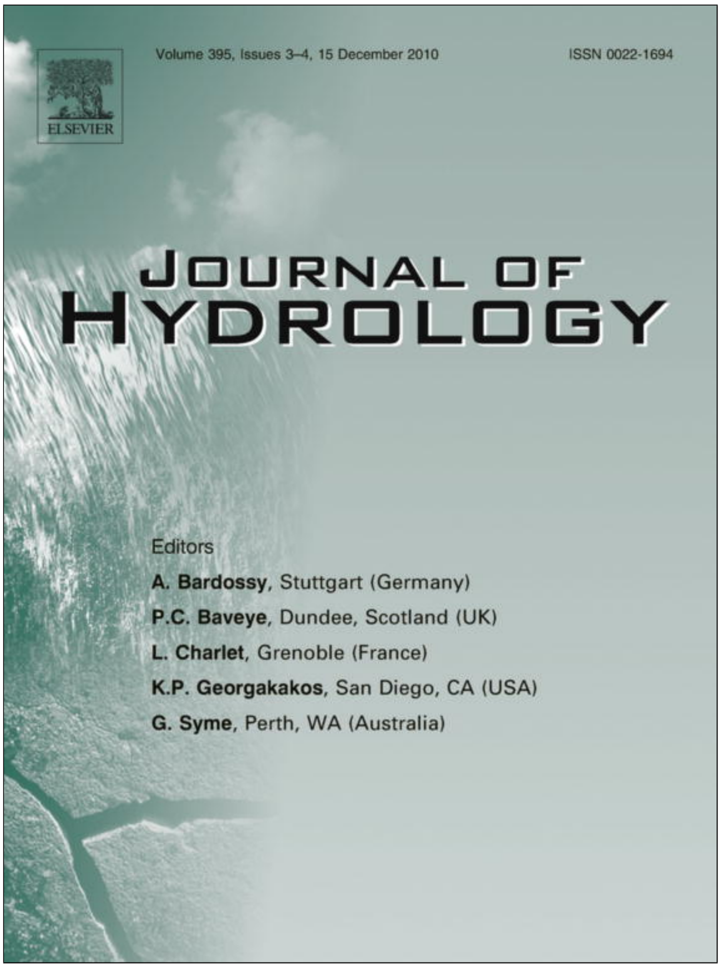

Volume 395, Issues 3–4, 15 December 2010 ISSN 0022-1694

JOURNAL OF HYDROLOGY

Journal of Hydrology 395 (2010) 190–198

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Forecasting of rainfall using ocean-atmospheric indices with a fuzzy neural technique

Gaurav Srivastava [a], Sudhindra N. Panda [b,*], Pratap Mondal [b], Junguo Liu [c]

[a] Department of Agricultural and Food Engineering, Indian Institute of Technology, Kharagpur 721 302, India
[b] School of Water Resources, Indian Institute of Technology, Kharagpur 721 302, India
[c] School of Nature Conservation, Beijing Forestry University, Qinghua East Road 35, Beijing Haidian District 100083, Beijing, China

ARTICLE INFO

Article history:
Received 11 March 2010
Received in revised form 20 July 2010
Accepted 20 October 2010

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Uwe Haberlandt, Associate Editor

Keywords:
Artificial Neural Network
Fuzzy-Ranking Algorithm
Sea Surface Temperature
Fuzzy c-mean clustering
Ocean-atmospheric predictor variables

SUMMARY

Forecasting of rainfall is imperative for rainfed agriculture of arid and semi-arid regions of the world where agriculture consumes nearly 80% of the total water demand. Fuzzy-Ranking Algorithm (FRA) is used to identify the significant input variables for rainfall forecast. A case study is carried out to forecast monthly rainfall in India with several ocean-atmospheric predictor variables. Three different scenarios of ocean-atmospheric predictor variables are used as a set of possible input variables for rainfall forecasting model: (1) two climate indices, i.e. Southern Oscillation Index (SOI) and Pacific Decadal Oscillation Index (PDOI); (2) Sea Surface Temperature anomalies (SSTa) in the 5° × 5° grid points in Indian Ocean; and (3) both the climate indices and SSTa. To generate a set of possible input variables for these scenarios, we use climatic indices and the SSTa data with different lags between 1 and 12 months. Nonlinear relationship between identified inputs and rainfall is captured with an Artificial Neural Network (ANN) technique. A new approach based on fuzzy c-mean clustering is proposed for dividing data into representative subsets for training, testing, and validation. The results show that this proposed approach overcomes the difficulty in determining optimal numbers of clusters associated with the data division technique of self-organized map. The ANN model developed with both the climate indices and SSTa shows the best performance for the forecast of the monthly August rainfall in India. Similar approach can be applied to forecast rainfall of any period at selected climatic regions of the world where significant relationship exists between the rainfall and climate indices.

1. Introduction

Climatic variability and its effect on human activity have been frequently discussed in literature (Fu et al., 2007; Khandekar and Neralla, 1984). One of the most crucial issues of global climatic variability is its effect on water resources. India's economy and societal infrastructures are finely tuned to the remarkable stability of the monsoon, with the consequence that vulnerability to small changes in monsoon rainfall is very high. In 2002, significant drop in the monsoon rains during July, results in a seasonal rainfall deficit of 19% and causes profound loss of agricultural production with a drop of over 3% in India's GDP. Hence, the prediction of the rainfall in Indian monsoon season remains an important concern (Parthasarathy et al., 1995). A number of researchers have attempted to forecast rainfall several months in advance using climatic indices such as Southern Oscillation Index (SOI), Pacific Decadal Oscillation Index (PDOI) and Northern Pacific Index (NPI) (Silverman and Dracup, 2000). The overall association of El Niño Southern Oscillation (ENSO), PDOI, and the Sea Surface Temperature (SST) with the rainfall patterns in different parts of India show broadly similar patterns (Roy et al., 2003). Both ENSO and PDOI have a significant negative effect on the rainfall in the peninsular region of India centered on the southeast coast and the northeastern region of India. On the other hand, the SST component exerts a positive influence in these two regions (Roy et al., 2003). These climatic indices and SST may play a vital role in rainfall forecast for India.

During rainy season between June and September, south-west monsoon contributes more than 75% of the annual rainfall of India (Singh, 2006). The perspective of Indian agriculture is very much dependent on onset/withdrawal of monsoon and depth of rainfall during rainy season. Average area under rainfed crops in India is about 75% of the total cultivable land (Roy et al., 2009). Rice is the major crop during rainy season cultivated both in rainfed and irrigated areas, covering nearly 90% of the total cultivated area. Critical growth stage of rainfed rice (includes booting, panicle initiation, flowering and milking) occurs between 49 (3rd August) and 78 days (1st September) after sowing of rice (16th June), which is

[*] Corresponding author. Tel.: +91 3222 282212.
E-mail addresses: gaurav.sun238@gmail.com (G. Srivastava), sudhindra.n.panda@gmail.com (S.N. Panda), water21water@yahoo.com (J. Liu).

0022-1694/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2010.10.025

normally coincides with the onset of effective monsoon (Panigrahi et al., 2002). Therefore, it is imperative to forecast the rainfall of August, which governs the rice production/productivity of Indian agriculture. Similar procedure can also be applied to forecast the rainfall of any other months.

Many researchers have already found the linkage between oceanic indexes and regional climate as well as weather. These oceanic indexes are measured as the difference of air pressure and Sea Surface Temperature (SST) between two places in an ocean. Changing the location of cold and warm water alters the path of the jet stream, which is responsible for rainfall as well as other meteorological parameters (Brabets and Walvoord, 2009). The empirical linkage between climatic indices (SOI, PDOI, and SST) and rainfall in India has been reasonably well simulated in various numerical models. The aforementioned indices can be applied to forecast rainfall of any months or seasons because climate indices are one of the best predictors of monsoon rainfall in India as practiced by the Indian Meteorological Department (IMD) (Rao, 1965; Shukla and Paolino, 1983).

Most of the research carried out in this area has used traditional statistical methods such as linear correlation or time series methods that develop relationship between climatic indices and rainfall in India (Maity and Kumar, 2007; Kumar et al., 2007). These methods assume a linear relationship between the independent variables and rainfall, whereas the relationships are likely nonlinear as the underlying processes are nonlinear. Apart from this, these methods are also not able to incorporate large number of input variables in the relationships because modeling such linear relationships with large number of input variables is computationally complex and lends on erroneous results. In the last few years, neural network and fuzzy system, whose applications are supported the universal functions approximation properties, have been widely used in nonlinear dynamic system modeling and forecasting (Cybenko, 1989; Wang and Mendel, 1992). The nonlinear modeling techniques such as Artificial Neural Network (ANN) and genetic algorithm avoid the need to reduce the input space in models. However, in practice, they tend to converge on local minima as the number of inputs and data points increase. In addition, these modeling techniques are computationally complex and normally require a lot of tuning to achieve good results. For these reasons, it is usually very difficult to use them to directly model nonlinear systems with a large number of inputs. Hence, it is necessary to reduce the size of the input space before modeling.

Linear techniques to reduce the number of input, such as principle component analysis (PCA) and partial least square cardinal components (PLS_CC) are well known, and they are computationally efficient and widely used. However, the significant input variables obtained from these techniques often fail to be used in nonlinear models like ANN (Lin et al., 1998). Partial mutual information (PMI) proposed by Sharma (2000), covers the nonlinear relationship between the rainfall and ocean-atmospheric variables and identifies the optimal combination of rainfall predictors. The main problem with PMI is that it works only as forward selection and the computation complexity increases with the number of possible input variable. For these reasons, Fuzzy-Ranking Algorithm (FRA) is a better option for nonlinear system rather than PCA or serial regression approaches.

Nonlinear relationship between the identified inputs and rainfall can be captured using the ANN. The ANN models perform best when they do not extrapolate beyond the extreme values of the data used for calibration (Minns and Hall, 1996; Tokar and Johnson, 1999; Liu et al., 2003). Consequently, in order to develop an accurate ANN model, the calibration data should contain all representative patterns that are present in the available data. For example, if the available data contains data samples (records) of extreme values that are excluded from the training set, the model cannot be expected to perform well, as the validation data will test the model's extrapolation ability, instead of its interpolation ability. When all of the patterns that are present in the available data are represented in the training set, the network trained with these training set shows the best generalization ability of the model. Thus the way that available data sets are divided into training, testing, and validation subsets, can have a significant influence on the performance of the neural network. In this study, three data division approaches are used, namely, random, self-organized map, and proposed fuzzy c-mean clustering approach.

In the present study, a new approach has been attempted to identify the significant predictors (input variables) for the forecast of the August rainfall in India. Initially Fuzzy-Ranking Algorithm (FRA) is used to identify the significant predictors for August rainfall in India; then three different approaches are applied to divide the data into representative subsets; finally, the August rainfall is forecasted with the identified inputs using an ANN technique. The ocean-atmospheric variables considered are climatic variability indices (i.e. SOI and PDOI), and Sea Surface Temperature anomaly (SSTa).

2. Data

SOI and PDOI were used to investigate the relationship between climatic variability and the precipitation in India. Similarly, SSTa data for $5^\circ \times 5^\circ$ grid points in the Indian Ocean were used in order to detect the possible effect of regional SST on precipitation. All the data used in this study are monthly based.

2.1. Rainfall in India

Rainfall in India has been recorded since January 1871. The monthly rainfall data used in this study (from the period of year 1901 to 2005) was obtained from the website of Indian Institute of Tropical Meteorology [http://www.tropmet.res.in/data.html]. Average annual rainfall of India is taken in the present study. A cubic root transformation was carried out in order to normalize the monthly rainfall data. Cube root transformation has very low impact on skewness and also it can represent the distribution of the original data set without any major change. The normalized data were standardized to a mean of zero and standard deviation of one, by subtracting the normalized mean and dividing by the normalized standard deviation for the period of year 1901–2005.

2.2. Southern Oscillation Index (SOI)

The SOI is an atmospheric see-saw process in the tropical Pacific sea level pressure between the eastern and western hemispheres associated with the El Niño and La Niña oceanographic features. The oscillation can be characterized by a simple index, or SOI. (Kawamura et al., 1998, 2002). The SOI was derived from monthly mean sea level pressure differences between Papeete, Tahiti (149.6°W, 17.5°S) and Darwin, Australia (130.9°E, 12.4°S). To calculate SOI, data on monthly mean sea level pressure were obtained for 140 years from January 1866 to December 2005 at Papeete and Darwin from Ropelewski and Jones (1987) and Allan et al. (1991).

2.3. Pacific Decadal Oscillation Index (PDOI)

PDOI is described as a long-lived pattern of Pacific climatic variability somewhat like El Niño. PDOI has two phases (warm and cool), and each phase persisted for 20–30 years in the 20th century. The PDOI data used in this study were obtained from the website of the Joint Institute for the Study of the Atmosphere and Ocean [http://tao.atmos.washington.edu/main.html].

2.4. Sea Surface Temperature anomalies (SSTa)

In this study data, SSTa were developed by Alexey Kaplan and his colleagues at the Lamont Doherty Earth Observatory (LDEO) of Columbia University in the USA (Kaplan et al., 1998). The Kaplan data on SSTa are based on the global Sea Surface Temperature record collected by the UK Meteorological Office, known as the MOHSST5 data in climatological literature. The available SSTa in the Indian Ocean (27.5°E and 32.5°S, 112.5°E and 32.5°S, 27.5°E and 32.5°N, 112.5°E and 32.5°N) for the period of January 1901–December 2005 were used for computation. The data were provided on the website of the International Research Institute for Climate Prediction [http://iri.columbia.edu/].

3. Methods

As the first step for developing a prediction model, FRA was applied to the possible input variables and the standardized August rainfall as described above to identify a significant input space for rainfall prediction. After identifying the significant input variables, they were utilized for forecasting August rainfall using ANN models.

3.1. Fuzzy-Ranking Algorithm (FRA)

Identification of significant input variables is one of the most important steps in the development of a prediction model. To capture the linear or nonlinear relationship between the model inputs and outputs, two-stage Fuzzy-Ranking Algorithm proposed by Lin et al. (1998) was used in this study. The fuzzy ranking process begins with the construction of fuzzy curves and surfaces for each input variable. Let for an output y there are n possible input variables, $x^1, x^2, \dots, x^n$. Each variable consists of M data points.

The single performance index for fuzzy curve (Pc^i) is given as

$$Pc^i = \frac{Py_c^i}{1 + Pv_c^i}, \tag{1}$$

where Py_c^i and Pv_c^i are the first stage and second stage performance indices for fuzzy curve respectively.

For fuzzy surface the single performance index ($Ps^{i,j}$) is defined as

$$Ps^{i,j} = \frac{Py_s^{i,j}}{1 + Pv_s^{i,j}}, \tag{2}$$

where $Py_s^{i,j}$ and $Pv_s^{i,j}$ are the first and second stage performance indices for fuzzy surface, respectively.

Once the fuzzy curves and surfaces have been generated, they are analyzed in order to determine which input variables are best able to predict the output variables. The FRA uses the performance index to rank the inputs. The performance index is a method that involves checking the mean square error between the fuzzy curve for the variable x^i and the output variable y. A small value of this performance index indicates that the variable is related to the output. A similar approach may also be taken for the fuzzy surfaces, which can also give information about whether the two variables are correlated. The FRA then normalizes the performance indices for the fuzzy curves and surfaces. This is carried out by computation of fuzzy curves and surfaces for a random variable generated by computer program. The performance index for the fuzzy curve of x^i is divided by the performance index of the simulated random variable in order to normalize it. Fig. 1 shows the flowchart of the FRA used in this study. The FRA applied in this study can be summarized in the following steps:

1. Add a test random variable R to the input set. Designate it as x^R.
2. Choose α, $0 < \alpha \leqslant 1$ (typically $0.99 < \alpha \leqslant 1$).
3. Generate fuzzy curve list and sort by their fuzzy curve performance index (Pc^i). The variable x^j with smallest valve of Pc^i is regarded as the most important input variable. Eliminate all variable other than the known random variable x^R, where $Pc^i/Pc^R > \alpha$ from additional consideration since they are apparently only randomly related to the output.
4. Use the most important variable from the last step, say x^j with remaining x^k, $k \neq j$, to generate fuzzy surface ($s^{i,j}$). The input variable x^m with the smallest fuzzy surface index ($Ps^{j,m}$) is regarded as the next most important. Eliminate all variable other than x^R where $Ps^{j,k}/Ps^{j,R} > alpha$ or $Ps^{j,k}/Pc^j > \alpha$ from additional consideration. X^m is selected for next significant variable.
5. Repeat step 4 until no more variables can be eliminated.

FRA was applied between August rainfall and three sets of inputs:

(a) Model (a): Climatic indices (SOI and PDOI) with lag 1–12 months.
(b) Model (b): SSTa with lag 1–12 months.
(c) Model (c): SOI, PDOI and SSTa with lag 1–12 months.

3.2. Data division approach

Three different approaches were followed for the division of data in training, testing, and validation sets for neural network.

1. Random approach.
2. Self-organized map (SOM) approach.
3. Proposed fuzzy c-mean clustering approach.

A new data division approach is proposed in this paper. The proposed data division approach is based on fuzzy c-means clustering. The fuzzy c-means clustering algorithm is based on the minimization of an objective function called c-means functional. It is defined by Dunn (1973) as:

$$J(X; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m \| x_k - v_i \|_A^2, \tag{3}$$

where $V = [v_1, v_2, v_3, \dots, v_c]$, $v_i \in R^n$ is a vector of cluster prototypes (centers), which have to be determined, and $D_{ikA}^2 = \|x_k - v_i\|_A^2 = (x_k - v_i)^T A (x_k - v_i)$ is a squared inner-product distance norm. Statistically, (3) can be seen as a measure of the total variance of x_k from v_i.

A fuzzy partition can be seen as a generalization of a hard partition, as it allows μ_{ik} attaining real values in [0, 1]. A $N \times c$ matrix $U = [\mu_{ik}]$ represents the fuzzy partitions; its conditions are given by:

$$\mu_{ij} \in [0, 1], 1 \leqslant i \leqslant N, 1 \leqslant k \leqslant c, \tag{4}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, 1 \leqslant i \leqslant N, \tag{5}$$

$$0 < \sum_{i=1}^{N} \mu_{ik} < N, 1 \leqslant k \leqslant c, \tag{6}$$

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} (D_{ikA}/D_{jkA})^{2/(m-1)}}, \tag{7}$$

where $1 \leqslant i \leqslant N$, $1 \leqslant k \leqslant c$, and

$$v_i = \frac{\sum_{k=1}^{N} \mu_{ik}^m x_k}{\sum_{k=1}^{N} \mu_{i,k}^m}, 1 \leqslant k \leqslant c, \tag{8}$$

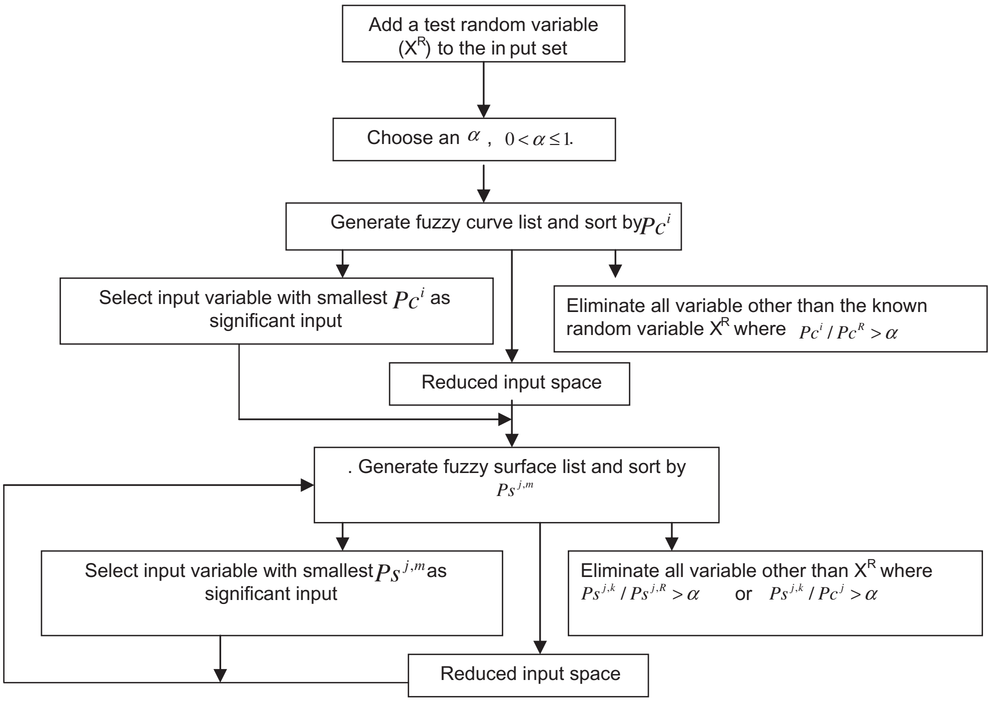
Add a test random variable (X^R) to the in put set
Choose an α , $0<\alpha\le 1.$
Generate fuzzy curve list and sort by Pc^i
Select input variable with smallest Pc^i as significant input
Eliminate all variable other than the known random variable X^R where $Pc^i / Pc^R > \alpha$
Reduced input space
. Generate fuzzy surface list and sort by $Ps^{j,m}$
Select input variable with smallest $Ps^{j,m}$ as significant input
Eliminate all variable other than X^R where $Ps^{j,k} / Ps^{j,R} > \alpha$ or $Ps^{j,k} / Pc^{j} > \alpha$
Reduced input space

Fig. 1. The flow diagram of Fuzzy-Ranking Algorithm.

where v_i is the cluster center. Once the clusters are formed the total information content is computed to identify the optimal numbers of clusters.

Let C_i^j be ith cluster at jth level (L_j). We measure the Net Information Gain (NIG) during the evolution from L_i to L_{i+1}. The gain or loss of information on cluster j from L_i to L_{i+1} is given by:

$$g_i = d_i \times M_i, \tag{9}$$

where d_i is the direction (increase or decrease); and M_i is the magnitude of change in information. If the offspring of cluster j overlap, information is deemed to have been lost and $d_i = -1$. In contrast, if offspring are clearly separated without overlap, information is deemed to have been gained and $d_i = 1$. The magnitude of information is measured using information theory.

$$M_j = -\sum_k p_k \ln p_k, \tag{10}$$

where k is the number of offspring of cluster j and p_k is the fraction of elements migrated from cluster j to k^{th} offspring. Total information content (I_i) is

$$I_i = \sum_{L_1}^{L_i} \sum_{j=1}^{i} g_j. \tag{11}$$

The level with largest information content is considered to be optimal and the number of cluster corresponding to that level is optimal. For optimal number of clusters the data set is divided into three subsets (training, testing, and validation subsets). For each cluster and each membership value interval (interval of 0.0–0.1; 0.1–0.2; ... ; 0.9–1) two data points (samples) are chosen, one is assign to testing set and the other one is assign to validation set. All the remaining samples are assigned to training set. If there are only two samples then one will be assigned to testing and the other one to training. In case there is only one sample then it has to be assigned to training set. This data division approach can be summarized in following steps:

1. Initial number of cluster is equal to 1.
2. The available data set are clustered using fuzzy c-mean clustering and the information content of the whole data set is computed.
3. Increase the number of clusters by 1 and repeat the step 2 until number of clusters reaches 50% of available data.
4. The level with maximum information content considered as being optimal and number of clusters corresponding to that level is optimal number of clusters.
5. For optimal number of clusters the data set is divided into three subsets (training, testing, and validation subsets). For each cluster and each membership value interval (interval of 0.0–0.1; 0.1–0.2; ... ; 0.9–1) two data points (samples) are chosen, one is assigned to testing set and the other one is assigned to validation set. All the remaining samples are assigned to training set. If there are only two samples then one will be assigned to testing and the other one to training. In case there is only one sample then it has to be assigned to training set.

3.3. Artificial Neural Network

Methods based on Artificial Neural Networks (ANN) have been used by several researchers in recent years in the area of precipitation estimation and forecasting. These are complex data driven tools that have been shown to act as "universal function approximators", and converge faster than other traditional approximators (Bishop, 1996; MacKay, 1992). In this study, nonlinear relationship between identified inputs and rainfall was captured using back-propagation ANN. A back-propagation network was developed using Matlab 7.1 for the rainfall forecasting.

3.4. Error statistics

Statistical parameters used to evaluate model forecasting against observed values were coefficient of determination (R^2) and root mean square error (RMSE) and average absolute percentage error (AAPE) which are defined below.

$$R^2 = \frac{\left[\sum_{i=1}^{N}\left(O_{\text{obs},i}-\bar{O}_{\text{obs}}\right)\left(O_{\text{fore},i}-\bar{O}_{\text{fore}}\right)\right]^2}{\left[\sum_{i=1}^{N}\left(O_{\text{obs},i}-\bar{O}_{\text{obs}}\right)^2\right]\left[\sum_{i=1}^{N}\left(O_{\text{fore},i}-\bar{O}_{\text{fore}}\right)^2\right]}, \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{obs},i}-y_{\text{fore},i}\right)^2}, \tag{13}$$

$$\text{AAPE} = \left|\left(1-\frac{1}{A_i}\sum_{k=1}^{K}\sum_{j=1}^{J}F_{ijk}W_{jk}\right)\right|, \tag{14}$$

where $O_{\text{fore},i}$ and $O_{\text{obs},i}$ are forecasted and observed values, respectively, for the i^{th} observation; N is the number of observations; and $\bar{O}_{\text{fore}}$ and $\bar{O}_{\text{obs}}$ are average forecasted and observed flows, respectively, for the prediction period; J is the number of forecast methods; K is the number of k-step-ahead forecasts; W_{jk} is the weight assigned to the k-step-ahead forecast generated by j forecasting method; F_{ijk} is the value of the k-step-ahead forecast generated by method j for time period i, and A_i is the actual value in time period i.

4. Results and discussion

4.1. Identification of significant inputs

Fuzzy-Ranking Algorithm (FRA) is applied to input sets as given in Section 3.1. The single stage fuzzy curve performance index values for Model (a) are shown in Table 1. In the first step after applying FRA criteria over initial input space SOI data with 4 months lag (SOI_{t-4}) corresponds to the smallest Pc^i/Pc^R was selected as first most significant input variable (Table 1). All the input variables having Pc^i/Pc^R greater than 1 were eliminated from the input space. Next step of FRA was applied to the remaining input space (Table 1).

Table 1
Fuzzy curve performance index values for Model (a).

Variables	Lag time (in months)	Pc_1^*	Pc_2^*	Pc_*	Pnc_*
SOI	4	0.938902	0.005917	0.933379	0.959045
PDOI	7	0.970981	0.033872	0.939169	0.964994
SOI	7	0.953178	0.013475	0.940505	0.966367
PDOI	11	0.950501	0.009172	0.941862	0.967761
SOI	8	0.967753	0.018341	0.950323	0.976455
PDOI	5	0.965512	0.013572	0.952584	0.978778
SOI	9	0.961516	0.00920	0.952751	0.97895
SOI	11	0.966011	0.012241	0.95433	0.980572
SOI	12	0.962723	0.004688	0.958231	0.98458
SOI	6	0.977675	0.019365	0.959101	0.985474
SOI	2	0.981535	0.022106	0.960307	0.986713
PDOI	3	0.972333	0.012075	0.960732	0.98715
SOI	10	0.975266	0.012757	0.962981	0.989461
SOI	5	0.98844	0.016974	0.971942	0.998668
Random		0.986409	0.013533	0.973238	1
PDOI	12	0.997469	0.023816	0.974266	1.001056
PDOI	4	0.981625	0.006947	0.974853	1.001659
PDOI	1	0.98764	0.012178	0.975757	1.002588
PDOI	10	0.985906	0.00883	0.977277	1.00415
SOI	3	0.981951	0.004552	0.977501	1.00438
PDOI	9	0.984656	0.006306	0.978486	1.005392
PDOI	6	0.982272	0.003583	0.978765	1.005679
PDOI	8	0.984293	0.004549	0.979836	1.006779
SOI	1	0.995594	0.013467	0.982365	1.009378
PDOI	2	0.994988	0.005044	0.989994	1.017217

Where Pc_1^* is the first stage fuzzy curve performance index; Pc_2^* is the second stage fuzzy curve performance index; Pc_* is a single performance index for fuzzy curve; and Pnc_* is a normalized single performance index for fuzzy curve.

Table 2
Fuzzy surface performance index values for reduced input space with SOI data with lag of 4 months for Model (a).

Variables	Lag time (in months)	Ps_1^*	Ps_2^*	Ps_*	Pns_*
PDOI	7	0.883133	0.041875	0.847638	0.982819
PDOI	3	0.898737	0.052629	0.853802	0.989966
SOI	5	0.908223	0.06051	0.856402	0.992981
SOI	6	0.894974	0.043808	0.857412	0.994152
SOI	7	0.897686	0.045048	0.85899	0.995981
PDOI	5	0.914017	0.059976	0.8623	0.999819
Random		0.920503	0.067304	0.862456	1
SOI	8	0.926093	0.072107	0.863807	1.001566
SOI	10	0.915365	0.050333	0.8715	1.010486
SOI	2	0.913727	0.048313	0.871617	1.010622
SOI	12	0.920664	0.047929	0.878556	1.018668
SOI	9	0.929132	0.055695	0.880114	1.020474
SOI	11	0.928127	0.048091	0.88554	1.026765
PDOI	11	0.954176	0.059135	0.900901	1.044576

Where Ps_1^* is the first stage fuzzy surface performance index; Ps_2^* is the second stage fuzzy surface performance index; Ps_* is a single performance index for fuzzy surface; and Pns_* is a normalized single performance index for fuzzy surface.

Table 3
Fuzzy surface performance index for reduced input space with PDOI data (lag 7 month) for Model (a).

Variables	Lag time (in months)	Ps_1^*	Ps_2^*	Ps_*	Pns_*
Random		0.916573	0.048482	0.87419	1
PDOI	3	0.920664	0.047929	0.878556	1.018668
SOI	5	0.929132	0.055695	0.880114	1.020474
SOI	6	0.928127	0.048091	0.88554	1.026765
SOI	7	0.908349	0.038102	0.875009	1.000937
PDOI	5	0.935446	0.03939	0.899995	1.029519

In the next step fuzzy surface performance index were computed. Table 2 shows fuzzy surface performance index corresponds to the remaining input variables after the first step for Model (a). PDOI data with 7 months lag (PDOI_{t-7}) corresponds to smallest $Ps^{j,k}/Ps^{j,R}$ was selected as the second most significant input variable. All the input variables having $Ps^{j,k}/Ps^{j,R}$ greater than 1 were eliminated from the input space.

Next, fuzzy surfaces were generated with the help of PDOI_{t-7} and the remaining input space. Table 3 shows the values of single performance index for fuzzy surface with PDOI_{t-7} data (lag 7 months).

At this stage none of the input space element satisfies the FRA criteria. Thus it means all the significant input variables have already being selected. So for August rainfall in India in Model (a), the significant input variables are SOI_{t-4} data and PDOI_{t-7} data. The same procedure was applied to identify the significant input variables of August rainfall in India for Model (b) and Model (c). The Identified significant input variables of August rainfall in India for these three models are shown in Table 4. Table 4 shows that for model (b) and (c) 16 and 14 input variables are identified as significant input variables. It shows only the first five significant input variables when the total number of significant inputs greater than 5. These identified inputs are utilized for the development of forecasting model using Artificial Neural Network.

4.2. Data division

4.2.1. Random approach

For this approach, 105 individual data points (rainfall and their corresponding significant input variables) were randomly divided into training, testing, and validation data sets. Seventy percent of

Table 4
Identified input variables (when the total number of identified inputs greater than 5, the first five identified inputs is shown).

Variable	Lag time (in months)	Location
Model (a) SOI and PDOI data with lag 1–12 months		
Total possible inputs = 24		
SOI	4	
PDOI	7	
Total identified inputs = 2		
Model (b) SSTa in Indian Ocean with lag 1–12 months		
Total possible inputs = 2280		
SSTa	1	22.5S, 32.5E
SSTa	2	32.5S, 87.5E
SSTa	3	32.5S, 72.5E
SSTa	9	32.5S, 32.5E
SSTa	4	22.5S, 57.5E
Total identified inputs = 16		
Model (c) SOI, PDOI and SSTa data with lag 1–12 months		
Total possible inputs = 2304		
SOI	4	
SSTa	1	22.5S, 32.5E
SSTa	2	32.5S, 87.5E
PDOI	7	
SSTa	6	17.5N, 37.5E
Total identified inputs = 14		

the input data points (73 data points) were used for training, 15% of the data points (16 data points) were used for testing and 15% of the data points (16 data points) were used for validation.

4.2.2. Self-organized map (SOM) approach

The self-organized map (SOM) was implemented for optimal data division using the MATLAB program. The inputs and their corresponding output of the predictive models were presented to the SOM as its input. A grid of 8 × 8 is chosen to ensure the maximum number of clusters are found from training data (Bowden et al., 2002). Thousand training iterations are used. The parameters of SOM are:

Learning parameter: lie between 0 and 1.

Start value = 0.9, End value = 0.1 and Decay function is Exponential.

σ for Gaussian neighborhood as percentage of map width: should lie between 0% and 100%.

Start value = 50%, End value = 1.0% and Decay function is Exponential.

Once the clusters are formed, three records from each cluster are sampled (one for each of training, testing, and validation). If a cluster contains only one record, then this record is placed in training set. If a cluster contains only two records, then one is placed in training and the other one is placed in the testing data set. After the selection for testing and validation data set all the remaining records are placed in the training data set.

This approach was applied to all the three models. As a result, in case of Model (a) total of 67 records were used as training, 25 were used as testing and 13 as validation. In case of Model (b) training set contains 70 records, testing contains 24, and validation contains 11 records. For Model (c) 64 records were used for training, 24 were used as testing, and 17 as validation set.

4.2.3. Fuzzy c-mean clustering approach

The proposed Fuzzy c-mean clustering approach for optimal division of data was applied to all three models by using a Fuzzy Clustering Toolbox for Matlab (http://www.fmt.vein.hu/softcomp/fclusttoolbox/). In case of Model (a) the information content of the data set was maximum, at level 16 with a value of 0.68. Thus the optimal number of clusters in case of Model (a) was equal to 16. As a result total 71 records were used as training, 19 were used as testing, and 15 were used as validation. For Model (b) the maximum information content was 0.56 at level 14. In this case the training contains 73 records, testing contains 18 records, and validation contains 14 records. For Model (c) the training set consists of 68 records, testing consists of 20 records, and validation consists of 17 records with information content of 0.62 at level 13.

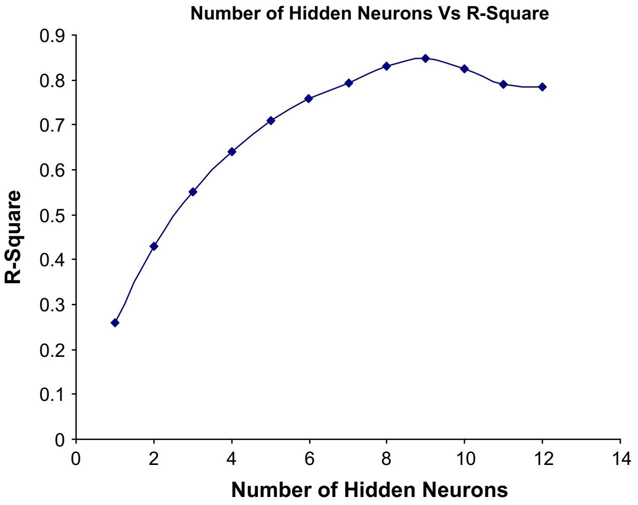

Fig. 2. Number of hidden neurons Vs. *R*-square values for Model (c) using FCM.

4.2.4. Forecast of August rainfall with ANN models

The numbers of nodes in input and output layers are fixed by the number of inputs and outputs, respectively. It is a common practice to fix the number of hidden layers in the network and then to chose the number of nodes in each of these hidden layers. It has been shown that only one hidden layer is required to approximate any continuous function, given that sufficient degree of freedom (i.e. connection weights) are provided (Cybenko, 1989). Hence only one hidden layer was utilized in this study and the number of hidden nodes increased by one at a time while computing the root mean square error (RMSE) for each network. When the reduction in the training RMSE becomes reasonably small, then number of hidden nodes are fixed. The identified inputs shown in Table 4 were used to develop a prediction model using ANN with 2 inputs for Model (a), 16 inputs for Model (b), and 14 inputs for Model (c). The networks were trained, tested and cross validated with different data sets (created by data division approaches described in Section 4.2). The Neural Network was created on the NNtool in Matlab 7.1. After the determination of the optimal network, cross validation with the validation set is employed to check the generalization ability of the model. Number of hidden neuron(s) against R^2 value, to find the optimal ANN structure during calibration of model (c) parameters using FCM, is shown in Fig. 2.

The coefficient of determination (R^2), RMSE and average absolute percentage error (AAPE) between observed and predicted rainfall for the training, testing, and validation sets for Models (a), (b) and (c) are given in Table 5. From this table, it can be seen that model (c) with proposed fuzzy c-mean clustering approach for data division showed the better performance than other two techniques. For self-organized map data division approach, Model (c) showed the better performance than other two techniques with RMSE ranging from 20.09 to 30.98 and AAPE ranging from 5.00% to 9.21%. The RMSE value for training sets in these three models has minimum value in Model (c) but the AAPE is the largest for training sets in Model (c).

Table 5
Statistical indices between observed and predicted rainfall for training, testing, and validation set for Models (a), (b) and (c).

Model	Data	Data division approach								
		Self-organized map			Random approach			Fuzzy c-mean clustering		
		R^2	RMSE (mm)	AAPE (%)	R^2	RMSE (mm)	AAPE (%)	R^2	RMSE (mm)	AAPE (%)
Model (a)	Training	0.7497	22.79	0.19	0.413	29.21	4.25	0.7693	17.29	2.25
	Testing	0.4416	36.42	8.58	0.3.456	37.15	9.19	0.4599	27.25	5.33
	Validation	0.3908	48.75	12.13	0.1901	49.87	15.74	0.4401	41.59	5.96
Model (b)	Training	0.695	20.64	1.97	0.507	23.14	3.36	0.726	20.46	2.29
	Testing	0.5239	33.91	5.7	0.291	38.3	7.12	0.5941	28.6	3.28
	Validation	0.356	39.72	9.27	0.1607	47.68	12.26	0.3637	27.68	2.14
Model (c)	Training	0.8288	20.09	9.21	0.569	23.65	10.13	0.8469	12.85	1.21
	Testing	0.5923	30.98	5.00	0.386	28.34	6.947	0.6038	28.34	3.81
	Validation	0.5406	30.47	6.05	0.3174	35.42	8.187	0.5669	29.92	5.01

Where RMSE is root mean square error; and AAPE is absolute average percentage error.

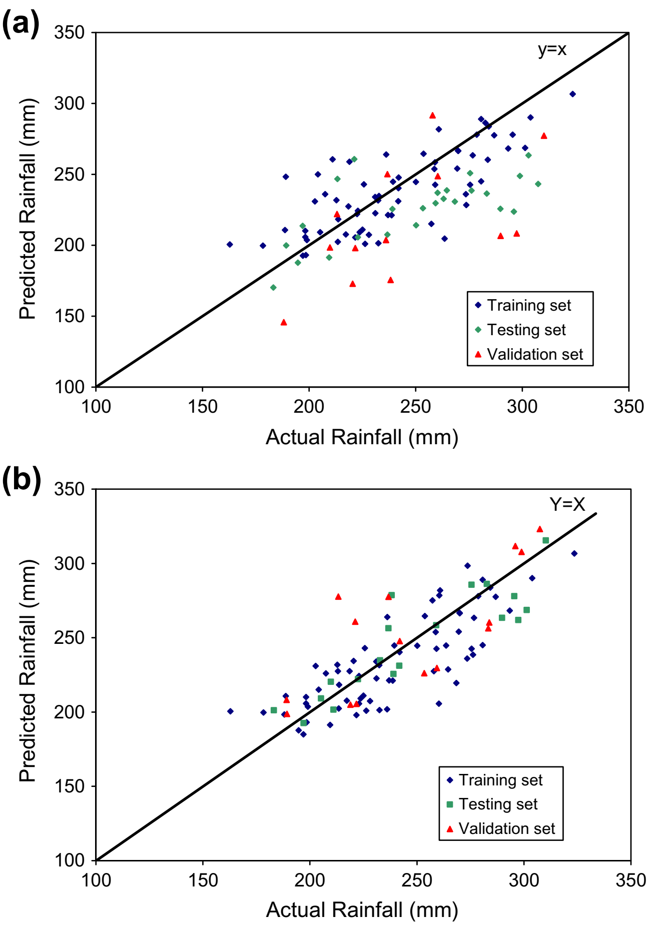

Fig. 3. 105-year forecasts for Model (a): (a) using SOM approach, (b) using fuzzy c-mean clustering approach.

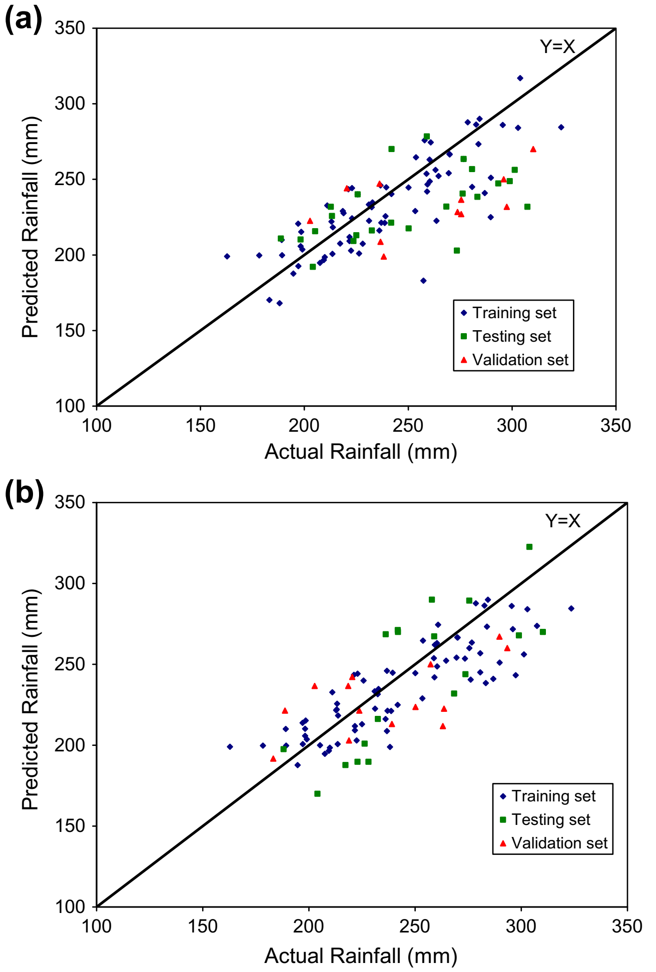

Fig. 4. 105-year forecasts for Model (b): (a) using SOM approach, (b) using fuzzy c-mean clustering approach.

Figs. 3a and b–5a and b shows the training, testing, and validation set forecast results for Model (a), Model (b), and Model (c) using SOM and proposed fuzzy c-mean clustering approach for data division, respectively. An ANN model developed for Model (a) using SOM data division approach was found overachieved for 170–220 mm range while for range of 280–330 mm the forecast results are underachieved as shown in Fig. 3. When an ANN model was developed for Model (c) using fuzzy c-mean clustering approach for data division, forecast shows more generalized results for validation set as the RMSE, AAPE are lowest for validation set and R^2 for validation set is highest than other models.

For proposed fuzzy c-mean clustering approach, Model (c) showed the best performance with RMSE ranging from 12.85 to 29.92 mm and AAPE ranging from1.21% to 5.01%. For Models (a) and (b), the proposed fuzzy c-mean clustering data division approach shows better performance than the data division approach of self-organized map.

For Model (a) RMSE ranges from 22.79 to 48.75 mm, and AAPE ranges from 0.19% to 12.13% for the data division approach of SOM, while for the proposed fuzzy c-mean clustering approach, RMSE ranges from 17.29 to 41.59 mm, and AAPE ranges from 2.25% to

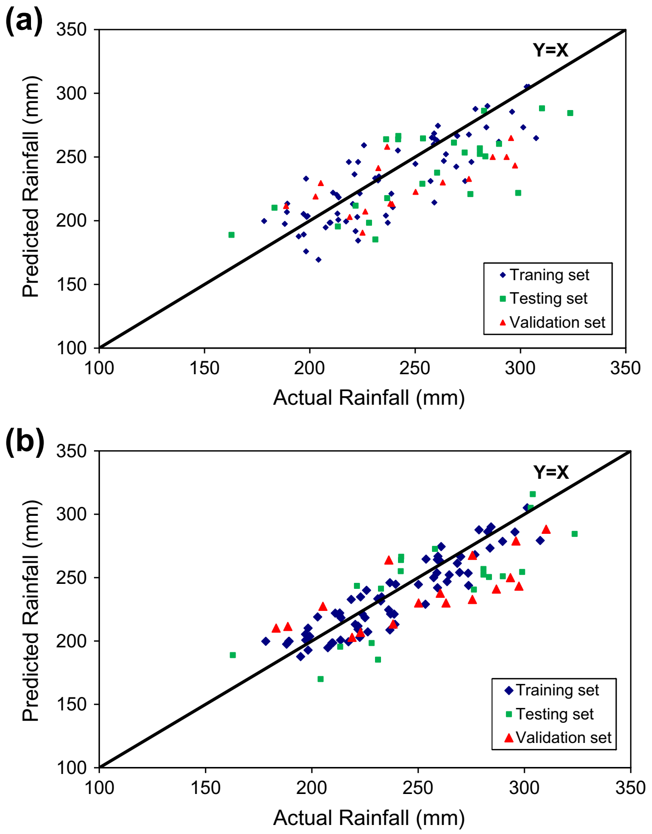

Fig. 5. 105-year forecasts for Model (c): (a) using SOM approach, (b) using fuzzy c-mean clustering approach.

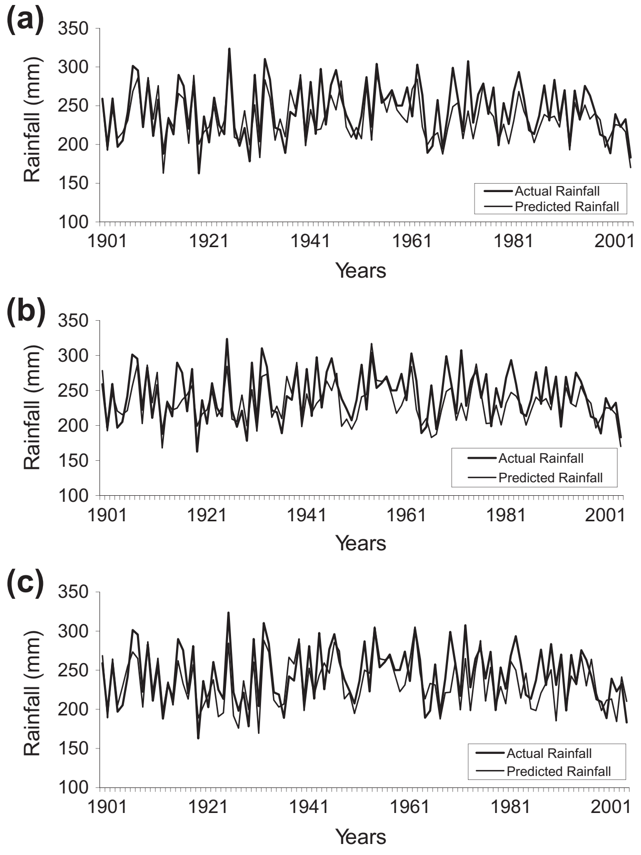

Fig. 6. 105-year forecast of August rainfall over India using SOM data division approach: (a) for Model (c); (b) for Model (b); and (c) for Model (a).

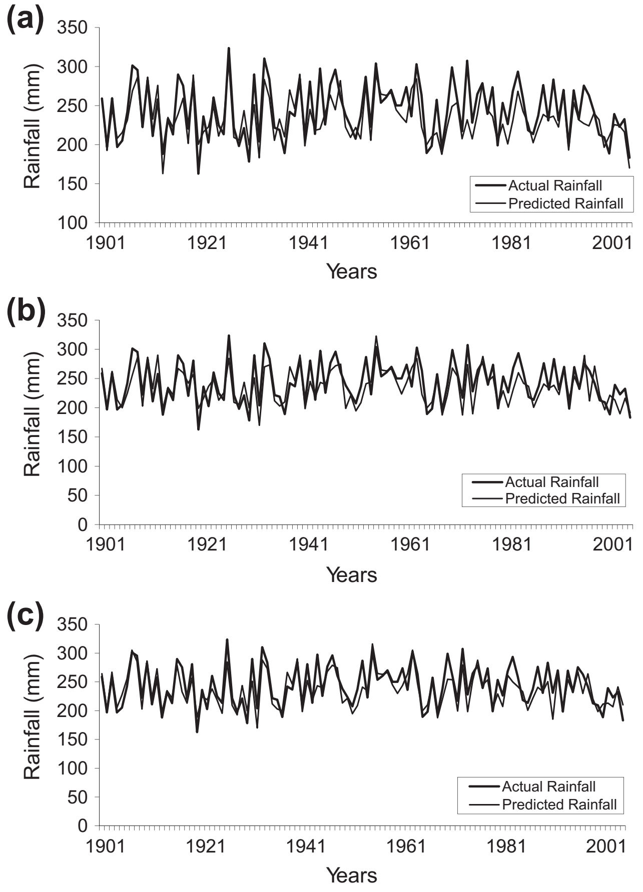

Fig. 7. 105-year forecast of August rainfall over India using Fuzzy c-mean clustering data division approach: (a) for Model (c); (b) for Model (b); and (c) for Model (a).

5.96%. Figs. 6 and 7 shows the 105-year forecast of August rainfall over India for Model (a), Model (b), and Model (c) using SOM and Fuzzy c-mean clustering data division approach, respectively. As shown in Figs. 6 and 7, in both of the data division approaches (SOM and Fuzzy c-mean clustering) predicted rainfall plot for Model (c) shows more generalized prediction pattern than the other models and based on model performance parameters (R^2, RMSE and AAPE), Model (c) provide better performance results.

5. Conclusions

For mean monthly August rainfall in India, regional Sea Surface Temperature or ocean climatic indices (SOI and PDOI) alone cannot give the best model description but combination of both of these is able to produce a better forecasting model. Model (c) (climatic indices and regional Sea Surface Temperature both considered as candidate predictors) showed the better performance with RMSE and AAPE ranges from 12.85 to 29.92 mm and 1.21% to 5.01%, respectively, when fuzzy c-mean clustering approach was used for data division in training, testing, and validation subsets. In this case, Model (c) with R^2 for validation set equal to 0.5669, showed a better generalization than the other models.

In case of random data division approach, there is a fair chance that all the data points used for training the network contains same information about the input–output relationship, whereas testing and validation set contains different information. Because the data division is completely random there is no guarantee that training set contains all the information about input–output relationship present in historical sample. While SOM and FCM both tries to capture all the information about input–output relationship in training set so that trained ANN model is became more generalized model. The main drawback of SOM approach is that there is no systematic procedure to get the optimal numbers of clusters. It is completely based on trial and error method, because of which it is possible that the clusters created by the SOM may contains some outliers, which will result into poor forecasting results. Whereas FCM approach provides a systematic procedure to determine the optimal number of clusters based on the maximization of information, which resolves the outlier problem and provide an optimal data division.

References

Allan, R.J., Nicholls, N., Jones, P.D., Butterworth, I.J., 1991. Further extension of the Tahiti-Darwin SOI, Early ENSO events and Darwin pressure. Journal of Climate 4, 743–749.

Bishop, C.M., 1996. Neural Networks for Pattern Recognition. Oxford University Press.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. Water Resources Research 38 (2), 1–11.

Brabets, T.P., Walvoord, M.A., 2009. Trends in streamflow in the Yukon River Basin from 1944 to 2005 and the influence of the Pacific Decadal Oscillation. Journal of Hydrology 371, 108–191.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control Signals and Systems 2, 303–314.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters. Journal of Cybernetics 3, 32–57.

Fu, G., Charles, S.P., Viney, N.R., Chen, S., Wu, J.Q., 2007. Impacts of climate variability on stream-flow in the Yellow River. Hydrological Processes 21, 3431–3439.

Kaplan, A., Cane, Y., Kuhsnir, A., Clement, A., Blument, M., Rajagopalan, B., 1998. Analyses of global sea surface temperature 1856–1991. Journal of Geophysical Research 103, 567–589.

Kawamura, A., McKerchar, A.I., Jinno, K., Eguchi, S., 2002. Statistical characteristics of Southern Oscillation Index and its barometric pressure data. Journal of Hydroscience and Hydraulic Engineering 20 (2), 41–49.

Kawamura, A., McKerchar, A.I., Spigel, R.H., Jinno, K., 1998. Chaotic characteristics of the Southern Oscillation Index time series. Journal of Hydrology 204, 168–181.

Khandekar, M.L., Neralla, V.R., 1984. On the relationship between the sea surface temperatures in the equatorial Pacific and the Indian monsoon rainfall. Geophysical Research Letter 11, 1137–1140.

Kumar, D.N., Reddy, M.J., Maity, R., 2007. Regional rainfall forecasting using large scale climate teleconnections and artificial intelligence techniques. Journal of Intelligence System, Freund & Pettman 16 (4), 307–322.

Lin, Y., Cunningham, G.A., Coggeshall, S.V., Jones, R.D., 1998. Nonlinear system input structure identification: two stage fuzzy curves and surfaces. IEEE Transaction on Systems, Man and Cybernetics-Part A: Systems and Humans 28 (5), 678–684.

Liu, J., Savenije, H.H.G., Xu, J., 2003. Forecast of water demand in Weinan City in China using WDF-ANN model. Physics and Chemistry of the Earth 28, 219–224.

Mackay, D.J.C., 1992. Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. Neural Computation 4.

Maity, R., Kumar, D.N., 2007. Hydroclimatic teleconnection between global sea surface temperature and rainfall over India at subdivisional monthly scale. Hydrological Processes 21 (14), 1802–1813.

Minns, A.W., Hall, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydrological Sciences Journal 41 (3), 399–417.

Panigrahi, B., Panda, S.N., Mull, R., 2002. Prediction of hydrological events for planning rainfed rice. Hydrological Sciences Journal 47 (3), 435–448.

Parthasarathy, B., Kothawale, D.R., Munot, A.A., 1995. Monthly and Seasonal Rainfall Series for All-India Homogeneous Regions and Meteorological Subdivisions, 1871–1994. Indian Institute of Tropical Meteorology, Pune, India.

Rao, K.N., 1965. Seasonal forecasting—India. WMO Technical Note No. 66, pp. 17–130 [WMO No. 162. TP. 79].

Ropelewski, C.F., Jones, P.D., 1987. An extension of the Tahiti-Darwin Southern Oscillation index. Monthly Weather Review 115, 2161–2165.

Roy, D., Panda, S.N., Panigrahi, B., 2009. Water balance simulation model for optimal sizing of on-farm reservoir in rainfed farming system. Computers and Electronics in Agriculture 65, 114–124.

Roy, S.S., Goodrich, G.B., Balling, R.C., 2003. Influence of El Niño/southern oscillation, Pacific decadal oscillation, and local sea-surface temperature anomalies on peak season monsoon precipitation in India. Climate Research 25, 171–178.

Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 1—a strategy for system predictor identification. Journal of Hydrology 239, 232–239.

Shukla, J., Paolino, D.A., 1983. The southern oscillation and long range forecasting of the summer monsoon rainfall over India. Monthly Weather Review 111, 1830–1837.

Silverman, D., Dracup, J.A., 2000. Artificial Neural networks and long-lead precipitation prediction in California. Journal of Applied Meteorology 39, 57–66.

Singh, C.V., 2006. Pattern characteristics of Indian monsoon rainfall using principal component analysis (PCA). Atmospheric Research 79, 317–326.

Tokar, A.S., Johnson, P.A., 1999. Rainfall-runoff modeling using artificial neural networks. Journal of Hydrologic Engineering 4 (3), 232–239.

Wang, l., Mendel, J.M., 1992. Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. IEEE Transaction on Neural networks 5, 807–814.