic studies (Leloup et al. 2007), and so on. The SOM technique is different from other statistical analysis tools like EOF and multiple regressions. In SOM, the clustering of each node (which has a specific pattern) is based on the Euclidean distance among the reference vectors associated with a node and the input data vector. Here, largest are the distances among any two reference vectors, more different are the two nodes and so are the patterns associated with the nodes. In EOF or extended EOF analysis, the data is classified in terms of variance. However, the "orthogonality" property of the EOF modes makes it logically unsuitable to analyze a quasiperiodic oscillation like the ISO. The shortcomings of EOF and various conventional techniques are discussed by Goulet and Duvel (2000). Also, in multiple regression (linear or nonlinear) a functional relationship is computed between predictor parameters and a predictand. However, in SOM the lattice is first chosen and then the patterns are obtained at each node through an iterative process without seeking (explicitly) any functional relationship between the parameters involved. In a way, the SOM technique is analogous to (but more complex than) nonparametric regression techniques (Heskes and Kappen 1995).

The aim of this study is to exploit SOM to identify a set of well-separated and distinguishable patterns from the input samples of dynamical parameters (defined in Table 1) so that the days associated with a given pattern represent a particular phase of the convectively coupled oscillation of rainfall (e.g., active, break, or normal situations and their transitions) without using the rainfall data. The basic steps for implementation of SOM in the present study are as follows:

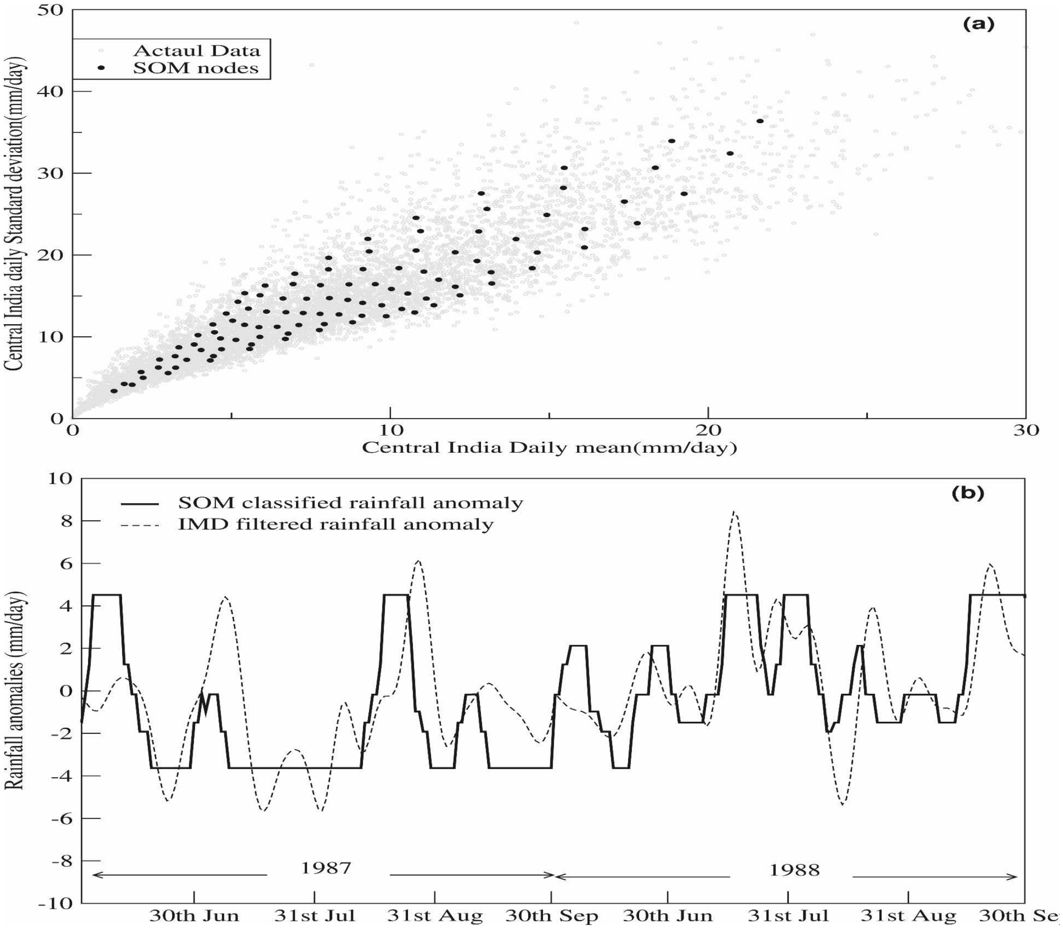

FIG. 3. (a) A simplified example of SOM clustering demonstrating the distribution of SOM nodes and data points in a 2D space. (b) The efficiency in capturing the rainfall patterns by SOM. The daily 10–80-day filtered area-averaged (15°–25°N, 70°–85°E) rainfall anomaly for June–September obtained from the IMD rainfall data is shown for two contrasting years, 1987 and 1988 (dashed line), together with the SOM-classified anomaly (solid line).

1) DECIDING THE NUMBER OF NODES

Distinct phases of ISO will be identified by a lattice of 3 × 3 SOM nodes. The choice of the 3 × 3 nodes was based on the fact that, on one hand, nodes should be kept a minimum and should have the least distortion and sufficiently low quantization error (a measure of error due to reduction in output dimension), while, on the other hand, it should produce maximum information of the important known phases of the oscillation (e.g., active states, break states, normal states). Let us consider the mean spatial patterns associated with active, break, and normal conditions (Krishnamurthy and Shukla 2000; Webster et al. 1998; Goswami and Ajaya Mohan 2001) as the "base" states (say A, B, and C). Each base state in turn can have "normal" (ensemble mean, A^0), "above normal" (A^+), and "below normal" (A^-) substates. Similar substates can be obtained for the states B and C. These substates of any base state could be characterized either by an east–west or north–south shift of the dominant spatial pattern or by an increase or a decrease of the intensity of the pattern due to the movement of the monsoon trough. Thus, a minimum of nine [$3A^{(0,+,-)}$, $3B^{(0,+,-)}$, and $3C^{(0,+,-)}$] states is required to have a more detailed idea of regional patterns and their transition from one phase to another (similar to a nonlinear curve fitting problem where it typically requires at least eight or nine points to trace a full nonlinear curve). Therefore, based on consideration of mathematical optimization and the physical requirement of identifying distinct patterns, a configuration of 3 × 3 states is chosen.

2) PREPARATION OF DATA

The standardized anomaly for the six circulation indices is now arranged to be used as input in the SOM routine. To determine whether a particular day (target

TABLE 2a. Values of area-averaged anomalies of rainfall from IMD data over central India for all 3 × 3 SOM nodes.

(1,3) −0.32	(2,3) −1.38	(3,3) −3.43
(1,2) 1.11	(2,2) 0.09	(3,2) −1.73
(1,1) 4.14	(2,1) 1.96	(3,1) −0.59

TABLE 2b. Mean days per ISO event (bold), percentage frequency of days (parentheses), and probability of no transition (braces) at each node over central India.

(1,3) **3** (10%) {68}	(2,3) **3** (9%) {63}	(3,3) **9** (22%) {88}
(1,2) **2** (8%) {57}	(2,2) **2** (5%) {47}	(3,2) **2** (8%) {59}
(1,1) **8** (20%) {87}	(2,1) **3** (8%) {63}	(3,1) **4** (10%) {74}

day) from each of the 122 days (starting from 1 June) and for each of 22 (1980–2001) yr is associated with a particular node of the 3×3 lattice, the target day, previous three days, and forward three days are considered. Thus we have data for seven days for each of the six indices, that is, $6 \times 7 = 42$ input values for any target day. Also for each target day we take the 1 May value of all six indices for the corresponding year of the target day (i.e., adding another six input values). The information for 1 May is added to make the reference vector "informed" (initialized) to a premonsoon condition of each variable for each year. Finally, the Julian day variation of six parameters is introduced as a variable (input value) according to (Cavazos 1999) $\sin[(2\pi t/365) - \pi/2]$, where t is the target day. It is introduced as a parameter to represent the annual cycle. Thus the input vector has 49 $(42 + 6 + 1)$ components (input values) for each target day. Similarly the associated reference vector has 49 weighing coefficients. Although in total there are $122 \times 22 = 2684$ days, for the training purpose we have selected 2074 samples collected from first 17 yr (17 yr × 122 days from 1 June to 30 September for the years 1980–1996). The NCEP–NCAR reanalysis dataset is sampled in a similar way.

3) RANDOM INITIALIZATION AND TRAINING

After determining the number of nodes and constructing the dataset, each reference vector of the nine nodes is initialized with some random values with the condition that none of the nine initial reference vectors are identical. The input vectors (having the identical dimension as the reference vector) are then broadcast parallel to each of the nodes. If the Euclidean distance between the input vector, $\mathbf{x}(n)$, and initial code vector at any of the nine nodes is minimum, it is the winning node. The code vector of this winner node is changed according to Eq. (1). The iteration is continued as many times as the total data record that we wish to train. This process is also repeated for many times (many training cycles) starting from a large number of neighbors and high learning rate until it is fine tuned to a single nearest neighbor and learning rate converging to zero. Thus, finally, the weight vectors for the nodes are arranged nonlinearly (because of the inclusion of neighborhood) into distinctly separated nodes. After this initialization and training of the reference vector (based on 17 yr of data) we classify the full sample (22 yr). Since each input vector has to be associated with a particular node, the corresponding target day will also be associated with that node. The dates clustered at each node are identified. If the summer monsoon ISO is a convectively coupled oscillation, the actual value for different variables (indices) on those dates clustered at a node corresponds to the commonality among various input parameters, and each pattern should be strongly related to a phase of the precipitation oscillation. In particular, one of the nodes should correspond to the active pattern, while another should correspond to the break pattern.

4. Results and discussions

a. Some basic statistics of ISO derived from SOM

Once we obtain the classification using the SOM algorithm, the dates from the 22 yr of ERA-40 data are collected at each node. To test whether these SOM nodes based on circulation data (without involving observed rainfall data) are related to organized rainfall anomalies, composite IMD rainfall anomalies averaged over central India (15°–25°N, 70°–85°E; hereafter CI) corresponding to the dates associated with each of the 3×3 nodes are shown in Table 2a. A strong positive/negative rainfall anomaly associated with the circulation states described by SOM nodes (1, 1)/(3, 3) testifies that these circulation states correspond to a strong active and an intense break condition, respectively. The values of area-averaged anomalies of rainfall corresponding to other nodes indicate that [(2, 1) and (1, 2)] represent less intense active states and [(2, 3) and (3, 2)] represent less acute break states, while near-neutral/normal states are represented by [(2, 2);(3, 1);(1, 3)]. In Table 2b we show the mean days per ISO event present in a node, the frequency of days clustered at each node in percentage (in parentheses), and the probability of no-transition (braces) from one node to the other. The three values are calculated from the 22 yr of ERA-40 data. Here the number of "events" is determined by counting the total number of times the data records are mapped consecutively to a particular node without any break. Mean days per event is defined by dividing the total number of days mapped onto a SOM node by the number of events counted for that node. Frequency of days is defined as the number of days clustered in a particular node divided by the total number of days used in the classification (22 yr × 122 days yr^{-1}). The probability of "no transition" (also expressed in percentage) is the probability that, when an input vector corresponding to a particular day is mapped to a node, the next day will be mapped again to the same node. Thus, for the active state (1, 1), out of the total projection onto the node, 87% of the cases are projected successively (i.e., without any break). Similarly, it is of the same order for the node (3, 3) and is lowest for (2, 2). This implies that, when a day is attached to an active (1, 1) or break (3, 3) node, the next day has the highest probability of clustering at the same node, and for the node (2, 2) the chance is least. Thus, SOM can give a simple visualization of the time evolution of the nodes and, hence, the ISO (the spatial propagation of ISO will be described in the subsequent sections). Further, it can be seen that mean days per event is highest for the active (8 days) and break (9 days) nodes and the corresponding percentage frequency of days clustered at these nodes is also higher. Assuming that one full cycle of ISO (active–break–active) is an episode, the total number of days per episode (obtained by summing the days at all nodes) is 36, which corresponds to the average periodicity of a low frequency ISO event. This implies that the "mean" repetition frequency of a break state (or an active state) is 36 days with unequal distribution of the days at each node. Thus, the above results verify the quantitative estimate of the ISO within a season available in various sources and allows for further application of the SOM to study the ISO.

As discussed above, if the large-scale circulation states clustered at the SOM nodes are associated with rainfall anomalies over CI, they could be used to construct the low frequency components of the latter. This is illustrated in Fig. 3b, where the area-averaged value of the 10–80-day filtered rainfall anomalies over CI from the IMD data is plotted together with the area-averaged value of the rainfall composite for each class for the years 1987 and 1988 that are mapped onto the SOM nodes. The filtering is done using a standard Lanczos filter (Duchon 1979). It may be seen that the rainfall anomaly for each class is actually following the rainfall anomaly from the observed data. A similar match is found for the other years (plot not shown). The temporal correlation for the rainfall mapped at each SOM node and the corresponding filtered rainfall for the 122 × 22 days is 0.58, which is significant at 99.9% confidence level. It is clear that the SOM technique, through the use of a number of large-scale circulation parameters, is able to capture the low-frequency subseasonal variability of rainfall over CI.

b. *Classification of precipitation states*

In this section, the strength of the SOM technique is further illustrated, and it is shown that the large-scale circulation-based SOM nodes not only give us the area-averaged rainfall anomaly over central India, but also provide a detailed spatial pattern of different phases of the rainfall oscillation. The composite rainfall anomalies over the Indian continent corresponding to the dates of each of the SOM nodes based on 22 yr of ERA-40 data are shown in Fig. 4. While the nodes (1, 1) and (3, 3) reproduce the well-known active and break patterns, respectively, with considerable regional details, the other nodes represent different phases of northward and eastward propagation of the dominant ISO mode. The northward and eastward propagation of the small positive anomaly in the southeastern corner of India in node (3, 3) can be seen if we follow the panels counterclockwise in Fig. 4. These composite rainfall anomalies in Fig. 4, identified from large-scale circulation parameters only, have good correspondence to the lag composites based on the rainfall index over central India using the IMD data.

The northward and eastward propagation of the rainfall anomaly over India is actually a part of the northward and eastward propagation of the large-scale rainband or the tropical convergence zone (TCZ) (Sikka and Gadgil 1980; Goswami 2005). Are the phases of rainfall oscillation identified by the large-scale circulation SOM nodes (Fig. 4) linked to the evolution of the TCZ? To examine this, composites of daily CMAP rainfall between 1980 and 2001 corresponding to the SOM nodes from ERA-40 circulation data are constructed and shown in Fig. 5. The composites in this figure clearly demonstrate that the large-scale circulation-based SOM nodes identify the evolutionary phases of the northward propagating rainband similar to ones obtained from rainfall data (see Goswami 2005; Waliser et al. 2003). Thus, the regional rainfall anomalies over India derived from the high-resolution rainfall data and identified by the SOM nodes (Fig. 4) are an integral part of the large-scale northward propagating rainband.

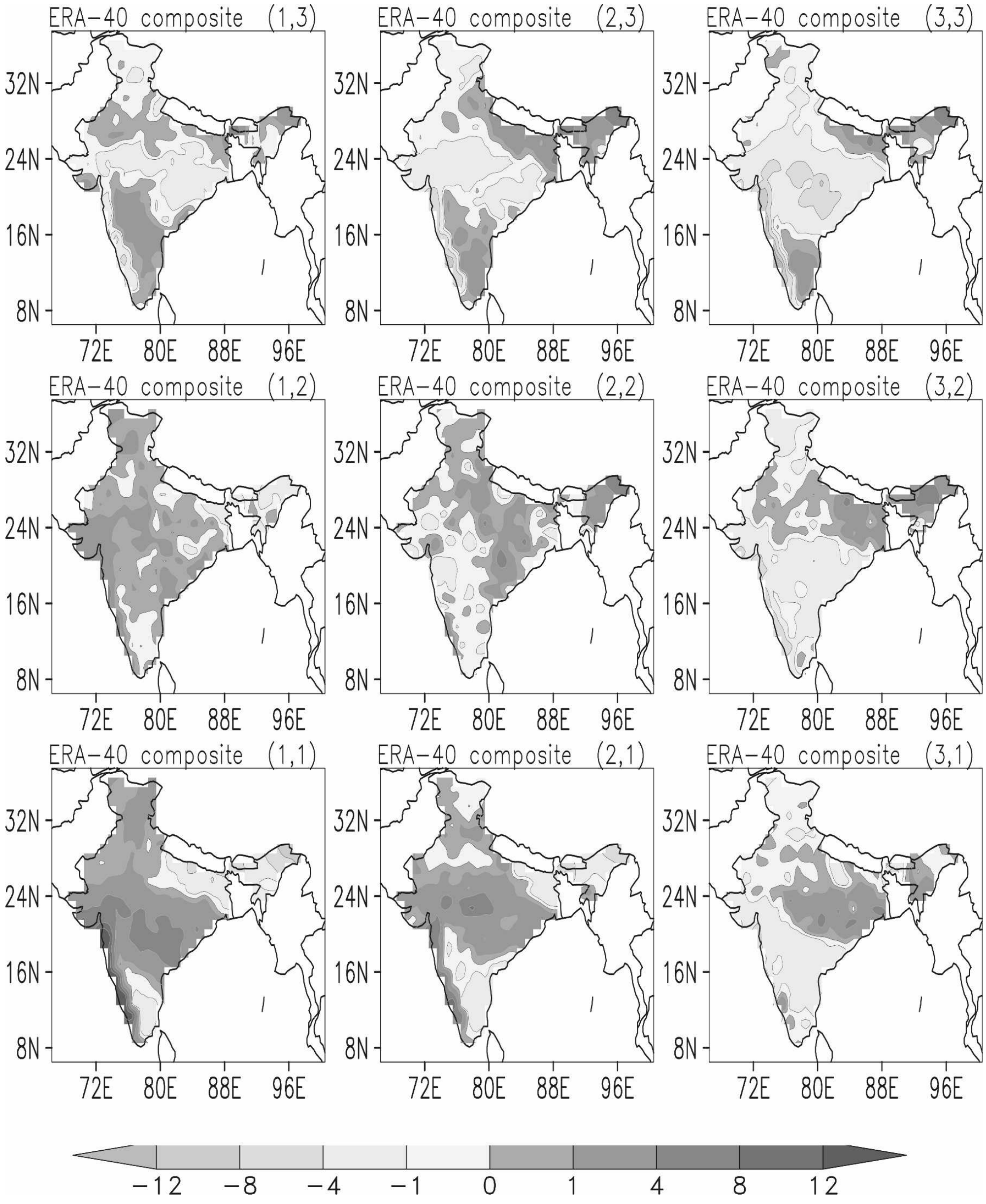

FIG. 4. The spatial distribution of anomalous precipitation (mm day^{-1}) associated with SOM-classified patterns, obtained by compositing the IMD daily rainfall anomaly corresponding to the days clustered at the respective SOM nodes. To get the anomaly, the daily long-term means are constructed from the 54 yr of IMD rainfall data. The states (1, 1) and (3, 3) are the most active and break nodes.

Hence, the SOM technique can identify the phases of summer monsoon ISO in rainfall through the use of only large-scale circulation parameters. It may, however, be argued that the specific humidity at 850 hPa over central India from ERA-40 (SH index), used as one of the indices may contain some information on observed rainfall. To test whether the SH index is crucial for identifying the phases of rainfall ISO (as in Fig. 4), we removed the SH index from the variables used for calculating SOM nodes. In place of the SH index, we introduced another circulation index, namely the kinetic energy (KE) of the low level jet (LLJ) defined by the KE of 850 hPa winds averaged over 5°–15°N, 50°–70°E (KELLJ). This index is also known to be related to the seasonal mean monsoon rainfall as well as ISO activity over India and the Bay of Bengal (Goswami and Xavier 2005; Ajayamohan and Goswami 2007). The SOM nodes were again calculated using 22 yr of ERA-40 data including the KELLJ index instead of the SH index. Composite rainfall anomalies over In-

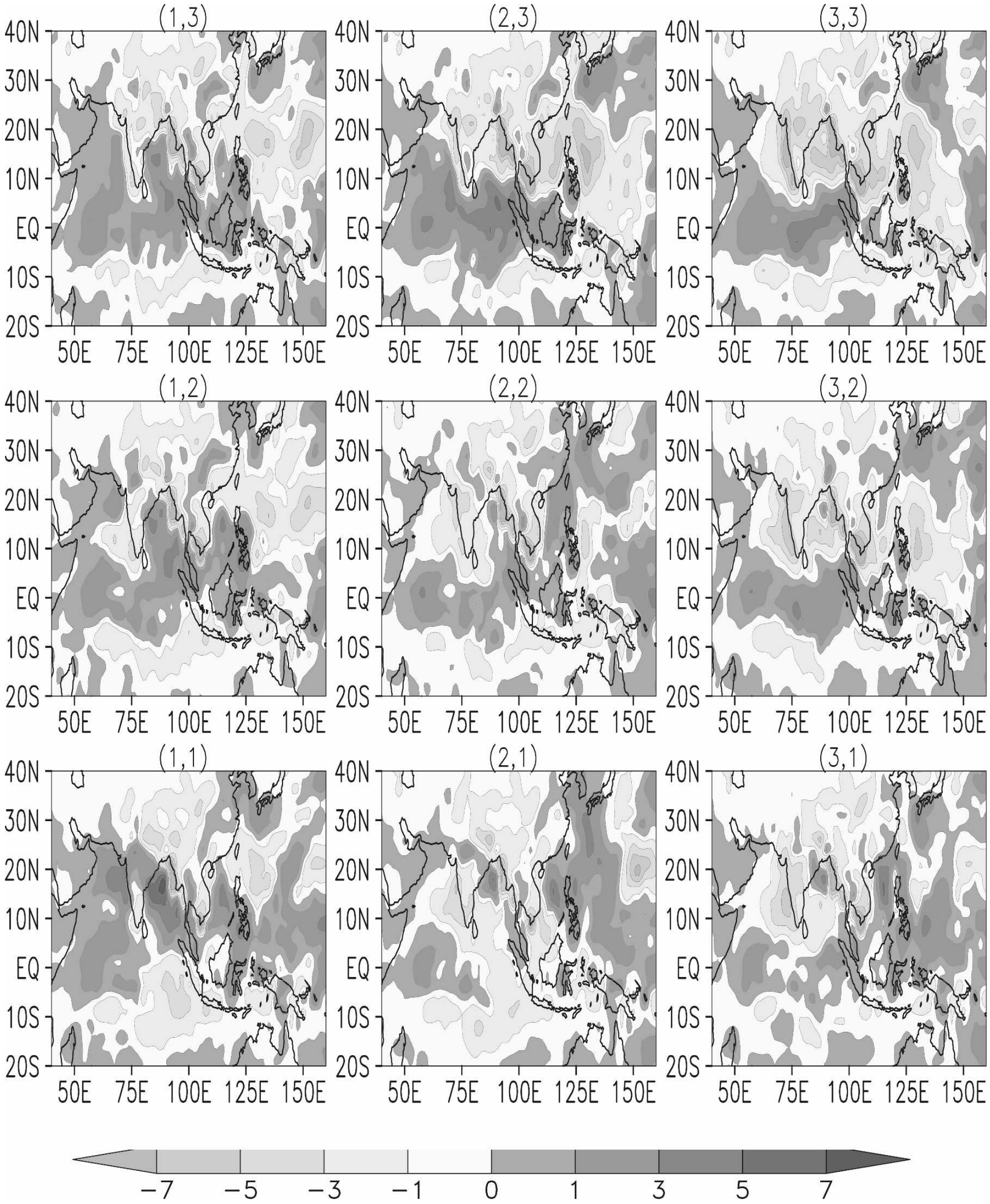

FIG. 5. As in Fig. 4 but constructed from CMAP global data. The CMAP pentad data is interpolated to daily values prior to construction of composites. It is clear that starting from any node [say, (3, 1)] there is a large-scale northward propagation of rainfall anomaly (follow the nodes counterclockwise).

dia constructed from the dates corresponding to SOM nodes thus derived are found to be almost identical to the ones shown in Fig. 4 (figure not shown). Thus, dynamical information (circulation indices) is sufficient to accurately identify the phases and evolution of the rainfall ISO. However, a minimum number of large-scale circulation variables are required to capture accurately the phases of the convectively coupled monsoon ISO. In this study we found that reducing the number of indices below six results in weakening of the active and break patterns [node (1, 1) or (3, 3) in Fig. 4].

How close are the patterns of rainfall anomalies corresponding to different SOM nodes to the phases of the rainfall ISO defined from rainfall itself? To test this, active and break composites obtained from the SOM nodes [(1, 1) and (3, 3) in Fig. 4] are compared with those obtained from rainfall index (PR index) in Fig. 6. The top (bottom) rightmost panel is the break (active) composite from 53 yr of IMD rainfall data (1951–2003) obtained from the composite number of days when the standardized anomaly of rainfall in CI (the PR index) is less (greater) than one standard deviation. Further, to

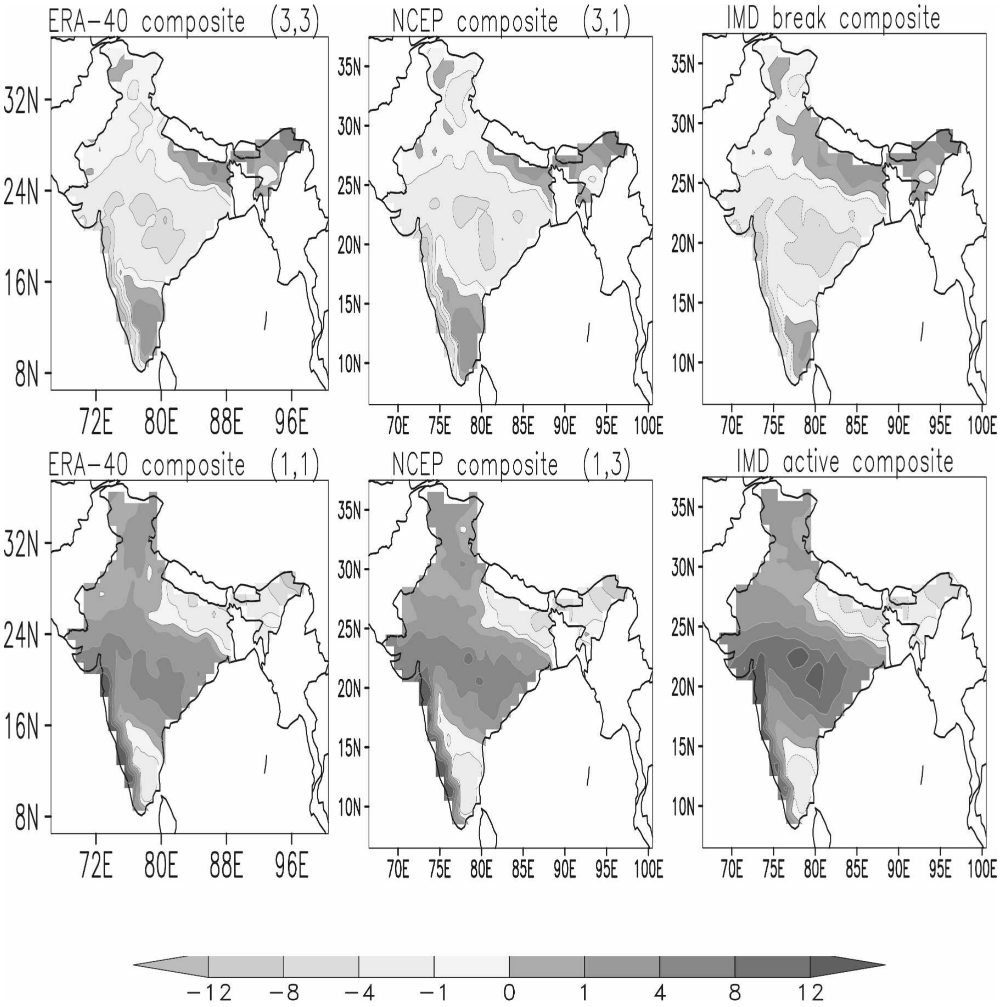

FIG. 6. Comparison of SOM-classified (top) driest and (bottom) wettest patterns of rainfall using the large-scale indices from (left) ERA-40 with those derived from (right) rainfall. (middle) SOM-classified pattern of rainfall using large-scale indices for NCEP–NCAR reanalysis.

test the robustness of the SOM technique, the SOM nodes are also constructed using the same large-scale indices shown in Table 1 from NCEP–NCAR reanalysis for the same period. The composite rainfall anomalies corresponding to the active and break nodes obtained from the NCEP–NCAR reanalysis circulation data are also shown in this figure (middle panels). The similarity between the patterns of rainfall derived from only dynamical inputs of the ERA-40 data and the NCEP–NCAR reanalysis data and those obtained from the precipitation index is striking. Correlations between different patterns are 0.9 or larger (Table 3). Two things are clear from the figure. First, both reanalysis datasets are dynamically consistent in capturing the active and break patterns. This implies that large-scale dynamics is equally well captured in both of the reanalysis data, at least in the intraseasonal scale. Second, active and break spatial patterns of rainfall can be accurately identified by the large-scale indices. The dates clustered at each node are reflections of the closeness (commonality) in the sequence of temporal evolution of these indices. This implies that these dates (at any node) represent similar phases in the time series and different nodes represent different phases of northward

TABLE 3. Spatial pattern correlation for the dry and wet patterns of rainfall classified by SOM from various datasets. The spatial correlation is defined for the region over the Indian landmass.

	ERA-40/ NCEP–NCAR	ERA-40/ IMD	NCEP–NCAR/ IMD
Dry pattern	0.97	0.93	0.94
Wet pattern	0.94	0.87	0.93

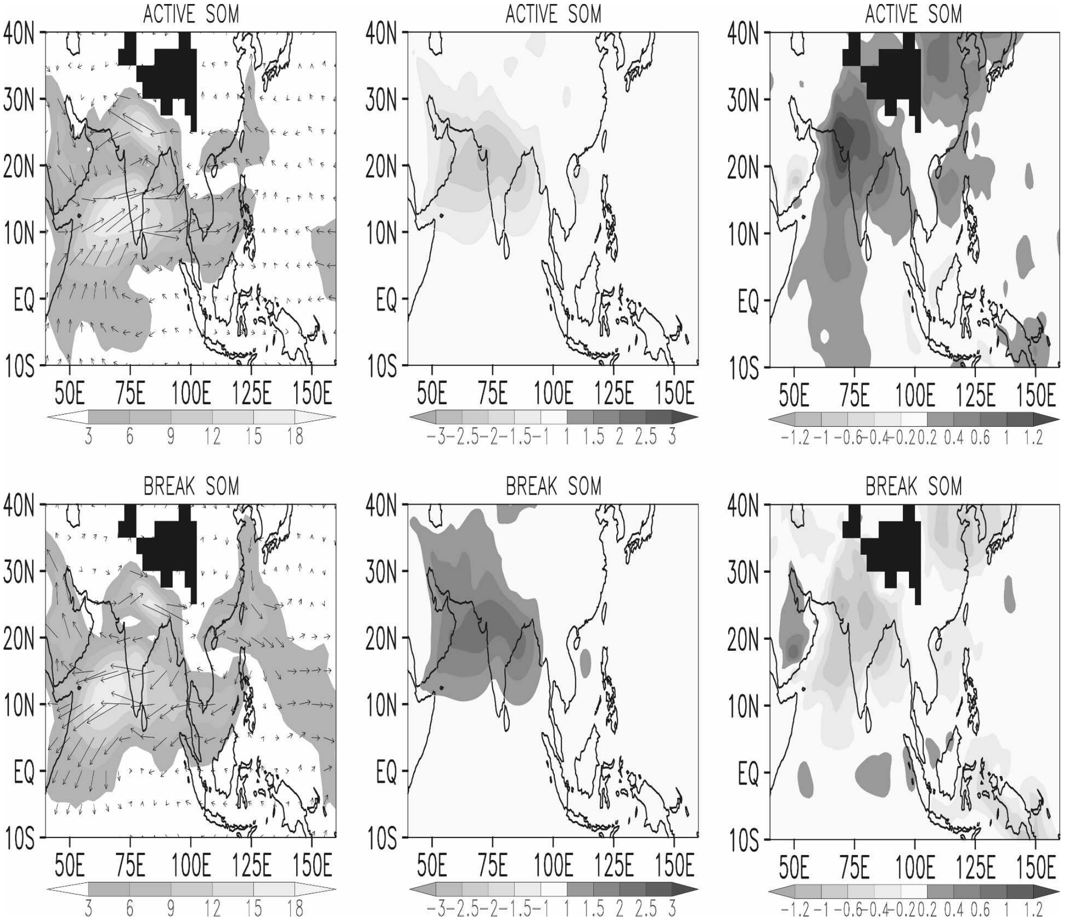

FIG. 7. Composite spatial plot of ERA-40 data for the (top) break and (bottom) active nodes for (left) wind (m s^{-1}) at 850 mb, (middle) mean sea level pressure (hPa), and (right) specific humidity (g kg^{-1}) at 850 mb. The composite technique is similar to Fig. 4. To get the anomaly, the daily long-term means are constructed from the 22 yr of ERA-40 data for all variables. The patterns of rainfall in Fig. 4 are well reflected in these parameters.

movement of the ISO pattern (rainband). Although the gross spatial feature matches quite well, the intensity of the active (break) composites from IMD rainfall data is at least 2–4 mm higher (lower) during the active (break) state than the SOM-captured intensity at the nodes from both ERA-40 and NCEP–NCAR reanalysis data over central India. This is due to the fact that the SOM technique based on large-scale dynamics from the ERA and NCEP–NCAR reanalysis data identifies only the low-frequency intraseasonal component of rainfall variability and is not expected to capture the extreme day-to-day precipitation fluctuation pattern, which is pronounced in the IMD daily data and results in larger composite values for active composites. However, the spatial pattern of rainfall from the large-scale dynamics and actual rainfall (using IMD data) during the active and break phases are close to each other, providing strong support for the convectively coupled nature of the monsoon ISO.

The large-scale patterns of some other dynamical variables associated with active and break phases identified by the SOM technique are noted in Fig. 7 where the composite anomalies of vector winds at 850 hPa, mean sea level pressure, and specific humidity at 850 hPa from dates collected at active and break SOM nodes [node (1, 1) and node (3, 3) in Fig. 4] using ERA-40 data are shown. This figure indicates that SOM-classified patterns of wind anomalies, sea level pressure, and humidity anomalies at these nodes are opposite to each other and are dynamically consistent with rainfall patterns at the same nodes. To get an idea as to how the six indices are configured at all of the different SOM nodes, the standardized anomaly of the six indices for days collected at the nodes is shown in Table 4. It may be noted that the values of the indices are nearly equal but of opposite sign at the active and break nodes [node (1, 1), and (3, 3)] and are arranged in a regular fashion similar to the strength of the rainfall (Table 2).

TABLE 4. Standardized anomalies of six indices for nine nodes obtained from the SOM classification.

Indices					
WY			GO		
0.71	0.19	−0.98	−0.09	−0.37	−0.69
0.85	0.30	−0.59	0.17	−0.09	−0.33
0.74	0.20	−0.20	0.72	0.63	0.28
WF			GP		
−0.09	−0.47	−0.96	−0.65	0.11	0.79
0.34	0.05	−0.23	−0.53	0.24	0.51
1.14	0.39	0.04	−0.86	−0.09	0.37
SH			MS		
−0.66	−0.65	−0.92	−0.45	0.30	1.03
0.14	0.18	0.03	−0.66	0.08	0.62
1.09	0.72	0.56	−1.14	−0.27	0.26

5. The different "shades" of the active and break states

In the previous section, the low-order SOM classification (e.g., 3 × 3) identifies nine distinct ensemble mean phases of the monsoon ISO. Such ensemble mean phases of ISO could be identified using other techniques as well (e.g., phase composite or lag regression with respect to a reference time series). The strength of the SOM technique lies in the fact that not only can it identify the ensemble mean phases, but it also can identify the event-to-event variability or different "shades" of any phase of the ISO together with their evolutionary history. This can be achieved by going over to a higher-order SOM classification.

As we introduce higher-order SOM classification with many more nodes, the SOM will identify that many patterns. However, in this case the patterns associated with the nodes may not be as well separated as in the lower 3 × 3 classification. Some of these patterns may be similar to one of the nine nodes of 3 × 3 classification, but differ slightly in their spatial structure and/or temporal evolution. In other words, with the increment in the number of nodes, we get other different shades of a particular pattern associated with one of the 3 × 3 previously classified nodes. We demonstrate the existence of different shades using a higher-order (9 × 9) SOM clustering. The classification is made on the basis of six large-scale dynamical parameters (as used before) taken from 54 yr of NCEP–NCAR reanalysis data (1951–2003). The other procedures are exactly similar to an earlier one using ERA-40 data. The area-averaged rainfall anomaly over central India is shown for the 9 × 9 nodes in Table 5, similar to Table 2a. We then calculate the correlations between patterns associated with each of the 3 × 3 nodes and those associated with each of the 9 × 9 nodes. This is shown in Fig. 8. Each of the nine panels in the figure corresponds to one of the 3 × 3 nodes and each panel consists of two circles. The largest outer circles (without shading) within a panel are 81 in number and arranged as 9 × 9 nodes. The spatial correlation is represented by the inner circles (filled or open), the magnitude of which is proportional to the area of the circle. The positive (negative) correlation is indicated by filled (open) circle. If the diameter of the inner circle (filled or open) is equal to the outer one, the correlation is +1 (−1). As can be seen, some nodes have spatial correlation very close to ±1. The spatial patterns for the three active and three break nodes having largest spatial correlation are plotted in Fig. 9. As earlier, the spatial plots are made by plotting the rainfall anomalies (IMD high-resolution data) from the dates clustered at each node. As can be seen, the top left and bottom left panels are identified with the active and break nodes of the 3 × 3 SOM nodes [node(1, 3) and node(3, 1)]; the other patterns are slightly different shades(variants) of the intense active and break patterns. Similarly, all of the 9 × 9 nodes, having different spatial correlations with one of the 3 × 3 nodes, are different shades of one of the 3 × 3 nodes and represent different spatial patterns of ISO. These different shades of rainfall are also associated with different spatial patterns of large-scale dynamical parameters. The spatial patterns of all the parameters used in the clustering (wind at 850 and 200

TABLE 5. Area-averaged rainfall anomalies over central India from the 9 × 9 nodes constructed from 54 yr of NCEP–NCAR reanalysis data (1951–2004). Positive and negative values are clustered in opposite corners, shown in bold.

1.03	1.39	1.77	1.93	1.65	1.27	**4.61**	**4.15**	**6.21**
−0.11	0.53	−0.52	0.08	−0.25	0.49	1.75	1.41	**4.38**
−1.40	−0.97	−0.97	−0.70	−1.22	0.41	0.62	1.81	**4.82**
−2.30	−2.45	−2.39	−1.97	−2.08	−1.07	0.76	**5.25**	**4.86**
−2.15	**−3.07**	**−2.61**	**−3.40**	−1.48	0.65	0.77	**2.63**	**3.59**
−2.93	**−4.16**	**−3.89**	**−3.41**	−2.86	−0.83	1.87	0.44	**3.21**
−2.99	**−3.85**	**−3.79**	**−2.73**	−2.50	0.88	−0.68	0.91	1.11
−3.04	**−3.94**	**−3.71**	**−2.37**	−0.70	1.31	0.43	0.31	0.52
−3.34	**−3.02**	**−2.97**	**−2.31**	−0.98	−0.57	−0.96	−0.87	−0.82

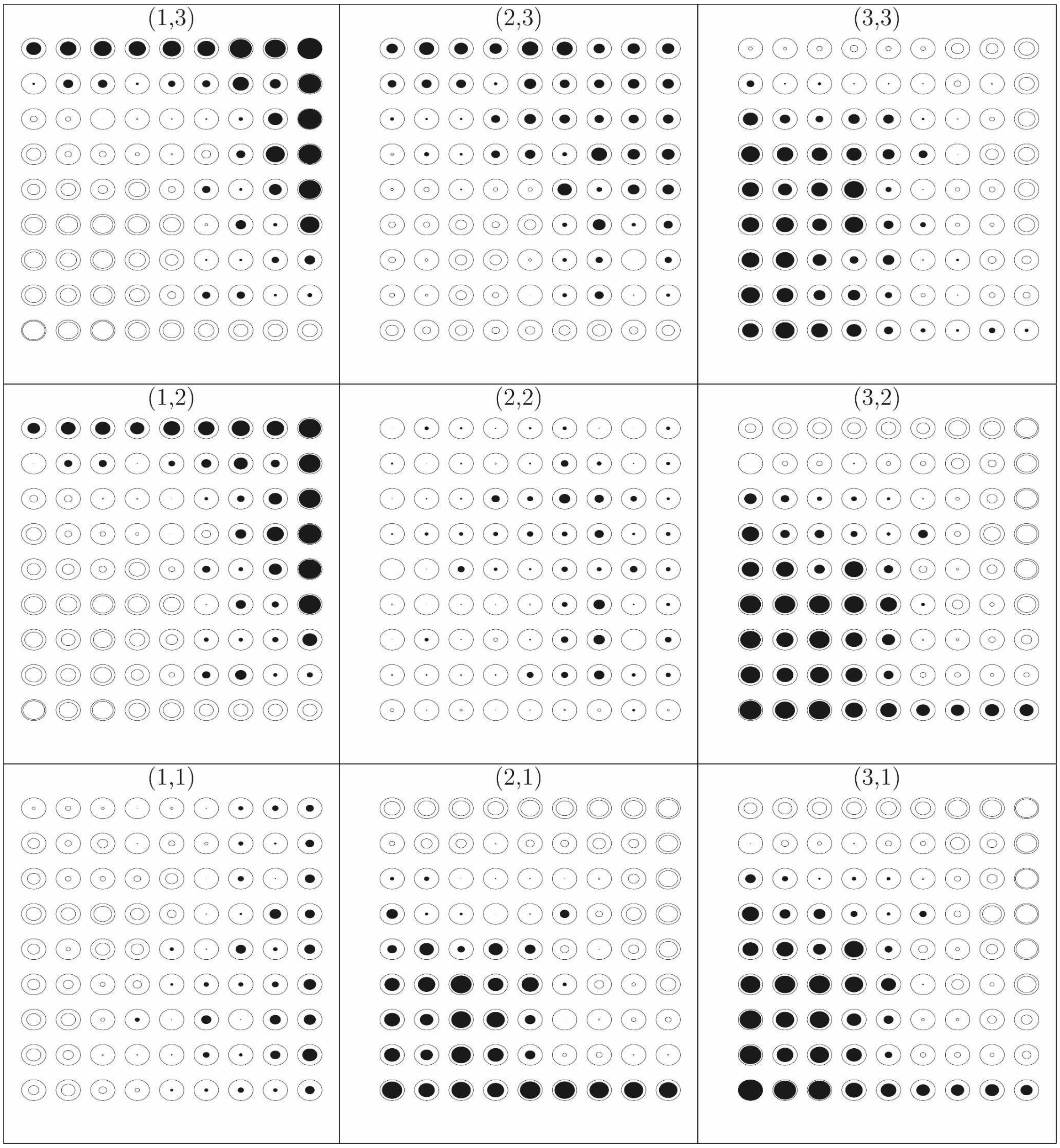

Fig. 8. Spatial anomaly correlation of each of the 3 × 3 nodes with each of the 9 × 9 nodes is shown. Each of the nine square panels represents one node of the 3 × 3 cluster, and each panel consists of 81 pairs of concentric circles arranged in a 9 × 9 array. The diameter of the filled or the open circles inside the larger circles is proportional to the strength of the anomaly correlation. Positive (negative) correlations are represented by filled (open) inner circles. Completely filled (open) circles have correlation +1 (−1).

mb, mean sea level pressure, geopotential height at 500 mb, specific humidity at 850 hPa) show differences in the spatial structures (plot not shown) for different shades. This implies that all of the large-scale parameters used in the study reflect different behavior of the rainfall ISO.

Can there be distinct shades of temporal evolution too? To answer this question, first we construct an evolutionary history of each of the 9 nodes of 3 × 3 classification. This is done by constructing a composite evolution of CI area-averaged rainfall anomaly from 30 days prior to 30 days after each of the clustered days at all the nodes. Similar evolutionary history of each of the 9 × 9 SOM nodes is also obtained by constructing acomposite around the clustered days at all 81 nodes. Like the spatial pattern correlations as computed in the previous paragraph, we make the temporal correlation for each of the 3 × 3 nodes with each of the 9 × 9

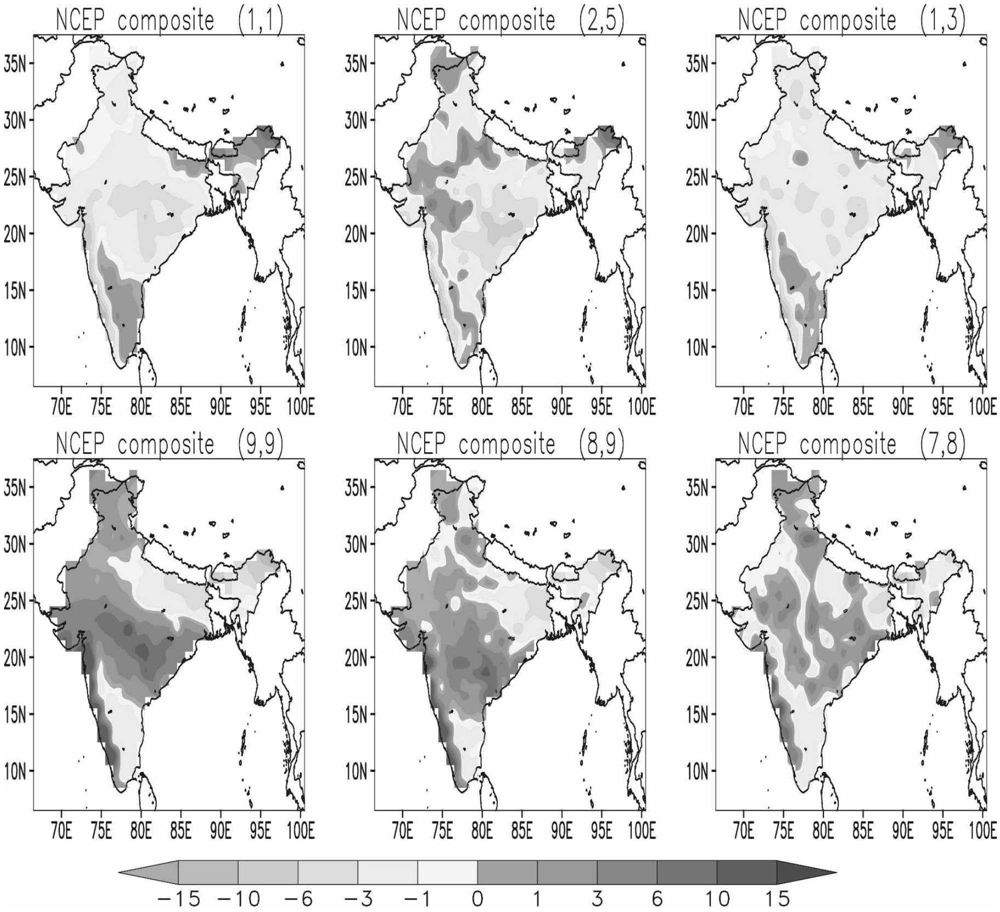

FIG. 9. The different spatial shades of rainfall active and break patterns. The top (bottom) left panel is the plot for most dry (active) node obtained from 9 × 9 classification. These two patterns have also the maximum anomaly pattern correlation with the driest and wettest patterns of the 3 × 3 classification. The other patterns are different shades of the most active and break patterns.

nodes. Not surprisingly, we get different temporal variants or shades similar to the spatial shades discussed in the previous paragraph (not shown). At least one of the nodes in 9 × 9 clustering has exactly similar temporal evolution with one of the 3 × 3 nodes. The other patterns are different shades of any of the 3 × 3 nodes.

Thus, using SOM we have identified different shades of active–break cycles (or ISO) that have different spatiotemporal evolution. This property of SOM is used to develop an analog prediction scheme and will be discussed in the next section.

6. A real-time extended-range prediction scheme for CI rainfall using the SOM

As mentioned in the introduction, skillful prediction of active and break spells of monsoon 3–4 weeks in advance would have tremendous utility for farmers and water resource managers. The potential for such extended-range prediction has been indicated in several recent studies (Goswami and Xavier 2003; Xavier and Goswami 2007; Webster and Hoyos 2004; Jones et al. 2004). Most of these studies use some form filtering of the data and suffer from the "end point" problem when applied to real-time prediction. Also, almost all studies of extended-range prediction of ISO phases use linear technique (e.g., regression) and hence capture only the ensemble-averaged phases of the dominant oscillation and fail to capture shades of different phases of the oscillation. The SOM technique being nonlinear and its ability to identify shades of different phases of the oscillation and their evolutionary history opens up a whole new possibility for extended-range prediction of active–break phases of the summer monsoon.

Here we present a scheme for the prediction of rainfall over central India four pentads in advance. The technique, based on SOM classification of pentad data of large-scale circulation indices (Table 1) constructed from NCEP–NCAR reanalysis, does not involve any filtering and, hence, is ideal for real-time prediction. For the training purpose, data for the years 1951–99 are used, while skill of the model is evaluated from hindcasts of independent IMD rainfall data during 2000–04. After a certain amount of experimentation, we decided to use SOM clustering with 15 × 15 nodes to include more shades of temporal evolution. Using data for the pentad 31 May–4 June, the first four-pentad lead forecast each year is made for the pentad starting with 20 June (i.e., 20–24 June, when the monsoon is well established over central India). In this manner, four-pentad lead forecast is made for 17 pentads during the summer season every year.

The prediction scheme is based on the following premise. The SOM classification on the training period extracts 15 × 15 patterns and their evolutionary history and stores them in the "reference" vectors. Time histories of the patterns are saved on the dates clustered at each node. For prediction from a given date, a "forecast" vector is created with current and past data for nine days for all the large-scale variables. This essentially contains the pattern and its evolutionary history at the initial time. If we could find an analog of this pattern and its evolution in the past from the reference vectors corresponding to different nodes, we could make a four-pentad prediction from the evolutionary history of the analog. In practice it is done as follows:

We normalized the data for each day and for each variable as

$$\frac{X(i) - X_{\min}}{X_{\max} - X_{\min}},$$

where X is any variable for ith day: $X_{\max}$ and $X_{\min}$ are the maximum and minimum values of the variable obtained for the period 1951–99. For a given day, the data for the day itself and the past nine days data of six large-scale dynamical indices are used to construct the SOM input vectors. Using these input vectors, SOM clustering is obtained for 15 × 15 nodes. Each node now consists of a reference vector (see section 2) identical in dimension to the input vectors and dates that depict individual shades of ISO. Similar to the input vectors constructed for the past data, forecast vectors are constructed for the starting day of the forecast and for the past nine days using the same six large-scale dynamical parameters. The Euclidian distance between the forecast vector and the reference vector attached to each of 15 × 15 nodes is calculated. The node for which this difference (distance) is least is considered the true analog of the forecast vector of that particular day. To account for uncertainty in the temporal evolution of the initial condition (forecast vector), we decided to construct ensemble mean forecasts based on a number of slightly poorer analogs in the neighborhood of the true analog. Through a series of experimentation, the optimum neighborhood criterion for four-pentad forecasts was found to be those reference vectors for which the difference between Euclidian distance lies within ~25% of the minimum distance. Since each node also has information of the dates clustered at the nodes, the four-pentad forecast for any day in the forecast period is the average of four-pentad value ahead of the days clustered at the closest node and the nearest neighbors. The amplitude of the forecasted rainfall for a given day of the year is corrected by a factor determined by the ratio of variance of observed rainfall over central India on that day to the variance of the four-pentad predicted rainfall for the same day during the training period. The observed variance for a particular day during June–September is determined from the period 1951–99. To make the forecast for the next day, the process is repeated. In this case some other node will be selected along with its nearest neighbors. This process is continued for all 17 pentads for each of the 5 yr selected for prediction (2000–04).The correlation between four-pentad lead forecasts and verifications of the area-averaged standardized CI rainfall anomaly for all of the 85 pentads (17 pentad yr^{-1} × 5 yr) is 0.55 with standardized rms error (RMSE) of 0.88 (Fig. 10a). This correlation is well above 99.99% confidence level.

While the correlation between four-pentad forecast and observed rainfall at the same lead time using six large-scale indices is highly significant, it still leaves a large fraction of variance unexplained. It appears that the six large-scale parameters are inadequate to delineate some aspects of the nonlinear convectively coupled monsoon ISO. It may be partly due to the fact that some of the patterns associated with some of the 15 × 15 nodes may not be quite distinct from each other. Such degeneracy may add random errors to the forecast. To reduce the contribution of random errors, we decided to adopt a multimodel ensemble strategy. To achieve this, we would like to construct a second SOM model with a set of different indices. For this purpose lag correlation between the central India rainfall time series and various circulation fields were carried out. From this exercise, three new indices are chosen that have strong correlation with the CI rainfall with a lag between 15 and 20 days. The indices are defined in Table 1 (last three indices). The SOM networks are then trained separately based on the past 19 days information for the same number of nodes, and the forecast rainfall is obtained in a similar way as the earlier one (i.e., based on the past nine days training).

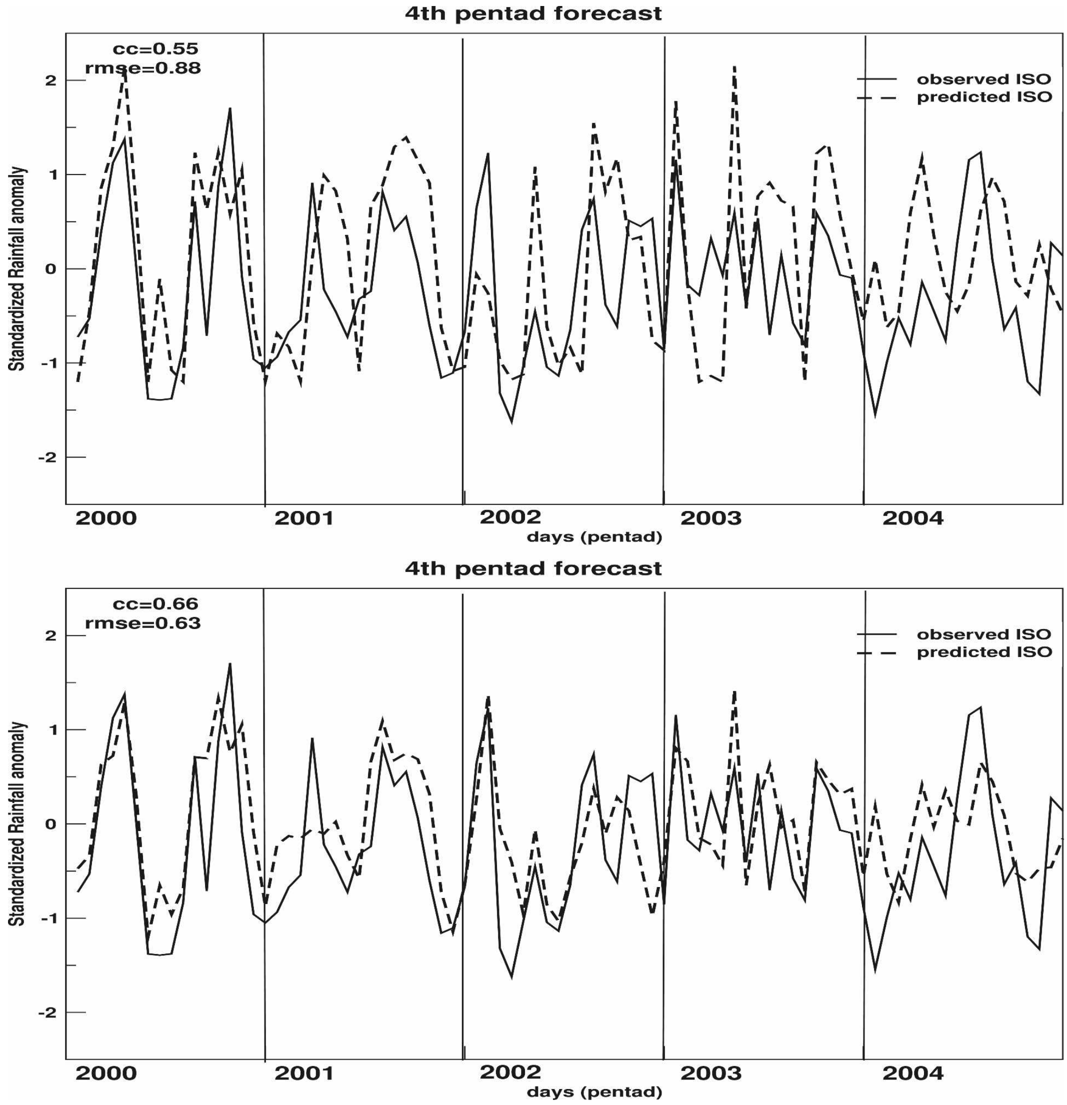

FIG. 10. (top) The four-pentad forecast for CI for the 85 pentads (17 pentad yr^{-1} × 5 yr) from 2000 to 2004. The first forecast for each year is for the pentad 20–24 Jun and the last pentad is for 8–12 Sep. This forecast is made using the SOM model with six dynamical parameters and the past nine day information. (bottom) Similar to (top) but showing the average forecast of the two-SOM model (i.e., one using the past 9-day information and one using the past 19-day information).

The two-model ensemble forecast rainfall is then obtained by averaging the two forecasts. The four-pentad forecast and verification are shown in Fig. 10b. The correlation between forecasts and verification for the 85 pentads is 0.66 and standardized RMSE is 0.63. Thus, the multimodel ensemble forecast strategy increases the forecast skill and reduces the RMSE. The extended drought in the years 2002 and 2004 is well captured using this forecast model.

From the previous example it is clear that the skill of this technique depends on the efficiency of SOM in picking up proper analogs. If a proper analog of an event in the forecast period exists in the past data, the future will be well forecast. Though Lorenz (1969) found that small errors grow rapidly, the prospect of a better rainfall forecast using an artificial neural network technique lies in the fact that it compensates the effect by efficiently separating the signal from the noise (Elsner and Tsonis 1992). We also wish to point out here that the choices of the indices in this study are partly made through a "frequentist" approach (e.g., the three indices used in the past 19 days training) and partly through a "Bayesian" approach (e.g., the six indices used in the past nine days training). There is always a chance of improving the forecast through some other choices of indices. In addition, forecasts for other regions of study may also be calculated. We conclude this section by saying that the above example of ISO forecasting may be considered an instance of application of SOM that laid the foundation of a more detailed study in applying SOM to ISO prediction. The above illustration is a logical extension of our identification of a nonlinear convectively coupled intraseasonal oscillation, isolating its various phases and signature of large-scale dynamics on the rainfall ISO using SOM during the summer monsoon over India.

7. Summary and conclusions

In this paper, a new innovative technique known as self-organizing map (SOM) has been introduced to study the monsoon ISO. Unlike the linear techniques used so far that could identify only the ensemble mean phases of the monsoon ISO, a nonlinear pattern recognition technique, SOM, is capable of identifying different shades of each ensemble mean phase, including their evolutionary history. This novel feature of the technique has opened up the possibility of a nonlinear extended-range prediction of monsoon rainfall.

We started with the recognition that the difficulty in predicting the active and break phases Indian summer monsoon is due to the event-to-event variability of the phases of intraseasonal rainfall oscillation. The event-to-event variability is a signature of nonlinearity of the oscillation. We hypothesize that the summer monsoon ISO is a convectively coupled oscillation, and hence it should be possible to identify the phases of rainfall oscillation by using large-scale circulation parameters. However, the relationship between rainfall and circulation being nonlinear, an effective method to isolate the commonality among the parameters such that they detect different phases of the nonlinear convectively coupled intraseasonal oscillation is needed. For this purpose, one needs to include a sufficiently large number of circulation parameters since the different shades of the convectively coupled intraseasonal oscillation of rainfall are reflected through a distinct common pattern among these parameters. Using many parameters to understand a single phenomenon, essentially, implies use of a dimensionality reduction and clustering algorithm. Thus, we use an unsupervised learning clustering algorithm called SOM to bring out the common features of active and break phases and their different shades.

Large-scale circulation indices (Webster and Yang 1992; Goswami et al. 1999; Wang and Fan 1999) have been successfully used in the past to represent the seasonal-mean monsoon and its interannual variability. In this study, we use the same large-scale circulation indices to describe the summer monsoon ISO. We use six daily large-scale dynamical indices as input parameters to the SOM algorithm and, from the dates associated with different SOM nodes, demonstrate that it captures the temporal evolution and the spatial patterns associated with different phases of the monsoon rainfall ISO (Figs. 4 and 5). The driest and wettest patterns of rainfall obtained from using only dynamical parameters (i.e., excluding the rainfall itself) through the use of the SOM algorithm resembles closely the composite active/break rainfall pattern obtained from using rainfall only (from the IMD rainfall). This also implies that the large-scale parameters used in this study are sufficient to determine the rainfall variability and hence are useful for predictions.

The ability of identifying distinct nonlinear phases of rainfall ISO using only large-scale circulation parameters through SOM and without involving any a priori criterion is noteworthy. This not only testifies to the strength of the SOM technique in bringing out the nonlinear coupled states but also establishes that the monsoon ISO is a nonlinear coupled oscillation. The novel feature of the methodology presented here is that it is independent of the datasets used. We demonstrate that the spatial structure and temporal evolution of the phases of the summer monsoon ISO in rainfall can be captured with equal fidelity by using large-scale circulation parameters from ERA-40 as well as NCEP–NCAR reanalysis datasets.

Even though we may be interested only in predicting the intraseasonal component of rainfall variability, predicting the same using rainfall time series is rather difficult due to the intrinsically large day-to-day variability in rainfall. Our demonstration using a nonlinear technique that a suitable combination of large-scale circulation variables are strongly linked to different phases of rainfall low-frequency intraseasonal oscillation (including some of the regional details) provides a framework for devising a prediction strategy for active and break phases without involving rainfall information. By using a 9 × 9 SOM classification instead of a 3 × 3 one, we also show that the large-scale circulation indices can identify different shades of each ensemble mean phase of the precipitation ISO. Based on this knowledge, a formal methodology for analog prediction scheme is proposed and demonstrated that a skillful four-pentad lead prediction of central India rainfall is possible using only circulation fields. The predicted rainfall captures the different phases of ISO and, as it does not involve any filtering, can be effectively used for real-time extended-range prediction of monsoon rainfall.

Acknowledgments. Financial support from the DOD/INDOMOD–SATCORE (ISP 1.5) project, DOD, Government of India, is acknowledged. We are grateful to the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

Ajayamohan, R. S., and B. N. Goswami, 2007: Dependence of simulation of boreal summer tropical intraseasonal oscillations on the simulation of seasonal mean. *J. Atmos. Sci.,* **64,** 460–478.

Ambroise, C., G. Seze, F. Badran, and S. Thiria, 2000: Hierarchical clustering of self-organizing maps for cloud classification. *Neurocomputing,* **30,** 47–52.

Annamalai, H., and J. M. Slingo, 2001: Active/break cycles: Diagnosis of the intraseasonal variability of the Asian summer monsoon. *Climate Dyn.,* **18,** 85–102.

——, and K. R. Sperber, 2005: Regional heat sources and the active and break phases of boreal summer intraseasonal (30–50 day) variability. *J. Atmos. Sci.,* **62,** 2726–2748.

Cavazos, T., 1999: Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas. *J. Climate,* **12,** 1506–1523.

Chatterjee, P., and B. N. Goswami, 2004: Structure, genesis and scale selection of the tropical quasi-biweekly mode. *Quart. J. Roy. Meteor. Soc.,* **130,** 1171–1194.

Chen, L., and J. Gasteiger, 1997: Knowledge discovery in reaction databases: Landscaping organic reactions by a self-organizing neural network. *J. Amer. Chem. Soc.,* **119,** 4033–4042.

Duchon, C. E., 1979: Lanczos filtering in one and two dimensions. *J. Appl. Meteor.,* **18,** 1016–1022.

Elsner, J. B., and A. A. Tsonis, 1992: Nonlinear prediction, chaos, and noise. *Bull. Amer. Meteor. Soc.,* **73,** 49–60.

Ferranti, L., J. M. Slingo, T. N. Palmer, and B. J. Hoskins, 1997: Relations between interannual and intraseasonal monsoon variability as diagnosed from AMIP integrations. *Quart. J. Roy. Meteor. Soc.,* **123,** 1323–1357.

Flatau, M. K., P. J. Flatau, and D. Rudnick, 2001: The dynamics of double monsoon onsets. *J. Climate,* **14,** 4130–4146.

Goswami, B. N., 2005: South Asian Monsoon. *Intraseasonal Variability in the Atmosphere–Ocean Climate System,* W. K.-M. Lau and D. E. Waliser, Eds., Springer-Praxis, 19–62.

——, and J. Shukla, 1984: Quasi-periodic oscillations in a symmetric general circulation model. *J. Atmos. Sci.,* **41,** 20–37.

——, and R. S. Ajaya Mohan, 2001: Intraseasonal oscillations and interannual variability of the Indian summer monsoon. *J. Climate,* **14,** 1180–1198.

——, and P. K. Xavier, 2003: Potential predictability and extended range prediction of Indian summer monsoon breaks. *Geophys. Res. Lett.,* **30,** 1966, doi:10.1029/2003GL017810.

——, and ——, 2005: Dynamics of "internal" interannual variability of the Indian summer monsoon in a GCM. *J. Geophys. Res.,* **110,** D24104, doi:10.1029/2005JD006042.

——, V. Krishnamurthy, and H. Annamalai, 1999: A broad-scale circulation index for the interannual variability of the Indian summer monsoon. *Quart. J. Roy. Meteor. Soc.,* **125,** 611–633.

Goulet, L., and J.-P. Duvel, 2000: A new approach to detect and characterize intermittent atmospheric oscillations: Application to the intraseasonal oscillation. *J. Atmos. Sci.,* **57,** 2397–2416.

Gutiérrez, J. M., A. S. Cofiño, R. Cano, and M. A. Rodríguez, 2004: Clustering methods for statistical downscaling in short-range weather forecasts. *Mon. Wea. Rev.,* **132,** 2169–2183.

——, R. Cano, A. S. Cofiño, and C. Sordo, 2005: Analysis and downscaling multi-model seasonal forecasts in Peru using self-organizing maps. *Tellus,* **57A,** 435–447.

Haykin, S. S., 1999: *Neural Networks: A Comprehensive Foundation.* 2nd ed. Prentice-Hall, 842 pp.

Heskes, T., and B. Kappen, 1995: Self-organization and nonparametric regression. *Proceedings of the International Conference on Artificial Neural Networks,* Vol. 1, F. Fogelman-Soulié and P. Gallinari, Eds., EC2 & Cie, 81–86.

Hewitson, B. C., and R. G. Crane, 2002: Self-organizing maps: Applications to synoptic climatology. *Climate Res.,* **22,** 13–26.

Jiang, X., T. Li, and B. Wang, 2004: Structures and mechanisms of the northward propagating boreal summer intraseasonal oscillation. *J. Climate,* **17,** 1022–1039.

Jones, C., L. M. V. Carvalho, R. W. Higgins, D. E. Waliser, and J.-K. E. Schemm, 2004: A statistical forecast model of tropical intraseasonal convective anomalies. *J. Climate,* **17,** 2078–2095.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Kohonen, T., 1990: The self-organizing map. *Proc. IEEE,* **78,** 1464–1480.

Krishnamurthy, V., and J. Shukla, 2000: Intraseasonal and interannual variability of rainfall over India. *J. Climate,* **13,** 4366–4377.

Krishnamurti, T. N., and D. Subrahmanyam, 1982: The 30–50 day mode at 850 mb during MONEX. *J. Atmos. Sci.,* **39,** 2088–2095.

——, P. K. Jayakumar, J. Sheng, N. Surgi, and A. Kumar, 1985: Divergent circulations on the 30 to 50 day time scale. *J. Atmos. Sci.,* **42,** 364–375.

Krishnan, R., C. Zhang, and M. Sugi, 2000: Dynamics of breaks in the Indian summer monsoon. *J. Atmos. Sci.,* **57,** 1354–1372.

Lawrence, D. M., and P. J. Webster, 2001: Interannual variations of the intraseasonal oscillation in the South Asian summer monsoon region. *J. Climate,* **14,** 2910–2922.

Leloup, J. A., Z. Lachkar, J.-P. Boulanger, and S. Thiria, 2007: Detecting decadal changes in ENSO using neural networks. *Climate Dyn.,* **28,** 147–162.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.,* **26,** 636–646.

Malmgren, B. A., and A. Winter, 1999: Climate zonation in Puerto Rico based on principal components analysis and an artificial neural network. *J. Climate,* **12,** 977–985.

Palakal, M. J., U. Murthy, S. K. Chittajallu, and D. Wong, 1995: Tonotopic representation of auditory responses using self-organizing maps. *Math. Comput. Modell.,* **22,** 7–21.

Rajeevan, M., J. Bhate, J. D. Kale, and B. Lal, 2006: High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells. *Curr. Sci.,* **91,** 296–306.

Ramamurthy, K., 1969: Monsoon of India: Some aspects of "break" in the Indian south west monsoon during July and August. Forecasting manual, Part IV, 18.3, India Meteorological Department, New Delhi, India, 13 pp.

Ramaswamy, C., 1962: Breaks in the Indian summer monsoon as a phenomenon of interaction between the easterly and subtropical westerly jet streams. *Tellus,* **14,** 337–349.

Sikka, D. R., and S. Gadgil, 1980: On the maximum cloud zone and the ITCZ over Indian longitudes during the southwest monsoon. *Mon. Wea. Rev.,* **108,** 1840–1853.

Sperber, K. R., J. M. Slingo, and H. Annamalai, 2000: Predictability and the relationship between subseasonal and interannual variability during the Asian summer monsoon. *Quart. J. Roy. Meteor. Soc.,* **126,** 2545–2574.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Waliser, D. E., and Coauthors, 2003: AGCM simulations of intraseasonal variability associated with the Asian summer monsoon. *Climate Dyn.,* **21,** 423–446.

Wang, B., 2005: Theory. *Intraseasonal Variability in the Atmosphere–Ocean Climate System,* W. K.-M. Lau and D. E. Waliser, Eds., Springer-Praxis, 307–360.

——, and Z. Fan, 1999: Choice of South Asian summer monsoon indices. *Bull. Amer. Meteor. Soc.,* **80,** 629–638.

Webster, P. J., and S. Yang, 1992: Monsoon and ENSO: Selectively interactive systems. *Quart. J. Roy. Meteor. Soc.,* **118,** 877–926.

——, and C. Hoyos, 2004: Prediction of monsoon rainfall and river discharge on 15–30-day time scales. *Bull. Amer. Meteor. Soc.,* **85,** 1745–1765.

——, V. O. Magaña, T. N. Palmer, J. Shukla, R. A. Tomas, M. Yanai, and T. Yasunari, 1998: Monsoons: Processes, predictability, and the prospects for prediction. *J. Geophys. Res.,* **103,** 14 451–14 510.

Xavier, P. K., and B. N. Goswami, 2007: An analog method for real-time forecasting of summer monsoon subseasonal variability. *Mon. Wea. Rev.,* **135,** 4149–4160.

Xie, P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.,* **78,** 2539–2558.

Yasunari, T., 1979: Cloudiness fluctuations associated with the Northern Hemisphere summer monsoon. *J. Meteor. Soc. Japan,* **57,** 227–242.