Hindawi Publishing Corporation
Advances in Meteorology
Volume 2016, Article ID 9031625, 12 pages
http://dx.doi.org/10.1155/2016/9031625

Research Article

Predictor-Year Subspace Clustering Based Ensemble Prediction of Indian Summer Monsoon

Moumita Saha,[1] Arun Chakraborty,[2] and Pabitra Mitra[1]

[1]*Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India*
[2]*Center for Oceans, Rivers, Atmosphere and Land Sciences, Indian Institute of Technology, Kharagpur, India*

Correspondence should be addressed to Moumita Saha; moumita.saha2012@gmail.com

Received 5 July 2016; Revised 17 August 2016; Accepted 24 August 2016

Academic Editor: Anthony R. Lupo

Forecasting the Indian summer monsoon is a challenging task due to its complex and nonlinear behavior. A large number of global climatic variables with varying interaction patterns over years influence monsoon. Various statistical and neural prediction models have been proposed for forecasting monsoon, but many of them fail to capture variability over years. The skill of predictor variables of monsoon also evolves over time. In this article, we propose a joint-clustering of monsoon years and predictors for understanding and predicting the monsoon. This is achieved by subspace clustering algorithm. It groups the years based on prevailing global climatic condition using statistical clustering technique and subsequently for each such group it identifies significant climatic predictor variables which assist in better prediction. Prediction model is designed to frame individual cluster using random forest of regression tree. Prediction of aggregate and regional monsoon is attempted. Mean absolute error of 5.2% is obtained for forecasting aggregate Indian summer monsoon. Errors in predicting the regional monsoons are also comparable in comparison to the high variation of regional precipitation. Proposed joint-clustering based ensemble model is observed to be superior to existing monsoon prediction models and it also surpasses general nonclustering based prediction models.

1. Introduction

Monsoon is one of the significant climatic phenomena for agricultural country like India. Indian summer monsoon rainfall (ISMR) which accounts for more than 70% of country's annual rainfall occurs during period of June to September. It is the source of fresh-water supply, assists in generation of hydroelectricity power, and nourishes the flora-fauna of the subcontinent. The variation in quantity and distribution of monsoon over the country is high. These add up more difficulties in understanding the phenomenon and make monsoon prediction more challenging. The state and dynamics of monsoon are influenced by multiple climatic predictors. Climatic predictors correspond to different climatic variables like sea surface temperature, sea level pressure, wind velocity, and so forth over the world whose variations have an impact on the climatic event of monsoon.

India meteorological department (IMD) has been one of the pioneer organizations in predicting summer monsoon of the country from 1886. Blanford was the first to forecast monsoon, studying the influence of varying thickness of Himalayas' snows on North-West India [1]. Blanford's success in forecasts in tenure of 1882–85 encouraged him to start operational long range forecast of monsoon for India. Walker [2] introduced statistical correlations between rainfall and different climatic variables for predicting Indian monsoon. Power regression model developed by Gowariker et al. [3] forecasted ISMR with good accuracy for long-period, but it failed to predict the drought condition of 2002 [4]. The model was revised by IMD to build two statistical models by Rajeevan et al. [4] to predict ISMR in two stages. Later, Rajeevan et al. [5] introduced ensemble models to overcome the limitations of the reference IMD operational model and they were proved to be superior to past IMD models. Different statistical techniques are developed based on number of statistical methods upon outputs of global circulation models [6, 7].

India meteorological department continually reassesses different climatic predictors and updates the prediction models for issuing better forecast of Indian monsoon [3–5] in

a year. Thus, proper selection of predictor plays a significant role in forecasting monsoon of India. DelSole and Shukla [8] selected number of predictors comparing the error variances of models with different predictor sets after initial screening out models providing poor forecast accuracy. DelSole and Shukla [9] also showed that skill of predicting monsoon from sea surface temperature with coupled atmosphere-ocean models was statistically significant, attributed to the fact that slowly evolving sea surface temperatures were primary source of predictability.

In addition to high variability in quantity of Indian monsoon over years, it is also observed that the set of influencing predictor variables of monsoon evolve with time. In view of mentioned two characteristics of monsoon, we propose joint clustering of monsoon years and predictors for better modeling of the phenomenon and more accurate prediction. This is attempted by clustering the monsoon years into homogeneous groups, such that the variations of climatic predictor variables in a group are least and selecting separate set of predictors for each cluster of monsoon years. Subspace clustering which clusters the years according to different subsets of predictor climatic variables as well as homogeneity of monsoon year patterns is used for this approach. Each cluster of years is represented by different predictor climatic variables set. We design model for each group having different monsoon years and predictor set using random forest for prediction of monsoon rainfall. Finally, ensemble of forecasts by models built for all the clusters are considered to present the final prediction of monsoon of the country, which overcomes the fragility of single prediction model in forecasting.

Sun and Chen [10] proposed a statistical downscaling approach which improves the precipitation prediction by selecting the predictors that are best predicted by the coupled general circulation model and have the most stable relationships with precipitation in terms of stable correlation. They have selected optimal set of predictors for each grid and have proceeded with its rainfall prediction. Our proposed schema also selects the initial set of predictors based on correlation study, but the main highlight of our approach lies in unsupervised learning of monsoon phenomenon and its predictors. The proposed approach simultaneously evaluates clusters of monsoon years which are alike in terms of their characteristics and it finds out optimal set of predictors for the group. The approach is data-driven, robust, and time-efficient. The salient features of the proposed method are as the following: (i) clustering approach, which helps in designing different models for different groups of years engrossing variability of monsoon over years, (ii) subspace clustering, which aids in selecting different predictor sets for different clusters which is essential for evolving models for better precision over time, (iii) random forest of regression tree which is used for building the prediction model which is capable of framing nonlinearity present in monsoon process, (iv) ensemble forecast of monsoon which is provided which assists in overcoming the drawbacks of single model, and (v) the approach which is independent of other numerical models for input prediction.

The article is composed of five sections which are organized in the following manner. Section 2 discusses the input predictor climatic variables and their sources. Motivation and methodology of proposed joint-clustering based approach for predicting monsoon are presented in Section 3. Prediction model and ensemble techniques are elaborated in Section 4. An evaluation of the proposed approach to prediction of aggregate and regional Indian summer monsoon is presented in Section 5 and finally the article is concluded in Section 6.

2. Predictor Climatic Variables and Their Preprocessing

Prediction of Indian summer monsoon rainfall (ISMR), occurring during June–September with good accuracy, is the main motivation of the work. Monsoon will be predicted for aggregate India as well as for four main homogeneous regions of India as partitioned by India meteorological department considering the distribution of rainfall. The initial set of input climatic predictors influencing Indian summer monsoon is selected by relying on physical understanding of monsoon phenomenon, the wind pattern flow, and various literature studies [5, 11, 12]. We have considered some of the input climatic predictor variables from the predictors used by India Meteorological Department (IMD is one of the primary organizations involved in monsoon prediction of the country) as proposed by Rajeevan et al. [5]. All the predictors (eight out of nine) except warm water volume over Pacific (due to unavailability of data) are considered. Eight monsoon predictors with same lead month as considered by IMD models are selected as input predictor variables considering their good forecasting skills for the phenomenon of monsoon.

In addition, seven more predictor variables are considered which are linked or teleconnected with the monsoon and may prove themselves as potential predictors influencing the phenomenon. These categories of variables include East Asia SST which is selected looking at its high correlation with Indian monsoon [13]. Das [14] illustrates the physical event of monsoon and describes the dynamics behind Indian monsoon, the flow of monsoon winds, and the geographical features influencing the monsoon. The study aids to add new climatic predictors, namely, Madagascar SLP, surface pressure of Tibetan low, and pressure gradient between Madagascar and Tibetan regions, which advects the monsoonal winds toward landmass responsible for rainfall. Equatorial Pacific Ocean SLP is chosen as one of the influencing factors for Indian monsoon. El-niño occurring in Equatorial Pacific Ocean motivated us to study correlation of SLP of that region with Indian rainfall and a good correlation is observed [12]. Indonesia SST [11] and North Central Pacific Ocean SLP are other two predictors considered for the study. Thus, the initial set consists of *fifteen* predictors which are considered for the proposed joint clustering based approach to prediction of summer monsoon of the country.

Rainfall considered to be predicted is of aggregate and four homogeneous regions of India. Rainfall data are obtained from India Meteorology Department, Pune (http://www.imdpune.gov.in/), for period 1948–2014.

(i) Aggregate India monsoon has long period average (LPA) of 877.3 mm with standard deviation (std) of 10%.

(ii) Central India monsoon has LPA of 976.4 mm with std of 14%.

(iii) North East India monsoon has LPA of 1324.6 mm with std of 11%.

(iv) North West India monsoon has LPA of 618.7 mm with std of 19%.

(v) South Peninsular India monsoon has LPA of 730.5 mm with std of 15%.

It is observed that standard deviation in rainfall of *regional* is higher than that of aggregate Indian summer monsoon, which adds more challenge and necessities in the regional predictions along with aggregate India monsoon prediction.

The sources of the input climatic predictor variables are described as the following.

(i) Sea level pressure (SLP), surface pressure (SP), zonal wind (WV) at 100 hPa, and air temperature are acquired from NCEP reanalysis data (http://www.noaa.gov/) [15] at a resolution of 2.5° × 2.5°.

(ii) Sea surface temperature (SST) is collected from NOAA_OI_SST_V2 (http://www.noaa.gov/) [16] at spatial resolution of 2° × 2°.

(iii) Niño refers to sea surface temperature anomaly over spatial coverage of 5°S to 5°N and 170°W to 120°W in Pacific Ocean acquired from National Center for Atmospheric Research [17].

All the mentioned predictors data are collected at monthly scale during 1948–2014 for the study. Time period of 1948–1994 is considered for designing and training the prediction models from the outcome clusters obtained by the proposed joint clustering of monsoon years and predictors. The approach is evaluated in terms of prediction accuracy for Indian summer monsoon during test period 1995–2014.

Data preprocessing consists of evaluation of anomaly data by calculating the deviation of variable value from long period average of the variable exclusively for each month (shown in (1)), followed by correlation study between monsoon and the climatic variables using Pearson correlation coefficient (2) considering variables' lead of *zero* to *twelve* months to consider the predictor month having highest correlation with monsoon.

$$\text{anomaly\_data}_m^y = X_m^y - \text{mean}\left(X_m\right), \tag{1}$$

where X_m^y denotes the climatic variable for month m of yth year and mean (X_m) is the average of the variable values over all the years under study for month m.

$$\gamma = \frac{\sum_{i=1}^{n}\left(z_i - \overline{z}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(z_i - \overline{z}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}, \tag{2}$$

where z_i and y_i represent the Indian monsoon and predictor's values at ith year, $\overline{z}$ and $\overline{y}$ are their corresponding mean, and n denotes the total number of years. Figure 1 shows the geographic regions of initial set of predictor variables considered for the proposed approach. The predictor climatic variables with their locations and highest correlated month are shown in Table 1.

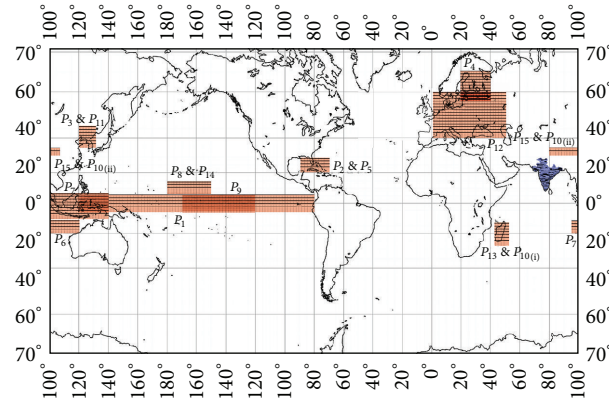

Figure 1: Geographical regions of climatic predictor variables influencing Indian monsoon (P_i represents climatic predictor variable i in Table 1).

3. Motivation and Methodology

Proposed joint clustering based approach for prediction of Indian summer monsoon is presented in this section. The schematic diagram of the approach is shown in Figure 2. The motivation behind the approach and details of the method for prediction of aggregate and regional monsoon of India are discussed.

3.1. Motivation. The motivation behind proposing joint predictor-year clustering based approach for forecasting summer monsoon of the country is twofold: (i) variation of rainfall quantity over years; (ii) prediction skills of predictors evolving with time.

Broadly monsoon years are categorized as drought (rainfall < −10% LPA), excess (rainfall > +10% LPA), or normal (−10% LPA ≤ rainfall ≤ +10% LPA). In addition, some monsoon years are associated with El-Niño or La-Niña event, which is anomalous warming or cooling of Equatorial Pacific Ocean, respectively. Some years are related to positive or negative Indian Ocean dipole event, which is event of greater warming of Western Indian Ocean than its Eastern part or vice versa. All these events influence Indian summer monsoon [12] and affect its distribution over years. Thus, it is beneficial to cluster the monsoon years into groups such that variation in quantity of rainfall within a cluster is less. In addition, it is advantageous because it is difficult for a single model to engross large variations present in monsoon phenomenon over years. We propose to develop separate models for every cluster which assists in better modeling of the phenomenon.

Furthermore, different sets of predictor variables influence the monsoon during different temporal spans. The phenomenon of monsoon is outcome of multiple physical

Table 1: Climatic predictor variables influencing Indian monsoon with geographic locations and correlated months (−1 represents previous year and 0 represents the same year).

Predictor variable	Predictor variable name	Location	Correlated months
P_1	Equatorial Pacific Ocean SLP anomaly	5°S–5°N, 120°E–80°W	Oct(−1) + Dec(−1)
P_2	North Atlantic Ocean SLP anomaly	20°N–30°N, 100°W–80°W	May(0)
P_3	East Asia SP anomaly	35°N–45°N, 120°E–130°E	Feb(0) + Mar(0)
P_4	North West Europe SP anomaly	55°N–65°N, 20°E–40°E	Oct(−1) + Dec(−1)
P_5	North Atlantic Ocean SST anomaly	20°N–30°N, 100°W–80°W	Dec(−1) + Jan(0)
P_6	Equatorial South Eastern Indian Ocean SST anomaly	20°S–10°S, 100°E–120°E	Feb(0) + Mar(0)
P_7	Indonesia SST anomaly	11°S–6°N, 95°E–142°E	April(0)
P_8	North Central Pacific zonal wind anomaly	5°N–15°N, 180°E–150°W	May(0)
P_9	Nino 3.4 SST anomaly	5°S–5°N, 170°W–120°W	MarAprMay(0) − DecJanFeb(−1)
P_{10}	Pressure gradient between Madagascar and Tibetan low	—	Feb(0) − Apr(0)
P_{11}	East Asia SST anomaly	35°N–45°N, 120°E–130°E	Dec(−1)
P_{12}	Europe land surface air temperature anomaly	40°N–60°N, 0°E–50°E	Nov(−1) + Feb(0)
P_{13}	Madagascar SLP anomaly	12°S–26°S, 43°E–51°E	Feb(0)
P_{14}	North Central Pacific Ocean SLP anomaly	5°N–15°N, 180°E–150°W	Oct(−1)
P_{15}	Tibetan low SP anomaly	29°N–35°N, 79°E–105°E	Apr(0)

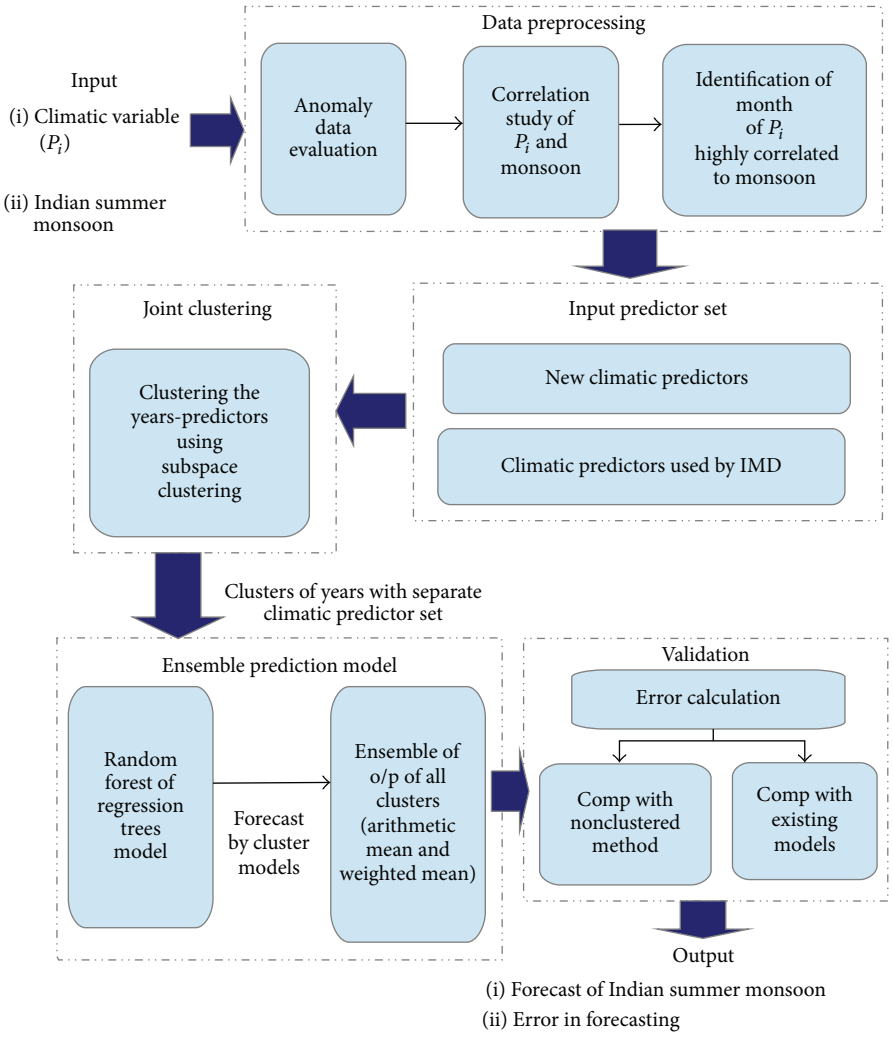

Figure 2: Proposed joint predictor-year clustering based approach to prediction of Indian summer monsoon rainfall.

climatic events and circulation of winds over the globe. Climatic predictor variables affecting Indian monsoon evolve with time to engross its complex behavior. As, for example, many literatures have elaborated on the weakening relationship between ENSO and Indian monsoon, ENSO had a huge impact on pattern of Indian monsoon rainfall in the past [18]. IMD has also rehashed its predictors for monsoon over time [3–5]. Thus, the predictor set needs to be upgraded to frame the nonlinearity of monsoon process over time. It is essential to select separate set of predictors for every cluster to design the prediction model with better precision.

These two objectives are fulfilled by proposed joint clustering of monsoon predictors and years for forecasting aggregate and regional Indian summer monsoon rainfall.

3.2. Joint Subspace Clustering of Monsoon Years and Predictors. Subspace clustering is used for our proposed approach of designing the prediction model for Indian monsoon. It seeks to find clusters in different subspaces within data [19]. Traditional clustering algorithms take all the dimensions to learn the pattern and cluster the data accordingly. However, in high-dimensional data, many attributes are irrelevant and they add negative effect to clustering. A subset of dimension is evaluated by subspace clustering algorithm to perform clustering by removing irrelevant dimensions.

Goal of joint clustering is bifold: firstly, it groups similar years in terms of climatic phenomenon, and finally, for each of these year groups, it simultaneously identifies a set of climatic predictor variables which will be efficient predictor set for those cluster years.

Projected clustering (PROCLUS) [20] method (a type of subspace clustering) is utilized to find a subset of climatic predictor variables for every cluster year. The clustering approach works in two steps: (i) locate the cluster centers and (ii) find appropriate set of dimensions in which each of the clusters exists.

This method extends the k-medoid algorithm by iteratively refining a full-space k-medoid clustering in top-down manner. It consists of three major steps, namely, initialization, iteration, and cluster refinement. The steps are described below.

(i) Initialization Phase. A set of potential medoids that are far apart are selected using a greedy algorithm. The purpose is to reduce set of points on which the hill-climbing method is to be executed.

(ii) Iteration Phase. Hill climbing approach is used to evaluate a good set of medoids. Random set of k medoids from the reduced dataset is chosen, which replaces bad medoids, and then the clustering quality is examined to check whether it has improved. Cluster quality is based on the average distance between instances and the nearest medoid. For each medoid, a set of dimensions are chosen whose average distances are smaller as compared to threshold value. After the subspaces have been selected for each medoid, the data instances are reassigned to their nearest medoids. Thus, subspace of possible cluster dimension and the instances to be assigned to corresponding clusters are evaluated.

(iii) Cluster Refinement Phase. The refinement phase is one pass over the data to improve the quality of clustering. It computes a new list of relevant dimensions for each medoid based on the clusters formed and reassigns points to medoids, removing outliers.

3.3. Cluster Based Ensemble Prediction of Monsoon. Prediction models are designed for every cluster obtained by subspace clustering. Each cluster has different monsoon years and different predictor variables. Prediction of monsoon is obtained from all the designed models for clusters. Ensemble techniques are used to combine the forecasts of different prediction models. Final prediction of Indian summer monsoon at aggregate and regional levels are provided as ensemble of forecasts obtained by models designed for every individual clusters.

3.4. Statistical Measures for Evaluating the Monsoon Prediction. Indian summer monsoon predictions at aggregate and regional levels are evaluated by different statistical measures. The monsoon predictions obtained by applying the proposed joint-clustering based approach are mainly evaluated in terms of mean absolute error in forecasting, which measures the mean value of deviation between forecast and actual precipitation. Additionally, the forecasts are also verified in terms of anomaly correlation coefficient, root mean square error, and prediction yields at different error categories to establish the efficacy of our proposed joint clustering based approach to prediction of monsoon. The measures are listed as following.

(i) Mean Absolute Error (MAE). Mean absolute error (MAE) represents the mean value of deviations between precipitation forecasts and actual rainfall values. It is defined by

$$\mathrm{MAE} = \frac{\sum_{t=1}^{n} \left| Y_i^t - X^t \right|}{n}, \tag{3}$$

where X^t and Y_i^t are the actual and predicted rainfall for tth test year by model for cluster i and n denotes the total number of test years.

(ii) Anomaly Correlation Coefficient (ACC). Anomaly correlation coefficient is the correlation between anomalies of forecasts and those of verifying values with the reference values, such as climatological values. In our case the climatological value is long period average (LPA) values of the rainfall. If the variation pattern of the anomalies of forecast is perfectly coincident with that of the anomalies of verifying value, ACC will take the maximum value of 1. On the other hand, if the variation pattern is completely reversed, ACC takes the minimum value of −1. ACC is defined in

$$\mathrm{ACC} = \frac{\sum_{t=1}^{n} \left(AX^t - \overline{AX} \right) \left(AY_i^t - \overline{AY} \right)}{\sqrt{\sum_{t=1}^{n} \left(AX^t - \overline{AX} \right)^2 \sum_{t=1}^{n} \left(AY_i^t - \overline{AY} \right)^2}}, \tag{4}$$

where AX^t and AY_i^t are the actual and predicted rainfall anomalies for tth test year by model built for cluster i; $\overline{AX}$

and $\overline{AY}$ are the mean of the anomalies; n is the total number of test years. The anomalies are defined as follows:

$$AX^t = X^t - L_{\text{rainfall}},$$
$$AY_i^t = Y_i^t - L_{\text{rainfall}}, \tag{5}$$

where X^t and Y_i^t represent the actual and forecast rainfall for tth test year by model built for cluster i and L_{rainfall} is the reference rainfall value which is calculated as long period average of rainfall values over time.

(iii) Root Mean Square Error (RMSE). Root mean square error calculates the differences between model predicted output and actual values of rainfall. They are a good measure to compare forecasting errors of various models. It is shown in

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} \left(Y_i^t - X^t\right)^2}{n}}. \tag{6}$$

(iv) Prediction Yield (PY). Prediction yields are evaluated at two different error categories (5% and 10% errors) to assess the overall prediction results by judging percent of predicted years within each allowed range of errors.

4. Monsoon Prediction Model and Ensemble Techniques

Prediction model designed for the clusters for forecasting monsoon and the ensemble techniques used to provide the final forecast of Indian monsoon are described.

4.1. Prediction Model Using Random Forest of Regression Trees. The prediction model used for framing the individual cluster obtained by subspace clustering is tree bagger [21]. It is the implementation of random forest of regression tree. Random forest bags an ensemble of decision trees for regression modeling [22]. Bagging stands for bootstrap aggregation. Every tree in the ensemble is grown on an independently drawn bootstrap replica of input data. Resampling is performed 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. To compute prediction of an ensemble of trees for new data instant, it takes a weighted average of predictions from individual trees:

$$\widehat{y}_{\text{bag}} = \frac{1}{\sum_{t=1}^{T} a_t I \ (t \in S)} \sum_{t=1}^{T} a_t \widehat{y}_t I \quad (t \in S), \tag{7}$$

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, and T is the total number of trees in the random forest.

We have used bagging algorithm for training the random forest of regression tree learners. This algorithm is chosen owing to reasons: (i) it uses bagging, a bootstrap aggregating technique for improving estimation, (ii) bagging aids in improving the predictive performance of underlying regression tree, (iii) tree ensembles deal with nonlinear features, and (iv) they can handle high dimensional data spaces as well as large number of training instances.

4.2. Ensemble Techniques. Three different weighted ensemble methods are adopted to present forecast of Indian summer monsoon from forecasts provided by individual model designed for different individual clusters.

(i) Simple Arithmetic Mean (Ensm Model 1). Equal weight is given to all the model's prediction designed for each cluster to evaluate the final forecast. It is shown in

$$\text{Predfinal}^t = \frac{1}{c} \sum_{i=1}^{c} P_i^t, \tag{8}$$

where Predfinal^t is the ensemble forecast presented for tth test year, P_i^t is the prediction for tth year given by model built for ith cluster, and c is the total number of clusters.

(ii) Weighed Ensemble in Linear Order (Ensm Model 2). Final forecast is given by weighted ensemble of the forecast by each cluster model, as shown in (9). The weight assigned to a cluster's model is inversely proportional to the distance of test year from respective cluster's center.

$$\text{Predfinal}^t = \sum_{i=1}^{c} W_i^t * P_i^t, \tag{9}$$

where

$$W_i^t = \left[\frac{1}{d_i^t}\right], \tag{10}$$

W_i^t is the weight given to the prediction of model (P_i^t) for tth test year designed for cluster i, and d_i^t is the euclidean distance of tth year from center of cluster i.

(iii) Weighted Ensemble in Quadratic Order (Ensm Model 3). Weight is calculated as exponential function of distance between test year and clusters' center. Forecast is given as weighted ensemble of forecasts using

$$W_i^t = \exp\left[\frac{1}{d_i^t}\right]. \tag{11}$$

5. Experimental Results and Discussions

Proposed joint clustering of monsoon years and predictors for designing ensemble prediction model is evaluated in terms of its efficiency in predicting Indian summer monsoon. Rainfall is considered as percentage of long period average value of rainfall. Monsoon is predicted at aggregate and regional divisions of India. India meteorological department has segregated Indian landmass into four homogeneous regions, namely, *Central, North East, North West*, and *South*

Table 2: Predictor climatic variables and count of years corresponding to clusters by subspace clustering. Details of climatic variables are described in Table 1.

Result set	Cluster 1		Cluster 2		Cluster 3	
	Number of years	Climatic variables	Number of years	Climatic variables	Number of years	Climatic variables
ResSet 1	26	P_7, P_8, P_{14}	5	$P_5, P_6, P_7, P_{14}, P_{15}$	11	P_1, P_5, P_7, P_{14}
ResSet 2	14	P_1, P_5, P_7, P_{14}	18	$P_2, P_5, P_6, P_7, P_{11}, P_{14}$	12	$P_1, P_7, P_8, P_{13}, P_{14}$
ResSet 3	26	$P_1, P_2, P_7, P_{13}, P_{14}$	10	$P_2, P_6, P_7, P_8, P_{10}, P_{11}, P_{14}$	10	$P_1, P_5, P_6, P_7, P_{12}, P_{14}$

Table 3: Mean absolute errors (%) for Indian summer monsoon prediction by models designed for individual cluster and ensemble models for test period 1995–2014.

Result set	Cluster 1	Cluster 2	Cluster 3	Ensm Model 1	Ensm Model 2	Ensm Model 3
ResSet 1	5.9	6.3	5.9	5.8	**5.7**	**5.7**
ResSet 2	6.7	5.7	5.7	5.5	**5.4**	**5.4**
ResSet 3	6.1	5.5	7.5	5.6	5.3	**5.2**

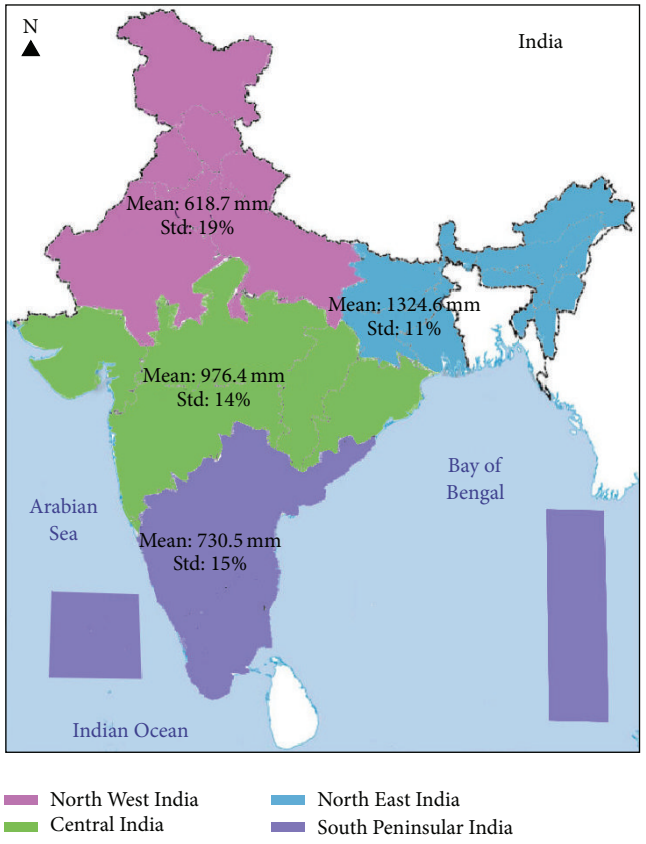

Figure 3: Four homogeneous regions of India as defined by India Meteorological Department.

Peninsular India depending upon the distribution of monsoon over the country. The homogeneous regions are shown in Figure 3.

5.1. Joint Subspace Clustering of Predictors and Years. Joint clustering is performed over monsoon years and predictors during training period 1948–1994. Proposed subspace clustering method (described in Section 3.2) groups the data into three clusters, considering best cluster quality. The algorithm groups the time series of monsoon predictors over year windows on different predictor parameter subspaces.

Each cluster is associated with optimal monsoon predictor variables, which are further used for building the prediction model. If some cluster lacks samples for training, we ignore that particular cluster from modeling and ensemble forecast given by other cluster models to predict Indian summer monsoon.

Table 2 shows the predictor variables for clusters along with number of years in the cluster as obtained by PROCLUS subspace clustering method. ResSet 1, ResSet 2, and ResSet 3 are obtained by subspace clustering specifying number of clusters as *three* and average dimension of parameter subspace as *four*, *five*, and *six*, respectively.

5.2. Prediction of Aggregate Indian Summer Monsoon. Indian summer monsoon is predicted using *random forest of regression tree* model designed for each cluster during test period 1995–2014. We study the prediction by models designed for all three individual clusters and finally combination of forecasts of all models designed from cluster using three weighted ensemble methods described in Section 4.2. Predictions are chiefly evaluated in terms of mean absolute errors in prediction (described in Section 3.4).

Mean absolute errors in prediction of aggregate Indian summer monsoon by models designed for clusters and ensemble of those predictions are presented in Table 3. The model reports mean absolute error of 5.7% for ResSet 1 by Ensm Model 2 and Ensm Model 3, which is superior to all individual clusters' model forecasts. ResSet 2 provides mean absolute error of 5.4% by Ensm Model 2 and Ensm Model 3; and ResSet 3 predicts aggregate monsoon with mean absolute error of 5.2%. Ensemble models for all *three* results sets predict monsoon at lead of *one* month in *May*, as predictors for cluster have at least lead of *one* month (Tables 1 and 2). Predictions attained are comparably better for forecasting complex phenomenon of monsoon. The results also show improvement of the ensemble approach over single model prediction.

Table 4: Forecast verification statistics for ensemble models during test period 1995–2014.

ResSet	Verification measures	Ensm Model 1	Ensm Model 2	Ensm Model 3
ResSet 1	ACC between actual and predicted rainfall anomalies	0.47	0.48	0.46
	RMSE for forecast (%)	7.6	7.5	7.4
	PY (%) at allowed error 5%	60	65	60
	PY (%) at allowed error 10%	85	85	85
ResSet 2	ACC between actual and predicted rainfall anomalies	0.50	0.51	0.52
	RMSE for forecast (%)	7.2	7.1	7.1
	PY (%) at allowed error 5%	55	55	55
	PY (%) at allowed error 10%	85	75	75
ResSet 3	ACC between actual and predicted rainfall anomalies	0.50	0.59	0.58
	RMSE for forecast (%)	8.3	7.8	7.8
	PY (%) at allowed error 5%	65	65	65
	PY (%) at allowed error 10%	75	75	75

Table 5: Mean absolute errors (%) for regional Indian summer monsoon by models designed for individual cluster and ensemble models for test period 1995–2014.

Regions	Cluster 1	Cluster 2	Cluster 3	Ensm Model 1	Ensm Model 2	Ensm Model 3
Central	9.4	7.6	10.2	8.2	7.5	**7.2**
North East	8.0	7.7	8.8	8.0	7.9	**7.8**
North West	12.3	10.6	15.3	11.5	12.0	**11.2**
South Peninsular	10.2	11.3	10.3	10.6	10.2	**9.8**

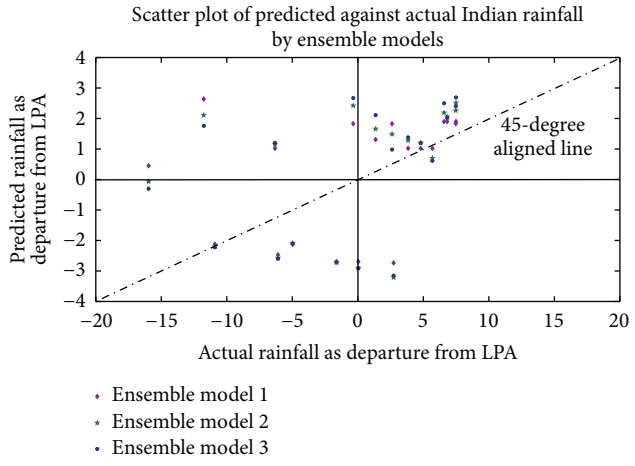

Figure 4: Scatter plot of actual against predicted rainfall by ensemble models during 1995–2014.

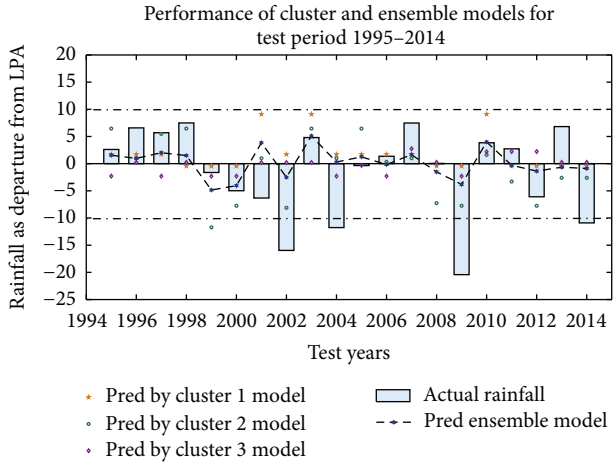

Figure 5: Forecasts of Indian summer monsoon by model designed for clusters and ensemble forecast during 1995–2014.

The other defined statistical measures as discussed in Section 3.4, for verifying the predicted rainfall by the ensemble models during test period 1995–2014, are shown in Table 4. Figure 4 shows the scatter plot of actual against predicted rainfall by ensemble models for ResSet 1. It is observed that most of the points lie in first or third quadrant which symbolizes that predicted rainfall has departure in same direction (positive or negative) from LPA as the actual rainfall. It is also noted that mostly the points are aligned to 45° *line* horizontally, which symbolize that the predicted rainfall is close to actual rainfall magnitudes.

Figure 5 shows actual and predicted rainfalls as departure from LPA by models designed for each cluster and ensemble forecast for ResSet 2 during test period 1995–2014. Actual rainfall is shown by bar and different symbols show the predicted rainfall by prediction models for cluster and ensemble prediction. The predicted rainfall follows the trend of actual rainfall. Moreover, the predicted rainfall shows same sign of anomaly (positive or negative from LPA rainfall) as actual for most of the test years. The model is unable to capture the extreme years properly. Perhaps, more physical knowledge about the phenomenon and other important

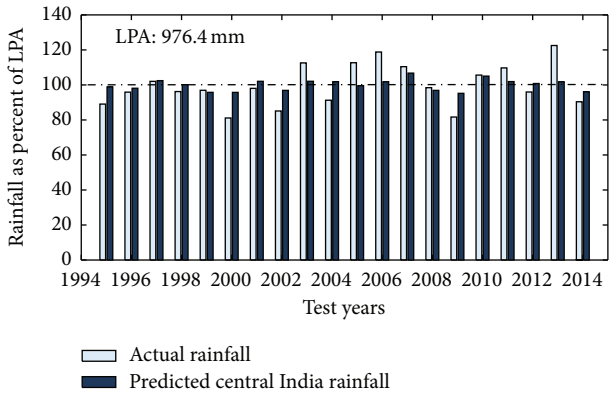

Figure 6: Forecast of Central India summer monsoon during 1995–2014.

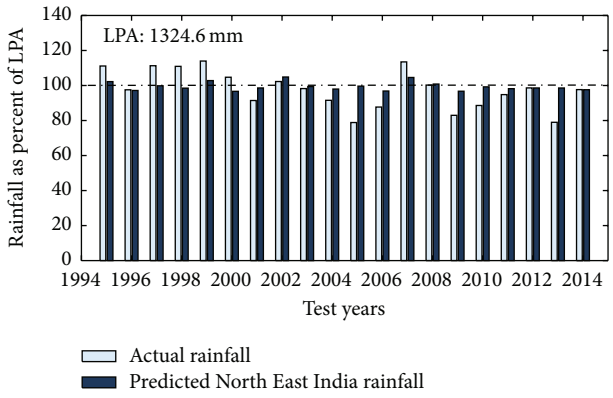

Figure 7: Forecast of North East India summer monsoon during 1995–2014.

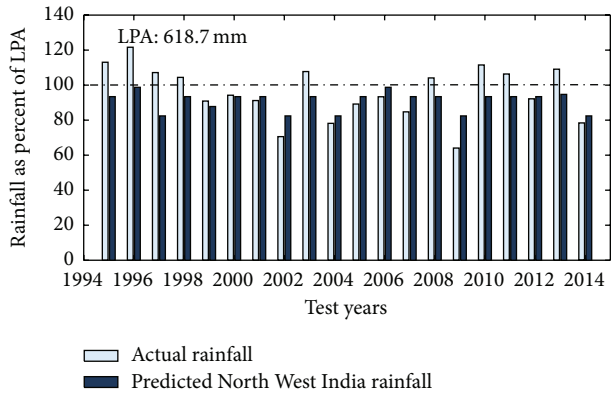

Figure 8: Forecast of North West India summer monsoon during 1995–2014.

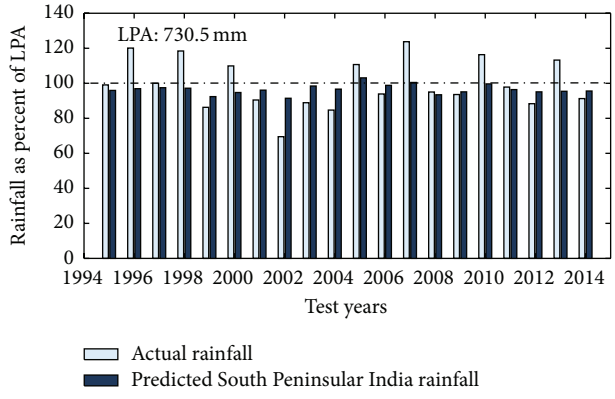

Figure 9: Forecast of South Peninsular India summer monsoon during 1995–2014.

predictors should be identified and used for even better prediction and capture of the extreme monsoon years.

5.3. Prediction of Regional India Summer Monsoon Rainfall. Predictions of Indian summer monsoon for four homogeneous regions are presented in term of mean absolute errors during test period of 1995–2014 with training of models for period 1948–1994 (3).

Standard deviation (std) of *Central*, *North East*, *North West*, and *South Peninsular* India rainfall is 14%, 11%, 19%, and 15% of long period average rainfall (LPA), respectively. Deviations of *regional* Indian summer monsoons are greater than aggregate Indian summer monsoon rainfall, which is 10% of LPA, making their forecasts more challenging. This increases the difficulties cumulatively in prediction of *regional* rainfalls.

Mean absolute errors in prediction of *Central*, *North East*, *North West*, and *South Peninsular* Indian rainfalls by the proposed joint-clustering based approach are shown in Table 5. We have provided the regional prediction with highest accuracy among the three result sets (described in Section 5.1).

The precipitation of *central* India region is predicted with mean absolute error of 7.2% owing to variation of 14% in rainfall of this region. *North East* India monsoon is forecasted with 7.8% error. Finally, *North West* and *South Peninsular* India rainfall is predicted with errors of 11.2% and 9.8%, respectively. These errors are comparable considering high variation of 19% and 15% of long period average rainfall at these regions.

Predicted regional rainfall is plotted against actual rainfall, where rainfalls are expressed as percentage of long period values. The plots are shown for *Central, North East, North West*, and *South Peninsular* India in Figures 6, 7, 8, and 9, respectively. It is noticed that the predicted rainfall follows the trends of the actual rainfall and does not report abnormally high error for any of the test years.

Regional predictions are also evaluated by the statistical measures, namely, anomaly correlation coefficient, root mean square error, and prediction yields (described in Section 3.4). These measures are presented for monsoon prediction of all four regions of India in Table 6. Anomaly correlation coefficient of 0.66 is observed between actual and forecasted precipitation of central India region which symbolizes good prediction of the region. *North West* India is predicted with least accuracy and it may be required to incorporate some local variables like humidity content and local wind velocity to predict monsoon of the region with better accuracy.

Table 6: Regional monsoon forecast verification statistics for ensemble models during test period 1995–2014.

Regions	Verification measures	Ensm Model 1	Ensm Model 2	Ensm Model 3
Central	ACC between actual and predicted rainfall anomalies	0.57	0.64	0.66
	RMSE for forecast (%)	10.2	9.8	9.7
	PY (%) at allowed error 5%	45	55	55
	PY (%) at allowed error 10%	75	75	75
North East	ACC between actual and predicted rainfall anomalies	0.45	0.50	0.53
	RMSE for forecast (%)	9.9	9.6	9.4
	PY (%) at allowed error 5%	40	55	55
	PY (%) at allowed error 10%	65	75	75
North West	ACC between actual and predicted rainfall anomalies	0.47	0.46	0.43
	RMSE for forecast (%)	13.6	13.3	13.4
	PY (%) at allowed error 5%	35	45	40
	PY (%) at allowed error 10%	60	65	60
South Peninsular	ACC between actual and predicted rainfall anomalies	0.54	0.58	0.59
	RMSE for forecast (%)	11.3	10.8	11.4
	PY (%) at allowed error 5%	45	55	50
	PY (%) at allowed error 10%	70	75	75

5.4. Comparison of Proposed Clustering Based Ensemble Model with Nonclustered General Approach. Proposed clustering based ensemble models are compared with nonclustered, nonensemble model built with all fifteen predictor climatic variables influencing monsoon and all years of monsoon under study.

Nonclustered approach provides mean absolute error of 7.3% for predicting aggregate Indian summer monsoon as compared to 5.2% error by proposed clustering based during test period 1995–2014. Monsoon of regions of India, namely, *Central, North East, North West*, and *South Peninsular,* is forecasted with errors of 10.1%, 9.3%, 13.2%, and 11.4, respectively, by nonclustered approach. Proposed joint-clustering based predictions are observed to be superior to nonclustering approach for all four homogeneous regions.

Anomaly correlation coefficient of 0.59 is observed between actual and predicted aggregate India rainfall by clustering method, while that for nonclustered approach is 0.46. Predicted regional monsoon by nonclustering approach shows poor anomaly correlation coefficient of 0.23, 0.31, 0.18, and 0.15, respectively, with the actual precipitation of the regions.

Root mean square error for aggregate monsoon prediction by nonclustered approach is 9.2%, while proposed subspace clustering based ensemble model outperforms nonclustered approach with 7.1% RMSE. Root mean square errors for regional monsoon are depicted as 13.6%, 11.9%, 15.9%, and 14.8%, respectively, by nonclustering approach. The performance of nonclustering approach is inferior to that of proposed clustering based approach for both aggregate and regional India monsoon.

Finally, prediction yields at 5% and 10% error categories by proposed ensemble model are 65% and 85% for aggregate India monsoon prediction, while those for nonclustered method are 40% and 60% only. All the results clearly depict the improvement in prediction by proposed joint-clustering based ensemble approach over nonclustered method of prediction.

5.5. Comparison of Proposed Clustering Based Ensemble Model against Existing Models. Predictions by the clustering based approach are compared with forecast by India Meteorological Department (IMD) models [4, 5]. It is compared with existing IMD operational power regression model [4] and pursuit projection regression (PPR) model [5]. We have compared prediction given by our approach during 2003–2014 with available IMD model's prediction. India meteorological department's operational model provides mean absolute error of 7.5% in predicting aggregate Indian monsoon. Additionally, currently running PPR model of IMD gives prediction in two phases: first in April (LRF1) and next in June (LRF2). LRF1 and LRF2 predict aggregate India monsoon with mean absolute errors of 7.1% and 6.5%, respectively. Proposed joint-clustering based ensemble models give mean absolute errors of 5.6%, 5.3%, and 5.2%, in the month of *May*. Comparison of prediction by proposed approach and IMD present models is shown in Figure 10.

The regional predictions by the proposed model are compared with predictions by IMD model [4, 5] during period 2004–2014. Predictions are shown in Table 7. IMD model predicts *Central* India region rainfall with mean absolute error of 12.2% whereas our proposed model has error of 8.3%. Similarly, for *North East*, *North West*, and *South Peninsular* India regions, our proposed model predicts with errors of 7.2%, 9.1%, and 7.6% in comparison to IMD model prediction with errors of 7.8%, 9.6%, and 8.9%, respectively. Proposed approach outperforms IMD model in prediction of monsoon for all four homogeneous regions of India. Thus, it can be concluded that prediction models with our proposed clustering based approach outperform all the IMD models [4, 5].

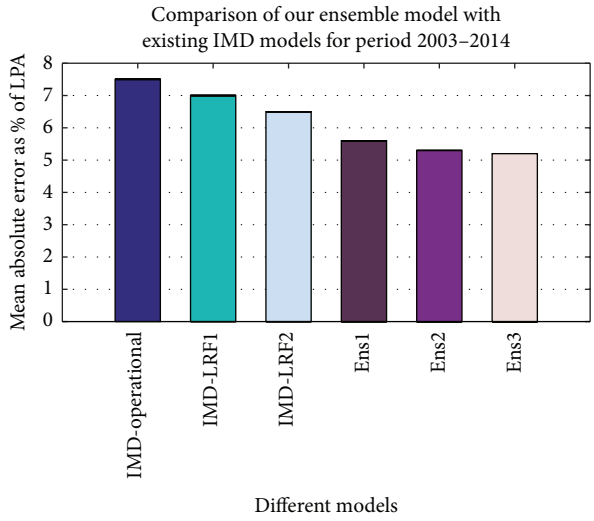

Figure 10: Comparison of prediction of Indian summer monsoon by proposed joint clustering based ensemble models with IMD existing models for period 2003–2014 [4, 5].

Table 7: Comparison between mean absolute errors (%) for prediction of regional Indian summer monsoon by proposed approach and IMD model [4, 5] during 2004–2014.

Regions	IMD prediction error	Proposed model prediction error
Central	12.2	**8.3**
North East	7.8	**7.2**
North West	9.6	**9.1**
South Peninsular	8.9	**7.6**

6. Conclusions

Prediction of Indian summer monsoon is critical due to complex underlying interaction of multiple global climatic variables. The predictor skills also drift and vary over years. The article attempts to address this by joint clustering of years and predictors of monsoon, utilizing different climatic variables as predictor sets for different clusters of years, and finally multimodel ensemble forecast is provided for aggregate and regional Indian summer monsoon. Different climatic predictors influencing summer monsoon of the country are identified and subspace clustering is performed for dual simultaneous grouping of predictors and years of monsoon. Ensemble forecast is provided by combining the forecasts by models designed using random forest of regression tree for every cluster.

Proposed joint clustering based ensemble model provides mean absolute error of 5.2% in prediction of aggregate Indian summer monsoon and it provides comparable errors for regional predictions. Proposed ensemble model performs superior to IMD's monsoon prediction model [5]. Prediction by our proposed approach also surpasses the prediction given by method without clustering and ensemble approaches. Thus, the joint predictor-year clustering based approach unknot a direction to frame the complex monsoon phenomenon in a better way to engross the variability in its quantity as well as predictors over time.

In the future, more climatic variables affecting Indian summer monsoon can be explored and prediction models can be improvised with various machine learning and statistical approaches to attain even better forecasting accuracy. Extremities of monsoon can also be focused to unravel the complexity of the phenomenon. It may be attempted to identify different spells of monsoon which will assist in identifying the extremes of monsoon. In addition to data-driven approach as proposed, physics behind the circulations involved in climatic systems should be focused and incorporated to clone the phenomenon in more efficient way.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

One of the authors, Professor Arun Chakraborty, James Rennel Fellow of MoES, greatly acknowledges the funding to carry out this research work from MoES, Government of India, under Samudra Gupta Chair Professor fund. The authors also want to acknowledge MHRD institute funding under application of artificial intelligence for societal needs.

References

[1] H. F. Blanford, "On the connexion of the Himalaya snowfall with dry winds and seasons of drought in India," *Proceedings of the Royal Society of London*, vol. 37, no. 232-234, pp. 3–22, 1884.

[2] G. Walker, *Correlation in Seasonal Variations of Weather, VIII: A Preliminary Study of World Weather*, vol. 24 of *Meteorological Office*, 1923.

[3] V. Gowariker, V. Thapliyal, S. Kulshrestha, G. Mandal, N. Sen-Roy, and D. Sikka, "A power regression model for long range forecast of southwest monsoon rainfall over India," *Mausam*, vol. 42, no. 2, pp. 125–130, 1991.

[4] M. Rajeevan, D. S. Pai, S. K. Dikshit, and R. R. Kelkar, "IMD's new operational models for long-range forecast of southwest monsoon rainfall over India and their verification for 2003," *Current Science*, vol. 86, no. 3, pp. 422–431, 2004.

[5] M. Rajeevan, D. S. Pai, R. Anil Kumar, and B. Lal, "New statistical models for long-range forecasting of southwest monsoon rainfall over India," *Climate Dynamics*, vol. 28, no. 7-8, pp. 813–828, 2007.

[6] A. Singh, M. A. Kulkarni, U. C. Mohanty, S. C. Kar, A. W. Robertson, and G. Mishra, "Prediction of Indian summer monsoon rainfall (ISMR) using canonical correlation analysis of global circulation model products," *Meteorological Applications*, vol. 19, no. 2, pp. 179–188, 2012.

[7] A. G. Turner and H. Annamalai, "Climate change and the South Asian summer monsoon," *Nature Climate Change*, vol. 2, no. 8, pp. 587–595, 2012.

[8] T. DelSole and J. Shukla, "Linear prediction of Indian monsoon rainfall," *Journal of Climate*, vol. 15, no. 24, pp. 3645–3658, 2002.

[9] T. DelSole and J. Shukla, "Climate models produce skillful predictions of Indian summer monsoon rainfall," *Geophysical Research Letters*, vol. 39, no. 9, Article ID L09703, 2012.

[10] J. Sun and H. Chen, "A statistical downscaling scheme to improve global precipitation forecasting," *Meteorology and Atmospheric Physics*, vol. 117, no. 3-4, pp. 87–102, 2012.

[11] A. Cherchi, S. Gualdi, S. Behera et al., "The influence of tropical Indian Ocean SST on the Indian summer monsoon," *Journal of Climate*, vol. 20, no. 13, pp. 3083–3105, 2007.

[12] K. K. Kumar, B. Rajagopalan, M. Hoerling, G. Bates, and M. Cane, "Unraveling the mystery of Indian monsoon failure during El Niño," *Science*, vol. 314, no. 5796, pp. 115–119, 2006.

[13] H. Wang and F. Xue, "The interannual variability of Somali Jet and its inuences on the inter-hemispheric water vapor transport and the East Asian summer rainfall," *Chinese Journal of Geophysics*, vol. 46, no. 1, pp. 11–20, 2003.

[14] P. K. Das, *The Monsoons*, National Book Trust India, New Delhi, India, 1988.

[15] E. Kalnay, M. Kanamitsu, R. Kistler et al., "The NCEP/NCAR 40-year reanalysis project," *Bulletin of the American Meteorological Society*, vol. 77, no. 3, pp. 437–471, 1996.

[16] R. W. Reynolds, N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, "An improved in situ and satellite SST analysis for climate," *Journal of Climate*, vol. 15, no. 13, pp. 1609–1625, 2002.

[17] E. M. Rasmusson and T. H. Carpenter, "Variations in tropical sea surface temperature and surface wind fields associated with the Southern oscillation/El Niño," *Monthly Weather Review*, vol. 110, no. 5, pp. 354–384, 1982.

[18] H. Wang and S. He, "Weakening relationship between East Asian winter monsoon and ENSO after mid-1970s," *Chinese Science Bulletin*, vol. 57, no. 27, pp. 3535–3540, 2012.

[19] L. Parsons, "Evaluating subspace clustering algorithms," in *Proceedings of the Workshop on Clustering High Dimensional Data and Its Applications, SIAM International Conference on Data Mining (SDM '04)*, pp. 48–56, 2004.

[20] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park, "Fast algorithms for projected clustering," *ACM SIGMOD Record*, vol. 28, pp. 61–72, 1999.

[21] MATLAB, *Statistics and Machine Learning Toolbox. MATLAB Version*, The MathWorks, Natick, Mass, USA, 2012.

[22] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.