Hindawi Publishing Corporation
Advances in Meteorology
Volume 2015, Article ID 329835, 12 pages
http://dx.doi.org/10.1155/2015/329835

Research Article

Fuzzy Clustering-Based Ensemble Approach to Predicting Indian Monsoon

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

[1]*Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, Paschim Medinipur, West Bengal 721302, India*
[2]*Centre for Oceans, Rivers, Atmosphere and Land Sciences, Indian Institute of Technology Kharagpur, Kharagpur, Paschim Medinipur, West Bengal 721302, India*

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

Received 2 January 2015; Revised 31 March 2015; Accepted 3 April 2015

Academic Editor: Xiaolong Jia

Indian monsoon is an important climatic phenomenon and a global climatic marker. Both statistical and numerical prediction schemes for Indian monsoon have been widely studied in literature. Statistical schemes are mainly based on regression or neural networks. However, the variability of monsoon is significant over the years and a single model is often inadequate. Meteorologists revise their models on different years based on prevailing global climatic incidents like El-Niño. These indices often have degree of severity associated with them. In this paper, we cluster the monsoon years based on their fuzzy degree of associativity to these climatic event patterns. Next, we develop individual prediction models for the year clusters. A weighted ensemble of these individual models is used to obtain the final forecast. The proposed method performs competitively with existing forecast models.

1. Introduction

Monsoon is a complex phenomenon of a climatic system. It is influenced by multiple climatic parameters and sea-atmosphere interactions. Prediction of monsoon is challenging due to large variability present in its patterns. Indian Meteorological Department (*IMD*) performs forecast of Indian summer monsoon rainfall (*ISMR*) since 1886. Indian monsoon forecast was initiated by Blanford [1] as early as 1882. The success of forecasts in span of 1882–1885 encouraged Blanford to design operational long range forecast model for monsoon in 1886. Subsequently, Walker [2] developed models studying the statistical correlations between rainfall and different global climate parameters. Thapliyal and Kulshrestha [3] introduce regression model in predicting south-west Indian monsoon rainfall. Gowariker et al. [4] propose power regression model for long-term forecast of monsoon, which provided accurate forecast for a long period, but failed to predict the extreme condition of 2002. In 2004, Rajeevan et al. [5] reassess different climatic parameters and introduce four new parameters to design statistical model for issuing long-range forecast of Indian monsoon. Succeeding in 2007, Rajeevan et al. [6] built models using ensemble multiple regression and pursuit projection regression to forecast Indian rainfall and proved to be superior to past *IMD* models. Schewe and Levermann [7] explain the change in distribution of Indian rainfall and also explain the reasons behind failure of monsoon in certain years. Wu et al. [8] propose a linear Markov model to predict short-term climate variability of East Asian monsoon. Fan et al. [9] develop two statistical prediction schemes for seasonal forecast of East Asian summer monsoon. The schemes take the direct outputs of the existing models and give better prediction of the summer monsoon.

Artificial neural networks (*ANN*) [10] are widely used in modelling the nonlinearity present in monsoon process. Sahai et al. [11] use *ANN* techniques with error backpropagation to forecast Indian summer monsoon rainfall. Hong [12]

predicts Indian summer monsoon utilizing recurrent neural network and also demonstrates successful employment of support vector machine in solving nonlinear regression and time series problems. Three different backpropagation neural learning rules, namely, momentum learning, conjugate gradient descent learning, and Levenberg-Marquardt learning, are used by S. Chattopadhyay and G. Chattopadhyay [13] to perform a comparative study of different neural network method to predict rainfall time series.

Presence of large variability in monsoon patterns makes it difficult for a single model to predict its distribution. A number of uncertainties including boundary condition, parameter, and structural uncertainties are involved in construction of these models. Thus, it remains fundamentally challenging to have a single model for prediction. Multimodel ensembles are proposed to overcome the weakness of single model, which combine the outcome of different models to produce efficient results [14, 15]. In addition, monsoon shows different characteristics over years. There exist groups of years where variation of climatic parameters and pattern of rainfall are similar. We use fuzzy clustering to cluster the similar years together and model them separately. The motivation behind using fuzzy clustering is that each year manifests a mixture of physical climatic events. We cannot hard cluster a year into a specific group; years have their membership of belongingness to every cluster. Fuzzy clustering is used to enclose the characteristics of different events being related to a year of study. We use the same set of climatic parameters as predictor set for every cluster but frame different models for each cluster.

A number of prediction models, namely, multiple regression (*MR*), multilayer perceptron (*MLP*), recurrent neural network (*RNN*), and generalized regression neural network (*GRNN*) models, are used for prediction of Indian monsoon for the year clusters. There exists viable reasons for using neural networks like *MLP*, *RNN*, and *GRNN* for modelling: (i) Indian monsoon is a complex process, which cannot be adequately modelled by linear models, (ii) nonlinearity in the time-series pattern can be well captured by neural network learning, (iii) climatic events are much closely related to near years parameters disturbance as compared to distant years, and neural network enables attaching weight to the year parameter in appropriate manner.

In this work, climatic parameters that are strongly correlated with Indian monsoon are identified at the onset, which is followed by fuzzy clustering of years into groups with degree of belongingness of each year to the clusters. Then we model each cluster with four types of models, namely, *MR*, *MLP*, *RNN*, and *GRNN*, to forecast rainfall. Weighted ensemble of forecasts given by respective models for each cluster is considered as final predicted rainfall. Analysis and comparisons are performed on aggregate Indian rainfall and finally, a meteorological interpretation of the obtained clusters is presented.

The paper is organised in the following manner. We discussed the details of data and predictor climatic parameters in Sections 2 and 3. Proposed clustering based approach, prediction model, and ensemble technique are presented in Section 4 with experimental results in Section 5. Meteorological significance is discussed in Section 6 and finally, conclusions are provided in Section 7.

2. Data Sets Used

We consider the annual Indian summer monsoon rainfall (*ISMR*), occurring in four months of June, July, August, and September. Annual *ISMR* is considered during period 1948–2013 for our study. The long period average (*LPA*) (1948–2013) of *ISMR* is 891.8 mm. *ISMR* is expressed as percentage of the *LPA* value. The data is obtained from Indian Institute of Tropical Meteorology, Pune (http://www.imdpune.gov.in/research/ncc/longrange/data/data.html) [16].

Predictor parameters sea level pressure (*SLP*) (http://www.esrl.noaa.gov/psd/gcos_wgsp/Gridded/data.noaa.erslp.html) and sea surface temperature (*SST*) (http://www.esrl.noaa.gov/psd/data/gridded/data.noaa.ersst.html) data are provided by the NOAA/OAR/ESRL/PSD, at spatial resolution of $2^\circ \times 2^\circ$ [17]. Surface pressure (*SP*) and zonal wind velocity (*WV*) data are collected from *NCEP* Reanalysis Derived data provided by the NOAA/OAR/ESRL PSD (http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.derived.surface.html) [18], available at resolution of $2.5^\circ \times 2.5^\circ$. Finally, Niño 3.4 data, which is the sea surface temperature anomaly for the spatial coverage of 5°S to 5°N and 170°W to 120°W in Pacific Ocean region is acquired from National Center for Atmospheric Research (http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml) [19]. All the above monthly data are considered for the period 1948–2013 in our study and analysis.

3. Global Climatic Parameters Influencing Indian Monsoon

Indian monsoon is strongly influenced by several global climatic parameters, occurring at places distant from Indian subcontinent. Identification of predictor parameters relies on physical understanding of monsoon event and wind pattern flow. We have selected the climatic parameters based on the parameters used by Indian meteorological department's models [5, 6], studying their correlation with Indian summer monsoon rainfall (*ISMR*) during our period of study (1948–2013). In the data preprocessing phase, climatic anomaly data are evaluated by calculating the deviation of parameter value from long-term average value of the parameter exclusively for each month, followed by correlation study between *ISMR* and the climatic parameters for a lag of zero to twelve months. We consider the best lagged predictor month having high correlation with *ISMR*. The predictor climatic parameters and their correlation values with Indian monsoon are shown in Table 1. Figure 1 shows the geographic location of climatic parameters influencing Indian monsoon.

Predictor Sets of Climatic Parameters. Based on the correlation with Indian monsoon, we have built five predictor sets for forecasting. Different combinations of the identified climatic parameters (Table 1) form the predictor sets. The predictor sets are shown in Table 2.

Table 1: Climatic parameters (CP) influencing Indian monsoon with geographical location, correlation values, and correlated month (0 signifies same years and −1 signifies previous year).

CP	CP name	Location	Correlation values	Correlated months
CP1	North Atlantic Ocean SST anomaly	20°N–30°N, 100°W–80°W	0.242	Jan (0)
CP2	North Atlantic Ocean surface pressure anomaly	20°N–30°N, 100°W–80°W	0.256	April (0)
CP3	East Asia SLP anomaly	35°N–45°N, 120°E–130°E	0.337	May (0)
CP4	East Asia surface pressure anomaly	35°N–45°N, 120°E–130°E	0.341	Mar (0)
CP5	Equatorial South Eastern Indian ocean SST anomaly	20°S–10°S, 100°E–120°E	0.200	Sept (−1)
CP6	Pressure gradient between Madagascar and Tibet	—	0.253	May (0)
CP7	Niño 3.4 SST anomaly	5°S–5°N, 170°W–120°W	0.311	Sept (−1)
CP8	Equatorial Pacific Ocean SLP anomaly	5°S–5°N, 120°E–80°W	0.272	Aug (−1)
CP9	North West Europe surface pressure anomaly	55°N–65°N, 20°E–40°E	0.183	Jan (0)
CP10	North Central Pacific zonal wind anomaly	5°N–15°N, 180°E–150°W	0.457	May (0)

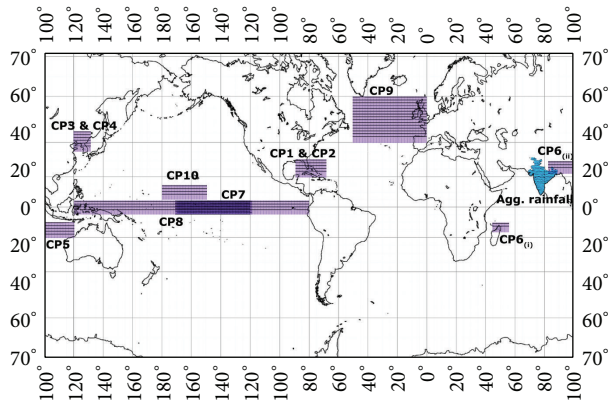

Figure 1: Climatic parameters over the globe governing Indian monsoon (purple patches signify the location of climatic parameters taken, and blue patch represents the Indian region); CP_i represents parameter i in Table 1.

Table 2: Predictor sets with climatic parameters.

Predictor sets	Climatic parameters
PredSet1	CP1, CP4, CP5, CP6
PredSet2	CP4, CP5, CP6, CP7
PredSet3	CP2, CP4, CP10
PredSet4	CP2, CP4, CP7, CP10
PredSet5	CP3, CP7, CP8, CP9

4. Methodology

We propose fuzzy clustering of monsoon years into groups followed by building models for each group separately and finally predicting Indian summer monsoon rainfall (*ISMR*) as weighted ensemble of forecasts provided by cluster models. The block diagram of the proposed fuzzy clustering-based approach to prediction of *ISMR* is shown in Figure 2. Detailed steps are described in the following subsections.

4.1. Motivation: Variability of Monsoon Patterns. Trends and distributions of monsoon vary to a large extent over years. It is thus necessary to group the years into clusters which have similar patterns of predictor climatic parameters affecting monsoon. The approach of clustering the years is effective as we can build separate models for each cluster. These cluster models will be more accurate as variation within cluster is less. Finally, ensemble of forecasts of these cluster models results in better prediction of Indian monsoon. As an example consider two clusters of years corresponding to strong El-Niño and North Atlantic Oscillation, respectively. A drought year has correlation with both events and hence might have significant degree of belongingness to both clusters.

4.2. Fuzzy Clustering of Monsoon Years. Fuzzy c-means clustering is used for grouping the similar years together. Fuzzy c-means (*FCM*) is a method of clustering which allows one instance of input to belong to more than one cluster with some membership of belongingness. *FCM* attempts to partition a set of N elements $Y = \{y_1, \ldots, y_n\}$ into a collection of c fuzzy clusters $C = \{\text{cen}_1, \ldots, \text{cen}_c\}$ and a partition matrix $W = w_{ij} \in [0, 1]$, $i = 1, \ldots, n$, $j = 1, \ldots, c$, where w_{ij} gives the degree of belongingness of element y_i to cluster with center cen_j.

FCM aims to minimize an objective function of (1). The update of partition matrix and centers occur in accordance with (2) and (3), respectively:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} w_{ij}^m \left\| y_i - \text{cen}_j \right\|^2, \quad 1 \le m \le \infty \tag{1}$$

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\left\| y_i - \text{cen}_j \right\| / \left\| y_i - \text{cen}_k \right\| \right)^{2/(m-1)}} \tag{2}$$

$$\text{cen}_j = \frac{\sum_{i=1}^{N} w_{ij}^m \cdot x_i}{\sum_{i=1}^{N} w_{ij}^m}, \tag{3}$$

where m denotes the level of cluster fuzziness.

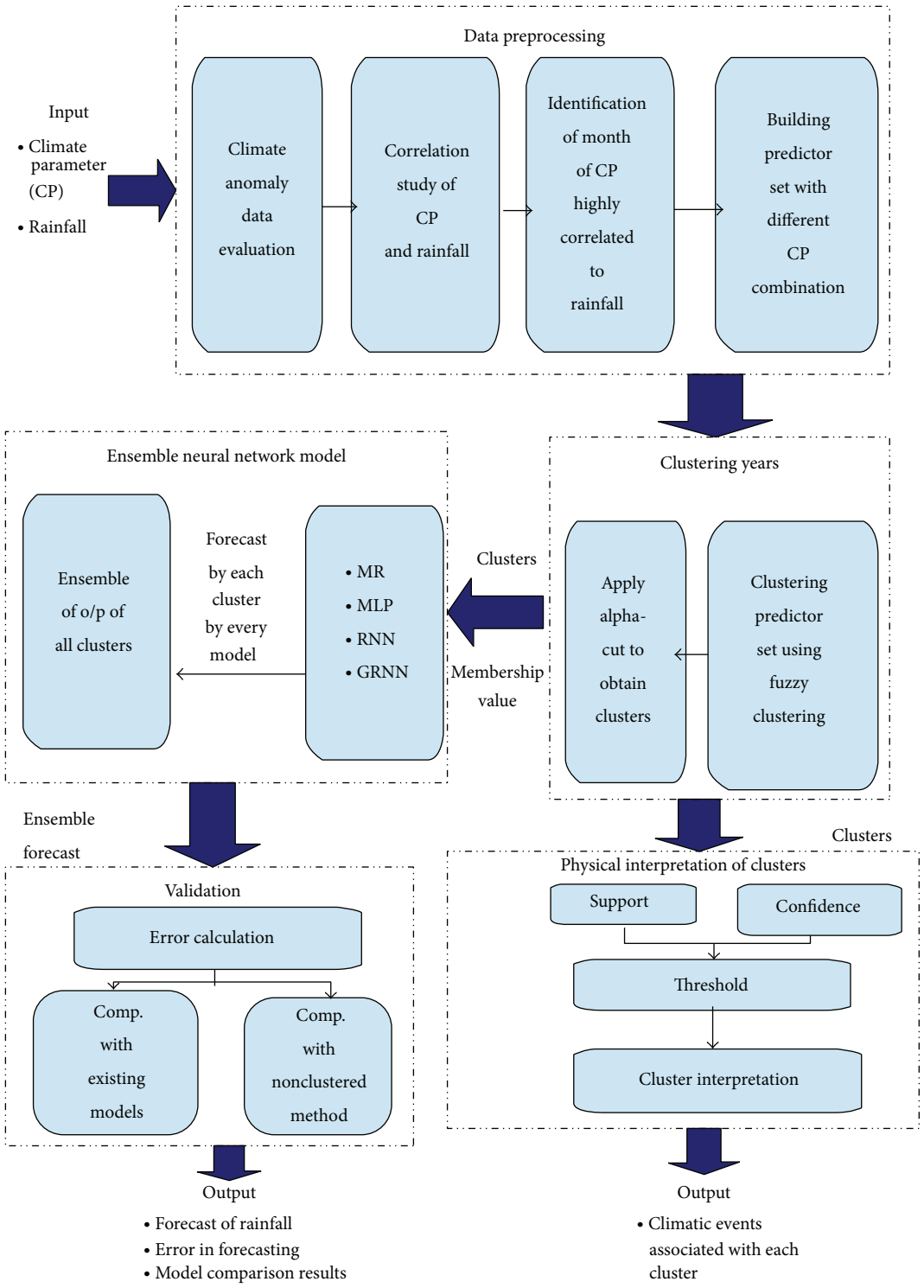

Figure 2: Proposed fuzzy clustering-based ensemble approach for prediction of Indian summer monsoon rainfall.

4.3. Prediction Models. Multiple regression and three models of artificial neural networks (*ANN*), namely, multilayer perceptron, recurrent neural network, and generalized regression neural network, are used to design prediction models for each cluster exclusively. Forecast of annual *ISMR* is provided by each cluster model separately and also by ensemble of all the clusters' model forecast. We describe below the models used.

4.3.1. Multiple Regression (MR). Multiple regression model is used to learn the relationship between several independent predictor variables (X_is) and a dependent variable (Y).

Table 3: Model parameter setting for MLP models.

Parameter set	Hidden layers	Training years	Training method
ParSet1	[3 5]	20	BFGS quasi-Newton backpropagation
ParSet2	[3 5 10]	15	Conjugate gradient backpropagation with Powell-Beale restarts
ParSet3	[5 10]	10	Scaled conjugate gradient backpropagation
ParSet4	[3 5]	15	Resilient backpropagation

Multiple regression model having p independent variables is shown in

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{4}$$

where x_{ij} is the ith observation of jth independent variable, where the first independent variable takes the value 1 for all i and ε represents the residual.

4.3.2. Multilayer Perceptron Neural Network (MLP). Multilayer perceptron neural network is a class of *ANN* where connections between the neurons do not form a directed cycle. In this network, the information propagates in only one direction, from input nodes, through hidden nodes, and to the output nodes. The independent and dependent variables constitute the input and output layers, respectively. Number of hidden layers with corresponding nodes must be determined empirically for each prediction task. Four different parameter sets are considered empirically for model designed to forecast *ISMR*, shown in Table 3.

4.3.3. Recurrent Neural Network (RNN). Recurrent neural network is a class of *ANN* which creates an internal state of the network to exhibit dynamic temporal behaviour. Climatic changes or events occurring in near or same time period are highly correlated. Similarly, rainfall patterns are more correlated to influencing factors in the near years as compared to the distant years. This phenomenon is well captured by *RNN* which gives weights in decreasing order to the values in near to distant years during training of network. Thus, it assists in modelling the system dynamics in much natural manner. Same set of climatic parameters as *MLP* network (Table 3) is considered with delay span of *2* units.

4.3.4. Generalized Regression Neural Network (GRNN). Generalized regression neural network is a variant of radial basis function network. *GRNN* has three layers of artificial neurons: input, hidden, and output. The hidden layer has radial basis neurons, while neurons in the output layer have linear transfer function. Output of radial basis neurons is the input scaled by the spread factor. Given p input-output pairs $x_i, y_j \in \Re^n \times \Re^1$, with n input variables and $i = 1, 2, \ldots, p$, y_j represents the output from each hidden unit. The *GRNN* output for a test point, $x \in \Re^n$, is described by

$$\widehat{y(x)} = \sum_{i=1}^{p} W_i y_i, \tag{5}$$

where

$$W_i = \frac{\exp\left(-\left\|x - x_i\right\|^2 / 2\sigma^2\right)}{\sum_{k=1}^{p} \exp\left(-\left\|x - x_k\right\|^2 / 2\sigma^2\right)}. \tag{6}$$

The reasons behind modelling using *GRNN* are (i) only one tunable design parameter (spread factor), (ii) one-pass algorithm (less time consuming), and (iii) accurately approximate functions from sparse data.

Optimal training year is ascertained for *MR* and *GRNN* models by varying training years from 5 to 30 and validating against least absolute error in prediction during validation period (1984–1993). A training of m years specifies that, for predicting rth year rainfall, available preceding m number of years $r-1, r-2, \ldots, r-m$ present in a particular cluster are considered for training.

4.4. Ensemble of Predictors. Complexity in monsoon process makes it difficult for a single model to predict rainfall accurately. We design separate models for each cluster of years obtained by fuzzy clustering using four predictors described in Section 4.3. Finally, annual *ISMR* is presented as weighted ensemble of forecasts of model designed for each cluster. Weight is taken as the fuzzy membership of belongingness of the test year in different clusters:

$$\text{Ensemble prediction}^t = \sum_{i=1}^{c} W_i^t \cdot P_i, \tag{7}$$

where P_i represents the prediction given by a model for cluster i, W_i^t is the fuzzy membership of tth test year to cluster i, and c is the total number of clusters.

4.5. Validation of Proposed Approach. The study is performed on data for the period 1948–2013. Fuzzy clustering is performed over the period to cluster it into *three* groups. The number of clusters is decided based on cluster quality. Separate prediction models are designed for all three clusters and ensemble of forecasts of these models is provided as predicted Indian summer monsoon rainfall. Test period 2001–2013 is considered to evaluate the forecasting skills of our proposed approach.

The forecast models for annual *ISMR* are chiefly evaluated in terms of mean absolute error. Other error statistics, namely, root mean square error, prediction yields, Pearson correlation, and Willmott index of agreement, are also evaluated to judge the efficacy of our proposed approach for prediction. They are described below.

(i) *Mean Absolute Error* (MAE). Mean absolute error for prediction of annual *ISMR* is calculated in the following way:

$$\text{MAE} = \frac{\sum_{i=1}^{N} |Y - X|}{N}, \tag{8}$$

where X and Y are the actual and predicted *ISMR* series for test period and N denotes the total number of test years.

(ii) *Root Mean Square Error* (RMSE). Root mean square error calculates the differences between model predicted output and actual values. They are a good measure to compare forecasting errors of various models:

$$\text{RMSE} = \sqrt{\frac{(Y - X)^2}{N}}. \tag{9}$$

(iii) *Prediction Yield* (PY). Prediction yields are evaluated at three different error categories (5%, 10%, and 15% errors) to assess the overall prediction results by judging percent of predicted years within each allowed range of errors.

(iv) *Pearson Correlation Coefficient* (PC). Pearson correlation coefficient measures the strength of linear association between actual and predicted values, where the value of 1 means a perfect positive correlation and the value of –1 means a perfect negative correlation:

$$\text{PC} = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}}, \tag{10}$$

where X and Y are the actual and predicted *ISMR* series for test period and $\overline{X}$ and $\overline{Y}$ are their corresponding mean.

(v) *Willmott Index of Agreement* (WI). Willmott index of agreement is a standardized measure of the degree of model prediction error. It varies between 0 and 1 with higher values indicating a better fit of the model for prediction:

$$\text{Index of agreement} = 1 - \frac{\sum_{i=1}^{N} \left|X_i - Y_i\right|^2}{\sum_{i=1}^{N} \left(\left|Y_i - \overline{X}\right| + \left|X_i - \overline{X}\right|\right)^2}. \tag{11}$$

5. Experimental Results and Analysis

In this section we present the evaluation of our proposed fuzzy clustering-based approach. We first present the results of fuzzy clustering of the monsoon years for different predictor sets. Forecasting skills are evaluated for all cluster and the ensemble model in terms of mean absolute errors for test period 2001–2013. In addition, other measures like root mean square errors in prediction, correlation between predicted and actual rainfall, prediction yields, and agreement index between actual and predicted rainfall are also estimated to establish the efficiency of our proposed approach to prediction of Indian summer monsoon rainfall.

Table 4: Cluster size (number of years) by fuzzy c-means clustering with α-cut of *0.3* over the period 1948–2013.

Predictor set	Cluster1	Cluster2	Cluster3
PredSet1	16	38	30
PredSet2	30	17	40
PredSet3	32	14	38
PredSet4	42	31	21
PredSet5	15	37	26

5.1. Clustering of Monsoon Years. Fuzzy clustering is performed over period 1948–2013 to cluster the data into *three* clusters. We have performed an α-cut, with value $\alpha = 0.3$ to assign the data instances to the clusters. The value is ascertained empirically such that the distribution of elements within clusters is regular. A data instance can be assigned to more than one cluster simultaneously. The cluster sizes are shown in Table 4 while considering various predictor sets.

5.2. Prediction Accuracy. We predict annual rainfall considering for all five predictor sets (Table 2) separately using four models, namely, *MR*, *MLP*, *RNN*, and *GRNN*. Test period is considered from 2001 to 2013.

5.2.1. Multiple Regression Model (MR). Multiple regression models are built for every cluster by ascertaining optimal training period for each predictor set. Optimal training period is evaluated by varying training years and validating them for least absolute error in prediction during validation period (1984–1993). Individual cluster based as well as weighted ensemble models are considered for prediction. Table 5 gives the mean absolute error for individual cluster based and ensemble models for test period 2001–2013. The model provides mean absolute error of 6.2% for *PredSet4* (Table 2). It is observed that the ensemble model outperforms all the single cluster models for every predictor set. Figure 3 shows the interannual variability of actual and ensemble predicted rainfall as percent of long period average (*LPA*).

5.2.2. Multilayer Perceptron Neural Network Model (MLP). Multilayer perceptron neural network model is designed with four different sets of parameters described in Table 2. Mean absolute errors of all cluster and ensemble models are shown in Table 6. *MLP* model reports an error of 4.0% for *PredSet4* (Table 2) with *MLP* parameters *ParSet1* (Table 3). The actual and predicted rainfall by models built for clusters and ensemble model is shown in Figure 4. Ensemble predicted rainfall closely follows actual rainfall.

5.2.3. Recurrent Neural Network Model (RNN). Mean absolute errors for prediction of annual rainfall by recurrent neural network model for the test period 2001–2013 are

Table 5: Mean absolute errors (%) for annual Indian summer monsoon rainfall prediction by individual MR cluster models and ensemble model for test period 2001–2013. Reports minimum error of *6.2*%.

Predictor set	Training years	Cluster1 error (%)	Cluster2 error (%)	Cluster3 error (%)	Ensemble error (%)
PredSet1	20	9.4	9.3	10.9	**8.6**
PredSet2	20	11.0	7.5	9.4	**8.3**
PredSet3	15	10.9	6.5	9.2	**6.7**
PredSet4	15	10.4	10.1	6.8	***6.2***
PredSet5	15	7.6	8.5	8.4	**7.9**

Table 6: Mean absolute errors (%) for annual Indian summer monsoon rainfall prediction by individual MLP cluster models and ensemble model for test period 2001–2013. Reports minimum error of *4.0*%.

Predictor set	Parameter set	Cluster1 error (%)	Cluster2 error (%)	Cluster3 error (%)	Ensemble error (%)
PredSet1	*ParSet4*	13.8	18.1	16.9	**8.2**
PredSet2	*ParSet3*	16.0	7.9	11.0	**5.2**
PredSet3	*ParSet1*	8.0	7.8	6.5	**6.5**
PredSet4	*ParSet1*	9.3	10.7	4.5	***4.0***
PredSet5	*ParSet1*	8.5	15.3	13.7	**11.0**

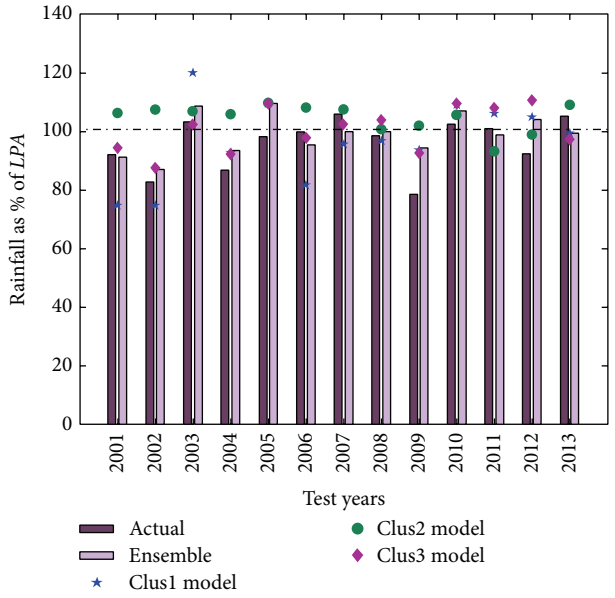

Figure 3: Performance of forecasts by proposed fuzzy clustering-based ensemble model and its respective three clusters models by *MR* for *PredSet4*. The *deep* and *light purple bars* represent the actual and predicted *ISMR* in terms of percent of *LPA*. The *symbols* represent forecasts given by individual cluster models. The results are shown for test period 2001–2013.

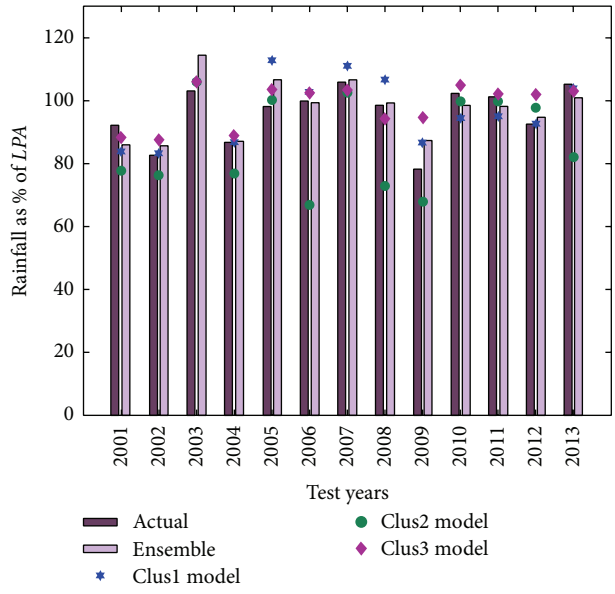

Figure 4: Performance of forecasts by proposed fuzzy clustering-based ensemble model and its respective three clusters models by *MLP* for *PredSet4*. The *deep* and *light purple bars* represent the actual and predicted *ISMR* in terms of percent of *LPA*. The *symbols* represent forecasts given by individual cluster models. The results are shown for test period 2001–2013.

presented in Table 7. *PredSet3* (Table 2) with *RNN* parameters *ParSet1* (Table 3) gives error of 5.1%. *RNN* gives weights in decreasing order of their distance from test year to the training years. The pattern of actual and ensemble predicted rainfall in terms of percentage of *LPA* is shown in Figure 5.

5.2.4. Generalized Regression Neural Network Model (GRNN). Generalized regression neural network ensemble and individual cluster models' errors in terms of mean absolute errors are presented in Table 8. The model reports an error of 6.1% for *PredSet3* (Table 2). Figure 6 shows the interannual variations of ensemble forecast of rainfall by *GRNN* ensemble model along with actual rainfall pattern in terms of percentage of *LPA* for period 2001–2013. It is observed that the predicted values are close to actual rainfall patterns. Prediction by models designed for clusters is shown by different symbols.

Table 7: Mean absolute errors (%) for annual Indian summer monsoon rainfall prediction by individual RNN cluster models and ensemble model for test period 2001–2013. Reports minimum error of *5.1*%.

Predictor set	Parameter set	Cluster1 error (%)	Cluster2 error (%)	Cluster3 error (%)	Ensemble error (%)
PredSet1	*ParSet1*	11.3	7.1	16.8	**7.0**
PredSet2	*ParSet1*	13.2	13.5	12.6	**8.5**
PredSet3	*ParSet1*	12.9	5.4	6.0	***5.1***
PredSet4	*ParSet1*	12.3	6.4	4.7	**5.9**
PredSet5	*ParSet2*	15.1	16.1	13.4	**8.8**

Table 8: Mean absolute errors (%) for annual Indian summer monsoon rainfall prediction by individual GRNN cluster models and ensemble model for test period 2001–2013. Reports minimum error of *6.1*%.

Predictor set	Training years	Cluster1 error (%)	Cluster2 error (%)	Cluster3 error (%)	Ensemble error (%)
PredSet1	20	10.0	7.6	7.6	**6.4**
PredSet2	30	7.1	8.9	7.6	**6.4**
PredSet3	20	5.8	9.2	6.0	***6.1***
PredSet4	20	6.3	6.6	7.2	**6.3**
PredSet5	25	7.1	9.4	11.9	**6.6**

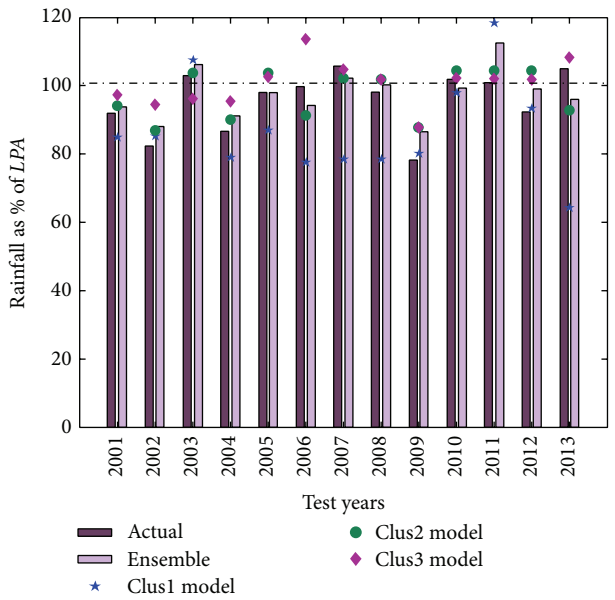

Figure 5: Performance of forecasts by proposed fuzzy clustering-based ensemble model and its respective three clusters models by *RNN* for *PredSet3*. The *deep* and *light purple bars* represent the actual and predicted *ISMR* in terms of percent of *LPA*. The *symbols* represent forecasts given by individual cluster models. The results are shown for test period 2001–2013.

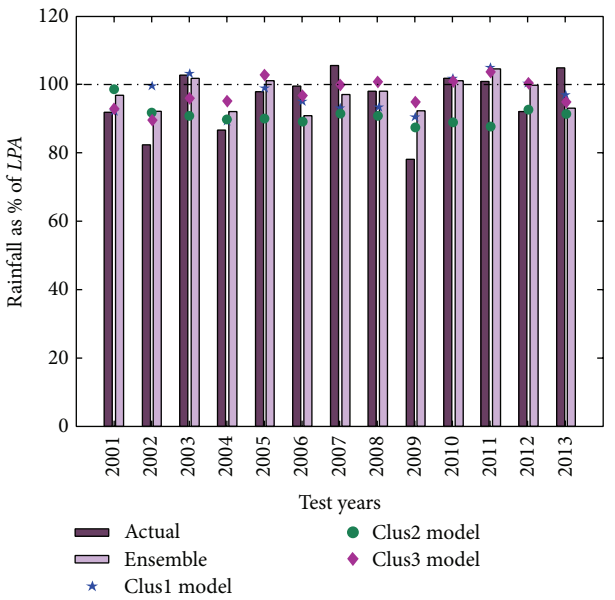

Figure 6: Performance of forecasts by proposed fuzzy clustering-based ensemble model and its respective three clusters models by *GRNN* for *PredSet3*. The *deep* and *light purple bars* represent the actual and predicted *ISMR* in terms of percent of *LPA*. The *symbols* represent forecasts given by individual cluster models. The results are shown for test period 2001–2013.

5.3. Statistical Measures for Validation of Proposed Approach. Next, we validate the models in terms of other accuracy measures besides mean absolute error. Table 9 shows different forecast verification statistics for ensemble models during test period 2001–2013. We summarize the observations below.

(i) *Root Mean Square Error (RMSE)*. *MLP* ensemble model gives *RMSE* of 5.3%, followed by *RNN* ensemble model with 6.4%. *GRNN* and *MR* models give *RMSE* of 7.4% and 8.4%, respectively.

(ii) *Prediction Yield (PY)*. *PY* for 5% error category of *MR*, *MLP*, *RNN*, and *GRNN* ensemble models is 46%, 69%, 53%, and 46%, respectively. They give prediction yield of 76%, 92%, 92%, and 84% for allowed error of 10% category. Finally at error category of 15%, *MR*, *MLP*, *RNN*, and *GRNN* ensemble models give yield of 92%, 100%, 92%, and 100%, respectively. Thus, none of the predicted years show abrupt deviation from corresponding actual rainfall pattern.

Table 9: Prediction evaluation statistics for ensemble models during test period 2001–2013 (Section 4.5).

Verification measures	MR	MLP	RNN	GRNN
RMSE for forecast (%)	8.4	**5.3**	6.4	7.4
PY (%) at allowed error 5%	46	**69**	53	46
PY (%) at allowed error 10%	76	**92**	**92**	84
PY (%) at allowed error 15%	92	**100**	92	**100**
PC between actual and predicted rainfall	0.61	**0.81**	0.71	0.49
WI between actual and predicted rainfall	0.71	**0.89**	0.81	0.62

Table 10: Comparison of absolute errors for rainfall prediction by proposed ensemble models (Ensml) with clustering (WC) approach to standard method with same models without clustering (NC) approach.

Predictor set	*MR*		*MLP*		*RNN*		*GRNN*	
	Tot. error (NC) (%)	Ensml error (WC) (%)	Tot. error (NC) (%)	Ensml error (WC) (%)	Tot. error (NC) (%)	Ensml error (WC) (%)	Tot. error (NC) (%)	Ensml error (WC) (%)
PredSet1	8.9	**8.6**	10.0	**8.2**	11.7	**7.0**	6.9	**6.4**
PredSet2	9.2	**8.2**	12.8	**5.2**	10.7	**8.5**	7.2	**6.4**
PredSet3	7.4	**6.7**	6.7	**6.5**	6.2	**5.1**	6.1	**6.1**
PredSet4	6.7	**6.2**	5.8	**4.0**	6.0	**5.5**	6.3	**6.3**
PredSet5	8.2	**7.9**	9.7	11.0	8.9	**8.8**	9.0	**6.7**

(iii) *Pearson Correlation (PC).* *PC* of 0.61, 0.81, 0.71, and 0.49 is observed for prediction by *MR*, *MLP*, *RNN*, and *GRNN* ensemble models, respectively. It is noticed that predicted rainfall by *MLP* ensemble model is highly correlated to actual values, while correlation for *GRNN* forecast is least.

(iv) *Willmott Index of Agreement (WI).* *WI* for *MR*, *MLP*, *RNN*, and *GRNN* ensemble models is 0.71, 0.89, 0.81, and 0.62, respectively. The index shows that the agreement between actual and predicted rainfall is high for *MLP* and *RNN* ensemble models.

All of the mentioned statistical measures (Table 9) as well as mean absolute error (Table 6) in prediction of monsoon ascertain *MLP* model to be the best among all four proposed models.

5.4. Comparison of Results

5.4.1. Comparison with State-of-the-Art Methods. Proposed fuzzy clustering-based ensemble prediction models are compared with the models used by Indian Meteorological Department (*IMD*). It is compared with existing 16-parameter power regression model [4] and Rajeevan et al. [5] 8- and 10-parameter models. Test period of seven years from 1996 to 2002 is considered. *IMD* models give root mean square errors of 10.8%, 7.6%, and 6.4%, respectively. The *MR*, *MLP*, *RNN*, and *GRNN* ensemble models give 6.0%, 3.4%, 4.4%, and 5.5% root mean square errors, respectively, outperforming all three *IMD* models. The results are shown as a bar graph in Figure 7.

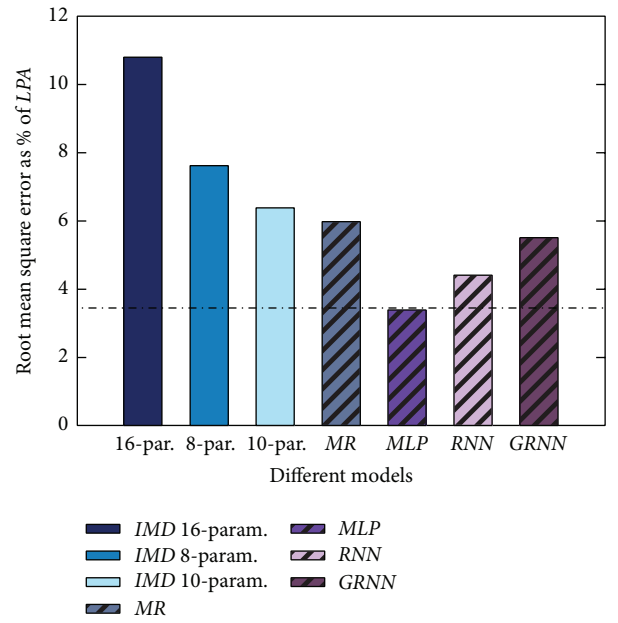

Figure 7: Comparison of *MR* (grey), *MLP* (purple), *RNN* (light purple), and *GRNN* (deep purple) models with *IMD* existing *16-param.* (deep blue), *10-param.* (blue), and *8-param.* (light blue) models for time period of 1996–2002 [4, 5]. Striped bars represent errors by our proposed models.

5.4.2. Improvement of Cluster-Based Models over Conventional Models. Ensemble model error obtained by combining all clusters' model output is compared with error obtained by same model (parameter), trained on the whole dataset without clustering. The mean absolute error for various models and predictor sets combinations are shown in Table 10. The result clearly depicts the improvement in prediction by clustering and ensemble method over nonclustered conventional method.

Table 11: Physical climatic events under study.

Climatic event	Number of years	Years associated with the event
Drought	13	1951, 1965, 1966, 1968, 1972, 1974, 1979, 1982, 1986, 1987, 2002, 2004, 2009
Flood	11	1953, 1956, 1958, 1959, 1961, 1964, 1970, 1975, 1983, 1988, 1994
El-Niño	23	1951, 1953, 1957, 1958, 1963, 1965, 1966, 1968, 1969, 1972, 1977, 1982, 1983, 1986, 1987, 1991, 1992, 1994, 1997, 2002, 2004, 2006, 2009
La-Niña	22	1950, 1954, 1955, 1956, 1964, 1970, 1971, 1973, 1974, 1975, 1984, 1985, 1988, 1989, 1995, 1998, 1999, 2000, 2007, 2008, 2010, 2011
Positive IOD	12	1957, 1961, 1963, 1967, 1972, 1977, 1982, 1983, 1994, 1997, 2006, 2007
Negative IOD	10	1958, 1960, 1964, 1971, 1974, 1975, 1989, 1992, 1993, 1996

Table 12: Threshold of support and confidence measures for associating obtained clusters with physical climatic events.

Predictor set	Support threshold	Confidence threshold
PredSet1	0.37	0.30
PredSet2	0.25	0.46
PredSet3	0.21	0.43
PredSet4	0.29	0.61
PredSet5	0.21	0.54

5.5. Prediction of the Year 2014. Annual Indian summer monsoon rainfall for the year of 2014 is 781.7 mm, which is 87.8% of *LPA* value. Proposed clustering-based ensemble *MR*, *MLP*, *RNN*, and *GRNN* models predict rainfall of 2014 as 96.1%, 80.3%, 80.0%, and 95.3% of *LPA*, respectively. Thus, proposed models show absolute error of 7.0% for forecasting rainfall of 2014.

6. Meteorological Analysis

Next, we try to visualize each cluster in terms of physical climatic events. The clusters obtained by fuzzy clustering are physically interpreted as being characterized by some global climatic events. The climatic events considered and studied during the time period 1948 to 2013 (period considered for clustering in our work) are El-Niño, La-Niña (http://ggweather.com/enso/oni.htm), positive and negative Indian ocean dipole (http://bom.gov.au/climate/IOD), drought, and flood, shown in Table 11.

Figure 8 shows the El-Niño and La-Niña years associated with drought, normal, and excess rainfall years during 1948–2013. The years having rainfall 10% above *LPA* are excess rainfall years and years having rainfall 10% below *LPA* are drought years. The El-Niño and La-Niña years are shown by color codes (*light green and green*) in the figure. The chart helps to visualize the cooccurrence of El-Niño and La-Niña events with extremities of *ISMR*.

6.1. Measuring Association between Climatic Events and ISMR. Support and confidence measures are considered to relate physical climatic event to the clusters generated by fuzzy clustering. They are defined below.

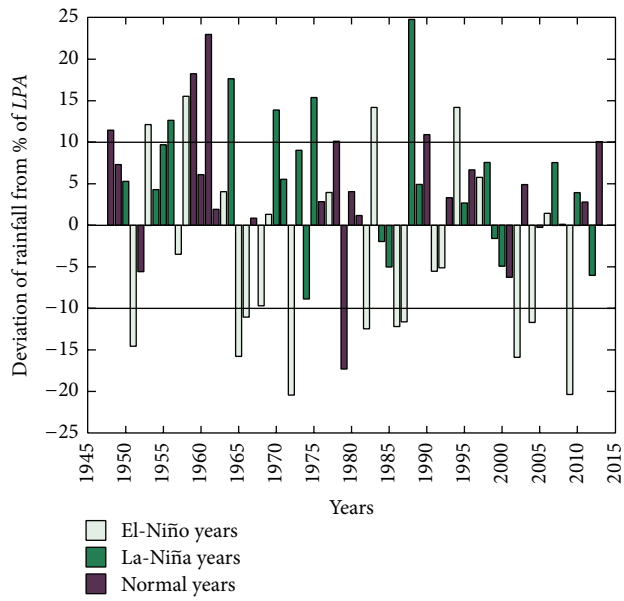

Figure 8: El-Niño (light green) and La-Niña (green) years association with drought (years below 10% of *LPA* rainfall), normal (years between +10% and −10% of *LPA* rainfall), and excess (years above 10% of *LPA* rainfall) years during period 1948–2013.

(i) *Support.* Support is defined as percentage of total number of years in the cluster corresponding to the climatic event:

$$\text{Support} = \frac{x_{ce}}{N}, \tag{12}$$

where x_{ce} denotes the number of years associated with a specific climatic event in the cluster and N is the total count of years in the cluster.

(ii) *Confidence.* Confidence is defined as percentage of years associated with the climatic event in the cluster to the total number of such event years:

$$\text{Confidence} = \frac{x_{ce}}{T_{ce}}, \tag{13}$$

where T_{ce} is the number of years associated with the climatic event during the period 1948–2013.

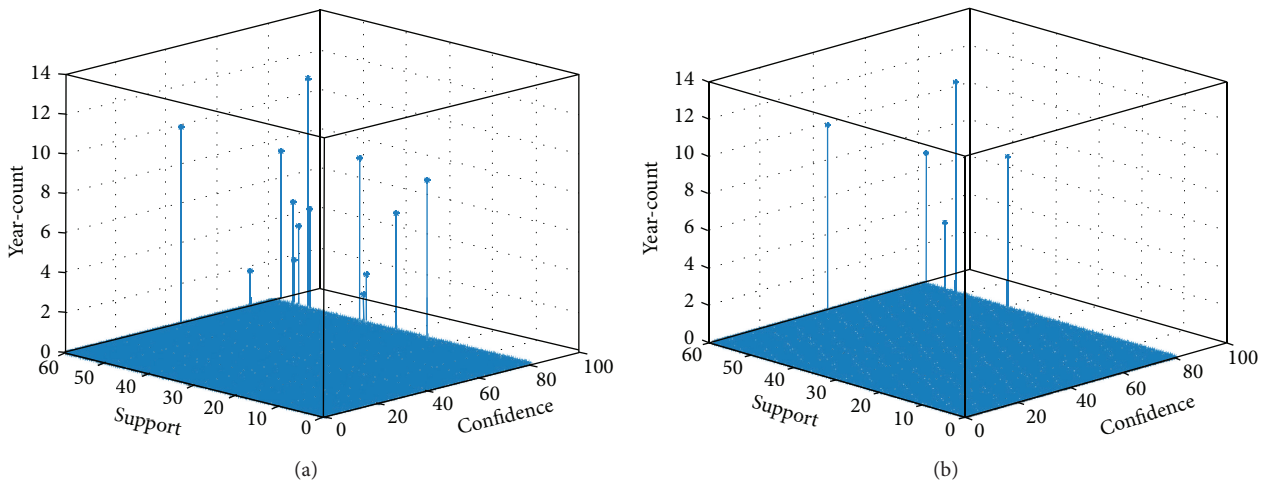

Figure 9: Histogram of the confidence and support measures as bins of year-count before (a) and after (b) thresholding for *PredSet1*.

Table 13: Identified physical climatic events being associated with clusters obtained by fuzzy clustering.

Predictor	Cluster1	Cluster2	Cluster3
PredSet1	Drought, El-Niño	La-Niña	La-Niña
PredSet2	Flood, La-Niña	Drought	Drought, El-Niño, La-Niña
PredSet3	El-Niño, positive IOD	Drought	Drought, El-Niño
PredSet4	La-Niña	Flood, La-Niña	Drought
PredSet5	—	Drought, El-Niño	Flood

We relate a cluster to a physical climatic event described in Table 11, if both support and confidence measures attain the corresponding thresholds. The thresholds are chosen in a way that 50% of years of study are under consideration. A low threshold compromises the importance of a climatic event being related to a particular cluster; on the other hand if even less number of years are taken, then threshold values should be high, which in turn will leave out most of the clusters. Therefore, as an optimal between the extremes, 50% of years are considered. Figure 9 shows histograms with confidence and support as bins of year-count for cases before and after threshold process, respectively, for predictors *PredSet1* (Table 2). The threshold values obtained for predictor sets are presented in Table 12. For each predictor set, we associate the clusters with physical climatic events, if they satisfy both support and confidence thresholds. The climatic events corresponding to cluster are shown in Table 13. Results establish coexistence of events of *La-Niña* and *flood*. It also puts light on high probability of occurrence of *El-Niño*, *drought*, and *positive IOD* events simultaneously.

7. Conclusion

Monsoon is an important phenomenon for economic development of agricultural-land like India. Large variability of monsoon over years makes prediction of rainfall a challenging task. The paper attempts to address this problem by clustering the years into similar groups and finally, multimodel ensemble forecast is provided for Indian summer monsoon rainfall.

Different climatic parameters with best correlated month value are identified and five different predictor sets are built for prediction of Indian monsoon. Four different models, namely, *MR*, *MLP*, *RNN*, and *GRNN*, are designed for each cluster exclusively. The final forecast is provided by weighted ensemble of forecasts by each cluster's model, where weight is considered as fuzzy membership of belongingness in each cluster. Multilayer perceptron ensemble model provides mean absolute error of 4.0% for prediction of annual rainfall, which is appreciable for forecasting complex monsoon process. Proposed fuzzy clustering-based ensemble approach surpasses the conventional approach. Performance of proposed clustering-based ensemble models is superior to existing *IMD*'s models [4, 5]. The error statistics also ascertain the superiority of multilayer perceptron model over other three proposed models. Lastly, in meteorological context the clusters are linked with global climatic events.

In the future, large number of climatic parameters influencing Indian monsoon can be explored and different predictor set can be used for different clusters of years to provide even better forecasting accuracy.

Conflict of Interests

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

Acknowledgment

This work is supported by RBU project through RESPOND program of ISRO through KCSTC, IIT Kharagpur.

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. T. Walker, "Correlation in seasonal variations of weather—IV, a further study of world weather," *Memoirs of the India Meteorological Department*, vol. 24, pp. 275–332, 1924.

[3] V. Thapliyal and S. M. Kulshrestha, "Recent models for long range forecasting of South-West monsoon rainfall in India," *Mausam*, vol. 43, no. 3, pp. 239–248, 1992.

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

[5] 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.

[6] M. Rajeevan, D. S. Pai, R. A. 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.

[7] J. Schewe and A. Levermann, "A statistically predictive model for future monsoon failure in India," *Environmental Research Letters*, vol. 7, no. 4, Article ID 044023, 2012.

[8] Q. Wu, Y. Yan, and D. Chen, "A linear markov model for east asian monsoon seasonal forecast," *Journal of Climate*, vol. 26, no. 14, pp. 5183–5195, 2013.

[9] K. Fan, Y. Liu, and H. Chen, "Improving the prediction of the east asian summer monsoon: new approaches," *Weather & Forecasting*, vol. 27, no. 4, pp. 1017–1030, 2012.

[10] F. Mekanik, M. A. Imteaz, S. Gato-Trinidad, and A. Elmahdi, "Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes," *Journal of Hydrology*, vol. 503, pp. 11–21, 2013.

[11] A. K. Sahai, M. K. Soman, and V. Satyan, "All India summer monsoon rainfall prediction using an artificial neural network," *Climate Dynamics*, vol. 16, no. 4, pp. 291–302, 2000.

[12] W.-C. Hong, "Rainfall forecasting by technological machine learning models," *Applied Mathematics and Computation*, vol. 200, no. 1, pp. 41–57, 2008.

[13] S. Chattopadhyay and G. Chattopadhyay, "Comparative study among different neural net learning algorithms applied to rainfall time series," *Meteorological Applications*, vol. 15, no. 2, pp. 273–280, 2008.

[14] N. Acharya, S. C. Kar, M. A. Kulkarni, U. C. Mohanty, and L. N. Sahoo, "Multi-model ensemble schemes for predicting northeast monsoon rainfall over peninsular India," *Journal of Earth System Science*, vol. 120, no. 5, pp. 795–805, 2011.

[15] V. R. Durai and R. Bhardwaj, "Improving precipitation forecasts skill over India using a multi-model ensemble technique," *Geofizika*, vol. 30, no. 2, pp. 119–141, 2013.

[16] B. Parthasarathy, A. A. Munot, and D. R. Kothawale, "Monthly and seasonal rainfall series for All-India homogeneous regions and meteorological subdivisions, 1871–1994," Tech. Rep. RR-065, Indian Institute of Tropical Meteorology, 1995.

[17] G. P. Compo, J. S. Whitaker, P. D. Sardeshmukh et al., "The twentieth century reanalysis project," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 654, pp. 1–28, 2011.

[18] 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.

[19] 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.