Deep learning for predicting the monsoon over the homogeneous regions of India

Article *in* Journal of Earth System Science · June 2017

DOI: 10.1007/s12040-017-0838-7

3 authors:

Moumita Saha
University of Colorado Boulder
28 PUBLICATIONS **336** CITATIONS

SEE PROFILE

Pabitra Mitra
Indian Institute of Technology Kharagpur
245 PUBLICATIONS **8,251** CITATIONS

SEE PROFILE

Ravi S. Nanjundiah
Indian Institute of Science Bangalore
198 PUBLICATIONS **4,240** CITATIONS

SEE PROFILE

J. Earth Syst. Sci. (2017) 126:54
DOI 10.1007/s12040-017-0838-7

Deep learning for predicting the monsoon over the homogeneous regions of India

Moumita Saha[1,4,*], Pabitra Mitra[1] and Ravi S Nanjundiah[2,3,4]

[1] *Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India.*
[2] *Divecha Centre for Climate Change, Indian Institute of Science, Bangalore 560 012, India.*
[3] *Indian Institute of Tropical Meteorology, Pune, Maharashtra 411 008, India.*
[4] *Present address: Centre for Atmospheric and Oceanic Sciences, Indian Institute of Science, Bangalore 560 012, India.*
**Corresponding author. e-mail: moumita.saha2012@gmail.com*

MS received 30 March 2016; revised 21 December 2016; accepted 3 February 2017; published online 12 June 2017

Indian monsoon varies in its nature over the geographical regions. Predicting the rainfall not just at the national level, but at the regional level is an important task. In this article, we used a deep neural network, namely, the stacked autoencoder to automatically identify climatic factors that are capable of predicting the rainfall over the homogeneous regions of India. An ensemble regression tree model is used for monsoon prediction using the identified climatic predictors. The proposed model provides forecast of the monsoon at a long lead time which supports the government to implement appropriate policies for the economic growth of the country. The monsoon of the central, north-east, north-west, and south-peninsular India regions are predicted with errors of 4.1%, 5.1%, 5.5%, and 6.4%, respectively. The identified predictors show high skill in predicting the regional monsoon having high variability. The proposed model is observed to be competitive with the state-of-the-art prediction models.

Keywords. Feature learning; stacked autoencoder; monsoon predictor; ensemble of regression trees; regional Indian summer monsoon.

1. Introduction

India being an agricultural country, the monsoon governs the pulse of life for its mankind along with the flora-fauna existing in the subcontinent. It has a large impact on agriculture, fresh-water renewal, the generation of hydro-electricity and altogether in the economy of the country. Understanding and predicting the monsoon and its variability are challenging. Variation of monsoon is significant over the spatial scale.

The Indian summer monsoon (ISM) prevailing between June and September contributes for more than 70% of Indian annual rainfall and thus, important for the gross agricultural production of the country. However, monsoon variation over different regions of the subcontinent is not uniform. Thus, it is significant to understand, analyze, and forecast the spatial variability of the Indian monsoon. Forecasting the regional variation of monsoon at an appropriate lead month will be beneficial for planning proper agricultural policies. India Meteorological Department (IMD) has segregated the Indian subcontinent into four different regions, namely, central, north-east, north-west and south-peninsular India (www.imdpune.gov.in) (regions are shown in figure 1).

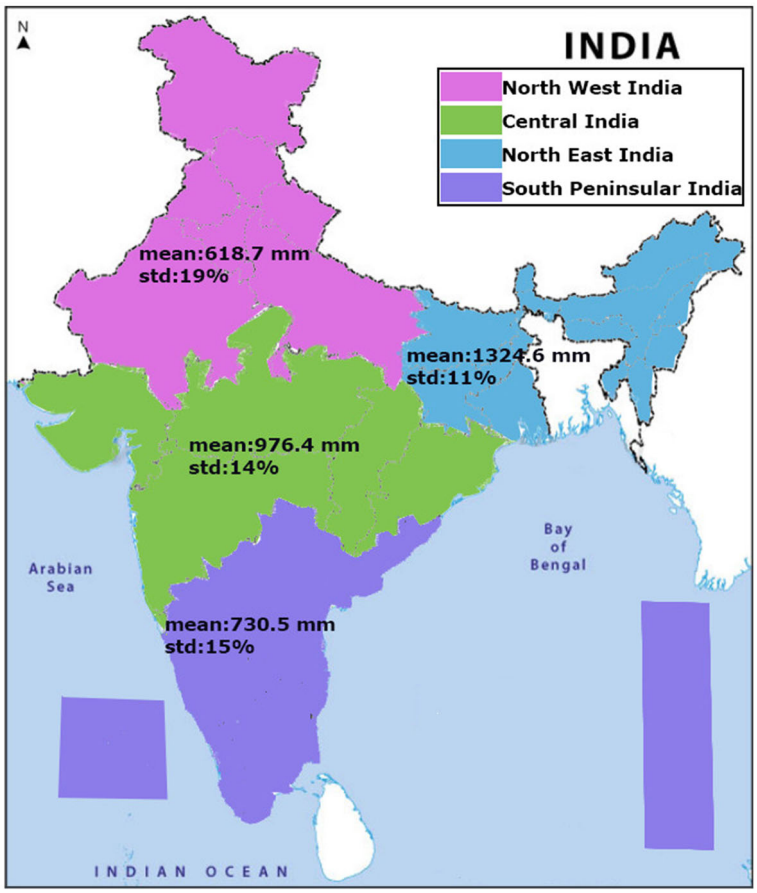

Figure 1. Four homogeneous regions of India as defined by India Meteorological Department.

The inter-annual variability at regional scales is higher compared to the national rainfall, and thus, it is more challenging to predict. Nair *et al.* (2013) have focused on regional scale Indian monsoon prediction along with the aggregate monsoon prediction using supervised principal component regression on the outcomes of the general circulation models. Regional prediction of the Indian subcontinent has also been attempted with the application of canonical correlation analysis (Sinha *et al.* 2013). Recently, Indian Institute of Tropical Meteorology, Pune has performed the aggregate Indian monsoon prediction along with the prediction of its regional sub-divisions using shared nearest neighbour clustering approach (Kakade and Kulkarni 2016). Their method has engrossed the non-linear connections between seasonal monsoon and cluster parameters along with the linear relations, and has provided promising prediction accuracies. All the four regions of India (as defined by IMD) could have different rainfall distributions as well as different influencing factors to a certain extent. The south-peninsular region receives rainfall during the onset of monsoon and a huge amount of orographic rainfall occurs due to the presence of Western Ghats. The north-east region being at the foothills of the Great Himalayan Ranges receives high rainfall. The north-west India usually receives less rainfall. Thus, owing to variations in monsoon patterns, it is highly necessary to predict the monsoon for different homogeneous regions separately for better framing of policies and improving the gross productivity of the country.

The inter-annual variability of the Indian summer monsoon is associated with the variation of different variables like sea level pressure, sea surface temperature, wind velocity, etc., over various geographical regions of the globe. They can be a good indicator of intensity of the monsoon. In addition, it is observed that the set of climatic predictors influencing the monsoon also evolve over time. IMD continually reassesses different climatic parameters and update the forecast model for delivering superior forecast (Gowariker *et al.* 1991; Rajeevan *et al.* 2004, 2007). A number of studies are performed for predicting

all-India summer monsoon, which include use of different artificial neural networks to forecast the monsoon with better accuracy (Saha and Mitra 2016). In comparison to the study of all-India monsoon, the exploration of regional India monsoon is limited in literature. We focus on identifying different set of predictors for the regional rainfalls of India for a better framing of their characteristics and finally, predicting the regional monsoon with good accuracy.

Conventionally, the selection of predictors of the monsoon are completely dependent on the study of global physical processes and prior meteorological knowledge (Delsole and Shukla 2012; Wang *et al.* 2015). However, we propose an unsupervised automated identification of predictors using deep learning methods. Stacked autoencoder (a deep neural network) is used for unsupervised feature learning, which helps in achieving the non-linear composition of climatic variables to identify the novel monsoon predictors. It considers the climatic variables all over the globe without any expert's supervision and propose the new predictors which are the combination of variables from different regions located at geographical distant locations. The single-layer autoencoder has proved its efficiency in predicting all-India summer monsoon with the newly discovered predictors by iterative training of architecture (Saha *et al.* 2016a). Stacked autoencoders are utilized for the prediction of all-India summer monsoon rainfall and the deep network shows its superiority over shallow single-layer autoencoder in forecasting the monsoon (Saha *et al.* 2016b). The prediction of early and late phases of Indian summer monsoon are also attempted using deep architecture of stacked autoencoder and the forecasts provided by the deep structure are comparable (Saha *et al.* 2016c).

We propose the use of stacked autoencoder for identification of predictors for regional monsoon and subsequently, predict the regional monsoon utilizing the identified predictors. The proposed approach differs from past methods using stacked autoencoder (Saha *et al.* 2016b, c) in the following ways: (i) present study is focussed on regional Indian summer monsoon, which have high variability and thus, more challenging to predict compared to all-India or early and late phases of monsoon, (ii) input climatic variables are different, it is observed that different predictor variables are significant for different categories of monsoon, (iii) an exclusive set of predictors are identified for each category of monsoon and subsequently, we have developed the prediction models individually for all four regions of the Indian monsoon, these models are different compared to the models developed for all-India monsoon (as an analogy, it can be said that regression models are used for different empirical problems, but separate models need to be developed for each problem. Similarly, in this case, the stacked autoencoder may be the backbone of our approaches for all-India (Saha *et al.* 2016b) and regional monsoon prediction, but the models developed are unique and different for each category of monsoon), (iv) prediction model developed for the regional India summer monsoon is the ensemble of regression trees with the bagging approach; the number of trees in the ensemble, the number of iterations, and other tunable parameters are set individually for each model developed for four regions. In conclusion, it can be said that though the technique would have been similar, development of model for each region are independent and they are built individually for all homogeneous regions of India.

We follow a two-step approach for the improvement of prediction of the regional monsoons. First we focus on the identification of new monsoon predictors by automated feature learning using the stacked autoencoder from climatic variables for the regional monsoon, and secondly, we used the ensemble of regression trees as the prediction model with identified new monsoon predictors to predict different regions of the monsoon (figure 2).

The proposed unsupervised learning of the predictors from input variables utilizing the stacked autoencoder assists in deriving new features. These features are selected based on their correlation with the regional monsoon, to model all the categories of monsoon distinctly. These are presented as the respective monsoon predictors. The monsoon predictors are an amalgamation of climatic variables from different locations of the world. The prediction model with the climatic predictors are utilized to forecast the seasonal monsoon over the Indian regions. This paper aims to highlight on how accurately the homogeneous sub-divisions of India are predicted or which sub-division has better skill in prediction with the newly identified predictors.

This paper is presented in the following manner. An overview of the climatic variables that are explored for the identification of new monsoon predictors is provided in section 2. The section also describes different regional monsoon categories to

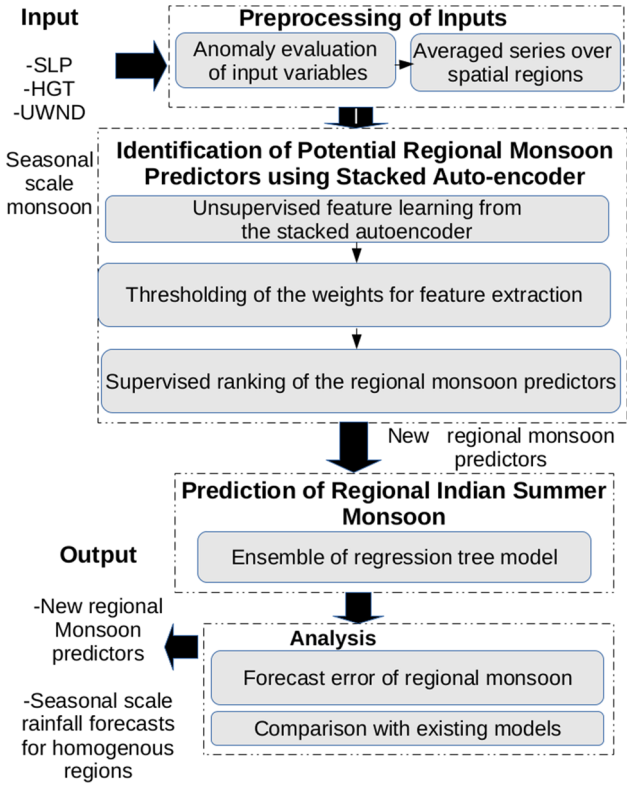

Figure 2. Identification of regional monsoon predictors using stacked autoencoder and the forecast of seasonal scale rainfall for the homogeneous regions of India.

be forecasted. Section 3 describes the basic architecture of stacked autoencoder and its working principle. Section 4 explains broadly the proposed automated feature learning for the identification of potential monsoon predictors. The prediction model for forecasting the monsoon is described in section 5. Section 6 focusses on the experimental results, which show the skills of the identified monsoon predictors in forecasting the regional rainfall of the country. Finally, the paper is concluded in section 8 with future scope of the proposed method.

2. Climatic variables considered

Climatic variables explored for feature learning, and thereby, identification of the potential monsoon predictors are described as follows.

- Sea level pressure (SLP) is a significant factor, whose spatial and temporal fluctuations over various regions of the world affects the monsoon circulation and rainfall (Rajeevan *et al.* 2007).
- India has the Great Himalayas and the huge Tibetan plateau as its prime boundaries. The winds blowing in the lower atmosphere are incapable of entering the region because of the high boundaries surrounding the Indian landmass. Thus, zonal-wind at the pressure level of 200 hecPascal (UWND) is taken into account. These winds can connect the Indian region to mid-latitudes and thus, influence the monsoon of the subcontinent.
- Last climatic input is the geo-potential height at 200 hecPascal pressure level (HGT). Gradients of 200 hecPascal of geo-potential height have an impact on the large scale circulation and thus affect the monsoon circulation and rainfall. All the above mentioned climatic variables are acquired from the NCEP reanalysis derived data, which are present at $2.5^\circ \times 2.5^\circ$ spatial resolution (Kalnay *et al.* 1996).

All data are considered for the period 1948–2015 on a monthly scale. The period between 1948 and 2000 is considered for feature learning, followed with identifying the monsoon predictors and training the prediction model designed using the potential monsoon predictors. A test-period of 2001–2014 is considered for judging the prediction skills of the identified predictors of the monsoon.

The analysis and prediction of the monsoon is considered for the one prevailing during June–September are described as following (www.imdpune.gov.in).

- The central India monsoon having the LPA of 976.4 mm with std of 14%.
- The north-east India monsoon having the LPA of 1324.6 mm with std of 11%.
- The north-west India monsoon having the LPA of 618.7 mm with std of 19%.
- The south-peninsular India monsoon having the LPA of 730.5 mm with std of 15%.

The regional monsoon is collected from India Meteorological Department situated at Pune (www.imdpune.gov.in) for 1948–2014. It is noticed that the variation in rainfall in the regional parts are higher than that of the aggregate Indian summer monsoon (variation of 10%), which adds more challenge and necessity in regional predictions as compared to the national prediction.

As a preprocessing step, anomaly data (denoted as $\text{anomalyData}^{\text{yr}}_{\text{mon}}$ for month mon of yr^{th} year) is derived from all the input variables by removing the monthly mean, as shown in equation (1).

$$\text{anomalyData}^{\text{yr}}_{\text{mon}} = \text{Var}^{\text{yr}}_{\text{mon}} - \text{mean}(\text{Var}_{\text{mon}}), \quad (1)$$

where $\mathrm{Var}_{\mathrm{mon}}^{\mathrm{yr}}$ is the value of climatic variable in the $\mathrm{mon}^{\mathrm{th}}$ month of the $\mathrm{yr}^{\mathrm{th}}$ year, and mean $(\mathrm{Var}_{\mathrm{mon}})$ denotes the mean of the $\mathrm{mon}^{\mathrm{th}}$ month of the climatic variable.

The anomaly of climatic variables are further processed in the following manner. Spatial rectangular regions of 20° longitude × 10° latitude is considered all over the world, which condenses to 324 rectangular regions ((360/20) longitudinal × (180/10) latitudinal).

A single time series is obtained by averaging all the series of a specific climatic variable within the cover of the rectangular region. The input features to the first layer of the stacked autoencoder are such averaged time-series of the selected regions. The number of input features (corresponding to inputs of the stacked autoencoder) for the climatic variable sea level pressure (SA_SLP), geo-potential height (SA_HGT) and u-wind (SA_UWND) are 324, each.

3. Overview of the deep stacked autoencoder network

Some discussions of the use of stacked autoencoder to climate studies are provided by Saha *et al.* (2016b, c). However, a brief overview of the model and its working is presented here. An autoencoder belongs to the category of artificial neural network. The architecture has the ability to learn the properties of input data space, and it is utilized for feature learning (Hinton and Salakhutdinov 2006; Baldi 2012). The simplest form of autoencoder, namely, the single-layer autoencoder has one hidden (or internal) layer in addition to an input and an output layer. Each layer consists of multiple nodes or neurons. The model fixes the targets (or output values) same as inputs. The architecture comprises of two parts: (i) the encoder (which operates between the input and internal layer), (ii) the decoder (which operates between the internal and output layers). This deep-learning based architecture provides a non-linear functionality by repeated training of both the parts (the encoder and decoder). The non-linear functionality is endorsed by the encoder, whereas the decoder is specialized in the reconstruction of input from the data form generated by the encoder. A single layer autoencoder based deep learning technique has been successfully used for the aggregate monsoon prediction of Indian subcontinent (Saha *et al.* 2016a).

Multiple single-layer autoencoders are combined to generate multi-layered architectures by using the output of autoencoder architecture from the antecedent layer as the input to the succeeding layer. The architecture of the stacked autoencoder is shown in figure 3. The stacked autoencoder assists in identifying new features at different levels of complexity. The deeper the layers in stacked autoencoder, more complicated are the discovered features. Thus, the features obtained from stacked autoencoders are more information-rich than that obtained from the single layer autoencoder and they can be more superior predictors of the monsoon.

Unsupervised pre-training is carried out considering a single layer at an instant, where all the layers are trained for minimizing the reconstruction errors for their input (shown in the left portion of figure 3). After all the layers are pre-trained, the deep architecture network is further tuned utilizing the gradient descent method. The method fixes the weights of all the layers in the deep network (shown in the right portion of figure 3). New features are evaluated from the hidden layers of deep learning-based stacked autoencoder architecture.

Formally, say $var \in R^m$ denotes input, the activation of neuron in all the internal layers is represented as hid_i^r, where i varies as $i = 1, \ldots, t$, where t is the count of neurons in the rth internal layer, the activation of the internal neurons are shown in equation (2).

$$hid^r(var) = fn\left(Wt^r \cdot var + bs^r\right), \tag{2}$$

where $fn\,(y) = (e^{2y} - 1)/(e^{2y} + 1)$ is the hyperbolic tangent activation function, $hid^r\,(var) \in R^t$ is the vector for the rth internal layer, and the weight matrix Wt^r has dimension as $(t \times m)$ from previous $(r-1)$th to the present rth internal layer, $bs^r \in R^t$ is the bias of present layer, the count of neurons in the antecedent and present internal layers are t and m, respectively.

For each layer and its succeeding layer (excluding the terminating internal layer and the output layer), the hyperbolic tangent function is used for activation of the neurons. The activation function for the last internal to the output layers is described by equation (3).

$$\widehat{var} = gn\left(Wt^s hid^s\,(var) + bs^s\right), \tag{3}$$

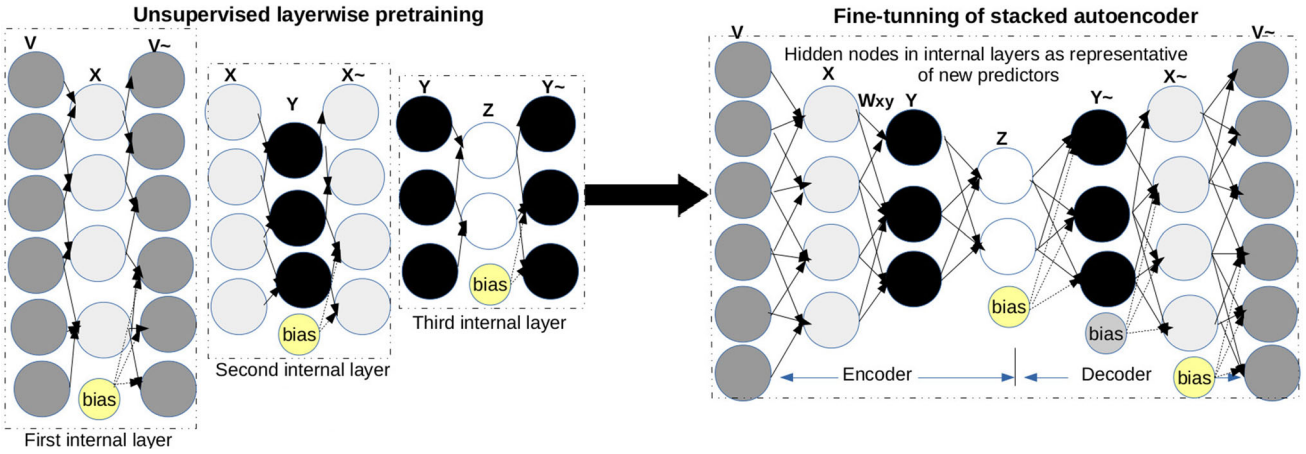

Figure 3. Architecture of the stacked autoencoder showing the pre-training and fine-tuning methods.

where $gn\,(y) = ((c * y) + d)$ denotes the linear activation function, c and d are the constants, $\widehat{var} \in R^m$ is the vector of neurons in the output layer, Wt^s is the matrix containing the weights corresponding to the neurons of the last internal to the output layer, and $bs^s \in R^m$ is the bias of the output layer.

For our problem, input features comprise of variables, namely, *SST*, *UWND* or *HGT* or combination of these variables (mentioned in section 4.1) and the potential monsoon predictors are obtained from the internal layers of the deep network.

4. Automated feature learning and the identification of predictors of the monsoon using stacked autoencoder

The proposed approach to automated feature learning using the stacked autoencoder for the identification of potential monsoon predictors and thereby the prediction of the monsoon in spatial domain (Indian regional monsoon) is shown in figure 2.

In this section, we will discuss the approach for the identification of monsoon predictors for the regional rainfall. The four homogeneous regions as partitioned by India Meteorological Department is shown in figure 1. The potential predictors for the different categories of monsoons are derived using a two-step approach.

The initial step is unsupervised without any information about the monsoon. This step involves the design of stacked autoencoder for climatic variables, followed by pre-training and optimizing the autoencoder, and finally, the threshold of weights to obtain potential predictors of the monsoon as a combination of climatic variables from different regions.

The second step is supervised with guidance of the monsoon distributions of all the regions. This step is distinct to the monsoon of all homogeneous regions. We will use the term *categories* or *regions* interchangeably in the rest of the article to denote *central, north-east, north-west,* and *south-peninsular* monsoons. For all the four monsoon regions (*central, north-east, north-west, south-peninsular*), a correlation between the potential predictors and distinct monsoon regions is studied considering different leads of the month. A ranked list of the monsoon predictors are selected and presented for all the four categories of monsoon in descending order of their correlation with the corresponding monsoon regions. These predictors are further utilized for the prediction of the *regional* monsoons.

4.1 *Architecture of the stacked autoencoder used for the identification of the monsoon predictor*

The stacked autoencoders are designed for deriving the potential monsoon predictors. The proposed approach starts with unsupervised learning (without any knowledge of the *regional* rainfall), so the architecture of autoencoder is same for all the categories of rainfall. Different stacked autoencoders are built considering different input climatic variables or their combinations.

The stacked autoencoder is created with a deepness of three layers. The first autoencoder of the deep network is fed with the features similar to climatic variables of the selected rectangular regions as mentioned in section 2. The

count of climatic features in the input (corresponding to the input neurons of the autoencoder in figure 3) to the number of features in internal layers (corresponding to the neurons of the internal layer of the autoencoder in figure 3) is taken as 15:1 for all the levels of the stacked autoencoder.

The features derived from the neurons in the internal layer denote complex features. These composite features can be the prototype for potential monsoon predictors. The internal (hidden) layer of the primary autoencoder is fed to the second architecture as input and the same pattern is ascertained for the third autoencoder of the deep structure.

The stacked autoencoders, namely, SA_SLP, SA_HGT and SA_UWND have architecture with their layers having number of neurons as [324 97 29 9 29 97 324]. The number of input features is 324, the internal layer of the primary autoencoder has 97 neurons, the second autoencoder has 29 neurons and finally, the third has 9 neurons. The rest architecture denotes the decoder component of the deep stacked autoencoder.

Besides these three architectures with individual climatic variable, there are two more stacked autoencoders, which are designed with the input as the combined features of SLP and HGT (SA_SLP_HGT); and combined features of SLP and UWND (SA_SLP_UWND). The stacked autoencoders SA_SLP_HGT and SA_SLP_UWND have their structures as [648 194 58 17 58 194 648], where each of climatic variables, namely, SLP, HGT and UWND have 324 input features.

4.2 *Identification of the monsoon predictors*

The identification of potential predictors for the *regional* Indian summer monsoon are performed by the initial unsupervised learning followed by supervised improvisation. The task is performed in the following steps, as shown in figure 4 are:

- Unsupervised feature learning by the stacked autoencoder;
- Thresholding for the feature extraction from the internal layers of the autoencoder;
- Supervised ranking of the predictors based on their correlation with the regional monsoon.

The first two steps are independent of any categories of rainfall, so they are same for all the *regional* monsoons. The final step is supervised and

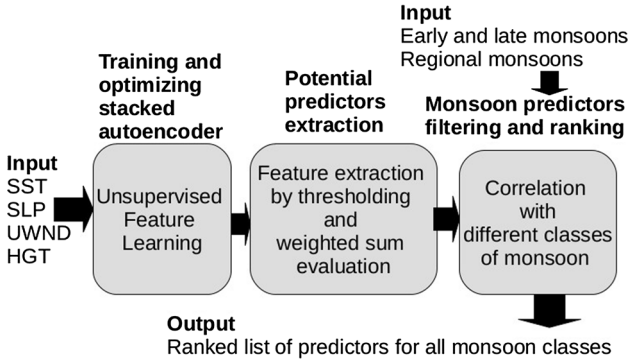

Figure 4. Identification of monsoon predictors for the regional monsoons.

the derived predictors from the stacked autoencoder are ranked distinctly for each of the four categories of rainfall, viz, central, north-east, north-west, and south-peninsular studying the correlation between the predictors and the respective monsoon categories.

4.2.1 *Unsupervised feature learning from the stacked autoencoder*

The unsupervised training of the stacked autoencoder results in the learning of new features in the deep hidden internal layers of architecture. Each autoencoder is pre-trained with the motivation of minimizing the reconstruction error of the input (the prime working principle of the autoencoder), followed by the final tuning of complete deep architecture, as described in section 3. The derived features from the internal layers denote the complex learned features, which correspond to the new monsoon predictors. The step is completely unsupervised without any consideration of monsoon values.

4.2.2 *Thresholding of the weights for feature extraction*

The stacked autoencoder is built to a deepness of three, which leads to learning of the property-endorsed features from the three hidden layers of network. Thresholding is performed over the weight matrix of all the hidden layers for acquiring the potential monsoon predictors from the complex learned features. The threshold is ascertained to engross all the input neurons that are influencial to the neuron in the internal layer and discard the effect of the remaining neurons.

The threshold is put by taking into consideration the weights which are larger than

two times of the standard deviation over the mean value of the weights. The threshold confirms the participation of not less than 15% of the input features in the building of new potential predictor. The weighted sum of the features being chosen after the thresholding denotes the potential climatic predictor.

Formally, if var_i denotes the input feature, $i = 1, \ldots, m$, and hid_j denotes the derived feature at internal layer, $j = 1, \ldots, h$, and the variable Wt_{ij} denotes the weight of input feature var_i for the derived feature hid_j. The potential monsoon predictor aligned with the derived feature hid_j is calculated using equation (4).

$$hid_j = \sum_{i=1}^{m} Wt_{ij} var_i, \quad \forall i, \quad | Wt_{ij} > threshold_j, \tag{4}$$

where $threshold_j$ denotes the chosen threshold value corresponding to the derived feature hid_j of a specified hidden layer.

4.2.3 *A supervised ranking of the monsoon predictors*

The predictors obtained from the learned features are ordered by consideration of their correlation with different regions of the monsoon, distinctly. The potential predictors are sorted in accordance with the magnitude of their correlation with the summer monsoon over the homogeneous regions.

The correlation with different categories of the monsoon are analyzed considering 1–12 months leads to find the best month with the highest correlation for potential predictor utilizing Pearson correlation (μ), as shown in equation (5). A lead of one month signifies the study of the correlation between the predictor in May and the *regional* monsoon (which initiates in June having total rainfall as an aggregate of June–September monsoon). In a similar way, three months lead signifies the predictor in March and 12 months lead means the predictor corresponding to June month of the preceding year correlating to the monsoon of the present year. The best lead month (one having highest μ value) of the identified monsoon predictors are considered for further evaluation. Thus, we obtain four ranked lists of the predictors for the four categories of monsoon (central, north-east, north-west, south-peninsular). These identified predictors are further used for the forecast of rainfall over the regions of India.

$$\mu = \frac{\sum_{i=1}^{n} \left(u_{mon}^{i} - \overline{u_{mon}}\right)\left(v_{mon}^{i} - \overline{v_{mon}}\right)}{\sqrt{\sum_{i=1}^{n} (u_{mon}^{i} - \overline{u_{mon}})^2} \sqrt{\sum_{i=1}^{n} (v_{mon}^{i} - \overline{v_{mon}})^2}}, \tag{5}$$

where u_{mon}^{i} and v_{mon}^{i} denote the regional India summer monsoon and climatic variables of the mon^{th} month at ith year, $\overline{u_{mon}}$ and $\overline{v_{mon}}$ are the corresponding mean for the mon^{th} month of the study period, and finally, the total count of years of study is denoted by n.

5. Ensemble of regression trees prediction model

The model used for forecasting the regional monsoon of India using the identified potential climatic predictors is the ensemble of regression trees (MATLAB 2012). The principle functioning involves melding of the outcomes of numerous weak learners into a single superior ensemble output. It assembles a number of trained models based on the regression tree and the input on which these models are learned (Loh 2008). The model forecasts the ensemble outcome by aggregating forecasts from the weak learner models.

The regression tree learners are trained utilizing the bagging algorithm. The ensemble size is selected empirically in a manner for keeping a balance between the accuracy and speed of the method. The ensemble size is ascertained as five for our case. A weighted average of forecasts from individual regression tree models are performed to provide the final output response from the ensemble model (equation 6).

$$\widehat{y}_{bag} = \frac{1}{\sum_{s=1}^{S} d_s P(s \in R)} \sum_{s=1}^{S} d_s \widehat{y}_s I(s \in R), \tag{6}$$

$\widehat{y}_s$ is the forecast by the tree s in the ensemble model, R denotes the collection of predictors of chosen trees that provide the forecast, $P(s \in R)$ is 1 when s is a member of the set R, or 0 in other case, d_s denotes the assigned weight of the sth tree, the weight of the tree is assigned according to the performance of the particular tree in the prediction over the validation set after training the model and S denotes the total count of trees in the ensemble.

6. Results and discussions

The prediction of rainfall in the regions of India during the south-west (SW) monsoon season is discussed. The proposed approach identifies the monsoon predictors for different regions of India for the summer monsoon period of the country. The acquired predictors are judged in terms of their efficiency in forecasting the different categories of monsoon.

The identified potential predictors are selected and ranked by a correlation study with the monsoon (described in section 4.2.3). For the prediction of the *regional* monsoon, four different ranked list of predictors are sorted in decreasing order of the correlation of potential predictors with each of the four categories of *regional* rainfalls. These ranked lists of monsoon predictors are made for all the monsoon categories distinctly for the individual features (SLP, UWND, and HGT) and for the combined features (SLP+UWND and SLP+HGT). For all the categories of monsoons (*regional*) and all the ranked list of predictors, the predictor sets are constructed considering the top correlated monsoon predictors of each climatic variable for predicting the seasonal rainfall on the regional scale (termed as *regional* monsoon).

The four regions of India – central, north-east, north-west and south-peninsular are analyzed in this section. The monsoon predictors obtained for *regional* rainfalls and their efficiency in the prediction of all four regions of India are explored in subsequent discussions.

6.1 *Geographical locations of the predictors of regional monsoon of India*

The spatial expansion of the identified monsoon predictors using sea level pressure for the central India regional rainfall is shown in figure 5. The evaluated monsoon predictors based on the geo-potential height for south-peninsular India are shown in figure 6, and for the north-west Indian monsoon, the predictors of u-wind are shown in figure 7. The new monsoon predictors are an amalgamation of climatic input variables from the spatial coverages at distant. Each single colour in the figure represents locations of climatic variable that are amalgamated to create a climatic predictor. Various colours in an area denote the indulgence of that area in formation of all those monsoon predictors, which are represented with different colours of the area. The location of the top five correlated predictors highlighted in the figure and they are presented in decreasing order of their correlation with regional monsoon category.

Different regions are combined with their corresponding contributing score weights to form the new potential monsoon predictor. These

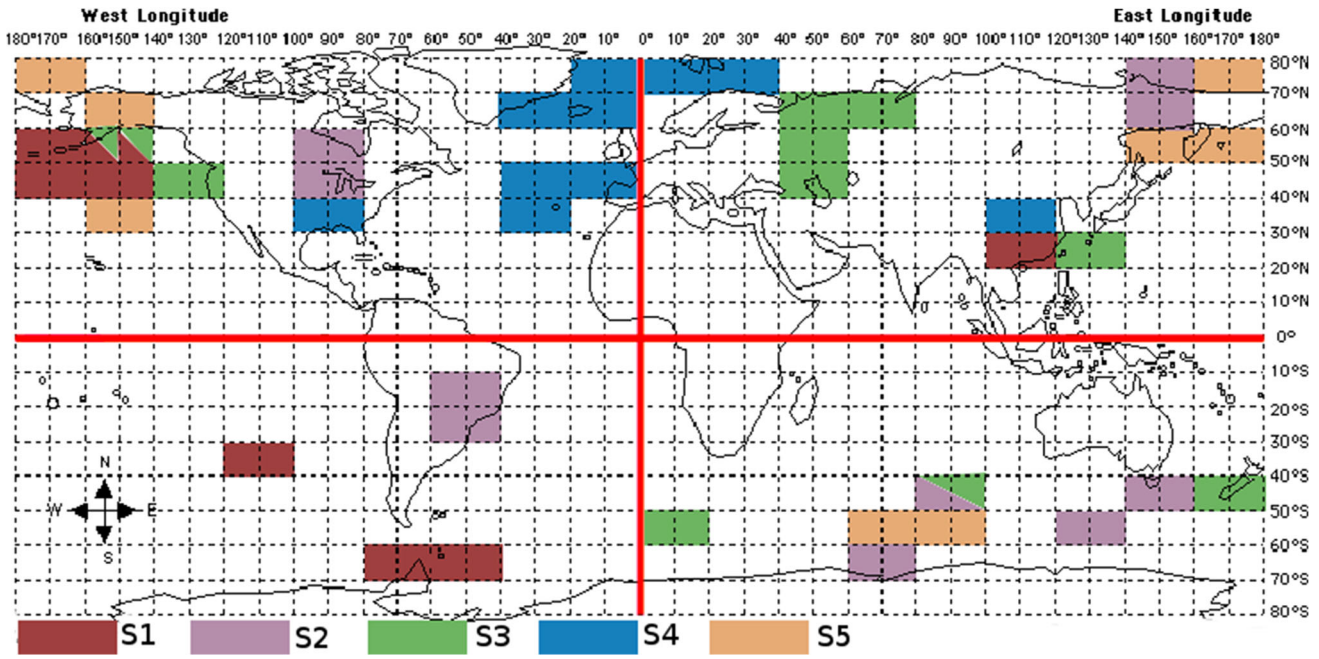

Figure 5. Spatial regions of the potential monsoon predictors of sea level pressure (S1–S5) for the central India summer monsoon.

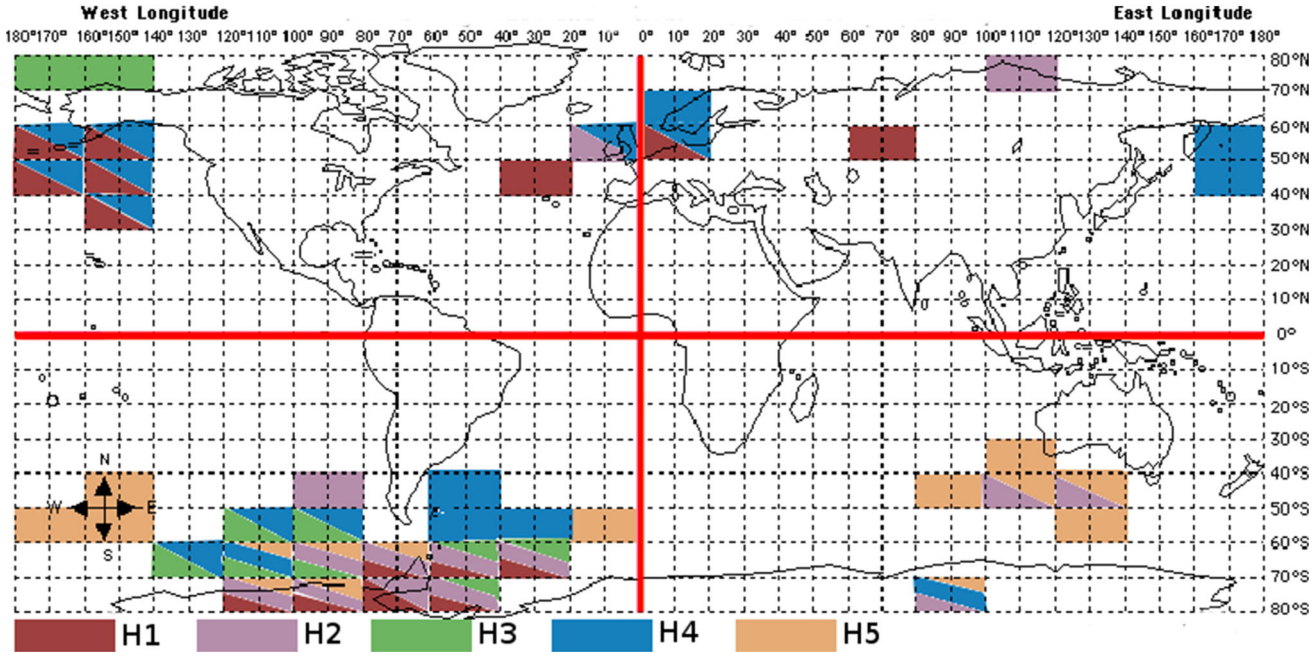

Figure 6. Spatial regions of the potential monsoon predictors of geo-potential height (H1–H5) for the south-peninsular Indian summer monsoon.

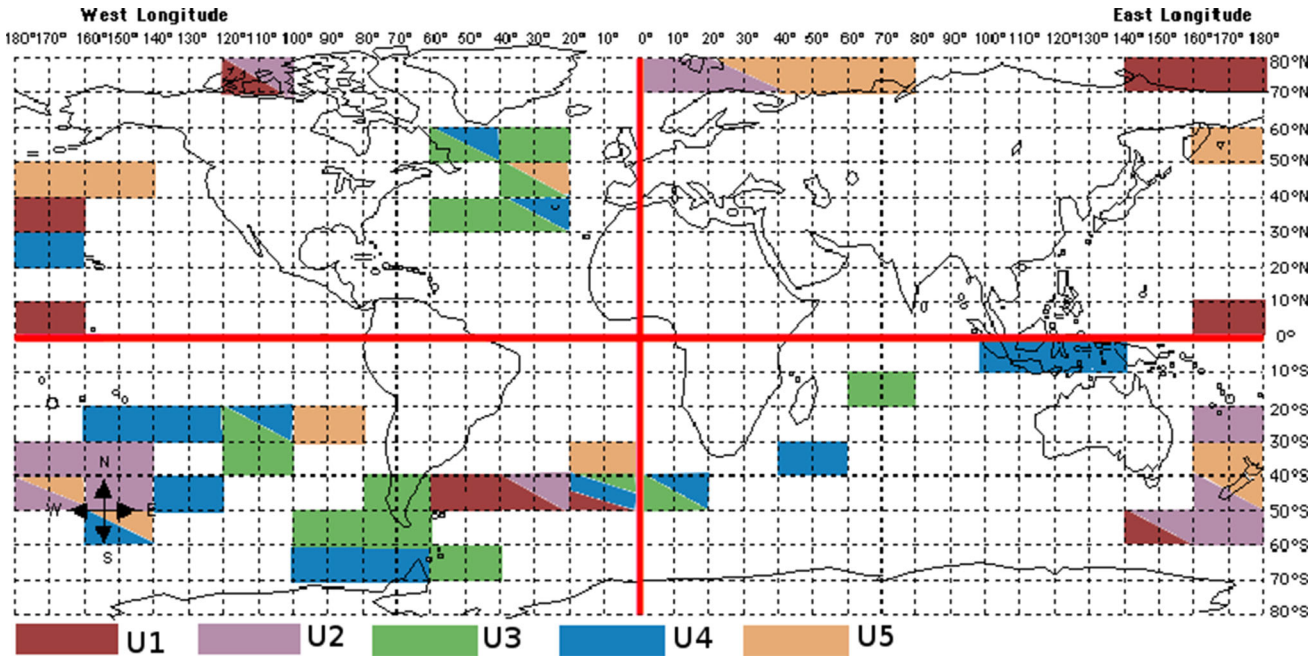

Figure 7. Spatial regions of the potential monsoon predictors of u-wind (U1–U5) for the north-west India summer monsoon.

weights are obtained from the stacked autoencoder, which combines the regions non-linearly using the tan-hyperbolic function and then the highly participating regions are selected following the thresholding method. Participation scores of the different regions in building up the new predictor of climatic variable sea level pressure for the central India region are shown in table 1.

Each score corresponding to the rectangular region of 20° longitude × 10° latitude as shown in figure 5 (e.g., there are eight such regions participating to form the first predictor pred1, shown by the brown colour in figure 5). The table also shows the correlation of the identified predictors with central India region monsoon and the corresponding correlated months. We notice that some

Table 1. *Identified potential monsoon predictors from SLP for the central India summer monsoon (−1 specifies predictors corresponding to the respective month of the previous year predicting the monsoon of the current year).*

Monsoon predictors	Weights of participating regions	Correlation values	Correlated months
Pred1	−0.19, −0.26,−0.31, −0.22, −0.24, −0.19, +0.20, −0.20	−0.37	August (−1)
Pred2	−0.28, −0.33, +0.26,−0.29, −0.29, −0.35, +0.27, +0.34, +0.32, +0.30, +0.36	−0.35	November (−1)
Pred3	+0.31,−0.25, −0.25, −0.23, −0.20, +0.20, −0.24, −0.30, +0.20, +0.38	−0.35	August (−1)
Pred4	−0.24, −0.20, −0.19,+0.19, −0.28, +0.26, +0.36, −0.26, −0.24, +0.27	−0.34	October (−1)
Pred5	−0.28, −0.27, −0.29, −0.22, −0.24, −0.25, −0.21, +0.23	−0.33	November (−1)

Table 2. *Mean absolute errors (%) for the forecast of the regional Indian summer monsoon with the identified predictors during the test period 2001–2014.*

Monsoon predictor variable	Central	North-east	North-west	South-peninsular
SLP	5.2	6.5	8.9	**6.4**
HGT	5.5	6.8	6.2	7.6
UWND	**4.1**	5.4	7.3	6.8
SLP+HGT	4.8	6.5	**5.5**	8.2
SLP+UWND	5.8	**5.1**	6.2	6.6

of the identified predictors have good correlations at long-lead times with regard to the SW monsoon season. We discuss the ability of these predictors in forecasting the monsoon on regional scales in the next section (at the present instant, we only have correlations, and high correlations do not guarantee accurate forecasts).

6.2 *Accuracies of the identified monsoon predictors for the four monsoonal regions*

The forecast of the Indian summer monsoon for the four homogeneous regions of India are presented in the term of mean absolute errors (equation 7) in forecasting monsoon during the test-period of 2001–2014. The period of 1948–2000 is considered for training the ensembled regression tree based prediction model.

The variation (std) of the central, north-east, north-west, and south-peninsular India rainfalls are 14%, 11%, 19%, and 15% of the long period average rainfall (LPA), respectively. The deviation of the *regional* Indian summer monsoons are greater than the annual Indian summer monsoon, which is 10% of the LPA, turning regional forecasts as highly challenging task. This increases the difficulties cumulatively in the prediction of the *regional* rainfalls.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \mid actual_i - predicted_i \mid}{n}, \qquad (7)$$

where $actual_i$ and $predicted_i$ denote the actual and predicted Indian monsoon for the ith year and the total count of test years is represented by n.

The mean absolute errors in the prediction of central, north-east, north-west, and south-peninsular Indian rainfalls by the predictor sets of distinct monsoon predictors of SLP, HGT, UWND and the combined predictors derived from SLP+HGT and SLP+UWND by the stacked autoencoders are shown in table 2.

The outcomes of the prediction of the *regional* monsoons are highlighted in the following discussions.

- Central India regional summer monsoon
  - The predictors derived from UWND perform superior to the other set of predictors in forecasting the central India monsoon with 4.1% mean absolute error at a lead of five months in January.
  - The other predictors from individual variable of SLP and HGT predict the monsoon with 5.2% and 5.5% errors in November of the previous year and April of the current year of rainfall, respectively.
  - The predictors introduced from the combined variables of SLP+HGT and SLP+UWND forecast monsoon with the errors of 4.8% and 5.8% in October of the previous year and April of the present year, respectively.
  - Strong correlation (μ of 0.91) is noted between the actual and predicted monsoon in January by UWND predictors for this region.
  - The errors are small as compared to the variation of 14% for the central India region monsoon.
- North-east India regional summer monsoon
  - The identified predictors of the combined variables of SLP+UWND forecast the monsoon of the north-east India region with a mean absolute error of 5.1% in March.
  - The predictors from individual variables of SLP, HGT, and UWND predict the monsoon of this homogeneous region with 6.5%, 6.8%, and 5.4% errors in the month of April, respectively.
  - Pearson correlation of 0.78 is observed for the predicted rainfall in March with actual rainfall of this region.
  - Prediction errors are considerable in contrast with the variation of 11% for this region.
- North-west India regional summer monsoon
  - A mean absolute error of 5.5% is acquired in forecasting the north-west India region monsoon by predictors of SLP+HGT at a lead of seven months in November.
  - The performance is superior in comparison with the high variation of 19% in the rainfall of this region.
  - The correlation between the predicted and actual rainfalls is 0.88 for the prediction presented in November.
  - The other identified predictors of SLP, HGT, UWND, and SLP+HGT forecast this regional rainfall with the errors of 8.9%, 6.2%, 7.3%, and 6.2% in the month of November, respectively.
- South-peninsular India regional summer monsoon
  - The monsoon of south peninsular India is forecasted with a mean absolute error of 6.4% in March by predictors obtained from SLP, in contrast to the variation of 15% in rainfall of this region.
  - Pearson correlation of 0.69 is noticed between the actual and predicted monsoon of the south-peninsular India region.
  - The predictors derived from the individual variable HGT, UWND predict monsoon with 7.6% and 6.8% errors at a long lead of 10 months in August of the previous year and in April of the current year, respectively.
  - Finally, the predictors from combined variables of SLP+HGT and SLP+UWND forecast the monsoon with 8.2% and 6.6% errors in February and April, respectively.
- The results suggest that the proposed method for identification of the new monsoon predictor provides a reasonable assessment of the subsequent regional monsoon much earlier than other existing techniques with promising accuracy.
- It is noted that the proposed model provides the forecast at a long lead times with a reasonable accuracy for the *regional* monsoons (e.g., the central India monsoon predicted in the month of (January) of the year, the north-west India monsoon in November of the previous year; and the north-east and south-peninsular can be predicted in March of the current year).
- The presented long-lead predictions are supportive for the higher authority to implement appropriate policies accordingly.

The variation of the predicted and actual rainfall from the *long period average* rainfall by the identified predictors for the central, north-east, north-west, and south-peninsular India are shown in figures 8, 9, 10, and 11 for the test-period 2001–2014, respectively.

The actual rainfall is shown with the bar and the predicted rainfall with different symbols for different climatic predictors in the figures. These figures show that the models are highly capable of forecasting the inter-annual variability of rainfall on the smaller regional scales. The predictions with the best accuracy are joined with a dotted line to highlight it in all the four cases.

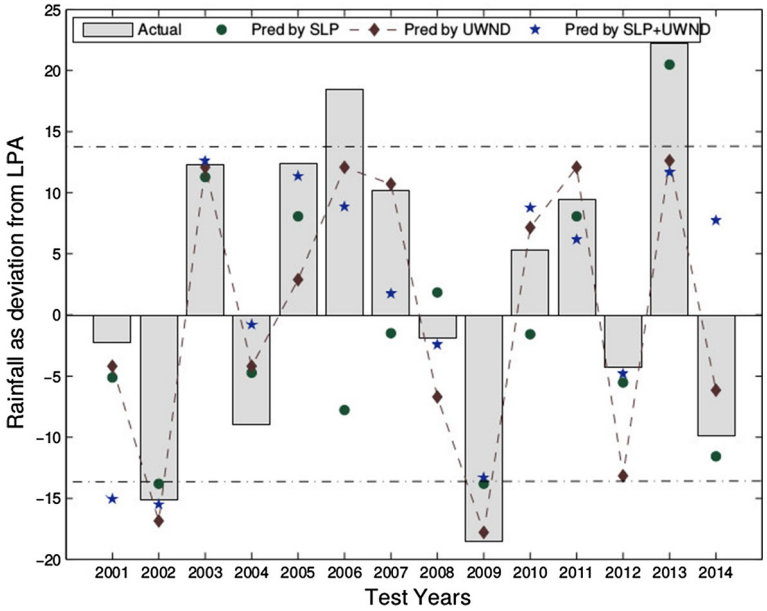

Figure 8. Forecast of the central India summer monsoon (June–September) by SLP, UWND and SLP+UWND during 2001–2014.

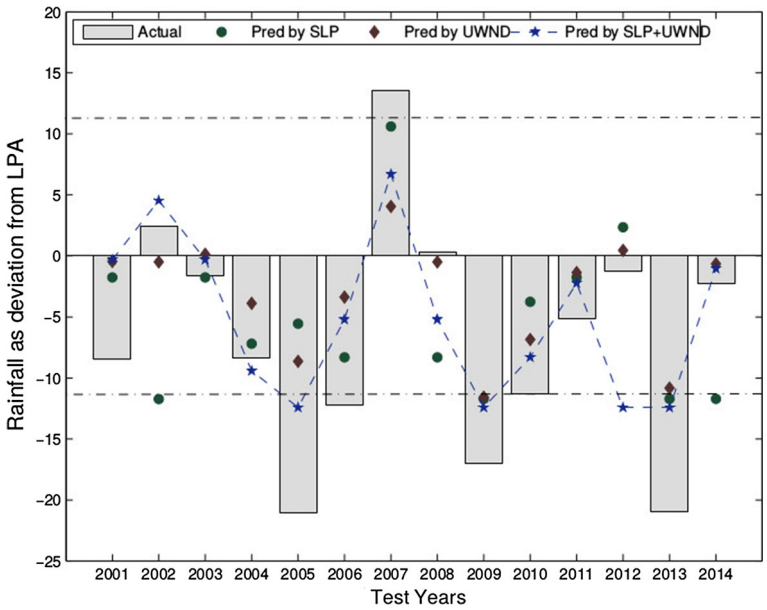

Figure 9. Forecast of the north-east India summer monsoon (June–September) by SLP, UWND and SLP+UWND during 2001–2014.

The figures show the performance of SLP, UWND, and SLP+UWND for the central and north-east India; and similarly, SLP, HGT, and SLP+HGT depict for the north-east and south-peninsular India.

For the central India regional rainfall, the trend of predicted rainfall completely follow the actual rainfall (figure 8). Even the extremes, including the droughts of 2002, 2009, and excess rainfalls during 2006, 2013 are also captured properly with the stacked autoencoder derived predictors for the central India monsoon.

Similarly, for the north-east India region, the predicted rainfall depicts the same sign of anomalies as the actual. The identified predictors correctly track the less rainfall years during 2005, 2009, and 2013, but it fails to capture the high rainfall of 2007 (figure 9).

The rainfall beyond the standard deviation during 2002, 2009, and 2014 are captured by the

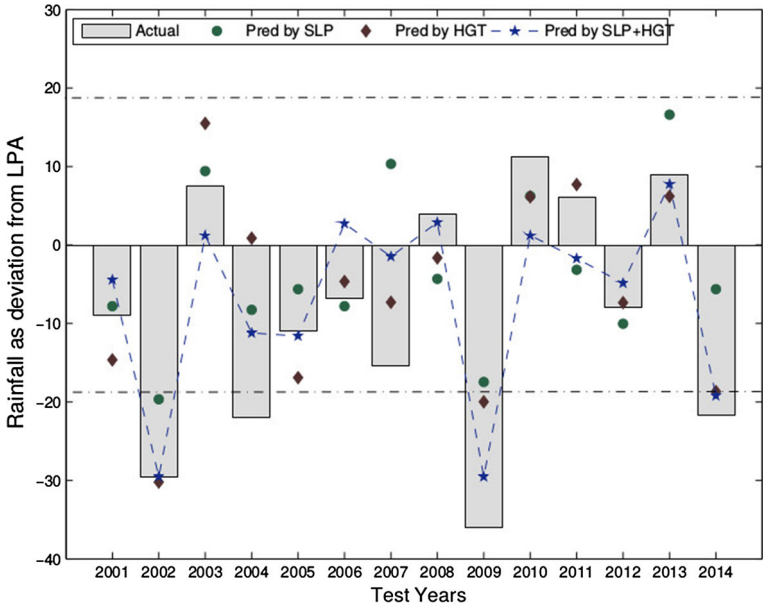

Figure 10. Forecast of the north-west India summer monsoon (June–September) by SLP, HGT and SLP+HGT during 2001–2014.

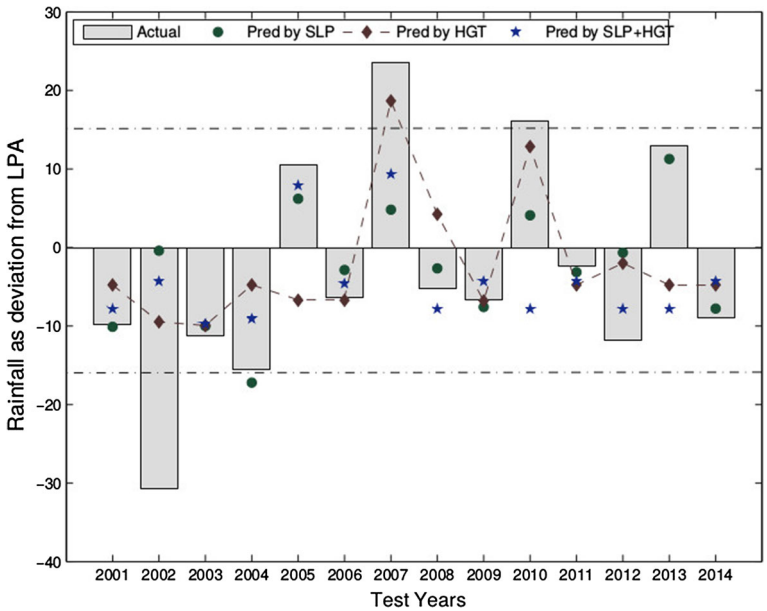

Figure 11. Forecast of the south-peninsular India summer monsoon (June–September) by SLP, HGT and SLP+HGT during 2001–2014.

identified predictors for the region of north-west India (figure 10).

Finally, for the monsoon of the south peninsular India region, excess rainfalls during 2007 and 2010 are detected properly, but the model fails to predict the low rainfall year of 2002 (figure 11).

Figure 12 shows the scatter plot of the predicted and actual monsoon by the identified monsoon predictors for the central India region. Mostly, plotted points are aligned at an angle of 45° from the horizontal, which indicates that the forecasted values are close to the observed rainfall. This highlights the efficiency of the identified predictors in forecasting the *regional* monsoon. The predictions by zonal wind velocity are seen to be more aligned to the 45° line from horizontal, which justify the high performance of the UWND predictors in predicting the *central* India monsoon.

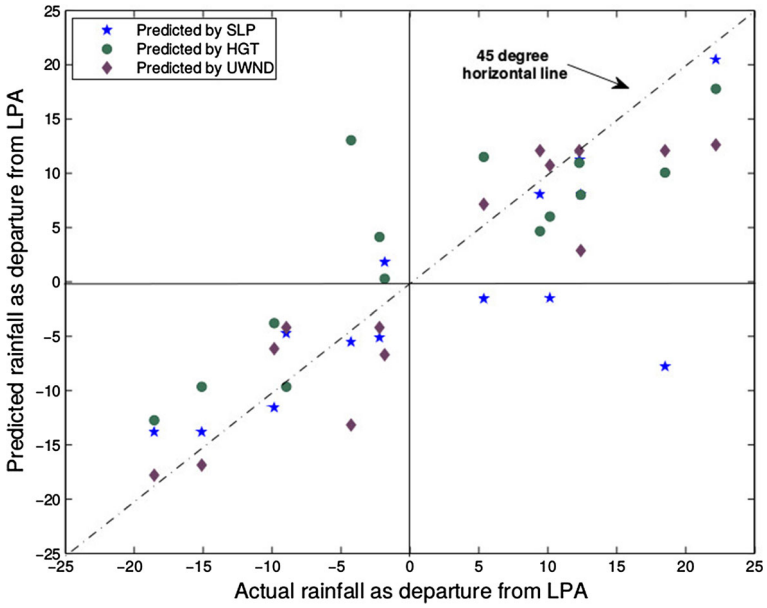

Figure 12. Actual against the predicted rainfall of the central India monsoon with the identified predictors of SLP, HGT, and UWND for the test-period 2001–2014.

6.3 *Comparison with the existing models*

India Meteorological Department started providing the prediction of the four regions of India since 2004. Forecasted regional monsoons by the identified predictors by learning through the stacked autoencoder are compared with the predictions of the IMD prediction model (Rajeevan *et al.* 2004, 2007) with the available IMD predictions (www.imd.gov.in) during the period 2004–2014. The comparison is performed with the IMD operational model because they are the only one which provides the prediction of the monsoon for the same regions as considered in our study. The predictions by our model and IMD model are shown in table 3.

The IMD model predicts the central India region rainfall with a mean absolute error of 12.2% whereas our proposed model have an error of 4.8%. Similarly, for the north-east, north-west, and south-peninsular India regions, our proposed model predicts with the errors of 5.4%, 6.1%, and 5.3% in comparison to the IMD model prediction with the errors of 7.8%, 9.6%, and 8.9%, respectively. The proposed model shows great promise as a forecasting tool.

6.4 *Predictions for the year 2016*

The regional Indian summer monsoon is predicted using the predictors derived from the stacked autoencoder-based approach for the year 2016. The central India monsoon is predicted by the identified predictors as 99% of the long period average rainfall. Similarly, the north-east and north-west India summer monsoons are forecasted as 106% and 88% of the LPA, respectively. Lastly, the south-peninsular India monsoon is predicted as 92% of the long period rainfall with our proposed method. India Meteorological Department reported the observed regional Indian monsoon rainfall of 2016 for the central, north-east, north-west, and south-peninsular India regions as 113%, 94%, 108%, and 113% of the LPA in its monsoon end-of-season report. Thus, our proposed approach forecasts the monsoon for the year with an error of 7% for the central and north-west India; and with no error for the south-peninsular India region. The predictors were unable to forecast the monsoon for the north-east region for this year, and they showed an error of 17% for the monsoon of this region. IMD model (Rajeevan *et al.* 2007) forecasts the monsoon of 2016 for the regional India as 113%, 94%, 108%, and 113% for the corresponding central, north-east, north-west, and south-peninsular India regions. IMD model forecasts the respective regional monsoon with the errors of 7%, 5%, 13%, and 21%, respectively.

Similar method utilizing the stacked-autoencoder (Saha *et al.* 2016b) with a different set of climatic variable predictors presented the forecast of the all-India summer monsoon for 2016 as 95.7% of the LPA, and IMD reported the observed rainfall for 2016 as 97% of the LPA.

Table 3. *Comparison between the prediction of the regional Indian summer monsoon by the proposed model and* IMD *model (Rajeevan et al. 2004, 2007) during the period 2004–2014 (by mean absolute errors (%)).*

Regional rainfall	IMD prediction error	Proposed model prediction error
Central India	12.2	**4.8**
North-east India	7.8	**5.4**
North-west India	9.6	**6.1**
South-peninsular India	8.9	**5.3**

7. Predictors identified for different homogenous regions of India

The monsoon predictors for all the four regions of India are identified utilizing a stacked autoencoder-based approach. We will discuss the predictors of monsoon based on climatic variables, namely, sea level pressure, geo-potential height, and u-wind, which are obtained with our proposed method in this section. The predictors are the amalgamation of variables from various geographical locations. There are four regional monsoons and three climatic variables, which require 12 figures to represent all the categories of monsoon for considered variables. We have presented three figures (figures 5, 6, and 7) out of 12 corresponding to three climatic variables (sea level pressure, geo-potential height, and u-wind).

The sea level pressure of regions of the Tibetan plateau, Bering Sea near Alaska and Southern Atlantic Ocean are combined to create a predictor influencing the summer monsoon of central India. The gradient of pressure between these regions may advect the monsoon winds towards the landmass (Das 1988). Another important predictor evaluates is the combination of sea level pressure of North America, South America (a part of Brazil and Argentina), and a region of the Southern Ocean below Australia. Other SLP-based predictors influencing the central India monsoon include the regions of North Atlantic Ocean (Hong *et al.* 2003), Norway, and north-western part of Russia. The sea-level pressure of the north-west Europe is another identified region, the same location was also considered by Rajeevan *et al.* (2007) as one of the predictor for the Indian summer monsoon. The north-east regional monsoon has a sea-level pressure based predictor as a combination of regions of north-east China and north-west Russia, and the Indian Ocean near Sumatra island. The pressure of discussed locations of east Asia is an important factor for prediction of the Indian monsoon (Rajeevan *et al.* 2004, 2007). The pressure differences between the mentioned regions assist the moisture-laid wind to flow over India. Other regions include Southern Pacific Ocean, Southern Atlantic Ocean, and Southern Ocean below Australia, which influence the monsoon of north-east India. The sea-level pressure of Southern Ocean below Australia was identified by Saha *et al.* (2016a) as an influencing factor of the Indian summer monsoon. A combination of sea level pressure of Brazil, the Bering Sea, and Southern Australia also forms a monsoon predictor. Southern Greenland and North Pacific Ocean also combine to produce a sea level pressure based predictor of the north-east monsoon. The North Pacific Ocean is known to be one of the influencing factor of the monsoon of Indian subcontinent (Cherchi and Navarra 2013).

The geo-potential height of north-western and central Russia above the Indian subcontinent and the region of Artic Ocean above the longitudinal stretch of Russia are ascertained as the predictors of north-west India. The same region of Artic Ocean over Russia was also identified by Kakade and Kulkarni (2016) for influencing north-west Indian monsoon. A combination of the regions of the North Pacific Ocean and South Atlantic Ocean is also evaluated as an important monsoon predictor. Regions of the South Pacific Ocean and North Pacific Ocean at south of Alaska is another important predictor for the north-west region of India. These are one of the newly-identified regions found to be influencing the regional Indian monsoon. The south-peninsular India is influenced by the geo-potential height of the amalgamation of locations of the South

Pacific and South Atlantic Ocean around the tail of South America. The North Pacific Ocean near Alaska and Finland–Sweden–Norway are noticed as other regions of interest. A region of central Russia with the same longitudinal stretch as India is also noted as an important predictor of the south-peninsular Indian monsoon (this region is same as identified by Kakade and Kulkarni 2016). The geo-potential height of the Southern Ocean surrounding the south-west and south Australia are also influencing monsoon predictors of south-peninsular India.

The u-wind of the North Atlantic Ocean (Hong *et al.* 2003), Equatorial Pacific Ocean (Gadgil *et al.* 2004), and Southern Ocean in the south of Africa are evaluated to be important for the central India monsoon. The u-wind over North Atlantic Ocean was evaluated as one of the important region influencing west-central Indian monsoon (Kakade and Kulkarni 2016). The u-wind flowing over the Artic Ocean above the north-west Russia and a region of Georgia and Uzbekistan are observed to be potential predictors of the monsoon. The regions over South Pacific and South Atlantic Ocean are noticed as influencing u-wind based predictors of the central India monsoon. The north-west Indian monsoon is influenced by the u-wind flowing over the Indian Ocean, regions of Indonesia and Malaysia, and a part of the Equatorial Pacific Ocean. Equatorial Pacific Ocean region corresponds to the locations of EQUINOO phenomenon, which is known to be important for the Indian monsoon (Gadgil *et al.* 2004). The other important regions include a combination of u-wind from the locations of the North and South Atlantic Ocean, the region of the South pacific Ocean, and North Pacific Ocean surrounding the Alaska region. These North Pacific Ocean surrounding the Alaska was identified as an influencing region for central north-east monsoon in literature (Kakade and Kulkarni 2016). The u-wind flowing in the south of the Madagascar region and a part of the Southern Ocean are also noticed as the monsoon predictors influencing the north-west region of India. The location of Madagascar is near the region of the Mascrene High which is known to be related to the Indian monsoon (Krishnamurti and Bhalme 1976). A more elaborate study and analysis of these regional monsoon predictors are required to be carried out to understand their role in regulating the regional monsoons of India.

8. Conclusions

Predicting the *regional* Indian summer monsoon is a challenging task. Here, we propose a deep neural network-based predictor identification method to improvise the accuracy of prediction of the regional monsoon. The approach comprises of the unsupervised feature learning from climatic variables followed by the threshold of weights, and finally, the supervised ranking and selection of the predictors based on the study of their correlations with the different categories of monsoons. The monsoon predictors are the amalgamation of input climatic variables from various geographical locations with the respective score of each location participating in the building of the predictor. The method helps in analyzing the climatic variables around the globe and subsequently, assists in identifying the new monsoon predictors which predict all the categories of the Indian summer monsoons with high accuracy.

We perform the prediction of the monsoon in spatial scale by predicting the four homogeneous regions of India. The predictions are provided at an appreciable lead which is encouraging and supportive for the government to impose appropriate strategies for the best economic growth of the country. An ensembled regression tree model with the identified monsoon predictors shows high skill in predicting all the categories of monsoon. The model predicts the *regional* rainfalls with the errors of 4.1%, 5.1%, 5.5%, and 6.4% for the central, north-east, north-west, and south-peninsular Indian regions, respectively. Despite of high variation of the *regional* rainfalls, the proposed approach with the identified predictors predicts the monsoon with high precision. The predictors are also successful in capturing the extremes of the monsoon. It needs to be highlighted that even at long-leads our proposed technique provides reasonable accuracy of the forecast and thus, seems promising as a forecasting tool.

The future scope of the work can comprise of the use of deep architecture as the convolution neural network which can be promising in extracting the complex predictors from its multiple layers of the small networks. The rectified linear units can increase the non-linear characteristics of the trained network assisting in building highly complicated and non-linear predictors of monsoon. Finally, the over-fitting of the network can be avoided using the dropout method, which speed-up

the process and assists in generalizing the network for evaluating more efficient predictors of the monsoon.

Acknowledgements

We gratefully acknowledge department of Computer Science and Engineering at Indian Institute of Technology Kharagpur, for providing all the supports for carrying out the work. We also deeply appreciate the support of Centre for Atmospheric and Oceanic Sciences at Indian Institute of Science Bangalore. Finally, we are thankful to our reviewers for their constructive suggestions.

References

Baldi P 2012 Autoencoders, unsupervised learning, and deep architectures; *ICML Unsupervised and Transfer Learning* **27** 37–50.

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

Das P K 1988 *The monsoons*; National Book Trust, India.

DelSole T and Shukla J 2012 Climate models produce skillful predictions of Indian summer monsoon rainfall; *Geophys. Res. Lett.* **39(9)** L09703.

Gadgil S, Vinayachandran P N, Francis P A and Gadgil S 2004 Extremes of the Indian summer monsoon rainfall, ENSO and equatorial Indian Ocean oscillation; *Geophys. Res. Lett.* **31(12)** L12213.

Gowariker V, Thapliyal V, Kulshrestha S M, Mandal G S, Sen Roy N and Sikka D R 1991 A power regression model for long range forecast of southwest monsoon rainfall over India; *Mausam* **42(2)** 125–130.

Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks; *Science* **313(5786)** 504–507.

Hong Y T, Hong B, Lin Q H, Zhu Y X, Shibata Y, Hirota M, Uchida M, Leng X T, Jiang H B and Xu H 2003 Correlation between Indian Ocean summer monsoon and North Atlantic climate during the Holocene; *Earth Planet. Sci. Lett.* **211(3)** 371–380.

Kakade S and Kulkarni A 2016 Prediction of summer monsoon rainfall over India and its homogeneous regions; **Meteorol. Appl. 23**(1) 1–13.

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 K C, Ropelewski C, Wang J, Jenne R and Joseph D 1996 The NCEP/NCAR 40-Year Reanalysis Project; *Bull. Am. Meteorol. Soc.* **77(3)** 437–471.

Krishnamurti T N and Bhalme H N 1976 Oscillations of a monsoon system. Part I. Observational aspects; *J. Atmos. Sci.* **33(10)** 1937–1954.

Loh W Y 2008 Classification and regression tree methods; *Encyclopedia of statistics in quality and reliability, Ruggeri*, (eds) Kenett and Faltin, pp. 315–323.

MATLAB 2012 Statistics and Machine Learning Toolbox; *MATLAB version 2012b, The MathWorks Inc., Natick, Massachusetts, US.*

Nair A, Mohanty U C and Acharya N 2013 Monthly prediction of rainfall over India and its homogeneous zones during monsoon season: A supervised principal component regression approach on general circulation model products; *Theor. Appl. Climatol.* **111(1–2)** 327–339.

Rajeevan M, Pai D S, Dikshit S K and Kelkar R R 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 D S, Kumar R A and Lal B 2007 New statistical models for long-range forecasting of southwest monsoon rainfall over India; *Climate Dyn.* **28(7–8)** 813–828.

Saha M, Mitra P and Nanjundiah R S 2016a Autoencoder-based identification of predictors of Indian monsoon; *Meteor. Atmos. Phys.* **128(5)** 613–628.

Saha M, Santara A, Mitra P, Chakraborty A, and Nanjundiah R S 2016b Stacked Autoencoder Based Identification of Monsoon Predictors for the Prediction of the Indian Summer Monsoon; *Theor. Appl. Climatol.* (**under review**).

Saha M, Mitra P and Nanjundiah R S 2016c Predictor Discovery for Early-late Indian Summer Monsoon Using Stacked Autoencoder; *Procedia Comp. Sc., ICCS* **80** 565–576.

Saha M and Mitra P 2016 Recurrent neural network based prediction of Indian summer monsoon using global climatic predictors; *IJCNN* 1523–1529.

Sinha P, Mohanty U C, Kar S C, Dash S K, Robertson A W and Tippett M K 2013 Seasonal prediction of the Indian summer monsoon rainfall using canonical correlation analysis of the NCMRWF global model products; *Int. J. Climatol.* **33(7)** 1601–1614.

Wang B, Xiang B, Li J, Webster P J, Rajeevan M N, Liu J and Ha K 2015 Rethinking Indian monsoon rainfall prediction in the context of recent global warming; *Nature Comm.* **6:7154**, doi:10.1038/ncomms8154.

Corresponding editor: Ashok Karumuri