# Identification of Indian monsoon predictors using climate network and density-based spatial clustering

**Article** *in* Meteorology and Atmospheric Physics · October 2019

DOI: 10.1007/s00703-018-0637-y

CITATIONS
7

READS
296

**2 authors:**

Moumita Saha
University of Colorado Boulder
**28** PUBLICATIONS **264** CITATIONS
SEE PROFILE

Pabitra Mitra
Indian Institute of Technology Kharagpur
**237** PUBLICATIONS **7,449** CITATIONS
SEE PROFILE

Meteorology and Atmospheric Physics
https://doi.org/10.1007/s00703-018-0637-y

ORIGINAL PAPER

# Identification of Indian monsoon predictors using climate network and density-based spatial clustering

**Moumita Saha**[1,2] · **Pabitra Mitra**[2]

Received: 28 August 2017 / Accepted: 23 August 2018


**Abstract**
The Indian summer monsoon is a complex climatic phenomenon with a large variability over the years. The climatic predictors affecting the phenomenon evolve with time, and consequently new predictors have gained importance. Several statistical approaches are being explored in the literature to identify the potential predictors influencing the Indian summer monsoon. A complex network paradigm involving climatic variables at the grids over the globe has been proposed for predictor identification and monsoon prediction. The approach initiates with the identification of communities in the climate network considering mutual similarity and the influence of climate variables of grids on the Indian summer monsoon. Spatial clustering is performed over the communities to identify the geographical regions of significance. The climatic predictors extracted from variables of these regions are evaluated in terms of their correlation with the monsoon as well as their forecasting skills in predicting the summer monsoon of the country. The newly identified predictors forecast monsoon with an error of 4.2%, which is significant for the prediction of the complex phenomenon of monsoon.

## 1 Introduction

Analyzing the climate dynamics as an interacting complex network yields valuable insights into several climatic phenomena. A multiple number of climatic predictors influence the state and dynamics of the climatic phenomenon. The monsoon is a prime and interesting climatic phenomenon that is widely studied (Rajeevan 2001; Gadgil 2003; Gadgil et al. 2005; Guhathakurta and Rajeevan 2008; Wang et al. 2015; Saha et al. 2016b; Saha and Mitra 2016). The dynamism of the monsoon phenomenon results from its dependence over a number of global climatic variables. The variation in the quantity and distribution of monsoon are high. In addition, the influencing predictors of monsoon also evolve over time. Thus, it is important to reconsider the monsoon predictors and explore different climatic variables over the world affecting the complex monsoon phenomenon. We concentrate our study on the Indian summer monsoon and in a complex network paradigm to explore and identify new climatic predictors influencing the phenomenon.

The use of climatic network in earth science is an emerging direction toward analyzing and understanding the climatic phenomena. Tsonis and Roebber (2004) suggested the concept of climatic network and represented the phenomena as a network of dynamic processes. They revealed that the overall dynamics result from interactions of two subsystems, one working in the higher latitudes and other in the tropics.

The climate networks are also built using complex networks-based concepts and they are utilized to figure out the interesting patterns present in the climatic system (Donges et al. 2009a, b). Steinhaeuser et al. (2011) proposed the analysis and modeling of the climatic events using a complex network-based approach. Clusters derived by a complex network approach are proven to be superior predictors than one obtained from the traditional clustering approach. Clustering methods are also used widely to detect the region of importance in the climatic network (Noor and Awan 2005; Steinbach et al. 2003). Steinhaeuser et al. (2010) detected the communities within the climatic system, and the approach was used for the potential predictors' identification in the

Responsible Editor: A.-P. Dimri.

✉ Moumita Saha
moumita.saha2012@gmail.com

Pabitra Mitra
pabitra@gmail.com

[1] Centre for Atmospheric and Oceanic Sciences, Indian Institute of Science, Bangalore, India

[2] Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India

Published online: 11 September 2018

climatic network. The new predictors elucidate reasons behind the changing climatic phenomenon and assist in analyzing the causes behind the phenomenon. Tsonis and Swanson (2008) have built networks for La-Niña and El-Niño. They have shown that the latter network (for El-Niño) is less stable and validated a better predictability of La-Niña event. Major climatic shifts as the transition between different equilibria of oscillators are explained using the climatic network (Tsonis et al. 2007).

The proposed work is focused in two main directions—(1) identification of new monsoon predictors utilizing community detection approach and density-based clustering, and (2) predicting the Indian summer monsoon (ISM) by the identified predictors.

In the proposed approach, climatic networks are built considering the spatial grids of the world as nodes of the network. The nodes are attributed with climatic variables and weighted edges are added by considering the similarity between the nodes. After the building of networks, communities are detected from the networks for identification of significant climatic regions. The community detection-based approach achieves higher performance in detecting similar groups as compared to the clustering method because unlike the clustering approach, the community detection method also focuses on the architecture of the network in addition to the attributes of the nodes. Finally, the density-based clustering is applied to the detected communities to obtain spatially localized regions, which are representative for the new monsoon predictors. The identified predictors are observed to be more correlated to the ISM than the existing predictors of the monsoon. Lastly, the prediction of Indian summer monsoon is performed utilizing the identified correlated monsoon predictors with ensemble regression model. The identified predictors establish their superiority in forecasting the Indian summer monsoon.

Section 2 of the article describes the data, the building of climatic networks, followed by the proposed predictor identification approach using the community detection and density-based clustering methods. The non-linear model for predicting the Indian summer monsoon is elaborated in Sect. 3. The concept of uncertainty and its association with the monsoon forecast is explored in Sect. 4. The detailed exploration of the monsoon predictors is provided with their predicting skills for the Indian monsoon in Sect. 5. Lastly, the article is concluded in Sect. 6.

## 2 Climatic network-based approach for identifying the predictors of Indian summer monsoon

The proposed method for the identification of predictors influencing the summer monsoon of the sub-continent is shown in Fig. 1. It elaborates all the steps followed in the approach to identify novel monsoon predictors, and finally forecasts the summer monsoon of the country.

### 2.1 Data sources and preprocessing techniques

The climatic variable considered are surface pressure (SP) and zonal wind at 850 hPa (UWND), which are the well-known influencing factors of the Indian monsoon phenomenon (Rajeevan et al. 2007; Saha and Mitra 2016; Saha et al. 2017). Surface pressure values and zonal wind values are accumulated from the NCEP reanalysis data NOAA/OAR/ESRL/ PSD (http://www.esrl.noaa.gov) (Kalnay et al. 1996), available at 2.5° × 2.5° resolution. Thus, considering the spatial resolution it boils down to 73 (180/2.5 + 1) latitudinal and 144 (360/2.5) longitudinal grids, which assemble to 10,512 nodes (73 × 144) in the climatic network built for the variable surface pressure (Net_SP) and zonal wind (Net_UWND).

The other climatic variable examined is sea surface temperature (SST), which has a high impact on the climatic phenomenon of monsoon (Rajeevan et al. 2004, 2007; Saha et al. 2016a, b; Saha and Mitra 2016). Sea surface temperature data are collected from NOAA_OI_SST_ V2 (http://www.esrl.noaa.gov) (Reynolds et al. 2002) at 2° × 2° resolution. We have considered the SST data at 4° × 4° grid points to reduce the computational overhead and this network of sea surface temperature (Net_SST) has 4050 (180/4 × 360/4) nodes. These are the initial grid location where sea surface temperature values are examined. Many of these locations

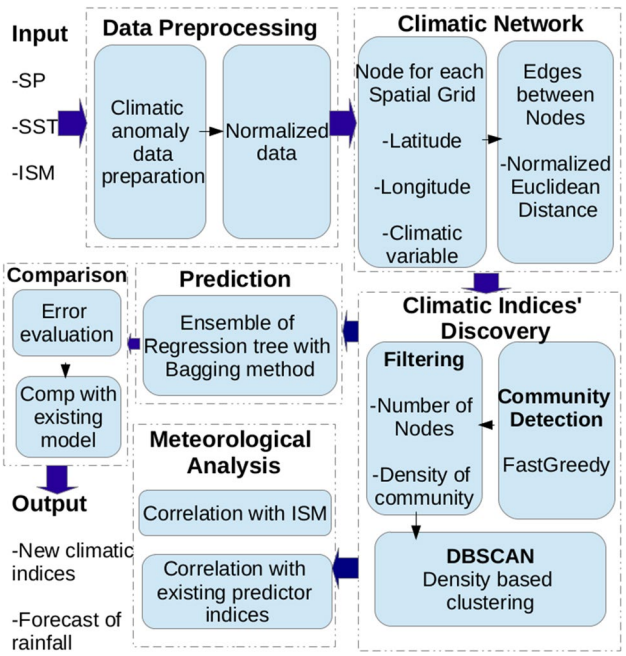


**Fig. 1** Climate network-based method for identifying monsoon predictors and predict the Indian summer monsoon

are over the land and values of sea surface temperature are not available over the land surfaces. Thus, the post-processing method includes the selection of grids over the sea with consideration of nodes having less than 20% as null values over time. The method also comprises the addition of links between the nodes considering the similarity measure. These are elaborated in Sect. 2.2. It is noted that the final networks for SP, UWND, and SST have fewer nodes as compared to the initial nodes. SP, UWND and SST data are examined for the period 1948–2018 on monthly scale for the study.

The prediction of the Indian summer monsoon (ISM), which accounts for total rainfall in June–September is the primary focus of the study. Rainfall data are collected from the India Meteorological Department (IMD: http://www.imdpune.gov.in), for the period 1948–2017. The long period average (LPA) rainfall over the span is 890.1 mm.

As a preprocessing step, the SP, UWND and SST anomaly values are evaluated by deducting the monthly mean from the respective month values of the variables (Eq. 1).

$$\text{anomalyData}_m^y = \text{realData}_m^y - \text{mean}(\text{realData}_m), \tag{1}$$

where $\text{realData}_m^y$ denotes the value of the variable for $m$th month of $y$th year. The mean ($\text{realData}_m$) signifies the average value of all years under study for the $m$th month.

## 2.2 Design of climatic network and link thresholding

The introductory step of the proposed approach involves the creation of climatic networks for variables, namely, surface pressure, zonal wind at 850 hPa, and sea surface temperature. The spatial grids at a resolution of 2.5° × 2.5° for SP and UWND, and 4° × 4° for SST over the world are considered as nodes in the respective networks. The network built for SP and UWND have 10512 nodes, and that for SST has 4050 nodes at the initial phase. The latitude, longitude, and the variable values over time at grid points characterize the nodes of the network. The values of the variable SST over the land surface are *null*. Such null nodes are eliminated from the network in the post-processing phase. The weighted edges are inserted considering the similarity between every node pair in terms of normalized euclidean distance (NED). The NED is calculated as shown in Eq. (2).

$$\begin{aligned}\text{NED}_{(n,m)} = &\left(\text{ED}_{(n,m)} - \left(\forall_{(x,y):x,y\in G,x\neq y} min\left(\text{ED}_{(x,y)}\right)\right)\right)\\ &\div\left[\left(\forall_{(x,y):x,y\in G,x\neq y}\max\left(\text{ED}_{(x,y)}\right)\right)\right.\\ &\left.-\left(\forall_{(x,y):x,y\in G,x\neq y}\min\left(\text{ED}_{(x,y)}\right)\right)\right],\end{aligned} \tag{2}$$

$$\text{ED}_{(n,m)} = \sqrt{\sum_{i=1}^{t}\left(n_i - m_i\right)^2},$$

where $(n, m)$ denotes an edge between the nodes $n$ and $m$, $G$ denotes the set of nodes, $\text{ED}_{(n,m)}$ denotes the Euclidean distance between the climatic variable's time series at nodes $n$ and $m$; and $t$ denotes the length of variable time series.

An edge is added between two nodes if the normalized Euclidean distance between nodes attains the threshold, computed in the following manner. The range of NED between all pairs of nodes is divided into 100 intervals and the occurrence frequency of weights in all the intervals are plotted. The sharp descent of the graph plot is ascertained as threshold and edges having NED less than the threshold (lesser the NED, closer are the nodes) are added (Fig. 2). The threshold NED values for Net_SP, Net_UWND, and Net_SST are 0.05, 0.03, and 0.02, respectively. We also varied the threshold around the ascertained threshold value and repeated the proposed approach. The changes in accuracy with the varying thresholds are observed to be comparable. Thus, we considered the ascertained threshold for building the edges of the climatic network.

Lastly, the networks are post-processed by removing the isolated nodes for building the connected network. After the post-processing and link addition, the final network built with the surface pressure variable has 9614 nodes and 196,606 edges, that for the zonal wind at 850 hPa has 7464 nodes and 10,416 edges, and finally, for the sea surface temperature variable has 1543 nodes and 1,094,514 edges. It is noted that the network for SST is immensely dense, which signifies that the variation of sea surface temperature between spatial grids is less as compared to that of sea level pressure.

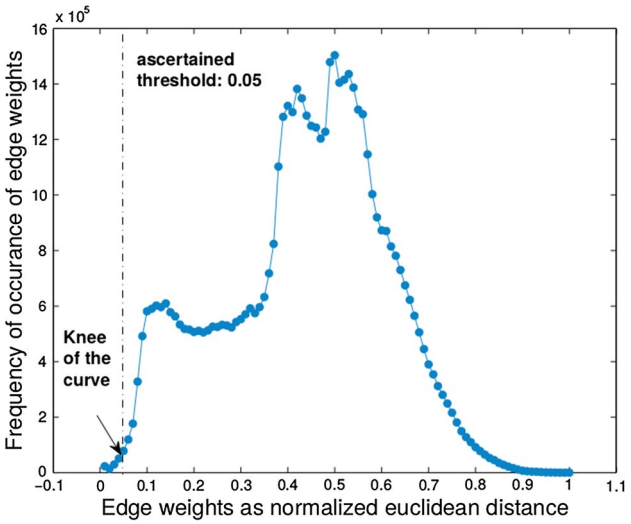


**Fig. 2** The frequency of edge weights at various intervals to calculate the threshold for surface pressure variable

### 2.3 Community detection followed by the density-based clustering for identifying the monsoon predictors

The proposed approach consists of three major steps, which are discussed in the following section.

(a) Identifying the communities of the climatic network.
(b) Filtering and the selection of the detected communities.
(c) Identifying the geographically localized regions of interest from the communities.

The fast-greedy community detection method (Clauset et al. 2004) is applied over the climatic network to detect communities. Communities aid to identify new potential climatic predictors influencing the Indian summer monsoon. The algorithm is selected considering the following properties—(1) utilization of the edge weights of the network, (2) high suitability for the intense and dense networks, and (3) computational efficiency in finding the communities within the network.

The fast-greedy is a hierarchical agglomeration method which optimizes the modularity of the network. It performs greedy optimization starting with the individual vertex being the community of single dimension. Any two different communities are constantly joined into a single community, whose combination produces the highest improvement in the modularity of the communities. The stopping criterion for the algorithm is the time when there is no further improvement in the modularity.

The communities by fast-greedy community detection are filtered by scrutinizing the density of nodes in the communities. This value is selected empirically.

The obtained communities may be sparsely located, which are processed to obtain geographically localized communities using density-based spatial clustering (DBSCAN) (Ester et al. 1996). The algorithm is used because it is a spatial clustering technique which aids in extracting a localized set of grids. Other supplementary reasons include—(1) number of clusters is not required a priori, (2) capability in detecting arbitrarily shaped clusters, (3) it is one scan, and (4) the approach is robust to outliers.

The latitude and longitude of grids in the communities are fed to DBSCAN to obtain a set of spatially localized dense clusters, which are representative for the potential monsoon predictors.

### 2.4 Identification of the climatic predictors from the clusters

The spatially localized clusters are considered to evaluate the new monsoon predictors. Each cluster consists of a number of grid points. For a specific cluster, the mean time series is evaluated over all the series of grids within the cluster. This mean time series represents the newly identified potential predictor. The evaluation of predictor variable is shown in Eq. (3). Thus, each cluster represents a newly identified potential monsoon predictor. A few identified predictors signify well-known monsoon predictors, symbolizing for the validation of our proposed method of identifying the monsoon predictors, while the others represent new localized geographical regions, which are significant for the phenomenon of the Indian monsoon.

$$\text{identified predictor} = \frac{\sum_{i=1}^{k} \left(P_i\right)}{k}, \tag{3}$$

where $P_i$ denotes climatic variable time series of the $i$th grid of localized cluster, and $k$ is the number of grid points within a localized cluster (representative of identified predictor).

## 3 Prediction model with identified monsoon predictors

Fitted ensembled regression tree model with bagging algorithm (RegTreeB) (MATLAB 2012) is used as the prediction model, which assembles a number of weak learner-trained models to provide the forecast. The model predicts the ensemble response by aggregating the predictions obtained from the trained weak learner regression models. The bagging method is utilized for building and training the regression tree weak learners of the ensemble model.

The prediction model is built using this algorithm for the following reasons—(1) it uses the bagging, a bootstrap aggregating method improvising estimation, (2) it assists in increasing the predictive ability of the underlying regression tree, and (3) the algorithm can work with a large number of training instances and high dimensional data.

## 4 Uncertainty associated with monsoon prediction

Ascertaining uncertainty involved in the forecast of the monsoon has a significance. Decision makers need to analyze the uncertainty involved in the monsoon to propose justifiable strategies. Uncertainty should be appreciably communicated or it may lead to a false certainty sense, improper decision-making, and overall reduced performance in the forecast. Uncertainty in the forecast arises from the probabilistic forecast of the phenomenon.

The uncertainty arises from different sources and it may be first-order or second-order uncertainty. The first-order uncertainty points toward the likelihood of a phenomenon occurring in accordance with a particular forecast or the risk

involved in it, whereas the second-order uncertainty is cited as 'uncertainty about uncertainty'. It results from how well the model has adapted to forecast or it highlights the model errors in execution. It is represented as a measure of reliability.

Taylor et al. (2015) propose propagating the uncertainty in the climate forecast in a preferable receiving format. They utilize surveys from user needs conducted on different organizations. Azad et al. (2015) showed that the uncertainty in predicting the monsoon of India is reduced. The performance is improved by treating the periodic and random part of the time series data separately with wavelet and neural network, respectively. We explain the uncertainty in the prediction of monsoon using measures like root mean square error, bias, correlation coefficient, and Willmott index, which define the uncertainty involved with the model or they give a measure of explaining how well the forecasts have satisfied the actual values.

## 5 Experimental results and analysis

The proposed climate network-based approach to identifying the monsoon predictors is judged by the measure of performance of identified predictors in forecasting the Indian monsoon.

### 5.1 Identified monsoon predictors

A correlation investigation of new monsoon predictors is performed with the prime monsoon period of India (i.e., the total rainfall during June–September) by considering a lead of 1–12 months. The lead months are considered to evaluate the best correlated month (a lead of one represents the month of May in the same year of predictor influencing monsoon of the year (monsoon starts in June), a lead of 2 represents April of the same year influencing the monsoon of the year, and finally a lead of 12 represents June of the previous year influencing the present year's monsoon). Pearson correlation ($\gamma$), shown in Eq. (4), is used for the purpose. The *best lead month* corresponds to the month of identified climatic predictor which has the highest correlation with the monsoon of India. The variable value of the best correlated month are used for further forecast. Top correlated identified predictors are filtered for all three variables, and are shown in Table 1. The table highlights the location of identified predictors along with their correlation values and the best correlated month with the Indian summer monsoon.

$$\gamma = \frac{\sum_{i=1}^{N}\left(X^i - \overline{X}\right)\left(Z_m^i - \overline{Z_m}\right)}{\sqrt{\sum_{i=1}^{N}\left(X^i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Z_m^i - \overline{Z_m}\right)^2}}, \tag{4}$$

where $X^i$ and $Z_m^i$ represent the Indian summer monsoon rainfall (total rainfalls for June–September) at $i$th year and identified climatic predictors of $m$th month at the $i$th year, $\overline{X}$ is the averaged monsoon rainfall and $\overline{Z_m}$ is the averaged climatic predictor for $m$th month, and $N$ is the total years. The identified predictors are ordered by their correlation with the Indian summer monsoon (the first predictor having the highest correlation and the last having the lowest). The identified predictors of surface pressure, sea surface temperature, and zonal wind are shown in Fig. 3a–c, respectively.

The identified predictors by the proposed climate network-based approach can be classified into two classes. Firstly, one class consists of the predictors belonging to regions which are well-known monsoon predictors. Identifying the established predictors supports our proposed method. Secondly, the ther class consists of predictors belonging to the new geographical regions whose impact on the Indian summer monsoon are not studied in literature. They are presented as a new set of monsoon predictors of the country.

Re-identification of the influencing monsoon predictors include surface pressure of the Equatorial South-Eastern Indian Ocean (S2), and the disturbance of this region affects the tropical climate (Achuthavarier et al. 2012). The region of the Pacific Ocean around Indonesia and Malaysia (S5) are studied to be influencing the summer monsoon of India (Rajeevan et al. 2004). The South-Eastern Equatorial Indian Ocean is shown to be tele-connected with the tropical Indian Ocean influencing the monsoon. It is notified that the change in sea surface temperature over the band stretching around the Equator over the Pacific Ocean (T2) influences the Indian summer monsoon (Rajagopalan and Molnar 2012). The sea surface temperature of the Equatorial East Pacific Ocean corresponds to the El Nĩno region, a known regulating factor for the Indian summer monsoon (Cherchi and Navarra 2013). Regarding 850 hPa zonal wind, Central North–South Pacific Ocean (U1) is also considered as an important monsoon predictor by the India Meteorology Department for predicting monsoon (Rajeevan et al. 2007). Moreover, North Pacific Ocean–Gulf-of-Alaska 850 hPa UWND (U3) and Equatorial North Pacific Ocean 850 hPa UWND (U4) shared similar regions of the North Pacific Ocean, which is a well-known monsoon predictor of India (Rajeevan et al. 2004, 2007). Finally, the Southern Indian Ocean 850 hPa UWND has evolved as an important predictor which is also used by Rajeevan et al. (2004) to forecast the Indian monsoon.

Newly identified climatic predictors include surface pressure of the Southern Indian Ocean (S1), the Tasmania–Southern Ocean (S3), and the region of Southern Ocean (S7). Other new predictors are the sea surface temperature of the Solomon Islands–Fiji–Pacific Ocean (T1), the Tasmania–South Indian Ocean (T3), the Philippine Sea (T5), the region of Southern Ocean (T6), and the South Pacific

**Table 1** Identified monsoon predictors (Pred.) for surface pressure, sea surface temperature, and zonal wind with geographical location, absolute correlation (Corr. value) and correlated month (Corr. month) (0 signifies the same year and − 1 signifies the previous year)

| | Pred. | Predictor name | Location | Corr. value | Corr. month |
|---|---|---|---|---|---|
| SP | $S_1$ | Southern Indian Ocean SP | 42°S–48°S, 50°E–80°E | 0.352 | Mar (0) |
| | $S_2$ | Equatorial South-Eastern Indian Ocean SP | 15°N–12°S, 75°E–100°E | 0.341 | May (0) |
| | $S_3$ | Tasmania–Southern Ocean SP | 45°S–48°S, 115°E–130°E | 0.333 | Sep (− 1) |
| | $S_4$ | Madagascar Equatorial South -Western Indian Ocean SP | 0°–10°S, 45°E–70°E | 0.298 | May (0) |
| | $S_5$ | Indonesia–Malaysia SP | 8°N–8°S, 102°E–132°E | 0.282 | May (0) |
| | $S_6$ | South-Eastern Equatorial Ocean SP | 25°S–38°S, 80°E–115°E | 0.243 | Sep (− 1) |
| | $S_7$ | Southern Ocean SP | 55°S–68°S, 40°E–60°E | 0.218 | Jul (− 1) |
| SST | $T_1$ | Solomon Islands–Fiji–Pacific Ocean SST | 2°S–15°S, 150°E–180°E | 0.347 | Oct (− 1) |
| | $T_2$ | North-West–Central Pacific Ocean SST | 28°N–13°S, 180°E–210°E | 0.274 | Dec (− 1) |
| | $T_3$ | Tasmania–Southern-Indian Ocean SST | 40°S–60°S, 140°E–150°E | 0.269 | Mar (0) |
| | $T_4$ | Equatorial East Pacific Ocean SST | 18°N–20°S, 210°E–260°E | 0.267 | Aug (− 1) |
| | $T_5$ | Philippine SST | 5°N–20°N, 130°E–140°E | 0.265 | Jun (− 1) |
| | $T_6$ | Southern Ocean SST | 40°S–68°S, 70°E–85°E | 0.263 | Jan (0) |
| | $T_7$ | South Pacific Ocean SST | 50°S–72°S, 230°E–260°E | 0.247 | Apr (0) |
| UWND | $U_1$ | Central North–South Pacific Ocean 850 hPa zonal wind | 30°N–35°S, 190°E–200°E | 0.437 | May (0) |
| | $U_2$ | Equatorial South Pacific Ocean 850 hPa UWND | 8°N–48°S, 225°E–240°E | 0.394 | May (0) |
| | $U_3$ | North Pacific Ocean–Gulf-of-Alaska 850 hPa UWND | 30°N–50°N, 194°E–214°E | 0.364 | May (0) |
| | $U_4$ | Equatorial North Pacific Ocean 850 hPa UWND | 30°N–10°S, 210°E–228°E | 0.356 | May (0) |
| | $U_5$ | United States–Mexico–Gulf-of-Mexico 850 hPa UWND | 45°N–0°, 250°E–282°E | 0.280 | Apr (0) |
| | $U_6$ | Southern Indian Ocean 850 hPa UWND | 35°S–55°S, 25°E–50°E | 0.279 | Sep (− 1) |
| | $U_7$ | South Pacific Ocean 850 hPa UWND | 25°S–75°S, 245°E–280°E | 0.267 | Apr (0) |
| | $U_8$ | North-Central Russia–China 850 hPa UWND | 35°N–65°S, 90°E–100°E | 0.255 | Feb (0) |

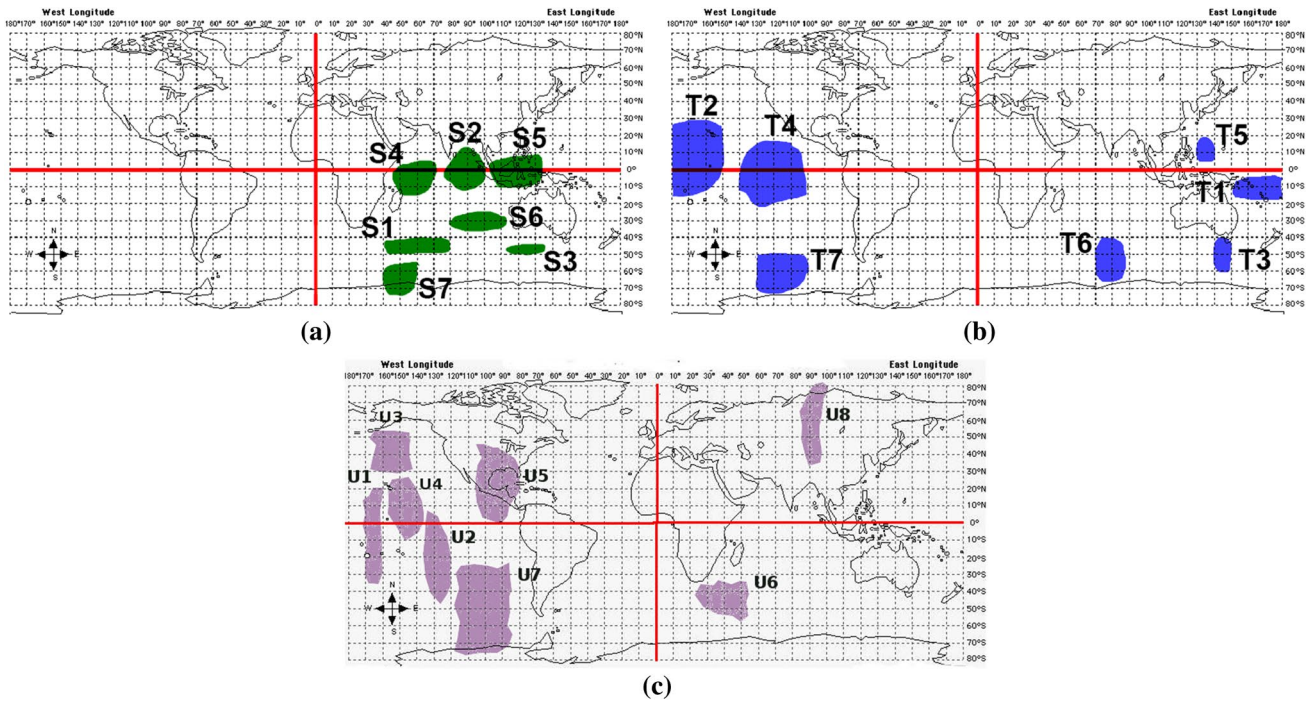


**Fig. 3** Identified climatic predictors for **a** surface pressure ($S_1$–$S_7$), **b** sea surface temperature ($T_1$–$T_7$), and **c** zonal wind ($U_1$–$U_8$). Monsoon predictors are arranged in accordance with their correlation with the summer monsoon of India (i.e., $S_1$ been most highly correlated and $S_7$ been the least)

Ocean (T7). Newly identified zonal wind-based predictors include Equatorial South Pacific Ocean 850 hPa UWND (U2), United States–Mexico–Gulf-of-Mexico 850 hPa UWND (U5), South Pacific Ocean 850 hPa UWND (U7), North-Central Russia–China 850 hPa UWND (U8). These regions are shown to correlate the Indian summer monsoon at different lead months (refer to Table 1).

We have presented the *top seven* climatic predictors for surface pressure and sea surface temperature variables, and the *top eight* for zonal wind variable, corresponding to regions obtained from the proposed approach. Reasons behind presenting these predictors are—(1) correlation values of other identified predictors with monsoon are lower compared to the presented set, (2) it is observed in literature that the predictor set with five–six predictors performs superiorly for the monsoon prediction (Rajeevan et al. 2007).

## 5.2 Prediction of monsoon with identified predictors

### 5.2.1 Predictor sets

The predictor sets are built considering the correlation of the identified predictors with the monsoon of India and their lead months of forecasting. Predictors are chosen in a way such that they forecast the monsoon in two different leads. D1_Y, D2_Y, D3_Y, and D4_Y denote the predictor sets (Y denotes either S for predictors of SP, U for predictors of UWND, T for that of SST, or S_U, S_T, U_T, and S_U_T for respectively combined predictors). Tables 2 and 3 show the identified predictors considered for individual predictor set along with the lead number of months, which possesses the best correlation with the monsoon. Finally, considering the lead months of the individual predictors, it declares the month for providing monsoon prediction.

### 5.2.2 Prediction performance

A non-linear model named ensemble regression tree with the bagging method (Sect. 3) is considered to forecast the monsoon. The prediction model is trained in two ways. The first method segregates the total period under the study into a separate set of training and test years, and the model is trained only once with the set of training instances, and tested over test instances. The second method uses the strategy of moving-window training. We calculate an optimal training period and, for every test year, the model is trained using instances of the preceding optimal number of years. Thus, for testing $t$ number of years, the model is required to be trained separately for all the cases (i.e., it is trained $t$ number of times).

In our first approach, the total period (1948–2017) is divided into an exclusive set of training and test set considering a 70–30 ratio. The model is trained for the period 1948–1994 and tested for 1995–2017. The prediction is evaluated by the mean absolute error (MAE), expressed as follows.

$$\text{MAE} = \frac{\sum_{i=1}^{N} \mid Y_i - X_i \mid}{N},$$

where $X_i$ and $Y_i$ are the predicted and observed monsoon for the $i$th year and $N$ denotes the total years.

The training errors for all the individual predictor with the ensemble regression tree model is presented in Table 4.

The prediction model and the identified predictors are evaluated over 23 years of the test period (1995–2017). The test errors are presented in terms of mean absolute errors. The predictions by individual predictor variable (SP, UWND, and SST) with static training period (first approach) are presented in Table 5 and those for combined identified predictors (SP + UWND, UWND + SST, SP + SST, and SP + UWND + SST) are shown in Table 6.

**Table 2** Predictor sets (Pred. sets) with the individual predictors of SP, UWND and SST for forecasting the Indian summer monsoon

| | Pred. sets | Identified predictors | Best lead month | Pred. month |
|---|---|---|---|---|
| SP | D1_S | $S_1, S_3, S_6, S_7$ | 3, 9, 9, 11 | March |
| | D2_S | $S_1, S_3, S_6$ | 3, 9, 9 | March |
| | D3_S | $S_1, S_2, S_3, S_4$ | 3, 1, 9, 1 | May |
| | D4_S | $S_1, S_2, S_3, S_4, S_5, S_6$ | 3, 1, 9, 1, 1, 9 | May |
| UWND | D1_U | $U_5, U_6, U_7, U_8$ | 2, 9, 2, 4 | April |
| | D2_U | $U_1, U_2, U_3, U_4, U_5, U_6$ | 1, 1, 1, 1, 2, 9 | May |
| | D3_U | $U_1, U_5, U_6, , U_7, U_8$ | 1, 2, 9, 2, 4 | May |
| | D4_U | $U_2, U_5, U_6, U_7, U_8$ | 1, 2, 9, 2, 4 | May |
| SST | D1_T | $T_1, T_2, T_3, T_4, T_5$ | 8, 6, 3, 10, 12 | March |
| | D2_T | $T_1, T_2, T_3, T_4, T_5, T_6$ | 8, 6, 3, 10, 12, 5 | March |
| | D3_T | $T_1, T_2, T_3, T_4, T_5, T_6, , T_7$ | 8, 6, 3, 10, 12, 5, 2 | April |
| | D4_T | $T_3, T_4, T_5, T_6, T_7$ | 3, 10, 12, 5, 2 | April |

**Table 3** Predictor sets (Pred. sets) with the combined predictors of SP + UWND, UWND + SST, SP + SST and SP + UWND + SST for forecasting the Indian summer monsoon

| | Pred. sets | Identified predictors | Best lead month | Pred. month |
|---|---|---|---|---|
| S + U | D1_S_U | $S_1, S_3, U_5, U_6, U_7, U_8$ | 3, 9, 2, 9, 2, 4 | April |
| | D2_S_U | $S_1, S_3, S_6, S_7, U_5, U_6, U_7, U_8$ | 3, 9, 9, 11, 2, 9, 2, 4 | April |
| | D3_S_U | $S_1, S_2, S_3, U_1, U_5, U_6, U_7, U_8$ | 3, 1, 9, 1, 2, 9, 2, 4 | May |
| | D4_S_U | $S_1, S_2, S_3, S_4, S_5, U_2, U_5, U_6, U_7, U_8$ | 3, 1, 9, 1, 1, 2, 9, 2, 4 | May |
| U + T | D1_U_T | $U_5, U_6, U_7, U_8, T_1, T_2, T_3$ | 2, 9, 2, 4, 8, 6, 2 | April |
| | D2_U_T | $U_5, U_6, U_7, U_8, T_4, T_5, T_6$ | 2, 9, 2, 4, 10, 12, 5 | April |
| | D3_U_T | $U_1, U_5, U_6, U_7, U_8, T_1, T_2, T_3, T_4$ | 1, 2, 9, 2, 4, 8, 6, 2 | May |
| | D4_U_T | $U_1, U_5, U_6, U_7, U_8, T_5, T_6, T_7$ | 1, 2, 9, 2, 4, 12, 5, 4 | May |
| S + T | D1_S_T | $S_1, S_3, S_6, T_1, T_2, T_3$ | 3, 9, 9, 8, 6, 3 | March |
| | D2_S_T | $S_1, S_3, S_6, S_7, T_1, T_2, T_3$ | 3, 9, 9, 11, 8,6, 3 | March |
| | D3_S_T | $S_1, S_3, S_6, T_3, T_4, T_5, T_6, T_7$ | 3, 9, 9, 3, 10, 12, 5, 2 | April |
| | D4_S_T | $S_1, S_3, S_6, S_7, T_3, T_4, T_7$ | 3, 9, 9, 11, 3,10, 2 | April |
| S + U + T | D1_S_U_T | $S_1, S_3, U_5, U_6, T_1, T_2$ | 3, 9, 2, 9, 8, 6 | March |
| | D2_S_U_T | $S_6, S_7, U_7, U_8, T_4, T_5$ | 9, 11, 2, 4, 10, 12 | April |
| | D3_S_U_T | $S_1, S_3, U_1, U_5, U_6, T_6, T_7$ | 3, 9, 1, 2, 9, 5, 2 | May |
| | D4_S_U_T | $S_5, S_6, U_2, U_5, U_6, T_1, T_2$ | 2, 9, 2, 9, 2, 8, 6 | May |

**Table 4** Training errors as mean absolute errors (%) for Indian monsoon prediction with SP, UWND, and SST for 1948–1994

| SP | | UWND | | SST | |
|---|---|---|---|---|---|
| Pred. set | Trng. error | Pred. set | Trng. error | Pred. set | Trng. error |
| D1_S | 6.0 | D1_U | 5.2 | D1_T | 5.5 |
| D2_S | 5.7 | D2_U | 5.1 | D2_T | 5.5 |
| D3_S | 5.3 | D3_U | 5.0 | D3_T | 5.3 |
| D4_S | 5.8 | D4_U | 5.4 | D4_T | 5.9 |

**Table 5** Mean absolute errors (%) for forecasting the Indian monsoon, with a static training span using individual predictors of SP, UWND, and SST for the test period 1995–2017

| SP | | UWND | | SST | |
|---|---|---|---|---|---|
| Pred. set | Pred. error | Pred. set | Pred. error | Pred. set | Pred. error |
| D1_S | 5.8 | D1_U | 5.7 | D1_T | 6.3 |
| D2_S | 5.8 | D2_U | 5.1 | D2_T | 5.8 |
| D3_S | 6.3 | D3_U | **4.9** | D3_T | 5.7 |
| D4_S | **5.3** | D4_U | 5.2 | D4_T | **5.6** |

Bold indicates the minimum error by every predictor variables

The second approach of moving-window training strategy is also utilized for predicting the Indian summer monsoon. In this method, if the number of training years is $n$, then for testing the $t$th test year, training years for the model is considered from $(t-1)$th to $(t-n-1)$th years (i.e., considering the number of training years as *ten*, and if we need to test for the year 1995, then the model is trained with the data of 1985–1994). Thus, the model needs to be trained individually for every test year with corresponding preceding training years. The training period is inspected from 5 to 45 years. The optimal period is observed as 20 years, which gives comparatively less error. Results for individual and combined variables for moving-window training strategy are shown in Tables 7 and 8. It is observed that the results are superior to that of the static training process as followed in our first approach. The reason underlying is that the moving-window training method can engross the pattern of variability of the close period of the test year.

Predictor set with identified predictors of surface pressure shows a mean absolute error of 4.6% in predicting the monsoon in May. The UWND predictors provide 4.4%

**Table 6** Mean absolute errors (%) for forecasting the Indian monsoon, with a static training span with the combined predictors of SP + UWND, UWND + SST, SP + SST, and SP + UWND + SST for the test period 1995–2017

| SP + UWND | | UWND + SST | | SP + SST | | SP + UWND + SST | |
|---|---|---|---|---|---|---|---|
| Pred. set | Pred. error | Pred. set | Pred. error | Pred. set | Pred. error | Pred. set | Pred. error |
| D1_S_U | 5.3 | D1_U_T | 6.0 | D1_S_T | 6.0 | D1_S_U_T | 5.3 |
| D2_S_U | 5.0 | D2_U_T | 5.7 | D2_S_T | 6.3 | D2_S_U_T | 6.2 |
| D3_S_U | **4.9** | D3_U_T | 5.6 | D3_S_T | **5.9** | D3_S_U_T | **5.1** |
| D4_S_U | **4.9** | D4_U_T | **4.6** | D4_S_T | 6.1 | D4_S_U_T | **5.1** |

Bold indicates the minimum error by every predictor variables

**Table 7** Mean absolute errors (%) for forecasting the Indian monsoon, with a moving-window training span with individual predictors of SP, UWND, and SST for the test period 1995–2017

| SP | | UWND | | SST | |
|---|---|---|---|---|---|
| Pred. set | Pred. error | Pred. set | Pred. error | Pred. set | Pred. error |
| D1_S | 4.7 | D1_U | **4.4** | D1_T | **5.1** |
| D2_S | 5.1 | D2_U | 4.6 | D2_T | 5.4 |
| D3_S | **4.6** | D3_U | 4.6 | D3_T | 5.2 |
| D4_S | 4.7 | D4_U | 4.8 | D4_T | 5.4 |

Bold indicates the minimum error by every predictor variables

error in April. The identified predictors of SST predict monsoon with 5.1% error in March.

Moreover, the predictor sets with combined predictors also forecast monsoon with good accuracy. Predictor set with surface pressure and zonal wind (SLP+UWND) predicts the monsoon in April with 4.4% error. Similarly, UWND + SST and SP + SST predictor sets predict Indian monsoon at 2 months lead with 4.5% and 4.9% errors, respectively. Finally, the predictor set built with all three variables (SP + UWND + SST) shows 4.2% error in forecasting the monsoon in May.

Figure 4a, b shows the actual and predicted rainfall as deviation from the long period average rainfall (LPA) for predictor sets with individual variable and combined variables, respectively. The result highlights that the predicted rainfall shows a similar deviation of rainfall (negative or positive departure from LPA) as actual for the majority of the test years.

All the extremes (drought—2002, 2004, 2009, 2014, 2015) are forecast with a negative anomaly from the LPA by the combined predictor sets. The predictors of surface pressure capture the drought year of 2014, and SP + SST predictors correctly capture the drought of 2009. For numerical models, even the direction of the deviation of predicted rainfall from LPA is incorrect in many years (Nanjundiah et al. 2013). Therefore, the identified predictors by the climate network-based method improve the accuracy of monsoon prediction of India.

**Table 8** Mean absolute errors (%) for forecasting the Indian monsoon, with a moving-window training span with combined predictors of SP + UWND, UWND + SST, SP + SST, and SP + UWND + SST for the test period 1995–2017

| SP + UWND | | UWND + SST | | SP + SST | | SP + UWND + SST | |
|---|---|---|---|---|---|---|---|
| Pred. set | Pred. error | Pred. set | Pred. error | Pred. set | Pred. error | Pred. set | Pred. error |
| D1_S_U | 4.7 | D1_U_T | **4.5** | D1_S_T | 5.0 | D1_S_U_T | 4.6 |
| D2_S_U | **4.4** | D2_U_T | 5.1 | D2_S_T | 5.0 | D2_S_U_T | 5.2 |
| D3_S_U | **4.4** | D3_U_T | **4.5** | D3_S_T | **4.9** | D3_S_U_T | **4.2** |
| D4_S_U | 4.6 | D4_U_T | 4.9 | D4_S_T | 5.3 | D4_S_U_T | 4.8 |

Bold indicates the minimum error by every predictor variables

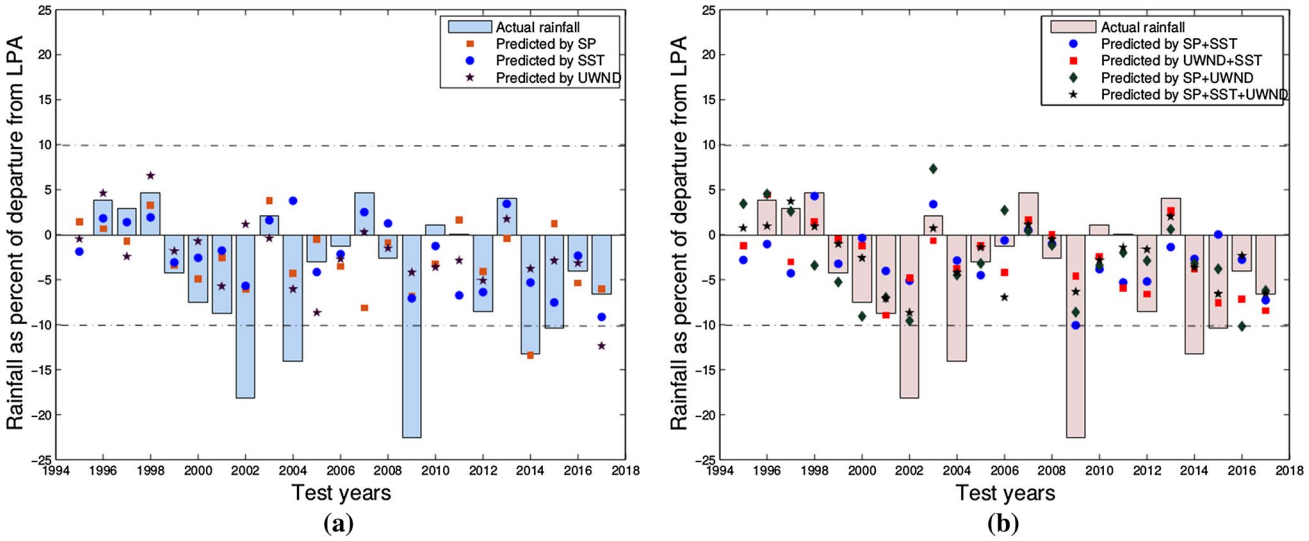


**Fig. 4** Forecasts by predictor set with the identified **a** individual predictors of SP (D4_S), UWND (D1_U), SST (D4_T), and **b** combined variables of SP + UWND (D1_S_U), UWND + SST (D1_U_T), SP + SST (D1_S_T), and SP + UWND + SST (D1_S_U_T) for the test period 1995–2017

#### 5.2.3 Other evaluation measures of prediction

We have also evaluated the performance of the predictors using different statistical measures. The measures and their corresponding results are elaborated in this section.

(a) Root mean square error (RMSE): calculates the variation of the model output against actual values.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_i - X_i\right)^2}{N}},$$

where $X_i$ and $Y_i$ are the observed and predicted monsoons for the $i$th year, and $N$ is the total years as defined earlier.

(b) Prediction yield (PY): is evaluated at three different error categories (5%, 10%, and 15%) to assess the overall prediction by judging the number of predicted years within the allowed errors range.

(c) Multiplicative bias (MB): is the ratio of the predicted to actual value; a closer value to 1 signifies good performance.

(d) Pearson correlation coefficient ($\gamma$): measures the strength of the linear association between the actual and predicted value (shown in Eq. 4).

(e) Willmott index of agreement (WI): is a measure calculating the degree of model prediction, with higher values indicating a better fit of model. It is shown in Eq. (5).

$$\text{Index of agreement} = 1 - \frac{\sum_{i=1}^{N} \mid X_i - Y_i \mid^2}{\sum_{i=1}^{N} (\mid Y_i - \overline{X} \mid + \mid X_i - \overline{X} \mid)^2}. \quad (5)$$

Table 9 elaborates the verification statistics for different predictor sets for forecasting the Indian summer monsoon. The monsoon prediction by combined predictors of surface pressure, zonal wind and sea surface temperature is observed to be superior as compared to the other predictors or their combinations.

We have also presented the skills of identified predictors by investigating their productivity in predicting the rainfall's negative or positive deviation from the LPA. A confusion matrix (Table 10) is used for the purpose. We have compared the predicted negative or positive deviation with the observed rainfall deviation from LPA.

True positive (TP) denotes the count of test years when both observed and predicted rainfall show positive deviation from LPA, true negative (TN) denotes the count of test years when both observed and predicted rainfall show negative deviation from LPA, false positive (FP) represents the count of test years where the observed rainfall shows negative deviation but it is predicted as positive deviation from LPA, and false negative (FN) represents the count of test years when the rainfall is predicted as negative deviation but the observed rainfall shows positive deviation from LPA. The related measures to the confusion matrix are defined as follows.

(a) Sensitivity: proportion of years that is correctly predicted as positive deviation from total observed positive deviation (TP/(TP + FN)).

(b) Specificity: proportion of years that is correctly predicted as negative deviation from total observed negative deviation (TN/(TN + FP)).

(c) Precision: proportion of positive deviation that is predicted correctly from the total number of predicted positive deviations (TP/(TP + FP)).

(d) Negative predictive value: proportion of negative deviation that is predicted correctly from the total number of predicted negative deviations (TN/(TN + FN)).

(e) Accuracy: proportion of years when it is correctly predicted to be the same as the observed deviation ((TP + TN)/(TP + TN + F + FN)).

**Table 10** Confusion matrix

| | Predicted | |
|---|---|---|
| | Positive | Negative |
| Observed | | |
| Positive | True positive (TP) | False negative (FN) |
| Negative | False positive (FP) | True negative (TN) |

**Table 9** Prediction evaluation measures for the model with the identified climatic predictors of SP, UWND, SST, SP + UWND, UWND + SST, SP + SST and SP + UWND + SST for the test period 1995–2017

| Verification measures | RMSE (%) | PY (%) at 5% | PY (%) at 10% | PY (%) at 15% | MB | $\gamma$ | WI |
|---|---|---|---|---|---|---|---|
| SP | 6.3 | 0.74 | 0.83 | 0.96 | 1.02 | 0.55 | 0.66 |
| UWND | 6.4 | 0.73 | 0.86 | 0.91 | 1.02 | 0.57 | 0.66 |
| SST | 7.0 | 0.60 | 0.91 | 0.91 | 1.02 | 0.41 | 0.57 |
| SP + UWND | 5.7 | 0.60 | 0.91 | **1.00** | 1.02 | 0.67 | 0.74 |
| UWND + SST | 6.2 | 0.69 | 0.86 | 0.95 | 1.02 | 0.57 | 0.62 |
| SP + SST | 6.4 | 0.60 | 0.78 | 1.0 | 1.02 | 0.57 | 0.61 |
| SP + UWND + SST | **5.6** | **0.78** | **0.95** | 0.95 | **1.02** | **0.76** | **0.71** |

Bold indicates the minimum error by every predictor variables

(f) F1 score: the harmonic mean of sensitivity and precision of the model ($(2 * TP)/(2 * TP + FP + FN)$).

The confusion matrix elaborating the correctly predicted number of positive and negative deviations from LPA as observed by all the predictor variables is presented in Table 11a–g. The observed number of positive and negative deviations from LPA rainfall during test period 1995–2017 are 8 and 15, respectively. The measures calculated from the confusion matrix to evaluate the performance of identified predictors in predicting correctly the positive or negative deviation rainfall are shown in Table 12.

**Table 11** Confusion matrix for monsoon prediction by (a) SP, (b) UWND, (c) SST, (d) SP + UWND, (e) UWND + SST, (f) SP + SST, (g) SP + UWND + SST predictors for the test period 1995–2017

| | Predicted | |
|---|---|---|
| | Pos. | Neg. |
| Observed | | |
| (a) | | |
| Pos. | 4 | 4 |
| Neg. | 2 | 13 |
| (b) | | |
| Pos. | 6 | 2 |
| Neg. | 2 | 13 |
| (c) | | |
| Pos. | 4 | 4 |
| Neg. | 1 | 14 |
| (d) | | |
| Pos. | 5 | 3 |
| Neg. | 2 | 13 |
| (e) | | |
| Pos. | 4 | 4 |
| Neg. | 1 | 14 |
| (f) | | |
| Pos. | 5 | 3 |
| Neg. | 1 | 14 |
| (g) | | |
| Pos. | 6 | 2 |
| Neg. | 1 | 14 |

#### 5.2.4 Uncertainty analysis

The uncertainty involved in monsoon prediction is explained in terms of *'Fraction of variance unexplained (FVU)'*. The measure is defined as the fraction of variance of dependent variable (the monsoon in our case) which cannot be explained or correctly predicted by the explanatory variables (identified monsoon predictors). FVU will be one if the identified predictors to not convey anything about the monsoon, and the prediction is said to be more accurate with less uncertainty as the FVU value approaches zero. The expression is shown in Eq. (6).

$$\mathrm{FVU} = \frac{\mathrm{VAR}_{\mathrm{error}}}{\mathrm{VAR}_{\mathrm{total}}} = \frac{\mathrm{SD}_{\mathrm{error}}/\mathrm{n}}{\mathrm{SD}_{\mathrm{total}}/\mathrm{n}} = \frac{\mathrm{SD}_{\mathrm{error}}}{\mathrm{SD}_{\mathrm{total}}}, \tag{6}$$

where

$$\mathrm{SD}_{\mathrm{error}} = \sum_{i=1}^{N} \left(X_i - Y_i\right)^2,$$

$$\mathrm{SD}_{\mathrm{total}} = \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2.$$

The terms are already defined in Sect. 5.2.3. The prediction provided by the surface pressure has FVU of 0.38 and that by the zonal wind and sea surface temperature provides FVU of 0.33, and 0.55, respectively. Lower values of the variable suggest that less fraction of variance remains unexplained, which symbolizes a good prediction of the monsoon.

#### 5.2.5 Comparisons with existing models

The prediction skill of monsoon predictors, which are identified by the proposed approach are investigated with the existing monsoon prediction models of India Meteorological Departments (IMDs). The models used by IMD are 16-parameter power regression model (Gowariker et al. 1991), 8-parameter and 10-parameter regression models (Rajeevan et al. 2004). The results are shown in Fig. 5.

The predictor sets with SP, SST, UWND, SP + SST, UWND + SST, SP + UWND, SP + UWND + SST provide 5.5%, 6.1%, 5.9%, 5.7%, 5.9%, 4.5% and 4.6% root mean square errors,

**Table 12** Evaluation for positive and negative deviation rainfall prediction with the identified climatic predictors of SP, UWND, SST, SP + UWND, UWND + SST, SP + SST and SP + UWND + SST for the test period 1995–2017

| Measures | Sensitivity | Specificity | Precision | Neg. Pred. Val. | Accuracy (%) | F1-Score |
|---|---|---|---|---|---|---|
| SP | 0.50 | 0.86 | 0.66 | 0.76 | 73.9 | 0.57 |
| UWND | **0.75** | 0.86 | 0.75 | 0.86 | 86.6 | 0.75 |
| SST | 0.50 | 0.93 | 0.80 | 0.77 | 78.2 | 0.61 |
| SP + UWND | 0.62 | 0.86 | 0.71 | 0.81 | 78.2 | 0.66 |
| UWND + SST | 0.50 | 0.93 | 0.80 | 0.77 | 78.2 | 0.61 |
| SP + SST | 0.62 | 0.93 | 0.83 | 0.82 | 82.6 | 0.71 |
| SP + UWND + SST | **0.75** | **0.93** | **0.85** | **0.87** | **86.9** | **0.80** |

Bold indicates the minimum error by every predictor variables

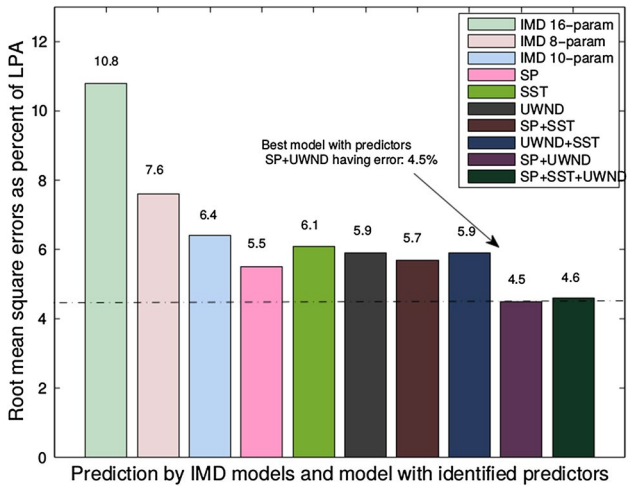


**Fig. 5** Root mean square errors in monsoon prediction by the predictors of SP, SST, UWND, SP + SST, UWND + SST, SP + UWND, SP + UWND + SST; and IMD's 16- (Gowariker et al. 1991), 10-, and 8-parameter models (Rajeevan et al. 2004) during 1996–2002

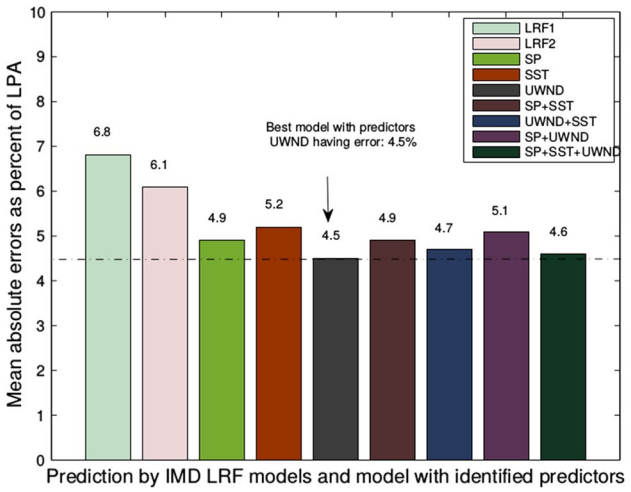


**Fig. 6** Mean absolute errors (%) in monsoon prediction by the predictors of SP, UWND, SST, SP + UWND, UWND + SST, SP + SST, SP + UWND + SST, and IMD's PPR model (LRF1 and LRF2) (Rajeevan et al. 2007) during 2003–2017

respectively in monsoon prediction, for the period 1996–2002 [period is considered to compare with existing forecasts by Rajeevan et al. (2004)]. The three IMD models predict monsoon with 10.8%, 6.4%, and 7.6% errors, respectively.

The prediction by the identified climatic predictors are also compared with the predictions provided by current IMD's model using pursuit projection regression (PPR) (Rajeevan et al. 2007). The monsoon predictions provided by the discovered predictors are compared during the period of 2003–2017 (available forecasts by IMD models). The pursuit projection regression model presents monsoon prediction in two intervals– first in April (LRF1) and the next in June (LRF2). The model predicts the monsoon with 6.8% and 6.1% mean absolute errors in April and June, respectively.

The identified predictors of SP, SST, and UWND provide 4.9%, 5.2%, and 4.5% errors, respectively. The combined predictors of SP + SST, UWND + SST, SP + UWND, and SP + UWND + SST provide errors of 4.9%, 4.7%, 5.1%, and 4.6%, respectively. Thus, the identified predictors by the proposed climate network-based approach are comparable with the monsoon models used by IMD (Gowariker et al. 1991; Rajeevan et al. 2004, 2007). The results are presented in Fig. 6 by a bar chart diagram.

### 5.3 Analysis based on correlation of monsoon predictors with the Indian summer monsoon

The Pearson correlation ($\gamma$) (Eq. (4)) between the identified monsoon predictors and the monsoon of India are explored with the same well-known predictors and the monsoon (Rajeevan et al. 2007). The important well-known monsoon predictors, as considered by India Meteorology Department, include the North Atlantic SST (NA_SST), the Equatorial South-Eastern Indian Ocean SST (ESE_IO_SST), the East Asia SP (EA_SP), the North Atlantic SP (NA_SP), the North-Central Pacific Ocean zonal wind (NC_PO_WV), and the North-West Europe SP (NW_Eu_SP). The identified predictors of SP, SST, and UWND having $\gamma$ of 0.35, 0.34, and 0.43 are comparable to the correlation of known IMD's predictors with the monsoon of India (shown in Fig. 7).

### 5.4 Monsoon of current year 2018

Indian monsoon for the current year 2018 is predicted. The identified predictors of SP, UWND, and SST forecast rainfall as 92.26%, 90.12%, and 92.15% of a long period average in May, March, and March, respectively. Additionally, the combined predictors of SP + UWND, UWND + SST, SP + SST, and SP + UWND + SST predict monsoon as 94.03%, 91.65%, 95.19%, and 95.22% in May, April, April, and May, respectively. Thus, averaging all the values, we present the Indian monsoon for 2018 as 92.94% of LPA rainfall, which is below normal for the current year.

## 6 Conclusions

The identification of monsoon predictors is always been a prime focus in earth science. In our work, community detection approach is used for identifying the monsoon predictors that are important for the monsoon of the

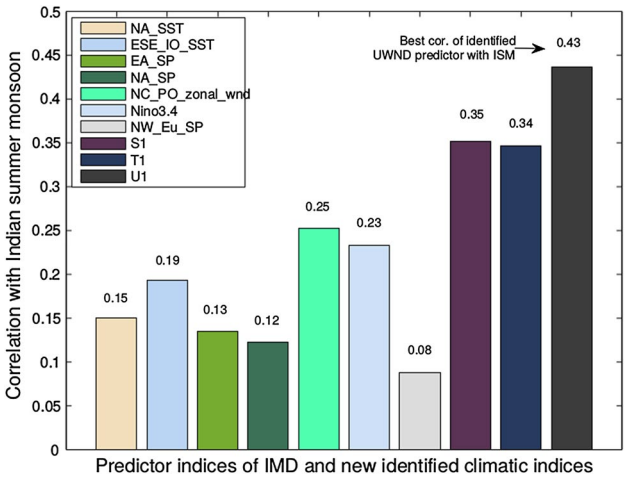


**Fig. 7** Correlation between the Indian monsoon and IMD's predictors as well as the identified predictors of surface pressure ($S_1$), sea surface temperature ($T_1$), and zonal wind ($U_1$)

subcontinent. The community detection method is followed by the density-based clustering method to obtain the localized geographical regions. These regions represent newly identified monsoon predictors. Some of the identified predictors correspond to known predictors of the monsoon, which validate the proposed predictors' identification approach, while some other new predictors are also found having high correlation with the Indian summer monsoon. The non-linear ensemble regression model, designed with identified monsoon predictors, was observed to be comparable to the IMD's existing models for forecasting the Indian monsoon.

The future scope of the work comprises the inclusion of more climatic variables and identification of the new predictors from an amalgamation of different variables. The focus will be on exploring the new climatic predictors which will be crucial to the summer monsoon and may prove as an even better estimator for the Indian summer monsoon.

**Acknowledgements** We would like to acknowledge the Department of Computer Science and Engineering at the Indian Institute of Technology Kharagpur for supporting the work. We would also like to express our gratitude toward the Centre for Atmospheric and Oceanic Sciences at the Indian Institute of Science Bangalore for assisting in completing the work in all possible ways.

## References

Achuthavarier D, Krishnamurthy V, Kirtman B, Huang B (2012) Role of the Indian Ocean in the ENSO-Indian summer monsoon teleconnection in the NCEP climate forecast system. J Clim 25(7):2490–2508

Azad S, Debnath S, Rajeevan M (2015) Analysing predictability in Indian monsoon rainfall: a data analytic approach. Environ Proc 2(4):717–727

Cherchi A, Navarra A (2013) Influence of ENSO and of the Indian Ocean Dipole on the Indian summer monsoon variability. Clim Dyn 41(1):81–103

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066,111

Donges J, Zou Y, Marwan N, Kurths J (2009) The backbone of the climate network. Europhys Lett 87(4):48,007

Donges JF, Zou Y, Marwan N, Kurths J (2009b) Complex networks in climate dynamics. Eur Phys J 174(1):157–179

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. KDD 96:226–231

Gadgil S (2003) The Indian monsoon and its variability. Annu Rev Earth Planet Sci 31(1):429–467

Gadgil S, Rajeevan M, Nanjundiah R (2005) Monsoon prediction—why yet another failure? Curr Sci 88(9):1389–1400

Gowariker V, Thapliyal V, Kulshrestha SM, Mandal GS, Sen Roy N, Sikka DR (1991) A power regression model for long range forecast of southwest monsoon rainfall over India. Mausam 42(2):125–130

Guhathakurta P, Rajeevan M (2008) Trends in the rainfall pattern over India. Int J Climatol 28(11):1453–1469

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77(3):437–471

MATLAB (2012) Statistics and machine learning toolbox. MATLAB version 2012b. The MathWorks Inc., Natick

Nanjundiah RS, Francis P, Ved M, Gadgil S (2013) Predicting the extremes of Indian summer monsoon rainfall with coupled ocean–atmosphere models. Curr Sci India 104(10):1380–1393

Noor M, Awan A (2005) Finding spatio-temporal patterns in climate data using clustering. In: Proceedings of the international conference Cyberworlds, IEEE, pp 8–15

Rajagopalan B, Molnar P (2012) Pacific ocean sea-surface temperature variability and predictability of rainfall in the early and late parts of the Indian summer monsoon season. Clim Dyn 39(6):1543–1557

Rajeevan M (2001) Prediction of Indian summer monsoon: status, problems and prospects. Curr Sci 81

Rajeevan M, Pai DS, Dikshit SK, Kelkar RR (2004) IMD's new operational models for long-range forecast of southwest monsoon rainfall over India and their verification for 2003. Curr Sci India 86(3):422–431

Rajeevan M, Pai DS, Kumar RA, Lal B (2007) New statistical models for long-range forecasting of southwest monsoon rainfall over India. Clim Dyn 28(7–8):813–828

Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W (2002) An improved in situ and satellite SST analysis for climate. J Clim 15:1609–1625

Saha M, Mitra P, Nanjundiah R (2016a) Predictor discovery for early-late Indian summer monsoon using stacked autoencoder. Proc Comp Sci 80:565–576

Saha M, Mitra P, Nanjundiah RS (2016b) Autoencoder-based identification of predictors of Indian monsoon. Meteor Atmos Phys 128(5):613–628

Saha M, Mitra P, Nanjundiah R (2017) Deep learning for predicting the monsoon over the homogeneous regions of India. J Earth Syst Sci 126(4):1–18

Saha M, Mitra P (2016) Recurrent neural network based prediction of Indian summer monsoon using global climatic predictors. In: Int. Jt. Conf. Neural Net., pp 1523–1529

Steinbach M, Tan P, Kumar V, Klooster S, Potter C (2003) Discovery of climate indices using clustering. In: Proc. ACM, ACM SIGKDD Int. Conf. KDD, pp 446–455

Steinhaeuser K, Chawla NV, Ganguly AR (2010) An exploration of climate data using complex networks. ACM SIGKDD Explor Nwsltr 12(1):25–32

Steinhaeuser K, Chawla NV, Ganguly AR (2011) Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. Stat Anal Data Min 4(5):497–511

Taylor AL, Dessai S, Bruine WBD (2015) Communicating uncertainty in seasonal and interannual climate forecasts in Europe. Phil Trans R Soc A 373(2055):20,140,454

Tsonis AA, Roebber PJ (2004) The architecture of the climate network. Phys A Stat Mech Appl 333:497–504

Tsonis A, Swanson K (2008) Topology and predictability of El Nino and La Nina networks. Phys Rev Lett 100(22):228502

Tsonis AA, Swanson K, Kravtsov S (2007) A new dynamical mechanism for major climate shifts. Geophys Res Lett 34(13):L13,705

Wang B, Xiang B, Li J, Webster PJ, Rajeevan M, Liu J, Ha K (2015) Rethinking Indian monsoon rainfall prediction in the context of recent global warming. Nature Commun 6:7154