Meteorol Atmos Phys (2016) 128:613–628
DOI 10.1007/s00703-016-0431-7

ORIGINAL PAPER

Autoencoder-based identification of predictors of Indian monsoon

Moumita Saha[1] · **Pabitra Mitra**[1] · **Ravi S. Nanjundiah**[2]

Received: 8 July 2015 / Accepted: 11 January 2016 / Published online: 2 February 2016

Abstract Prediction of Indian summer monsoon uses a number of climatic variables that are historically known to provide a high skill. However, relationships between predictors and predictand could be complex and also change with time. The present work attempts to use a machine learning technique to identify new predictors for forecasting the Indian monsoon. A neural network-based non-linear dimensionality reduction technique, namely, the sparse autoencoder is used for this purpose. It extracts a number of new predictors that have prediction skills higher than the existing ones. Two non-linear ensemble prediction models of regression tree and bagged decision tree are designed with identified monsoon predictors and are shown to be superior in terms of prediction accuracy. Proposed model shows mean absolute error of 4.5 % in predicting the Indian summer monsoon rainfall. Lastly, geographical distribution of the new monsoon predictors and their characteristics are discussed.

Responsible Editor: J. T. Fasullo.

✉ Moumita Saha
moumita.saha2012@gmail.com

Pabitra Mitra
pabitra@gmail.com

Ravi S. Nanjundiah
ravi@caos.iisc.ernet.in

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

[2] Centre for Atmospheric and Oceanic Sciences, Divecha Centre for Climate Change, Indian Institute of Science, Bangalore, India

1 Introduction

Predicting complex climatic phenomena such as the Indian summer monsoon is a challenging task. Geographical features, wind flow directions, and multiple sea-atmospheric interactions influence the strength of the monsoon. Knowing the variability of monsoons is important (even though for the entire region as a whole the standard deviation is just 10 % of mean) and the relationship with predictors changes with time, it is vital to revise existing predictors and introduce new predictors affecting the monsoon. We focus on automated identification of predictors important for the Indian summer monsoon rainfall (ISMR).

India Meteorological Department (IMD) had been forecasting Indian summer monsoon rainfall since 1886. Forecast of Indian monsoon was initiated by Blanford (1884) as early as in 1882. The success of forecasts in span of 1882–1885 encouraged Blanford to design an operational long-range forecast model for monsoon in 1886. Subsequently, Walker (1924) developed models based on the statistical correlations between rainfall and different global climate variables. Thapliyal and Kulshrestha (1992) introduced regression model in predicting south-west Indian monsoon rainfall. Gowariker et al. (1991) proposed power regression model for long-term forecast of monsoon, which provided accurate forecast for a long period, but failed to predict the extreme condition of 2002. In 2004, Rajeevan et al. (2004) reassessed different climatic variables and introduce four new variables to design a statistical model for issuing long-range forecast of Indian monsoon. Rajeevan et al. (2007) built models using ensemble multiple regression and pursuit projection regression to forecast Indian rainfall which proved to be superior to past IMD models. Gadgil et al. (2005) analyzed

the forecasts issued by IMD using statistical models and found that a drought had never been predicted in IMD's forecasts.

Identification of new predictors influencing climatic phenomena has been a prime focus in climate field. In practice, predictors influencing an phenomenon are selected by studying the physical processes and utilizing meteorological experiences. We have used sparse autoencoder, a neural network-based dimensionality reduction technique for identification of monsoon predictors. Autoencoder is widely used in feature reduction and discovery (Hinton and Salakhutdinov 2006). The ability of autoencoder techniques to efficiently handle large amount of data makes them attractive tools for use in climate studies. Salient features of our proposed approach to identification of monsoon predictors are as follows—(i) sparse autoencoder is used for reducing the dimensionality of the features considered from all grids of the world to important grids' feature influencing monsoon, (ii) autoencoder can combine the climatic variables in linear and non-linear fashion from geographically distant regions to built new monsoon predictors, (iii) advanced machine learning prediction models are used to predict monsoon using identified predictors, and (iv) combination of predictors from two or more distinct variables is considered and such predictor sets show better skill than predictors from individual variable.

Many studies have been conducted to determine potential predictors for the Indian Summer Monsoon. Kumar et al. (2011) used singular value decomposition-based graphical technique to analyze patterns of ten climatic variables. Boriah et al. (2004) utilized clustering of climatic variables viz. sea surface temperature and sea level pressure to identify predictors for land temperature. Sap and Awan (2005) used kernel k-means algorithm with spatial constraint to identify the spatio-temporal patterns in the climate system. Similar nearest neighbors-based clustering approach had been used for detection of new predictors, which were validated against known predictors and shown to overcome limitations of PCA and SVD approaches (Steinbach et al. 2003).

Selection of monsoon predictors played a significant role in forecasting Indian summer monsoon rainfall. DelSole and Shukla (2002) selected number of predictors comparing the error variances of models with different predictor sets after initial screening out of models providing poor forecast accuracy. It produced better forecasts of Indian monsoon, than model with all predictors or model with the most favorable statistics. DelSole and Shukla (2012) also showed that skill of predicting ISMR from sea surface temperature obtained from coupled atmosphere-ocean models was statistically significant, attributed to the fact that slowly evolving sea surface temperatures were primary source of predictability. Wang et al. (2015) specified that the failure of prediction models for forecasting Indian monsoon was mainly due to their inability to capture new predictability sources like central-Pacific El Nino-Southern Oscillation, deepening of the Asian Low and strengthening of North and South Pacific Highs during boreal spring. Hence with changing environment, new predictors and influencing phenomena should be ascertained to update the prediction models for better performance.

Liu et al. (2015) utilized autoencoder architecture for weather forecasting. Their study determined key features from a large dataset of hourly temperature and wind velocity data by layer-wise feature granulation and used them for predicting temperature. Song et al. (2013) have shown that autoencoder techniques determine highly non-linear functions from complex data and these functions are effective and stable. Li and Yang (2013) proposed a hybrid strategy for non-linear feature reduction using autoencoder and forecasted wind power utilizing sparse Bayesian regression model. In our approach, autoencoder assists in performing non-linear combination of climatic variables of different geographical locations to identify new monsoon predictors.

The purpose of our work is twofold—(i) identification of new monsoon predictors using autoencoder from climatic variables, namely, air temperature, sea surface temperature, and sea level pressure, (ii) utilization of identified monsoon predictors for forecasting Indian summer monsoon rainfall (ISMR), which acts as validation of our proposed predictor identification approach.

The approach initiates with considering grids of 20° longitude $\times 10°$ latitude encompassing the globe. All the time series of climatic variables, namely, air temperature, sea surface temperature, and sea level pressure within the specified grids are averaged to obtain a single time series, which represents the particular grid's variable. Time series corresponding to grid points act as inputs to autoencoder. New monsoon predictors are obtained from autoencoder as non-linear combination of input variables. The suitability of the predictor so obtained is determined by thresholding and correlation analysis. Identified monsoon predictors are compared with existing predictors of Indian monsoon for validation and are evaluated in terms of their forecasting skills for Indian monsoon. Prediction models with newly identified monsoon predictors show superiority in monsoon prediction over existing prediction models.

The article is organized as follows. Section 2 describes the basic architecture of autoencoder. Section 3 explains broadly the proposed approach of monsoon predictor identification using autoencoder. Models used for predicting monsoon rainfall are described in Sect. 4. Section 5 elaborates the experimental results and analysis of identified predictors on the ground of their forecasting skills for

Indian monsoon. Meteorological characteristics of identified monsoon predictors are explained in Sect. 6 and finally, the article concludes in Sect. 7.

2 Single-layer autoencoder architecture

An autoencoder is an artificial neural network used to learn compressed, complex characteristics of data and are used for the purpose of dimensionality reduction (Baldi 2012). Single-layer autoencoder has only one hidden layer. Autoencoder sets its target values equal to the input values. It consists of two components—(i) encoder (works in input-to-hidden layer), (ii) decoder (works in hidden to output layer). It provides a non-linear mapping function by iteratively training the encoder and the decoder. The encoder adopts the non-linear mapping function, and the decoder performs data reconstruction from the representation generated by the encoder. The process continues iteratively and guarantees that the mapping function is stable and effective to represent the original data.

Formally, say $x \in R^n$ represents an input, the activation of each neuron in the hidden layer, h_i, for $i = 1, \ldots, m$ is shown in Eq. 1.

$$h(x) = f\left(W_{inp}x + b_{inp}\right), \tag{1}$$

where $f(z) = \frac{e^{2z}-1}{e^{2z}+1}$ is the non-linear hyperbolic tangent activation function applied component-wise, $h(x) \in R^m$ is the vector of neuron activation, W_{inp} is the $(m \times n)$ weight matrix from input-to-hidden layer, and $b_{inp} \in R^m$ is bias vector. The network output is shown in Eq. 2.

$$\widehat{x} = g(W_{hid}h(x) + b_{hid}), \tag{2}$$

where $g()$ denotes a linear function and $\widehat{x} \in R^n$ is a vector of output values, W_{hid} is the $(n \times m)$ weight matrix from hidden to output layer, and $b_{hid} \in R^n$ is bias vector. For our problem, the input comprises the climatic variables like air temperature, sea surface temperature or sea level pressure over the world grids and potential predictors are obtained as output from hidden layer as non-linear combination of input climatic variable at different geographical regions. Thus, input and output to autoencoder are climatic variables and their combination, respectively, where autoencoder is assisting in dimensional reduction or complex feature building.

Given a set of q input instances x_i, $i = 1, \ldots, q$, the weight matrices W_{inp} and W_{hid} and the bias vector b_{inp} and b_{hid} are updated using gradient back-propagation algorithm to minimize the reconstruction error $\sum_{i=1}^{q} \| x_i - \widehat{x} \|^2$. The architecture of the autoencoder is shown in Fig. 1.

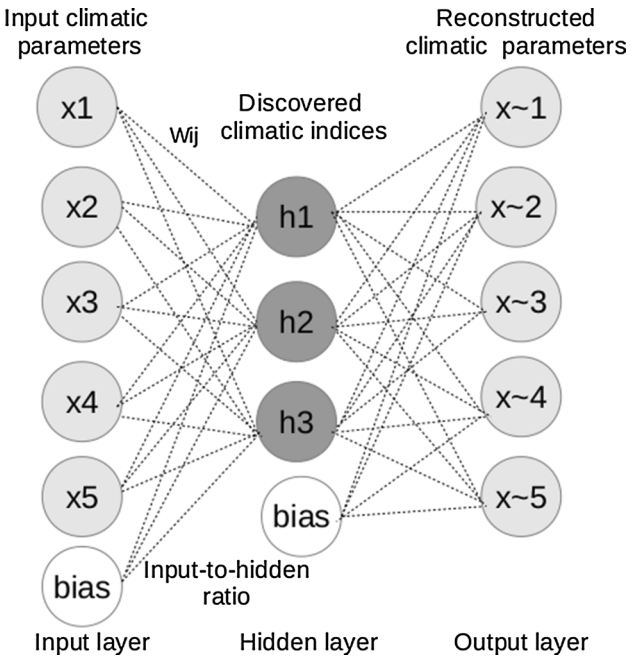

Fig. 1 Architecture of autoencoder

3 Identification of monsoon predictor using autoencoder

Autoencoder is used for identification of new monsoon predictors significant for Indian monsoon and these potential monsoon predictors are utilized for forecasting the monsoon. The block diagram of our proposed approach is shown in Fig. 2.

3.1 Preprocessing: initialization of climatic variables

The proposed approach of predictor identification initiates with preprocessing of input climatic variables. We consider three climatic variables, namely, air temperature (AT), sea surface temperature (SST), and sea level pressure (SLP) for our task. Initially, the world is divided into spatial rectangular grids of dimension 20° longitude $\times$10° latitude, which sum up to 324 grids [(360/20) $\times$ (180/10)]. As a preprocessing step, the grids having less than 20 % values as Null (eg. some grids near the poles have most of their values as Null) are considered for further analysis. All the time-series values of climatic variable within the spatial cover of grid are averaged to obtain a single time series of climatic variable representing that particular grid. Each such averaged time series of the selected grids are considered as input nodes of the autoencoder. After preprocessing, the number of input nodes in autoencoder designed for climatic variable air temperature (Aut_AT) is 136, that

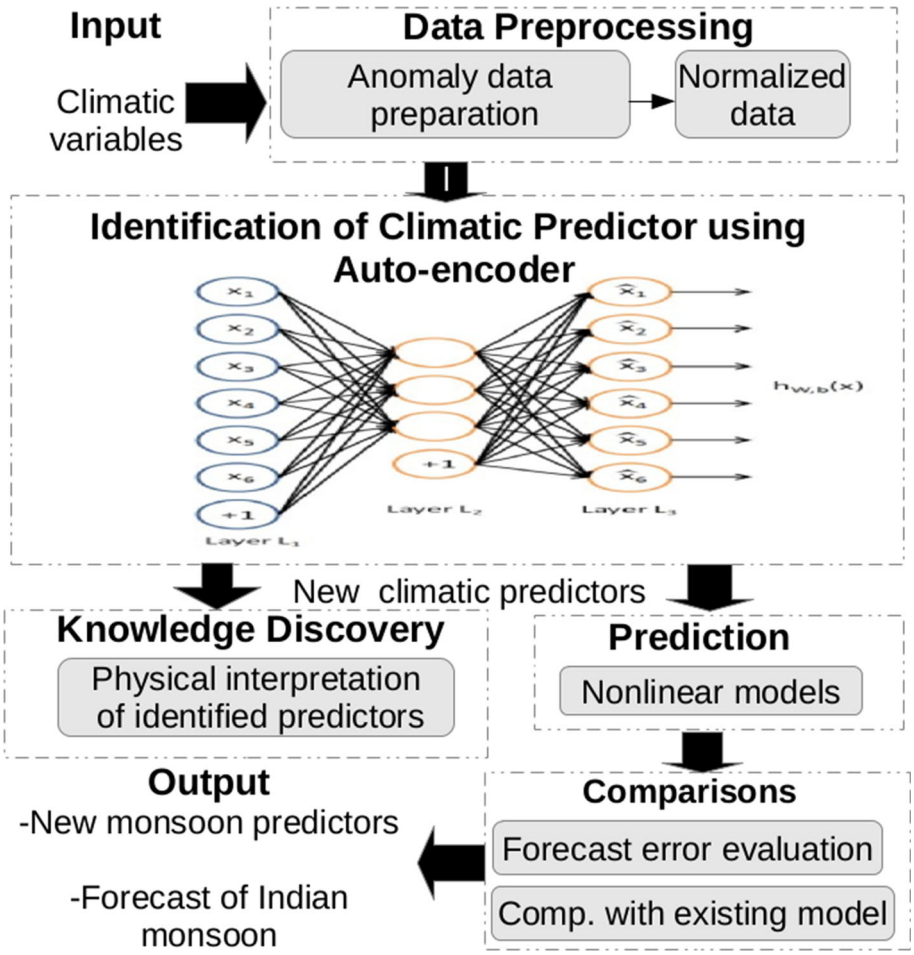

Fig. 2 Block diagram of proposed approach to identification of monsoon predictor using autoencoder and prediction of Indian monsoon

for sea surface temperature (Aut_SST) is 137, and 102 for sea level pressure (Aut_SLP).

3.2 Monsoon predictor identification: new monsoon predictor identification using autoencoder

Identification of predictors is performed in three major steps—(i) firstly, non-linear mapping of input variables to reduced composite features, which will act as representative for new monsoon predictors, (ii) predictor selection based on threshold of weight matrix, and (iii) filtering by correlation study of predictors with Indian monsoon.

3.2.1 Non-linear predictor mapping

Separate autoencoders are designed for all three climatic variables. We used single-layer autoencoder for our approach. Input layer consists of nodes corresponding to climatic variables of selected grid points. Input-to-hidden layer ratio is ascertained as 15 %. The nodes in the hidden layer represent the composite features which are representative of potential monsoon predictors. The autoencoders, Aut_AT and AUT_SST, have 20 nodes in the hidden layer, and the autoencoder Aut_SLP has 15 nodes in the hidden layer. Input data are considered for time period 1901–1993 for training the autoencoder. All the data are at monthly scale, thus they sum up to 1116 (93 years $\times$ 12 months) training instances. Autoencoder is trained iteratively using gradient descent back-propagation algorithm to reduce the reconstruction error of the output nodes from the input nodes. The process is continued until the reconstruction error does not reduce further or it gets saturated. The optimized weight and bias matrices are obtained at the end of training the autoencoder.

3.2.2 Post-training thresholding of weights

After the autoencoder being trained using the training instances, post-processing is performed to obtain the new monsoon predictors. We examine the weight matrix corresponding to input-to-hidden layer. For each hidden node that represents potential predictor, we divide the range of weight into ten equal intervals and plot the frequency of occurrence of weight in each interval. The knee of the plot

(i.e the sharp fall of the curve) is considered as the threshold weight and we further consider only the input nodes whose weights attain the threshold for evaluating the potential predictors. Threshold is chosen as the knee point of the curve in the notion that the input nodes having weighted edge greater than this threshold have greater contribution in the value of hidden nodes and nodes having less weight values than the threshold are neglected as their contributions are less. Following this approach, it leads to incorporation of only highly influencing input nodes in identification of new monsoon predictors (from hidden nodes), while ignoring the rest. Figure 3 shows the threshold for an monsoon predictor for climatic variable air temperature. Threshold value of 0.11 is ascertained for this monsoon predictor from the knee of the threshold curve.

Finally, new monsoon predictors are evaluated as weighted sum of input nodes having weights greater than the ascertained threshold.

Formally, say x_i represents input node, $i = 1, \ldots, n$, and h_j represents hidden node, $j = 1, \ldots, m$, and W_{ij} represents the weight of input node x_i for corresponding hidden node h_j. A particular monsoon predictor corresponding to hidden node h_j is evaluated as:

$$h_j = \sum_{i=1}^{n} W_{ij} x_i, \quad \text{for all } i, \text{ such that } W_{ij} > threshold_j$$

where $threshold_j$ represents the threshold value chosen for hidden node h_j.

3.3 Postprocessing: monsoon predictor selection

Hidden node values as weighted sum of thresholded input nodes represent the newly identified monsoon predictors. A correlation study of the identified predictors and Indian summer monsoon rainfall is performed to determine predictors important for Indian monsoon according to top correlation values. High correlation values are observed for identified predictors. Detailed correlation results and spatial location of identified predictors are presented in Sect. 5.

4 Forecasting models with identified monsoon predictors for Indian summer monsoon rainfall

Identified monsoon predictors are used to forecast Indian summer monsoon rainfall (ISMR). We built different predictor sets taking identified monsoon predictors for variables AT, SST, and SLP, exclusively. We concentrate on prediction of cumulative ISMR occurring during months of June, July, August, and September. Test period from 1994 to 2014 is considered for evaluating the forecasting efficiency of newly identified monsoon predictors. Two models (described in following sections) with identified monsoon predictors as input variables are built to forecast rainfall. The reasons for selecting the models are as follows—(i) model uses bagging technique which is a bootstrap aggregating technique for improving estimation (Breiman 1996), (ii) bagging aids in improving the

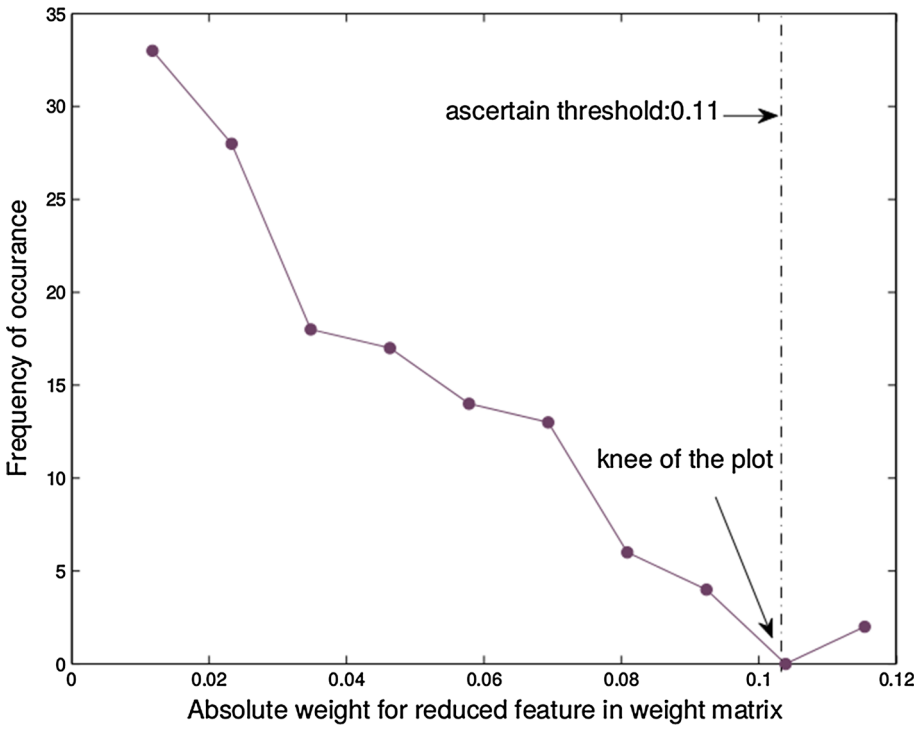

Fig. 3 Frequency of occurrence of weight values at different weight interval to ascertain the threshold from knee of the plot for climatic variable air temperature

predictive performance of underlying regression tree, (iii) tree ensembles can deal with non-linear features, and (iv) they can handle high-dimensional data spaces as well as large number of training instances.

4.1 Fitted ensemble of regression tree with Bagging algorithm (RegTreeB)

- The main principle of this model is melding results from many weak learners into one high-quality ensemble prediction.
- It combines a set of trained weak regression tree learner models and data on which these learners are trained (Loh 2008).
- Predicts ensemble response for new data by aggregating predictions from the weak learners.
- Bagging algorithm is used for training the regression tree learners. Bagging is a type of ensemble learning and it generally constructs deep trees to estimate the generalization error. To bag a weak learner such as a regression tree on a dataset, many bootstrap replicas of this dataset are generated and regression trees are grown on these replicas. Each bootstrap replica is obtained by randomly selecting n observations out of n with replacement, where n is the size of dataset. An average over predictions from individual trees are performed to present the final predicted response of trained ensemble.
- The size of ensemble is chosen empirically in a way that it keeps balance between speed and accuracy. To set an appropriate size, it is started with few members to several in an ensemble, training the ensemble, and then checking the ensemble quality until adding more members does not improve ensemble quality.
- Model is expressed in functional form as following (MATLAB 2012).

  $$model = RegTreeB(Pred, Out, Algo, numb),$$

  where, Pred represents data matrix with each row representing one instance and each column contains one predictor value (new identified monsoon predictor), Out is a numeric column vector with the same number of rows as Pred, with corresponding rainfall values, Algo represents the bagging technique in our case, numb is the number of weak learners for ensemble process.

4.2 Ensemble of bagged decision tree (DecTreeB)

- The model bags an ensemble of decision trees for regression modeling (Liaw and Wiener 2002).
- Bagging stands for bootstrap aggregation. Every tree in the ensemble is grown on an independently drawn bootstrap replica of input data. Observations not included in this replica are "out of bag" for the specific tree. Re-sampling is usually done by bootstrapping observations. In addition, every tree in the ensemble randomly selects predictors for decision splits, which improve the accuracy of bagged trees.
- It relies on the regression tree functionality for growing individual trees. Regression tree accepts the number of features selected at random for each decision split.
- Another important parameter is the number of predictors selected at random for every decision split. This random selection is made for every split, and every deep tree involves many splits. It is generally considered as one-third of predictors used for regression.
- To compute prediction of an ensemble of trees for new data instant, it takes an weighted average of predictions from individual trees (MATLAB 2012).

  $$\widehat{y}_{bag} = \frac{1}{\sum_{t=1}^{T} a_t I(t \in S)} \sum_{t} T a_t \widehat{y}_t I(t \in S),$$

  where $\widehat{y}_t$ is the prediction from tree t in the ensemble, S is the set of predictors of selected trees that comprise the prediction, $I(t \in S)$ is 1 if t is in the set S, and 0 otherwise, a_t is the weight of tree t.

For both the prediction models, number of training years is varied from ten to sixty years and optimal training period is evaluated in terms of least error in forecasting for validation period 1984–1993. Figure 4 shows the mean absolute errors with varying training years. Optimal training year is obtained as 25 years. It is the number of training years considered for predicting the rainfall of subsequent year (eg. training of 25 years from 1969 to 1993 is performed for forecasting 1994 rainfall, and other test year are predicted in same manner).

5 Experimental results and discussions

Proposed approach to identification of monsoon predictor using autoencoder is evaluated in terms of their efficiency in predicting annual Indian summer monsoon rainfall and their characteristics in climate domain.

5.1 Data sets used for the work

Sea level pressure (SLP) and air temperature (AT) are collected from PSD gridded datasets provided by NOAA/OAR/ESRL/PSD (www.esrl.noaa.gov/ psd/) (Compo et al. 2011). Sea surface temperature (SST) is acquired from NOAA Extended Reconstructed V3 Data provided by the

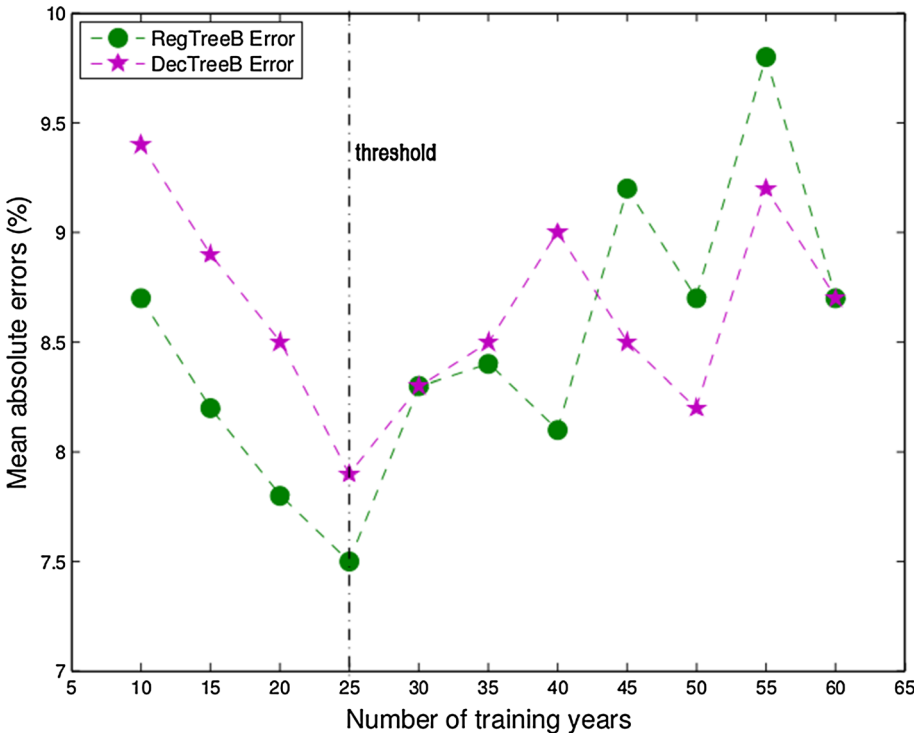

Fig. 4 Scatter plot of mean absolute errors of forecasts by two prediction models with identified monsoon predictors of air temperature for training period of different lengths (10–60 years). Mean absolute error is computed for validation period 1984–1993. Optimal training period is obtained as *twenty-five*

NOAA/OAR/ESRL/ PSD (www.esrl.noaa.gov/psd/) (Xue et al. 2003; Smith et al. 2008). All the above monthly data are available at spatial resolution of $2^\circ \times 2^\circ$ and considered for period 1901–2015. Annual Indian summer monsoon rainfall quantity (ISMR) (i.e. collective rainfall occurring in months of June, July, August, and September) is collected from India Meteorological Department (http://www.imdpune.gov.in/research/ncc/longrange/data/data.html) for time period 1901–2014. The long period average (LPA) (1901–2014) of rainfall is 898.3 mm with standard deviation of 92.7 mm. Rainfall is expressed as percentage of the LPA value.

As a preprocessing step, SLP, SST, and AT data are converted to monthly anomaly data by subtracting the monthly mean from the corresponding data values.

$$Climatic\ anomaly_m = X_m - mean(X_m),$$

Here, X_m denotes the climatic variable value for month m and mean (X_m) is the average of the variable values over all the years under study for month m.

5.2 Autoencoder-based identified monsoon predictors

The autoencoders for air temperature (Aut_AT), sea surface temperature (Aut_SST), and sea level pressure (Aut_SLP) have number of input nodes as 136, 137, and 102, respectively. Architecture of autoencoder is designed considering input-to-hidden layer ratio as 15 %. All the autoencoders are trained with respective climatic data (air temperature, sea surface temperature or sea level pressure) for training period of ninety-three years from 1901 to 1993. New monsoon predictors are identified from hidden nodes of respective autoencoders, denoted by CI_AT, CI_SST, and CI_SLP. A number of potential predictors for variables air temperature, sea surface temperature, and sea level pressure are 20, 20, and 15, respectively. Identified monsoon predictors are ranked considering their correlation with ISMR and finally they are utilized for forecasting Indian monsoon.

5.3 Geographical locations of identified monsoon predictors

Identified monsoon predictors using the proposed method for climatic variables air temperature, sea surface temperature, and sea level pressure are shown in Figs. 5, 6, and 7, respectively. For every variables, top six highly correlated monsoon predictors with ISMR are presented.

It is observed that the obtained monsoon predictors are not geographically localized but they are combination of climatic variables situated at different spatial locations. The combination of distinct geographical regions having different time leads forms the new potential predictors. Each color represents the geographical location of climatic variable which is combined to form a monsoon predictor. Monsoon predictors are ranked according to their correlation with Indian monsoon and presented in the same order

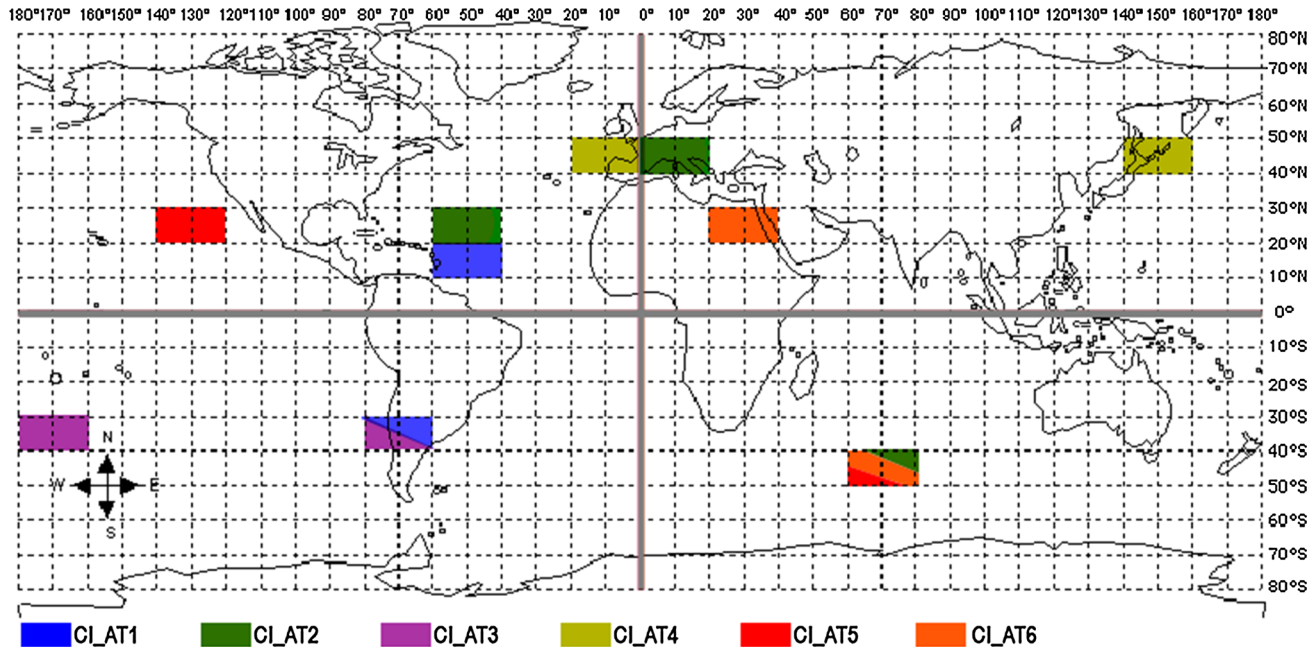

Fig. 5 Geographical locations of identified monsoon predictors for climatic variable air temperature. Climatic variable corresponding to same *colored geographical regions* combined to represent an monsoon predictor (CI_AT1-CI_AT6 represent six identified air temperature-based monsoon predictors). *Different colors in the same region* represent participation of that region in all those monsoon predictors

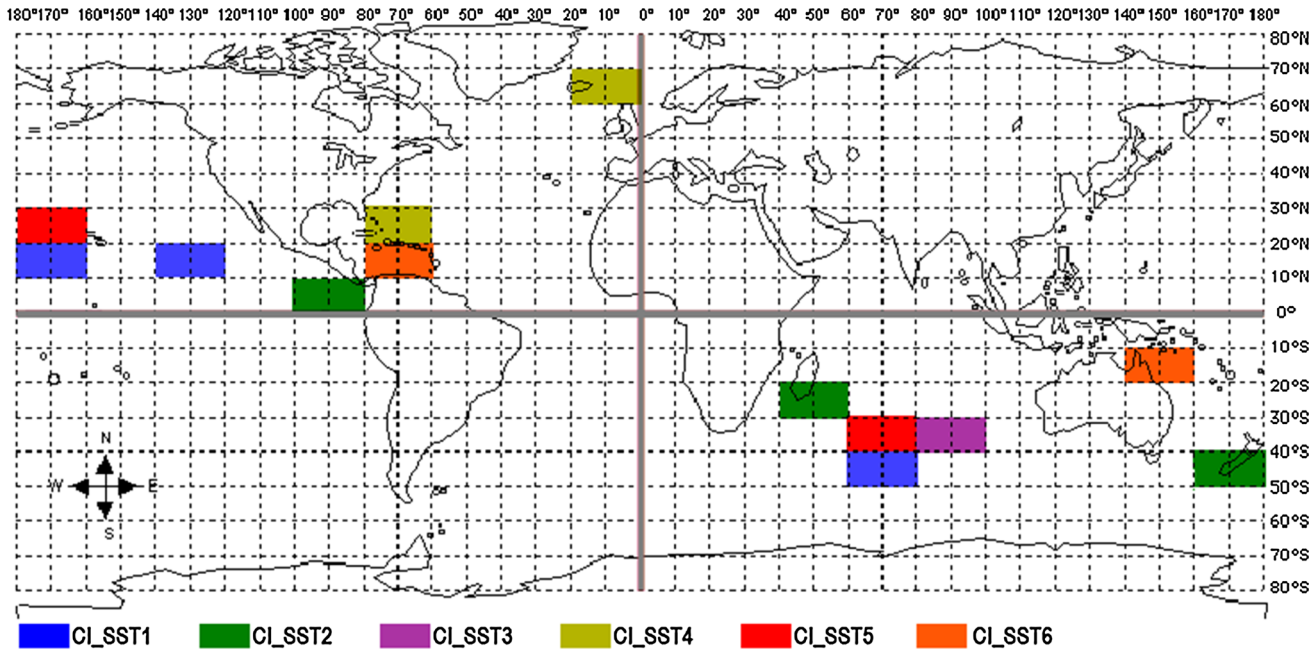

Fig. 6 Geographical locations of identified monsoon predictors for climatic variable sea surface temperature. Climatic variable corresponding to same *colored geographical regions* combined to represent an monsoon predictor (CI_SST1-CI_SST6 represent six identified sea surface temperature-based monsoon predictors)

(eg. CI_AT1 has the highest and CI_AT6 has the lowest correlation with monsoon).

Correlation of top six predictors along with the correlated month (as shown in Figs. 5, 6, and 7) with Indian summer monsoon rainfall is shown in Table 1. Some identified monsoon predictors are mapped to location of known established monsoon predictors, which act as validation of our proposed approach of predictor identification. It is noted that the earliest signal comes from SST (the slowly varying component of climate) while signal from

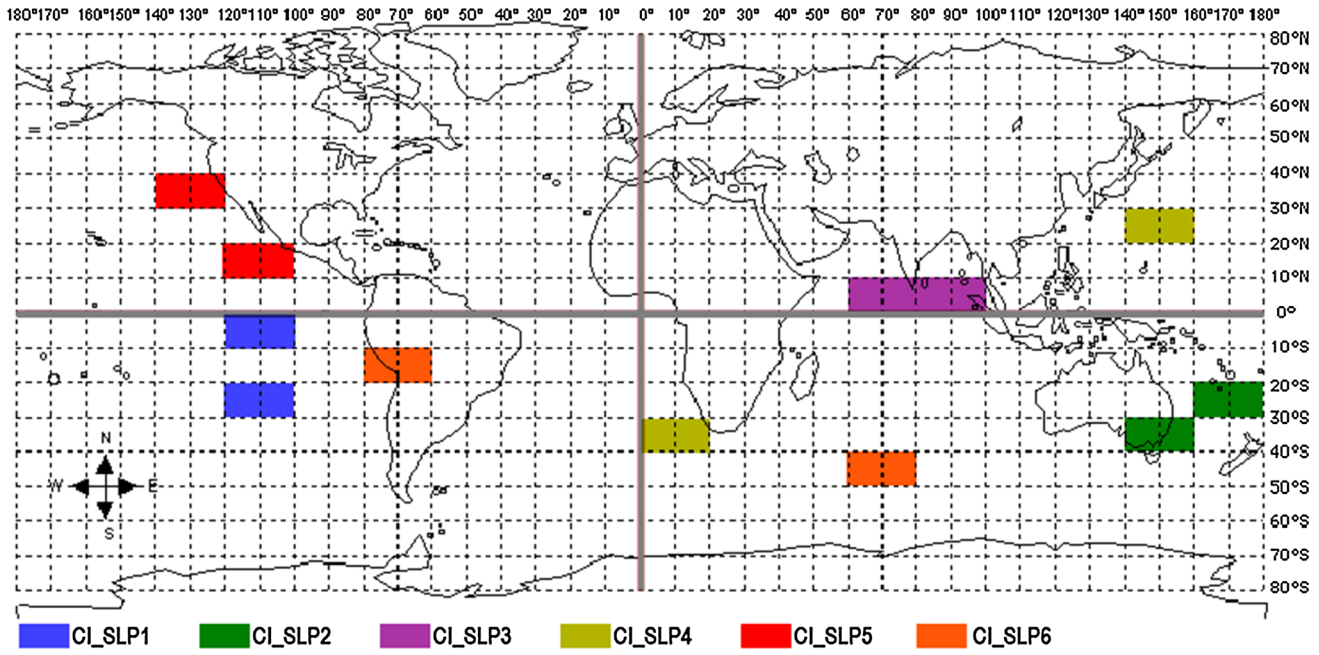

Fig. 7 Geographical locations of identified monsoon predictors for climatic variable sea level pressure. Climatic variable corresponding to same *colored geographical regions* combined to represent an monsoon predictor (CI_SLP1-CI_SLP6 represent six identified sea level pressure-based monsoon predictors)

Table 1 Correlation of top *six* new monsoon predictors with ISMR along with correlated month for climatic variables AT, SST, and SLP

Autoencoder for climatic variable	Correlation values	Corresponding correlated month
Aut_AT	+0.36, −0.32, +0.29, +0.28, +0.26, +0.26	Apr, May, Apr, May, May, May
Aut_SST	−0.35, +0.28, +0.23, +0.22, +0.21, −0.20	May, Jan, Jan, Mar, Apr, Apr
Aut_SLP	+0.34, +0.30, +0.30, +0.24, −0.22, +0.20	May, May, May, May, Mar, May

SLP (an atmospheric signal, the faster varying component) gives signal about Indian summer monsoon rainfall later.

5.4 Prediction skills of identified monsoon predictors

Different predictor sets are built from the identified potential predictors for forecasting annual Indian summer monsoon rainfall. Monsoon quantity is expressed as percentage of long period average (LPA) value of rainfall. Forecasting skill of the identified predictors is provided in terms of mean absolute error in prediction of monsoon during test period (1994–2014).

Four predictor sets are built using identified monsoon predictors for all climatic variables (AT, SST, SLP). Identified monsoon predictors are ranked according to their correlation with Indian monsoon exclusively for three variables. Predictor sets P1_X, P2_X, P3_X, and P4_X are built with top 3, 4, 5, and 6 identified monsoon predictors from the ranked list, respectively (*X* in nomenclature denotes either of AT, SST, or SLP). Predictor set built for individual climatic variables is used as input to the prediction models (described in Sect. 4) to forecast annual Indian summer monsoon. Mean absolute errors for predictor sets are shown in Table 2. The bold values in table show the least mean absolute errors provided by identified predictors of different climatic variables in prediction of Indian summer monsoon.

Identified monsoon predictors of air temperature show mean absolute error of 4.5 % by P4_AT, which is built with top six identified monsoon predictors of air temperature, by RegTreeB model. P3_SST predictor set based on sea surface temperature gives mean absolute error of 5.3 % by RegTreeB model. Predictor set with identified sea level pressure monsoon predictors (P1_SLP) gives mean absolute error of 5.4 %.

Figure 8 shows predicted rainfall by identified monsoon predictors of air temperature against actual rainfall for test period 1994–2014. It is observed that magnitude of predicted forecasts is close to actual rainfall values in most of the test years. Pearson correlation of 0.72, 0.44, and 0.50 is observed between actual and predicted rainfall by identified monsoon predictors of AT, SST, and SLP, respectively.

Figure 9 shows actual and predicted rainfall as departure from LPA value of rainfall. It shows that predicted rainfall follows the same trend as the actual rainfall, even predicted rainfall follows the same sign (positive or negative from departure from LPA value) as the actual rainfall in most of the test years. In terms of positive or negative anomaly from the LPA rainfall, it is observed that among twenty test years, fourteen years show same sign of anomaly as actual rainfall and two forecasts are on border line for prediction performed with identified monsoon predictors of air temperature. The count of correct direction of anomaly of rainfall prediction is fifteen for forecast by monsoon predictors of sea level pressure, and it is little low with value twelve by monsoon predictors of sea surface temperature, but five of rest test years are predicted as near to zero departure from LPA rainfall (in border line).

Extreme years during the test period are also predicted with same anomaly sign. All the drought years (2002, 2004, 2009, 2014) are predicted correctly with same sign (negative anomaly from mean) by the identified monsoon predictors of AT, SST, and SLP (except year 2002 by monsoon predictor of SLP). It is noted that the magnitude of deficit is very well captured during 2009 by monsoon predictors of sea level pressure. For numerical models, it is observed that even the sign of the anomaly is incorrect in many test years (Nanjundiah et al. 2013), thus comparatively the identified monsoon predictors improve the prediction accuracy of Indian monsoon. It can be concluded that identified monsoon predictors are effective in forecasting Indian monsoon, which signify the success of our proposed autoencoder-based approach to identification of potential monsoon predictor.

We also framed four predictor sets with combination of identified predictors of different variables to forecast Indian monsoon. Combined predictor sets are (i) comb1—comprises top three highly correlated identified monsoon predictors of each AT and SLP, (ii) comb2—comprises top three predictors of each AT and SST, (iii) comb3—comprises top three predictors of each SST and SLP, and (iv)

Table 2 Mean absolute errors (%) for annual Indian summer monsoon rainfall prediction for dataset built with individual identified monsoon predictors for variables AT, SST, and SLP during test period 1994–2014

Predictor sets	RegTreeB model	DecTreeB model
P1_AT	5.5	4.8
P2_AT	**4.5**	**4.6**
P3_AT	4.8	4.7
P4_AT	**4.5**	**4.6**
P1_SST	5.6	5.7
P2_SST	5.6	5.9
P3_SST	**5.3**	**5.6**
P4_SST	6.2	6.2
P1_SLP	**5.4**	**5.8**
P2_SLP	5.9	5.9
P3_SLP	6.6	5.9
P4_SLP	6.1	6.4

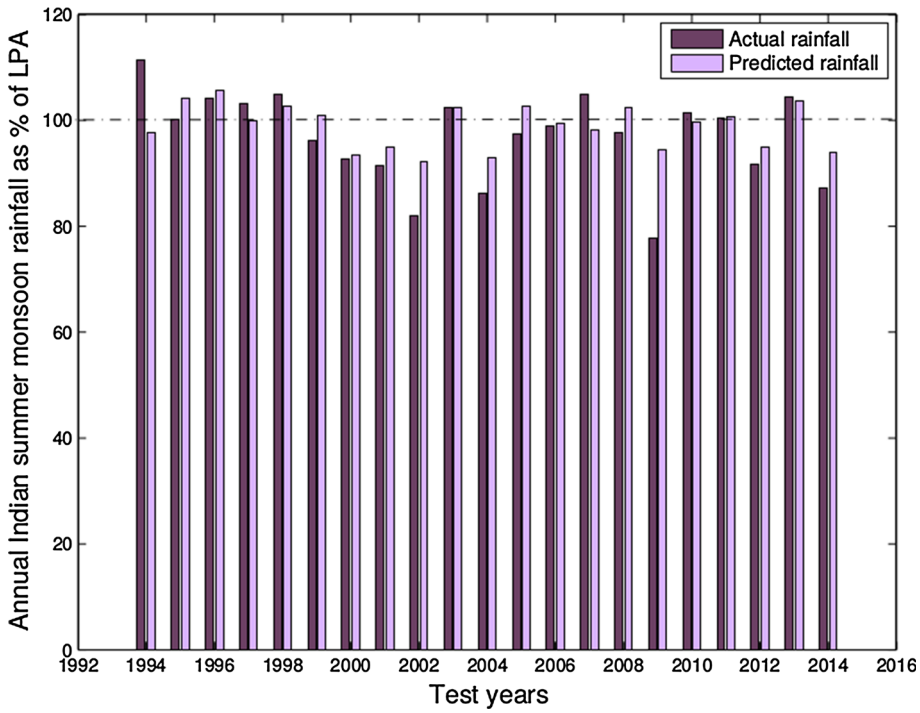

Fig. 8 Performance of forecasts by identified monsoon predictors of air temperature (AT) for test period 1994–2014. *Dark* and *light bars* represent the actual and predicted ISMR

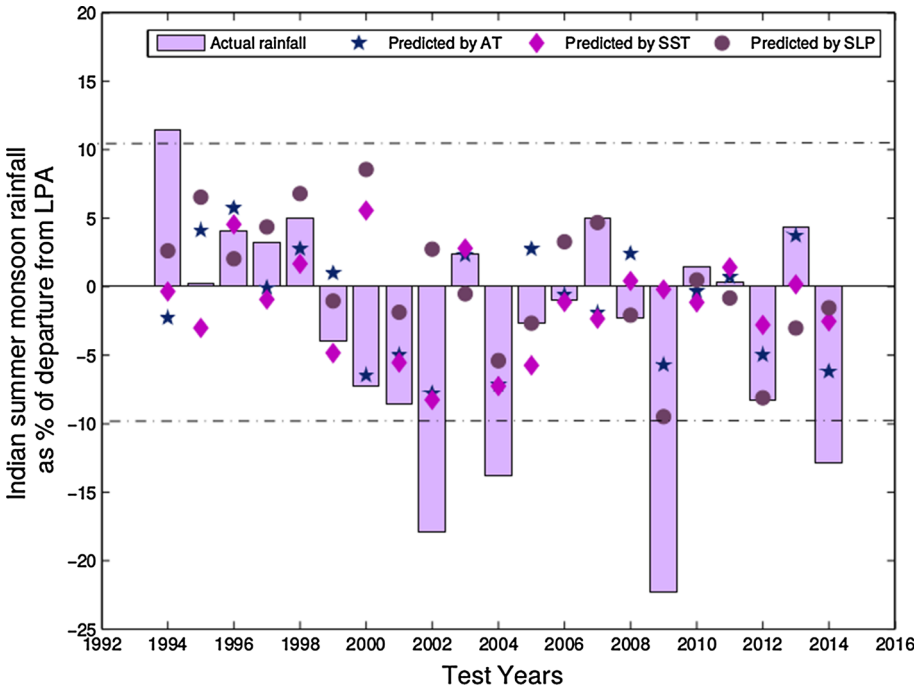

Fig. 9 Performance of forecasts by identified monsoon predictors of AT, SST, and SLP for test period 1994–2014. *Bar* represents the actual ISMR, and the *symbols* represent forecasts given by models with identified monsoon predictors

comb4—comprises top two predictors of each AT, SST and SLP. Predictor set with with air temperature and sea level pressure-based monsoon predictors (comb2) shows best performance in forecasting with mean absolute error of 4.8 %. Table 3 shows the forecasting skills by the combined predictor sets. They show superior performance than the predictor sets of individual SST and SLP, but they show less accuracy than individual set of AT. Figure 10 shows the prediction by the considered combined predictors sets during the test period 1994–2014.

Identified monsoon predictors are compared with existing India meteorological department's (IMD) models (Gowariker et al. 1991; Rajeevan et al. 2004). It is compared with existing 16-parameter power regression model (Gowariker et al. 1991), 8 and 10-parameter IMD models (Rajeevan et al. 2004). We have compared prediction given by the identified monsoon predictors in time period 1996–2002 with IMD model's prediction specified in article by Rajeevan et al. (2004). IMD models give root mean square errors of 10.8, 7.6, and 6.4 %, respectively. Predictor sets with proposed new monsoon predictors of AT, SST, and SLP give root mean square errors of 4.7, 6.0, and 6.2 %, respectively. Results are shown in Fig. 12.

Prediction by our proposed models with identified monsoon predictors is also compared with recent three predictions given by different India meteorological department's (IMD) models—(i) one prediction from IMD operational power regression model (Rajeevan et al. 2004), (ii) two predictions from IMD current pursuit projection regression (PPR) model (Rajeevan et al. 2007) in two different lag. We have compared prediction given by the identified monsoon predictors in time period 2003–2014 with available IMD model's prediction. IMD operational model gives mean absolute error of 7.5 %. Current running PPR model of IMD gives prediction in two phases—first in April (LRF1) and next in June (LRF2). LRF1 and LRF2 predict Indian monsoon with mean absolute errors of 7.1 and 6.5 %, respectively.

Predictor sets with identified new monsoon predictors of AT, SST, and SLP give mean absolute errors of 4.4, 5.6, and 4.4 %, in the month of May, respectively. Thus, it can be concluded that prediction models with our identified monsoon predictors outperform all IMD models (Gowariker et al. 1991; Rajeevan et al. 2004, 2007) to a great extent. Results are shown in Fig. 12.

An elaborate comparison of the predictions by IMD models and prediction by identified monsoon predictors of

Table 3 Mean absolute errors (%) for annual Indian summer monsoon rainfall prediction for dataset built with combined identified monsoon predictors for variables AT, SST, and SLP during test period 1994–2014

Combined predictor sets	Predictors	RegTreeB model	DecTreeB model
comb1	AT+SLP	5.1	5.3
comb2	AT+SST	**4.8**	5.2
comb3	SST+SLP	5.0	**5.0**
comb4	AT+SST+SLP	5.2	5.2

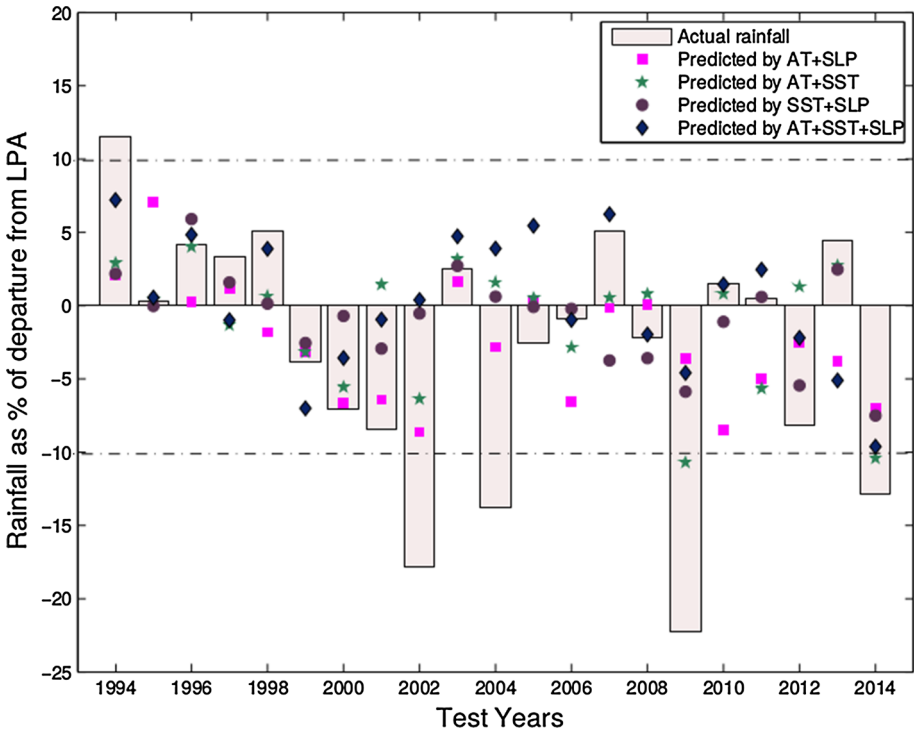

Fig. 10 Performance of forecasts by combined identified monsoon predictors of AT, SST, and SLP for test period 1994–2014. *Bar* represents the actual ISMR, and the *symbols* represent forecasts given by models with combined identified monsoon predictors

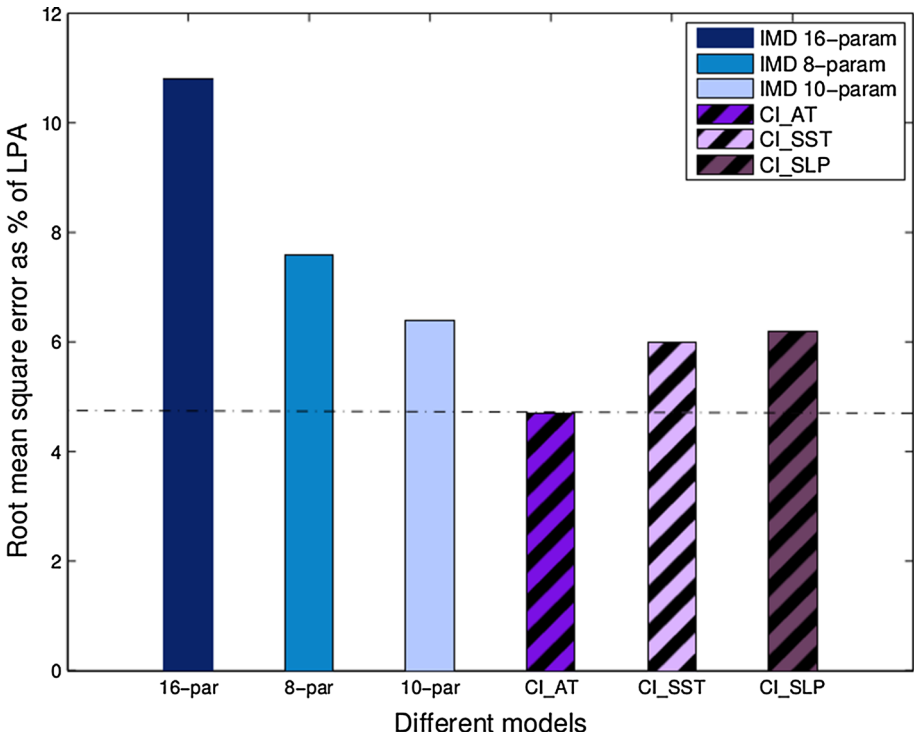

Fig. 11 Comparison of prediction by identified monsoon predictors of air temperature CI_AT, sea surface temperature CI_SST, and sea level pressure CI_SLP models with IMD existing 16-parameter (Gowariker et al. 1991), 8-parameter, and 10-parameter (Rajeevan et al. 2004) models for period 1996–2002

air temperature for all test years against actual rainfall is shown in Fig. 13. It is noted that prediction by identified monsoon predictors is nearer to the actual rainfall as compared to the prediction made by the IMD models. Identified monsoon predictors also detect the extremes during the period.

Prediction of the year 2015: Annual Indian summer monsoon rainfall for the year 2015 is forecasted using the identified monsoon predictors. We consider the predictor sets with identified predictors which have least mean absolute errors during test period to predict rainfall of 2015. Predictor sets with identified monsoon predictors of

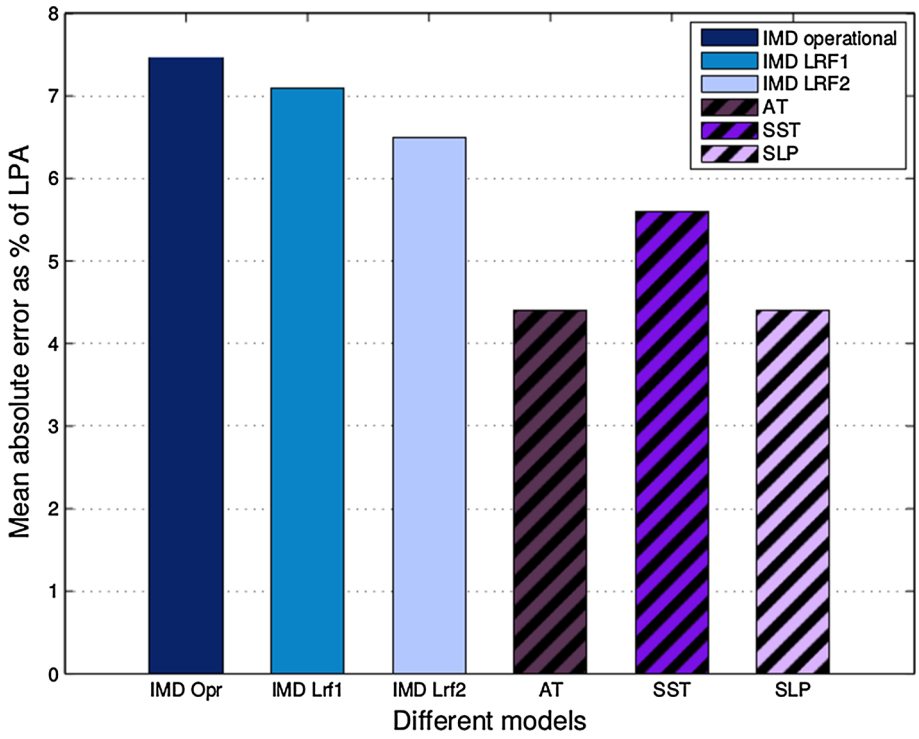

Fig. 12 Comparison of prediction by identified individual monsoon predictors of AT, SST, SLP models with IMD existing operational (Rajeevan et al. 2004) and PPR (LRF1 and LRF2) (Rajeevan et al. 2007) models for period 2003–2014

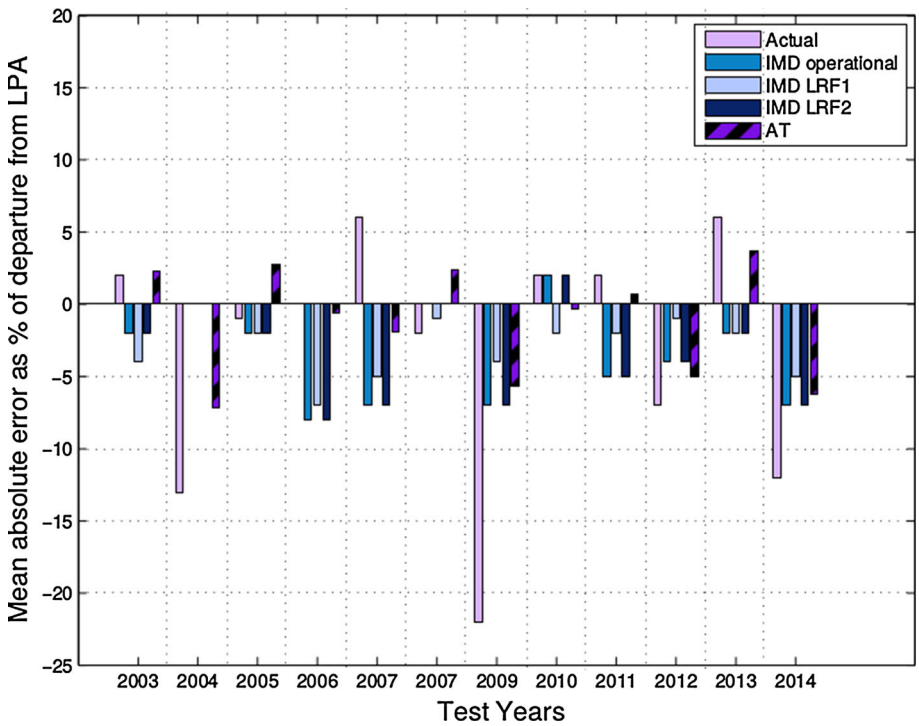

Fig. 13 Prediction by identified monsoon predictors of air temperature and IMD existing models (Rajeevan et al. 2004, 2007) against actual rainfall for period 2003–2014

air temperature, sea surface temperature, and sea level pressure (P4_AT, P3_SST, and P1_SLP) forecast rainfalls as 788.8, 847.8, and 818.4 mm, respectively. Predicted rainfalls are 87.9, 94.4, and 91.2 % of long period average rainfall. All the predictor sets proposed below normal rainfall for the current year and also specify high chance of occurrence of drought phenomenon. IMD has downgraded its forecast of Indian summer monsoon for the current year 2015 from 93 % of LPA to 88 % of LPA in second updated forecast in June after the first forecast in April (www.imd.gov.in). Currently, IMD has reported on 1st October, 2015 the actual Indian summer monsoon

rainfall for 2015 as 14 % deficit from LPA. Prediction by the identified monsoon predictors of air temperature is fully aligned to actual Indian summer monsoon rainfall for the current year which predicts it as 12.1 % deficit from LPA.

6 Meteorological interpretations of identified monsoon predictors

Identified monsoon predictors for climatic variable air temperature, sea surface temperature, and sea level pressure (Figs. 5, 6, and 7 in Sect. 5) are categorized into two classes—(i) predictors coinciding with geographical regions already known important for Indian monsoon process, and (ii) predictors corresponding to some new regions whose influence over Indian monsoon is not studied in the past. Both categories of identified monsoon predictors are discussed in the following sections.

6.1 Recapture of known monsoon predictors

Recapturing known monsoon predictors corresponding to geographical regions known to be important for Indian monsoon validates our proposed autoencoder-based approach for monsoon predictor identification.

- *South Equatorial Pacific Ocean SLP* (Fig. 7: CI_SLP1) [0°S–40°S, 100°W–120°W; May]: The region of south equatorial Pacific Ocean is found to be an influential predictor of Indian monsoon (Cherchi and Navarra 2013) correlating Indian monsoon [(correlation (μ) of +0.34].
- *South Equatorial Indian Ocean SLP* (Fig. 7: CI_SLP3) [10°N–0°N, 60°E–100°E; May]: High sea level pressure in this region (having μ +0.30) leads to difference in pressure with Central Asia landmass low, which could lead to stronger winds and hence higher moisture advection from ocean to landmass. This region also encompasses a significant part of the Indian Ocean dipole region and its atmospheric component the equatorial Indian Ocean oscillation (EQUINOO). EQUINOO is known to have a strong association with Indian Summer Monsoon (Gadgil et al. 2004). Goswami and Ajayamohan (2001) have also shown that convection over the equatorial Indian Ocean is inversely related to strength of Indian monsoon on inter-annual timescales.
- *Peru-South Eastern Equatorial Indian Ocean SLP* (Fig. 7: CI_SLP6) [10°S–20°S, 60°W–80°W; 40°S–50°S, 60°E–80°E; May]: Integration of sea level pressure of regions of Peru and South Eastern Indian Ocean (Schott et al. 2009) builds a monsoon predictor, having μ of +0.20 with monsoon.
- *North Pacific Ocean SST* (Fig. 6: CI_SST1) [10°N–20°N, 120°W–180°W; May]: Sea surface temperature in North Pacific Ocean is correlated (−0.35) to Indian monsoon (Cherchi and Navarra 2013).
- *Madagascar-South Eastern Indian Ocean-New Zealand-North Equatorial Pacific Ocean SST* (Fig. 6: CI_SST2) [20°S–30°S, 40°E–60°E; 40°S–50°S, 160°E–180°E; 0°N–10°N, 80°W–100°W; January]: Anomaly in sea surface temperature of these regions (Li et al. 2008) build up a strong monsoon predictor influencing Indian monsoon with correlation of +0.28 with the phenomenon. The Madagascar region is in the vicinity of the Mascrene High which is known to be associated with the Indian summer monsoon (Krishnamurti and Bhalme 1976). North Equatorial Pacific encompasses parts of the Nino 1 and Nino 2 regions. ENSO is known to be significantly associated with Indian summer monsoon rainfall.
- *North Western Pacific Ocean-South Eastern Indian Ocean SST* (Fig. 6: CI_SST5) [20°N–30°N, 160°W–180°W; 30°S–40°S, 60°E–80°E; April]: Combination of sea surface temperature of this two regions is an important predictor for monsoon rainfall over India (μ is +0.21).
- *West Europe-South Eastern Indian Ocean AT* (Fig. 5: CI_AT2) [40°N–50°N, 0°E–20°E; 40°S–50°S, 60°E–80°*E*; May]: Landmass of West Europe and South Eastern Indian Ocean combines to form an air temperature-based monsoon predictor, shows correlation of −0.32 with Indian monsoon.
- *North East Africa-South Eastern Indian Ocean AT* (Fig. 5: CI_AT6) [20°N–30°N, 20°E–40°E; 40°S–50°S, 60°E–80°E; May]: Air temperature in North East Africa and South Eastern Indian Ocean (Schott et al. 2009) regions represents a predictor important for Indian monsoon (μ is +0.26).

6.2 Identification of new monsoon predictors

New monsoon predictors are estimated using our proposed approach, which are associated to new geographical regions are mentioned in this section.

- *Tasman Sea-Southern Australia SLP* (Fig. 7: CI_SLP2) [30°S–40°*S*, 140°E–160°E; 20°S–30°S, 160°E–180°*E*; May]: Amalgamation of sea level pressure of Tasman Sea with Southern Australia and its coastal region in non-linear manner is a representative for new monsoon predictor, having good correlation (+0.30) with Indian monsoon.

- *South Atlantic Ocean-North Pacific Ocean SLP* (Fig. 7: CI_SLP4) [30°S–40°S, 0°E–20°E; 20°–*N*50°N, 140°E–160°E; May]: Pressure gradient between this two regions is evaluated as an important sea level pressure-based monsoon predictor having μ of +0.24 with Indian monsoon, which could influence winds into Indian Landmass.
- *North Western Pacific Ocean-Coastal Mexico SLP* (Fig. 7: CI_SLP5) [40°N–50°N, 120°W–140°W; 20°N–30°N, 100°W–120°W; March]: This region is found to be tele-connected with Indian region and sea level pressure anomaly over the region shows a correlation of −0.22 with ISMR.
- *South Eastern Indian Ocean SST* (Fig. 6: CI_SST3) [30°S–40°S, 80°E–100°E; January]: Sea surface temperature of the region is found to be an important region influencing Indian monsoon (Li et al. 2008)(μ of +0.23).
- *Norwegian Sea-Cuba-North West Atlantic Ocean Sea Surface Temperature* (Fig. 6: CI_SST4) [60°N–70°N, 0°W–20°W; 20°N–30°N, 60°W–80°W; March]: Sea surface temperature over regions of Norwegian Sea and Cuba with its adjacent North West Atlantic Ocean (Hong et al. 2003) is non-linearly combined to acquire a potential predictor by our proposed approach, having μ as +0.22.
- *Caribbean Sea-Coral Sea SST* (Fig. 6: CI_SST6) [10°N–20°N, 60°W–80°W; 10°S–20°S, 140°E–160°E; April]: These regions jointly evaluated as a predictor influencing Indian monsoon with correlation of -0.20.
- *Argentina-Western North Atlantic Ocean AT* (Fig. 5: CI_AT1) [30°S–40°S, 60°W–80°W; 10°N–20°N, 40°W–60°W; April]: Argentina along with North Atlantic Ocean (Kucharski et al. , 2008) on the south-east of Cuba is acquired as a predictor influencing Indian monsoon (μ of +0.36).
- *South Western Pacific Ocean-Argentina AT* (Fig. 5: CI_AT3) [30°S–40°S, 160°W–180°W; 30°S–40°S, 60°W–80°W; April]: Regions of South Western Pacific Ocean and Argentina combine to form a non-linear monsoon predictor, highly correlated (+0.29) to Indian monsoon.
- *Spain-Japan AT* (Fig. 5: CI_AT4) [40°N–50°N, 0°*W*–20°W; 40°N–50°N, 140°E–160°E; May]: Air temperature over Spain with its surrounding North Atlantic Ocean and Japan with its neighboring North Pacific Ocean represents a predictor important for Indian monsoon(μ of +0.28).
- *Eastern North Pacific Ocean-South Eastern Indian Ocean AT* (Fig. 5: CI_AT5) [20°N–30°, 120°W–140°W; 40°S–50°S, 60°E–80°E; May]: Amalgamation of air temperature in regions of Eastern North Pacific Ocean and South Eastern Indian Ocean is obtained as a potential monsoon predictor correlated (+0.23) to Indian summer monsoon.

7 Conclusions

Identification of new global predictors for Indian monsoon is attempted using autoencoder, which assists in capturing complex and non-linear monsoon predictors. Proposed autoencoder-based approach assists in determining potential predictors as non-linear combination of climatic variables of different geographical locations. The approach helps in exploring climatic variables over the world and thereby identifying new predictors which forecast Indian monsoon with high accuracy. Some of the identified monsoon predictors resemble already existing monsoon predictors important for Indian summer monsoon rainfall, which stand as validation of our proposed approach of monsoon predictor identification. Non-linear models are designed with identified monsoon predictors, to evaluate their forecasting skills for Indian monsoon. Mean absolute errors of 4.5, 5.3, 5.4 % are obtained by predictor sets built with identified monsoon predictors of air temperature, sea surface temperature, and sea level pressure, respectively. Prediction by our identified monsoon predictors also captures the extremes. Our model with identified monsoon predictors shows same sign of anomaly for all four drought years during the test period 1994–2014. Finally, characteristics of newly identified monsoon predictors are also explored.

The future directions of our work include utilization of deep neural network like stacked autoencoder for identification of more composite monsoon predictors and represent them at different regional level. Other deep neural architecture like convolutional neural network can also be used to extract non-linear predictors, from combination of different climatic variables from multiple layers of small neural collections of the network. In addition, rectified linear units can be used, which is an non-saturating activation function, which increases the non-linear property of the network and thus assists in extracting more complex and non-linear predictors. Lastly, dropout method can be included for training of the network which speed up the process and prevent over-fitting of network resulting in generalizing the network and assisting in identification of more efficient predictors of monsoon. Finally, the significance of the identified predictors in climate domain from background of physical climatic process and their influence in different climatic phenomenon can also be explored.

References

Baldi P (2012) Autoencoders, unsupervised learning, and deep architectures. ICML Unsupervised Transf Learn 27:37–50

Blanford HF (1884) On the connexion of the Himalaya snowfall with dry winds and seasons of drought in India. Proc R Soc Lond 37(232–234):3–22

Boriah S, Simon G, Naorem M, Steinbach M, Kumar V, Klooster S, Potter C (2004) Predicting land temperature using ocean data. In: Proceedings of the knowledge discovery in databases KDD, Citeseer

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

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

Compo G, Whitaker J, Sardeshmukh P, Matsui N, Allan R, Yin X, Gleason B, Vose R, Rutledge G, Bessemoulin P, Bronnimann S, Brunet M, Crouthamel R, Grant A, Groisman P, Jones P, Kruk M, Kruger A, Marshall G, Maugeri M, Mok H, Nordli O, Ross T, Trigo R, Wang X, Woodruff S, Worley S (2011) The twentieth century reanalysis project. Q J R Meteor Soc 137(654):1–28

DelSole T, Shukla J (2002) Linear prediction of Indian monsoon rainfall. J Clim 15:3645–3658

DelSole T, Shukla J (2012) Climate models produce skillful predictions of Indian summer monsoon rainfall. Geophys Res Lett 39(9):L09–703

Gadgil S, Vinayachandran P, Francis P, Gadgil S (2004) Extremes of the Indian summer monsoon rainfall, ENSO and equatorial Indian ocean oscillation. Geophys Res Lett 31(12):L12213

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

Goswami BN, Ajayamohan R (2001) Intraseasonal oscillations and interannual variability of the Indian summer monsoon. J Clim 14(6):1180–1198

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

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

Hong YT, Hong B, Lin QH, Zhu YX, Shibata Y, Hirota M, Uchida M, Leng XT, Jiang HB, Xu H (2003) Correlation between Indian ocean summer monsoon and North Atlantic climate during the Holocene. Earth Planet Sci Lett 211(3):371–380

Krishnamurti TN, Bhalme H (1976) Oscillations of a monsoon system. Part I. Observational aspects. J Atmos Sci 33(10): 1937–1954

Kucharski F, Bracco A, Yoo JH, Molteni F (2008) Atlantic forced component of the Indian monsoon interannual variability. Geophys Res Lett 35(4):L04706

Kumar N, Nasser M, Sarker SC (2011) A new singular value decomposition based robust graphical clustering technique and its application in climatic data. J Geogr Geol 3(1):227–238

Li S, Lu J, Huang G, Hu K (2008) Tropical Indian Ocean basin warming and East Asian summer monsoon: a multiple AGCM study. J Clim 21(22):6080–6088

Li Y, Yang R (2013) A hybrid algorithm combining auto-encoder network with sparse Bayesian regression optimized by artificial bee colony for short-term wind power forecasting. Przeglkad Elektrotechniczny 89(2a):223–228

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22

Liu JNK, Hu Y, He Y, Chan PW, Lai L (2015) Deep neural network modeling for big data weather forecasting. In: Information granularity, big data, and computational intelligence, Springer, pp 389–408

Loh WY (2008) Classification and regression tree methods. Encyclo of statistics in quality and reliability, pp 315–323

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

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

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 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

Sap MNM, Awan AM (2005) Finding spatio-temporal patterns in climate data using clustering. In: Proceedings of the 2005 international conference cyberworlds, IEEE, pp 8–15

Schott FA, Xie SP, McCreary JP (2009) Indian Ocean circulation and climate variability. Rev Geophys 47(1):RG1002

Smith TM, Reynolds R, Peterson T, Lawrimore J (2008) Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). J Clim 21(10):2283–2296

Song C, Liu F, Huang Y, Wang L, Tan T (2013) Auto-encoder based data clustering. In: Progress in pattern recognition, image analysis, computer vision and applied, Springer, pp 117–124

Steinbach M, Tan PN, Kumar V, Klooster S, Potter C (2003) Discovery of climate indices using clustering. Proceedings of ACM, ACM SIGKDD, pp 446–455

Thapliyal V, Kulshrestha S (1992) Recent models for long range forecasting of south-west monsoon rainfall in India. Mausam 43(3):239–248

Walker G (1924) Correlation in seasonal variations of weather-IV, a further study of world weather. Mem India Meteorol Dept 24:275–332

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

Xue Y, Smith T, Reynolds R (2003) Interdecadal changes of 30-yr SST normals during 1871–2000. J Climate 16:1601–1612