


## RESEARCH ARTICLE

# Long-range forecast of Indian summer monsoon rainfall using an artificial neural network model

Rajashree Acharya  | Jayanti Pal | Debanjana Das | Sutapa Chaudhuri

Department of Atmospheric Sciences, University of Calcutta, Kolkata, India

### Correspondence

Sutapa Chaudhuri, Department of Atmospheric Sciences, University of Calcutta, 51/2, Hazra Road, Kolkata—700 019, India.  
Email: sutapa.chaudhuri@gmail.com

This study develops an artificial neural network (ANN) model with a nonlinear perceptron rule for use in the long-range forecasting (LRF) of Indian summer monsoon rainfall (ISMR). In developing the model, two predictor sets are adopted from the India Meteorological Department (IMD), SET-I and SET-II, to prepare the input matrix of the model, while the output is ISMR. The data used were collected over the period 1980–2017. The model is trained with input data from 1980 to 2012, and the skill of the model is estimated by validating the model output with observation during the period 2013–2017. The result reveals that that second-stage forecast is better than first-stage forecast due to the incorporation of a North Atlantic sea surface pressure anomaly and a North Central Pacific zonal wind anomaly at 850 hPa in the input matrix. The study further reveals that the multilayer perceptron (MLP) model with a back-propagation algorithm is best among the ANN models used in the study. The prediction capability of the ANN model is also checked by comparing it with a multiple nonlinear regression (MNLR) model developed with the two predictor sets. The robustness of the prediction accuracy is estimated by computing Willmott's index for each of the ANN and MNLR models.

### KEYWORDS

ANN, ISMR, LRF, MNLR, Willmott's index

## INTRODUCTION

The Indian economy is largely dependent on its agricultural production, which itself is greatly dependent on the Indian summer monsoon rainfall (ISMR) during June–September. About 65% of the total cultivated land in India is under the influence of a rainfed agricultural system (Swaminathan, 1998). The precise long-range forecast (LRF) of ISMR is, therefore, very useful for better planning of finance, power and water resources. The large variability at different spatial (districts, states and so on) and temporal (monthly, bimonthly and so on) scales makes it difficult to predict ISMR accurately with sufficient lead time. Researchers have used various approaches to study and predict seasonal and intra-seasonal rainfall and the associated variability (Walker, 1914, 1923; Gowardker *et al.*, 1989, 1991; Navone and Ciccato, 1994; Guhathakurta *et al.*, 1999; Rajeevan *et al.*,

2000; Rajeevan and McPhaden, 2004; Guhathakurta, 2006; Pai and Rajeevan, 2006; Chaudhuri *et al.*, 2015). The operational forecast of the India Meteorological Department (IMD) was initially based on the 16 parameters for power regression and parametric models (Gowardker *et al.*, 1989, 1991). The forecasts issued during these periods were qualitatively correct. However, the mean forecast errors were more than the mean errors of the forecasts based on climatology alone. Ineffective prediction in 2002 and 2004 of ISMR turned out to be a serious concern among the meteorological community of India (Gadgil *et al.*, 2005). The failure in the forecast of ISMR for 2002 necessitated a critical evaluation on the 16-parameter power regression and parametric models. In 2003, two new models (8 and 10 parameter models) were introduced for the operational forecast by the IMD. Further, a two-stage forecasting strategy was also adopted with the provision for a forecast update by the end

of June or the first week of July (Rajeevan and McPhaden, 2004). Forecasts for the seasonal ISMR for the country as a whole are issued in two stages according to this new strategy. The first-stage forecast is issued in mid-April and an update or second-stage forecast is issued by the end of June. The 2003 operational forecasts for ISMR based on these new models were found to be correct. The IMD further explored new statistical methods with the objective for an improved forecast. Researchers tried to apply an artificial neural network (ANN) in the ISMR forecast. Various kinds of neural networks are currently available. However, only a feed-forward neural network (FFNN) and a back-propagation neural network (BPNN) are usually used in an ISMR forecast (Sahai *et al.*, 2000; Chakraverty and Gupta, 2007; Aksoy and Dahamshe, 2008; Singh and Borah, 2013). The neural network technique has a strong potential for pattern recognition and signal processing problems, and it has the ability to predict (Bose and Liang, 1995; Gardner and Dorling, 1998; Haykin, 1999; Chaudhuri, 2010; Chaudhuri *et al.*, 2013, 2014). Venkatesan *et al.* (1997) used the neural network technique to predict the monsoon rainfall of India by using a few predictors to compare the results with linear regression techniques. They found that the model based on the neural network technique performed better.

The present study uses the neural network technique to develop an LRF model for a two-stage forecast of ISMR using nine predictors. These predictors have also been used by the IMD in its operational forecast. The present study uses six predictors in each stage forecast following the work of the IMD. First, three predictors are common in both stages of the forecast, while others are different. Since some of these predictors are intercorrelated, principal component analysis (PCA) is implemented for the two sets of data. The PCA helped to select the pertinent inputs for the ANN model. The skill of the ANN model is checked with a multiple nonlinear regression (MNL) model developed using the same predictor sets. The consistency of the prediction accuracy is estimated through Willmott's index (Willmott, 1982) for both ANN and MNL models.

## 1 | DATA AND METHODOLOGY

### 1.1 | Data

The purpose of the present study is the LRF of accumulated rainfall (mm) for all India during the summer monsoon season. The ISMR series of the IMD used in this study was constructed from a network of around 2000 rain gauges distributed all over India; the data used are from the period 1980–2017. The data set is separated into two parts for modelling and simulation purposes: (a) a training data set for the period 1980–2012; and (b) a test data set for the period 2012–2017. Two sets of data are used in this study. The first set (Table 1) is used for the first-stage forecast issued in mid-April; and the second set (Table 2) is used for the second-stage forecast issued by the end of June. The monthly data set of various parameters is used, with predictors such as sea surface temperature (SST), mean sea level pressure (MSLP), zonal wind at a 850 hPa pressure level, station level surface air temperature, and monthly mean warm water volume (WWV) over the Pacific Ocean (Meinen and McPhaden, 2000; McPhaden, 2003; Rajeevan and McPhaden, 2004). The SST data obtained from the monthly National Oceanic and Atmospheric Administration (NOAA) Extended Reconstructed Global Sea Surface Temperature version 2 (ERSST.v2) are  $2^\circ \times 2^\circ$ , latitude–longitude grid resolution (Smith and Reynolds, 2004), which were obtained from the National Climatic Data Centre (NCDC) in Asheville, NC, USA (<http://www.ncdc.noaa.gov/oa/climate/research/sst/sst.html>). The monthly surface sea level pressure and 850 hPa zonal wind data of the NCEP/NCAR reanalysis (Kalnay *et al.*, 1996) were used from the NOAA-Cooperative Institute for Research In Environmental Sciences (CIRES) Climate Diagnostics Centre in Boulder, CO, USA (<http://www.cdc.noaa.gov/>). The spatial resolution of these data is  $2.5^\circ \times 2.5^\circ$  latitude and longitude. The WWV data were available on real time at [http://www.pmel.noaa.gov/tao/el\\_nino/wwv/](http://www.pmel.noaa.gov/tao/el_nino/wwv/), which was based on upper ocean temperature field analysis (McPhaden *et al.*, 1998). The monthly land–surface–air temperatures of five stations from Europe are another

TABLE 1 Details of predictors SET-1

Number	Parameter	Period	Spatial domain
A1	North Atlantic sea surface temperature anomaly (NASSTA)	December–January	20 °N–30 °N, 100 °W–80 °W
A2	Equatorial South East Indian Ocean sea surface temperature anomaly (ESEIOSSTA)	February–March	20S–10 °S, 100 °E–120 °E
A3	East Asia surface pressure anomaly (EASPA)	February–March	35 °N–45 °N, 120 °E–130 °E
A4	Europe surface–land–atmospheric temperature anomaly (ESLATA)	January	Five stations
A5	North West Europe surface pressure anomaly tendency (NWESPAT)	December–February (0)— September–November (–1)	65 °N–7 °S, 20 °E–40 °E
A6	WWV anomaly	February–March	5 °S–5 °N, 120 °E–80 °W

TABLE 2 Details of predictors SET-2

Number	Parameter	Period	Spatial domain
J1	North Atlantic sea surface temperature anomaly (NASSTA)	December–January	20 °N–30 °N, 100 °W–80 °W
J2	Equatorial South East Indian Ocean sea surface temperature anomaly (ESEIOSSTA)	February–March	20 °S–10 °S, 100 °E–120 °E
J3	East Asia surface pressure anomaly (EASPA)	February–March	35 °N–45 °N, 120 °E–130 °E
J4	NINO3.4 SST anomaly tendency	March–May (0)—December–February (0)	Five stations
J5	North Atlantic sea surface pressure anomaly	May	65 °N–75 °N, 20 °E–40 °E
J6	North Central Pacific zonal wind anomaly at 850 hPa (NCPZWA-850)	May	5 °S–5 °N, 120 °E–80 °W

predictor obtained from *World Climatic Data* published by the NCDC at <https://www.ncdc.noaa.gov/sotc/global/201513> (Rajeevan *et al.*, 2007). These stations were Orland, Oslo/Gendermon, Ostursund/Froson, Karlstad and De Bilt. The monthly mean NINO3.4 index data were used as another predictor, which was obtained from the Climate Prediction Centre, NOAA (<http://www.cpc.ncep.noaa.gov/>). Details of the predictors in SET-I and SET-II are summarized in Tables 1 and 2. These include the time period and spatial domain for each predictor.

## 1.2 | Methodology

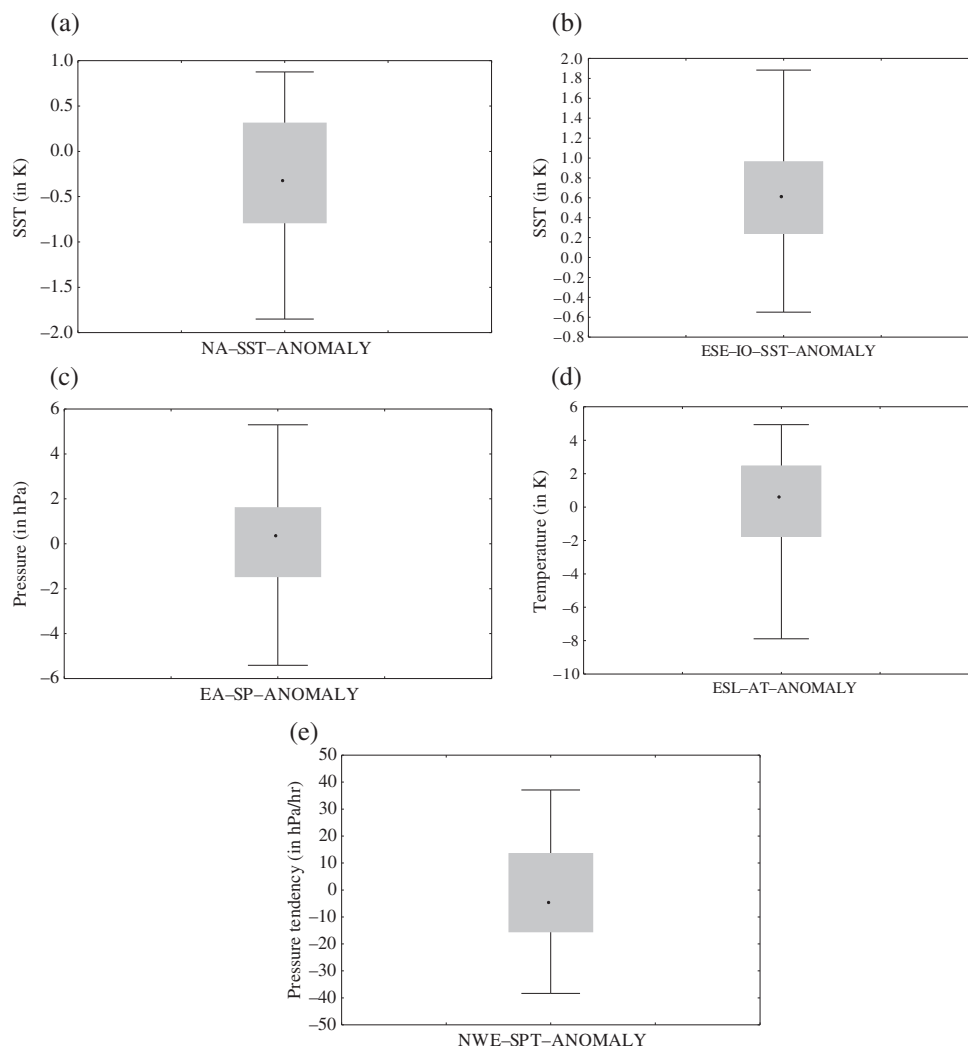
The predictors were derived from the raw data collected from different sources. The SST predictors were derived as the simple arithmetic average of the monthly ERSST.v2 anomalies over the respective geographical region. The periods used for averaging are given in Tables 1 and 2. The pressure and wind predictors were similarly derived from the NCEP reanalysis data. The land–surface–air temperature anomaly over North West Europe was computed as the average of the surface air temperature anomalies of the five land stations in Europe (Orland, Oslo/Gendermon, Ostursund/Froson, Karlstad and De Bilt). The seasonal tendency in the NINO3.4 anomaly index was computed by subtracting monthly anomalies averaged over the winter, December–February, from that averaged over the spring season, March–May. The WWV anomaly over the Pacific (February–March) was derived from the monthly averages of the WWV computed between 5 °N and 5 °S integrated across the Pacific basin, including all ocean areas

between 80 °W and 120 °E (Rajeevan and McPhaden, 2004). The lower boundary for the integration is the depth of the 20°C isotherm, which is located in the middle of the upper thermocline. All anomalies were computed using the climatological base period 1980–2017. The intercorrelation between the predictor and its variability was assessed to understand their characteristics (Table 3 and Figures 1–4). Owing to the presence of the interrelation between predictors, principal component analysis (PCA) was implemented (Figures 5 and 6). Hsieh and Tang (1998) have recommended the PCA method as a prefilter for compressing the large spatial data field in order to train the neural network. PCA is a useful tool for transforming the original variables into a set of uncorrelated variables. It consists of finding an orthogonal transformation of the original variables to a new set of uncorrelated variables, called principal components or empirical orthogonal functions, which are derived in decreasing order of importance. The components are linear combinations of the original variable, but are orthogonal and uncorrelated to each other. Another advantage of the PCA is that it reduces the dimensionality of the data as the first few components will account for most of the variation in the original data, and by truncating the insignificant modes, the noise level also is reduced. PCA was carried out for 38 years. It was analysed based on a correlation between ISMR and the components along with the variable contribution to the components (Table 3). Based on this analysis, a suitable predictor was chosen by evaluating its physical linkage and degree of relationship with the summer monsoon rainfall. Thereafter, those predictors were used as an

TABLE 3 Correlation between different predictors

	A1/J1	A2/J2	A3/J3	A4	A5	A6	J4	J5	J6	ISMR
A1/J1	1.00	−0.24	−0.08	−0.12	−0.14	−0.04	0.21	−0.02	0.30	−0.18
A2/J2		1.00	0.22	<b>0.46</b>	0.00	−0.30	<b>−0.36</b>	−0.06	<b>−0.52</b>	<b>0.33</b>
A3/J3			1.00	0.05	0.07	0.10	−0.21	−0.19	−0.07	<b>0.36</b>
A4				1.00	<b>−0.60</b>	−0.07	0.05	−0.23	<b>−0.43</b>	0.30
A5					1.00	0.04	−0.09	0.34	0.04	−0.22
A6						1.00	<b>0.53</b>	−0.02	0.25	−0.28
J4							1.00	−0.22	0.09	−0.29
J5								1.00	0.00	<b>−0.38</b>
J6									1.00	<b>−0.43</b>
ISMR										1.00

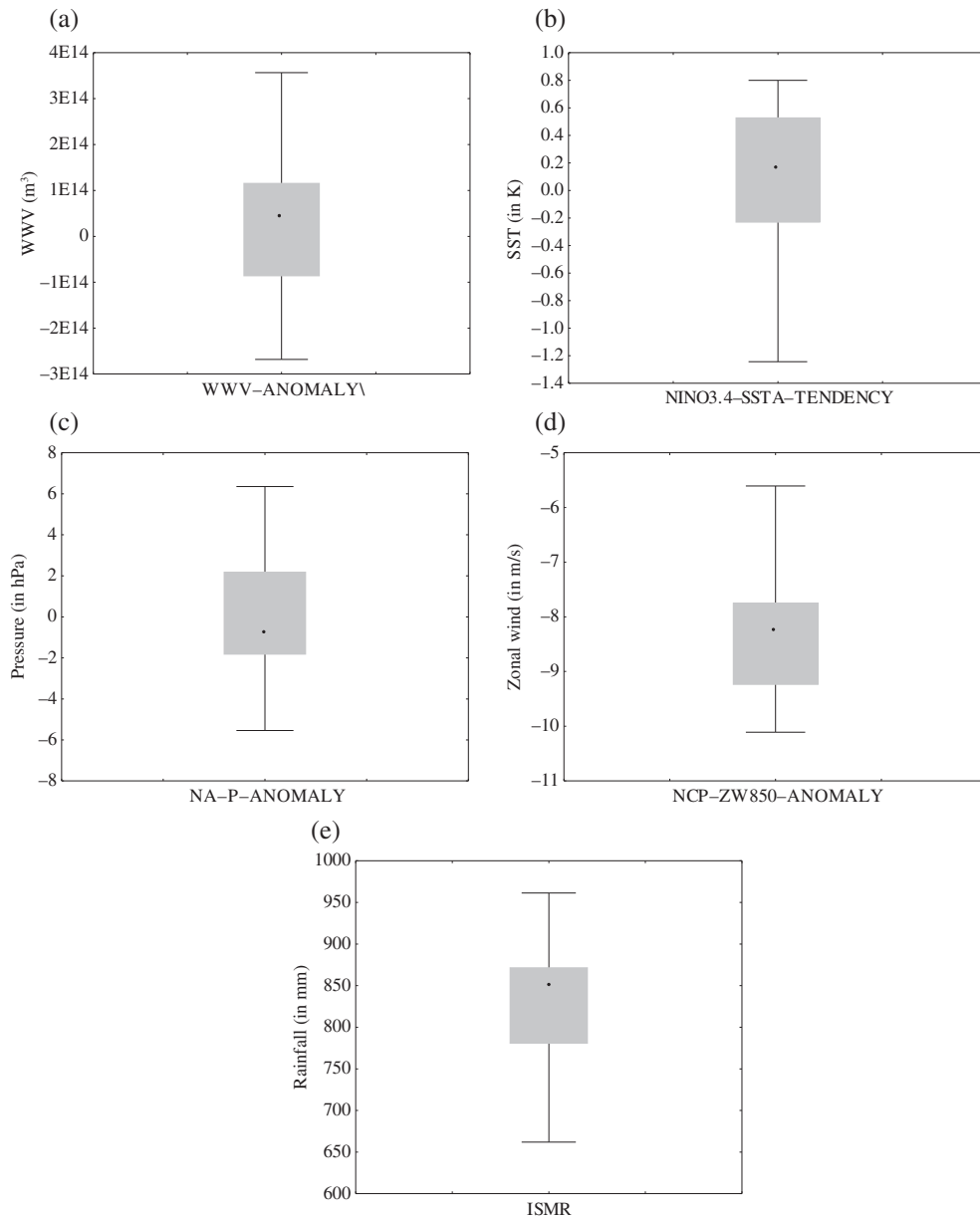
Notes: Bold values are statistically significant at the 95% confidence level.  
ISMR: Indian summer monsoon rainfall.



**FIGURE 1** Variability in (a) the North Atlantic sea surface temperature anomaly (NASSTA); (b) the Equatorial South East Indian Ocean sea surface temperature anomaly (ESEIOSSTA); (c) the East Asia surface pressure (EA-SP) anomaly; (d) the Europe surface–land–atmospheric temperature anomaly (ESLATA); and (e) North West Europe surface pressure anomaly tendency (NWESPAT)

input matrix in the ANN (Hornik, 1991) (Tables 4 and 5). The ANN is composed of single- or multiple-layer neurons (Figure 9). For complex problems, a multilayer perceptron (MLP) is the best model as it overcomes the drawback of the single-layer perceptron by adding more hidden layers. In a feed-forward MLP network, the inputs signals are multiplied by the connection weights and first summed and then directed to a transfer function to give the output for that neuron. The transfer function (purelin, hardlim, sigmoid, logistic) executes on the weighted sum of the neuron's inputs. A trained ANN can be applied to predict the result of the new independent input data set. The neural connections/weights need to be adjusted during the lifetime of a neuron. Several neural network architectures are discussed in the literature (Muller and Reinhardt, 1991; Bose and Liang, 1998) based on the training algorithm. The training algorithms back-propagation (BP), Levenberg–Marquardt (LM) and conjugate-gradient (CG) are mostly used to solve nonlinear problems (Guhathakurta *et al.*, 1999; Chaudhuri *et al.*, 2016; Singh, 2018). The BPNN is one of the most significant developments in the area of ANN

(Bryson and Ho, 1969; Rumelhart *et al.*, 1986). The main advantage of using a BPNN with a multilayer feed-forward (MLFF) neural network is that it can optimize the output error obtained by differentiating the calculated output and target output with some adjustment of weights. In the BPNN, each piece of information is sent back again in the reverse direction to optimize the error rate. It can be trained under the process of three phases: (1) using the FFNN for the training process of input information, various adjustments of weights and nodes are made in this phase; (2) to calculate the error; and (3) to update the weights. The performance of the neural network mostly depends on the number of layers defined, the number of nodes used in each layer and the number of connectivities among the nodes (Wilson *et al.*, 2002). The number of nodes in the hidden layer should be fewer than the input parameters (Singh, 2018). In the present study, the MLP, radial basis function networks (RBFN) and generalized regression neural network (GRNN) are analysed with a suitable input to select the best network for the LRF of ISMR (Tables 4 and 5). The best MLP model was compared with the MNLR model.



**FIGURE 2** Variability in (a) the warm water volume (WWV) anomaly; (b) the NINO3.4 sea surface temperature (SST) anomaly; (c) the North Atlantic pressure (NA-P) anomaly; (d) the North Central Pacific zonal wind anomaly at 850 hPa (NCPZWA-850); and (e) Indian summer monsoon rainfall (ISMR)

Finally, the performance of the best MLP and MNL models was evaluated using the MAE, PE, RMSE and Willmott's index (Willmott, 1982).

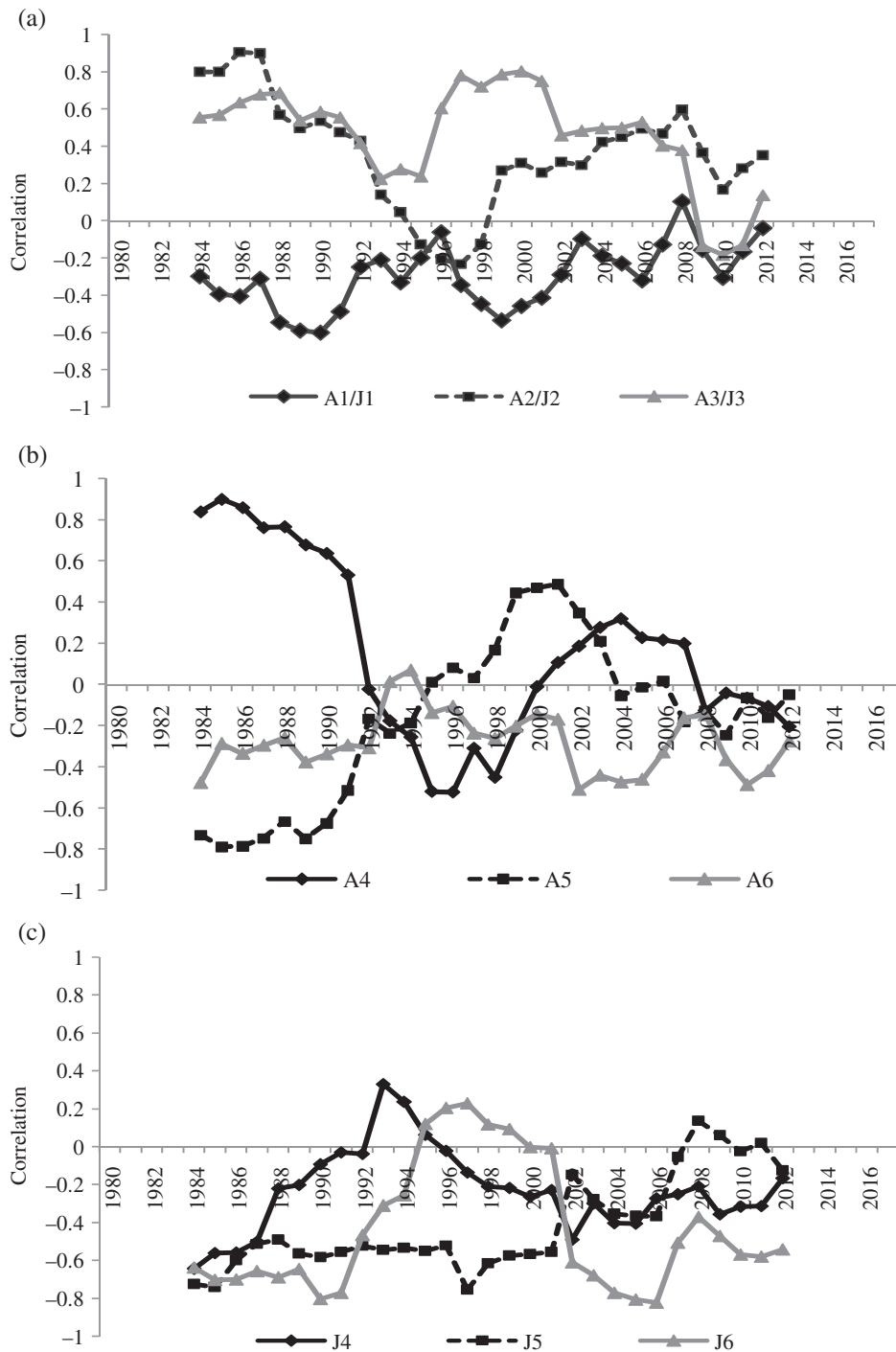
## 2 | IMPLEMENTATION

The primary analysis evaluates inter-annual variability of ISMR with selected predictors (Tables 1 and 2). The inter-correlation between predictors was also analysed to assess the inherent behaviour of the parameters (Table 3). The PCA was implemented before constructing the neural network model (Figures 5 and 6). The selection of the input variables of the ANN was based on PCA analysis (Tables 4 and 5). The entire data set was separated into training and test (validation) sets. The first 33 years, from 1980 to 2012, are used

to train the neural network models; the remaining five years, from 2013 to 2017, are used to test the model's performance. Next, the sigmoid function was used as the activation function to compress the output values to be in the range between 0 and 1 (Hagan *et al.*, 1996). A sigmoid function was non-linear and had a smooth gradient. As it compressed the data between 0 and 1, the activation bounds were in a range. Scaling was performed to avoid the asymptotes of the sigmoid function (Comrie, 1997):

$$Z_i = 0.1 + \left( 0.9 * \frac{(X_i - X_{\min})}{(X_{\max} - X_{\min})} \right)$$

where  $X_i$  is the  $i$ -th input value; and  $X_{\min}$  and  $X_{\max}$  are the minimums and maximums of the input matrix respectively. Various model architectures with a BP algorithm consisting of different hidden layers were subsequently constructed to



**FIGURE 3** The 10-point moving window correlation between Indian summer monsoon rainfall (ISMR) and the nine predictors

find the best model for a precise forecast. The neural network was trained until the train error was minimized and the computed output was nearly equal to the desired output. The entire step of the training process was then halted. A different training algorithm was evaluated with the same network structure after obtaining the suitable network architecture. The performance of the proposed model was evaluated with the help of mean absolute error (MAE), root mean square error (RMSE) and prediction error (PE) and the evaluation of algorithm was performed with mean square error (MSE) (Figures 6 and 7).

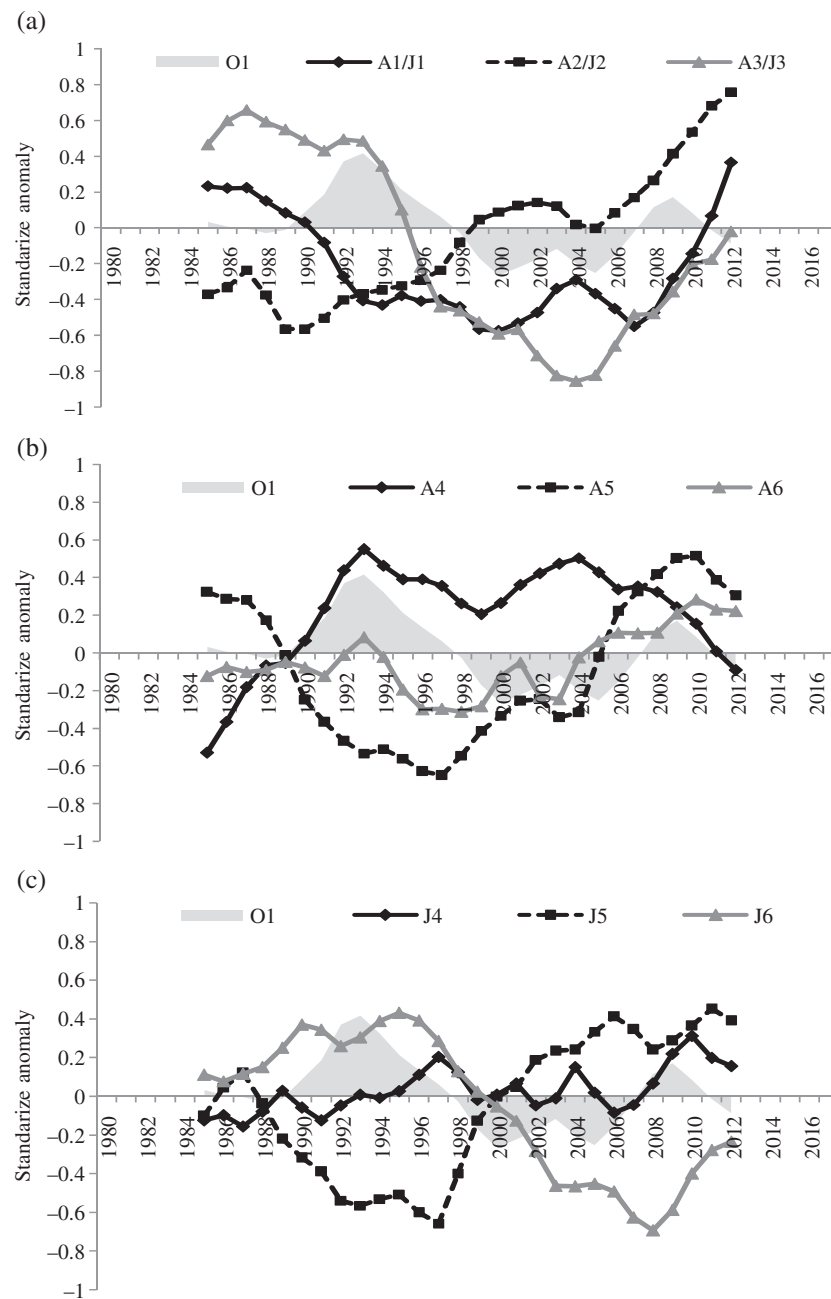
The MAE in each set was computed as:

$$MAE = \langle |y_a - y_p| \rangle$$

where  $\langle \rangle$  implies the average over all input data;  $y_a$  is the actual value of the output; and  $y_p$  is the predicted value.

RMSE and MSE were computed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_p - y_a)^2}{n}}$$



**FIGURE 4** Standardized anomalies in different predictors for the long-range forecasting (LRF) of Indian summer monsoon rainfall (ISMR)

$$MSE = \frac{\sum_{i=1}^n (y_p - y_a)^2}{n}$$

The quality of the prediction was obtained from the performance with the test set of data. PE was computed as:

$$PE = \frac{\langle |y_a - y_p| \rangle}{\langle y_a \rangle}$$

where  $\langle \rangle$  implies the average over all inputs for every input data set (Patterson, 1996). The results were validated with observation during the period 2013–2017.

An MNLR model was framed for both stages of the forecast. Regression analysis showed a relation between input

and output. The relation between input and output for the first-stage forecast is:

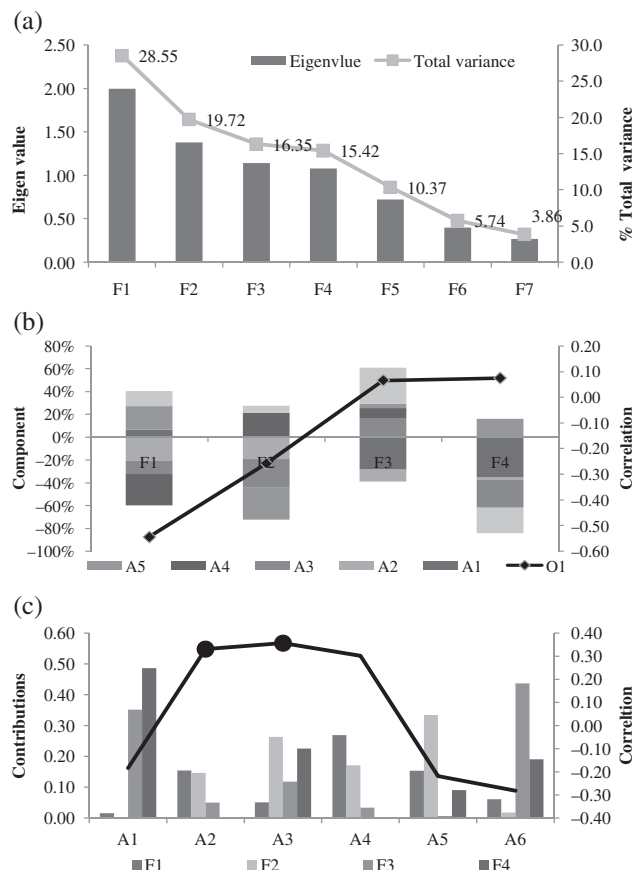
$$O1 = 0.38 - 0.62 * A1 + 0.90 * A1^2 + 0.45 * A2 - 0.03 * A2^2$$

while for second-stage forecast, the relation is:

$$\begin{aligned} O1 = & 0.63 - 0.39 * J2 + 0.55 * J2^2 + 0.49 * J3 \\ & - 0.07 * J3^2 - 0.69 * J5 + 0.26 * J5^2 \\ & + 0.35 * J6 - 0.54 * J6^2 \end{aligned}$$

The performance of the best MLP and MNLR models was evaluated using the MAE, RMSE, MSE, PE and Willmott's index. Willmott's index is defined as:





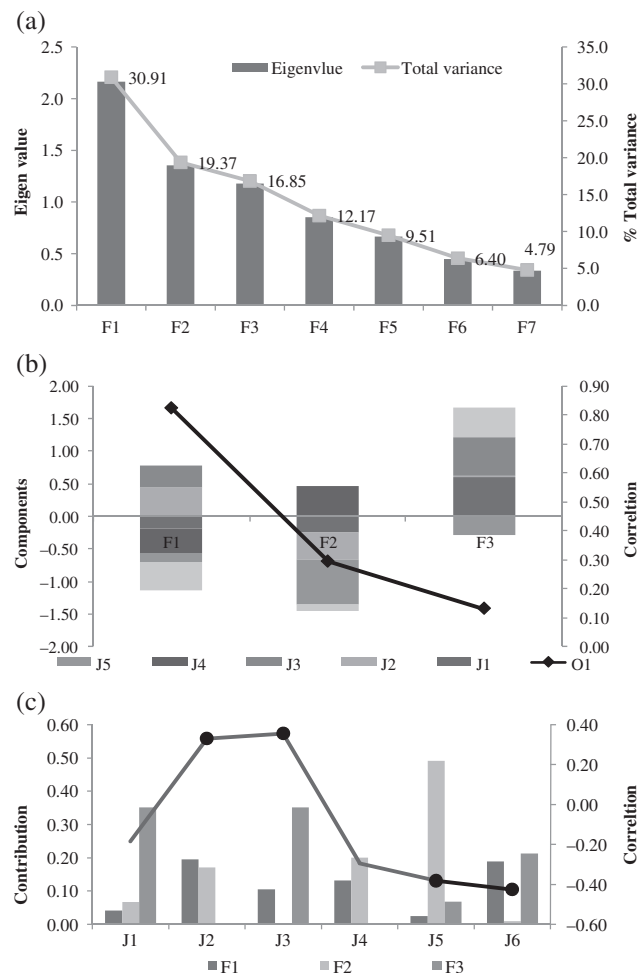
**FIGURE 5** Principal component analysis (PCA) for the first-stage long-range forecasting (LRF) of Indian summer monsoon rainfall (ISMR) provides (a) an eigenvalue and total variance explained by the principal component; (b) four components of six variables along with correlation between components and the target variable; and (c) the contribution of variables to the four components along with the correlation between variables and the target

$$d = \left[ \frac{\sum_{i=1}^N (P_i - O_i)^2}{\sum_{i=1}^N (|P_i - O| + |O_i - O|)^2} \right], 0 < d < 1$$

where  $P_i$  is predicted value;  $O_i$  is the observed and  $O$  is the average of the observed value; and  $N$  is the number of observations. The better predictive model would be identified as that having a low error and higher value of  $d$ .

### 3 | RESULTS AND DISCUSSION

Variability in each predictor was assessed using Box-Whisker plots (Tables 1 and 2). The combined North Atlantic sea surface temperature anomaly (NASSTA) of December–January was observed to be confined between  $-0.77$  and  $0.40$  (Figure 1a) during the study period (1980–2017), while the combined Equatorial South East Indian Ocean sea surface temperature anomaly (ESEIOSSTA) of February–March was observed to vary between  $0.37$  and  $1.15$  (Figure 1b). The East Asia surface pressure anomaly (EASPA) of February–March varied between  $-1.39$  and  $1.61$  (Figure 1c). The Europe surface–



**FIGURE 6** Principal component analysis (PCA) for the second-stage long-range forecasting (LRF) of Indian summer monsoon rainfall (ISMR) provides (a) an eigenvalue and total variance explained by the principal component; (b) four components of the six variables along with correlation between components and the target variable; and (c) the contribution of variables to the four components along with correlation between variables and the target

land–atmospheric temperature anomaly (ESLATA) obtained for five stations in Europe during January was observed to be confined between  $-2.27$  and  $2.44$  (Figure 1d). The North West Europe surface pressure anomaly tendency (NWESPAT) between the preceding

**TABLE 4** Architectures of the neural network model for a first-stage forecast of Indian summer monsoon rainfall (ISMR) using the three input predictors from SET-I

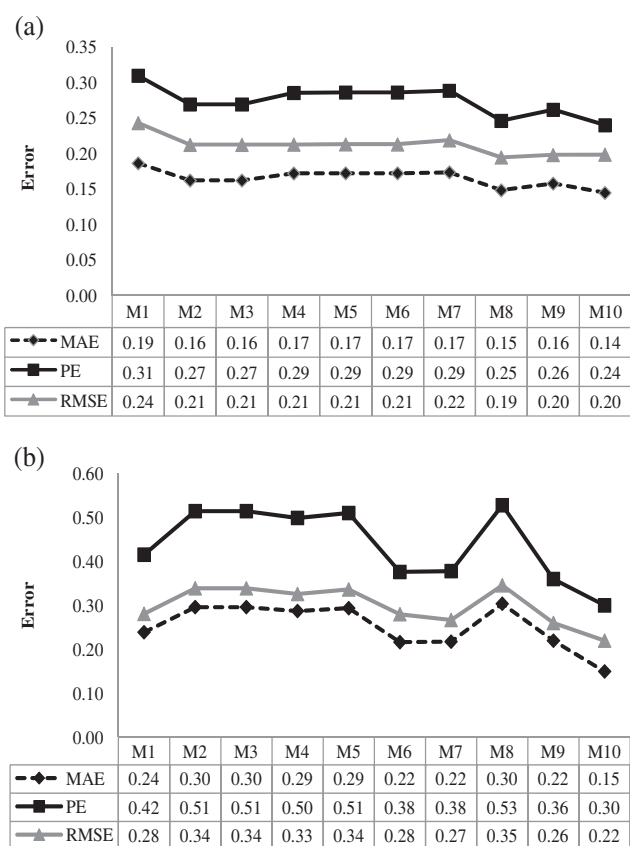
	Model	Total Error
M1	RBF 2:2-1-1:1	4.177018
M2	RBF 2:2-2-1:1	2.681379
M3	RBF 2:2-2-1:1	2.681379
M4	GRNN 2:2-17-2-1:1	3.688133
M5	GRNN 2:2-17-2-1:1	3.700567
M6	GRNN 2:2-17-2-1:1	3.696958
M7	MLP 2:2-1-1:1	0.175250
M8	MLP 2:2-2-1:1	0.085137
M9	MLP 2:2-2-2-1:1	0.134869
M10	MLP 2:2-2-2-1:1	0.079410



**TABLE 5** Architectures of the neural network model for a second-stage forecast of Indian summer monsoon rainfall (ISMR) using three input predictors from SET-II

	Model	Total Error
M1	GRNN 4:4-17-2-1:1	2.369766
M2	GRNN 4:4-17-2-1:1	2.467176
M3	GRNN 4:4-17-2-1:1	2.428040
M4	RBF 4:4-2-1:1	3.609108
M5	RBF 4:4-3-1:1	3.546851
M6	RBF 4:4-3-1:1	3.546851
M7	MLP 4:4-1-1:1	0.172251
M8	MLP 4:4-2-1:1	0.102001
M9	MLP 4:4-4-4-1:1	0.186264
M10	MLP 4:4-4-3-1:1	0.079128

post-monsoon season and the present winter season varied between  $-13.30$  and  $10.75$  (Figure 1e). The WWV anomaly of February–March was observed to vary between  $-8.60\text{E}13$  and  $1.20\text{E}14$  (Figure 2a). The SST anomaly tendency over NINO3.4 varied between  $-0.23$  and  $0.53$  (Figure 2b), while the surface pressure anomaly over the North Atlantic was observed to be confined between  $-1.77$  and  $2.27$  (Figure 2c). The North Central Pacific zonal wind anomaly at 850 hPa (NCPZWA-850) was observed to vary

**FIGURE 7** Evaluation of 10 neural networks using the mean absolute error (MAE), prediction error (PE) and root mean square error (RMSE) during (a) training with data for the period 1980–2012; and (b) validation with data for the period 2012–2017 for the first-stage long-range forecast of Indian summer monsoon rainfall (ISMR)

between  $-9.16$  and  $-7.16$  (Figure 2d). The accumulated rainfall (mm) during the Indian summer monsoon (ISM) was observed to be confined to between 778.20 and 871.00 (Figure 2e). The interrelation between predictors and ISMR was analysed by linear correlation (Table 3). The correlation was evaluated for 38 years (1980–2017). The correlations were statistically significant at the 95% confidence level. Note that ESEIOSSTA is negatively correlated with the NINO3.4 SST anomaly tendency and NCPZWA-850. The ESLATA of January was also negatively correlated with the NWESPAT, while the NWE-SPAT was positively correlated with the NCPZWA-850. WWV anomaly was positively correlated with the NINO3.4 SST anomaly tendency. The correlation between predictors and ISMR showed a significant correlation with the ESEIOSSTA, EASPA, NASSTA and NCPZWA-850. Other correlations were found to lose their significance in the present decade. A positive correlation between ESEIOSSTA and ISMR that depicts the higher SST anomaly over the Indian Ocean results in a higher evaporation which leads to a high moisture content in the atmosphere over the Indian Ocean, which is subsequently transported towards the Indian subcontinent through a low-level jet. The pressure anomalies over the East Asian region during February–March were associated with relative shifting of the climatological low-pressure area over the North Pacific to its east and the high-pressure area over middle Asia to its west. The shifting of the low pressure of the North Pacific towards the west lowers the pressure of East Asia. This modifies the circulation pattern and the typhoon tracks over the North Pacific, which may influence summer monsoon performance (Saha *et al.*, 1981). The negative relationship between the surface MSLP anomaly over the North Atlantic (May) and ISMR can be explained through the variation in the strength of the subtropical level climatological high-pressure area over the Atlantic. The variation in the strength or position of the climatological high in the North Atlantic may be associated with the North Atlantic Ocean (NAO). The negative correlation between the NCPZWA-850 and ISMR is observed when the zonal wind at 850 hPa over the North Central Pacific weakens the convection zone over the Central Pacific and shifts eastward, which subsequently pulls down all moisture over the Western Indian Ocean which leads to a weakening of the ISM. Figure 3 shows a 10-point moving window correlation between ISMR and nine predictors considered in the present study. The magnitude of correlation  $> 0.3$  is statistically significant at the 95% confidence level. Note that A1/J1 is mostly correlated negatively throughout the period, but this correlation is neither stable nor significant, while A3/J3 is correlated positively with ISMR, and this correlation is quite stable as well as significant throughout the period. The A2/J2 shows a positive and stable correlation with ISMR, except for a short tenure

(1995–1998) (Figure 3a). The correlation between A4 and ISMR is highly positive during the early decade, which suddenly drops to negative during 1994–1999 (Figure 3b). In the recent decade, the association between A4 and ISMR becomes insignificant. The negative correlation between A5 and ISMR becomes positive after 1994 and 2004 and it loses its significance. The correlation between A6 and ISMR is negative throughout the period (Figure 3b). The correlations between ISMR and J4–J6 are negative during most of the period of study (Figure 3c). In the present decade, the correlations of J4 and J5 with ISMR are observed to lose their statistical significance, while J6 shows a significant negative association (Figure 3c). To understand how these correlations affect the relation of predictors with ISMR, the 11-point moving average is analysed, which provides an insight into the decadal variation (Figure 4). A positive epoch of a standardized anomaly depicts a higher-than-normal deviation, while a negative epoch depicts the reverse condition (Figure 4). The 11-point moving average of ISMR depicts a positive epoch early in the decade, followed by a negative epoch. In the recent decade, a positive epoch of ISMR was identified. A1/J1 shows a positive epoch followed by a negative epoch (Figure 4a). The positive epoch is observed to start in the recent decade. The negative epoch followed by a positive epoch in A2/J2 was observed. An increasing trend was also noted in A2/J2 in the present decade. The positive epoch followed by a negative epoch in A3/J3 variation was noted for the present study period. The association between ISMR and A3/J3 can be explained through correlation (Figure 4a). A4 is observed to possess a positive epoch during most of the study period; however, in the present decade, it showed a decreasing trend and turned towards a negative epoch (Figure 4b). A5 shows the opposite variation from A4. A4 has a negative epoch and becomes positive in 2006. The decadal variation of A6 is less than other predictors. In the present decade it has a positive epoch (Figure 4b). The decadal variation of J4 is less and random and fluctuates between  $-0.2$  and  $0.3$  (Figure 4c). Initially, early in the decade, J5 shows a negative epoch. It turned towards a positive epoch after 2000 (Figure 4c). The PCA was implemented with two sets of predictors used for a two-stage forecast along with a target output (Figures 5 and 6). Though both sets of predictors contain three common predictors, another three predictors are different in the two sets. The covariance matrixes for the two sets are therefore different. The PCA analysis with Set I predictors provides seven principal components (Figure 5). The variance explained by each component along with its eigenvalue depicts a gradual decrease from the first to the seventh component (Figure 5a). Though the first five components have shown 90.19% variance, the eigenvalue of the first four components is  $> 1$ . The first four components are, therefore, considered for further analysis. The six predictors of Set I in the first four components and correlation of the four components with ISMR (target output) were analysed to understand the relation between

target output and the four components (Figure 5b). The correlation between components and target output depicts that the first two components are correlated with a magnitude  $> 0.2$ . The maximum contribution towards the first, second, third and fourth components is observed for A4, A5, A6 and A1 respectively (Figure 5c). The contribution of the predictors for the first two components needs to be analysed as it is highly correlated with ISMR. Note that the average contribution of the predictor for the first component is 0.117. Thus, for the first component, only those parameters are considered that have a higher contribution than average. It is observed that A2, A4 and A5 have contributions of 0.154, 0.269 and 0.153 respectively. The average contribution for the second components is about 0.156. Thus, A3–A5 have a higher-than-average contribution. The correlation between ISMR and the predictor shows that A2 and A3 have a significant positive correlation with ISMR. Thus, for the first-stage forecast model, A2 and A3 were considered as the suitable inputs. The PCA analysis with Set II predictors of the second-stage forecast also provides seven principal components (Figure 6). The first four components show 88% variance; however, the eigenvalue of the first three components is  $> 1$ . The first three components are, therefore, considered for further analysis (Figure 6a). The six predictors of Set II in the first three components and the correlation of these three components with ISMR (the target output) are analysed to understand the relation between the target output and the four components (Figure 6b). Similarly, for the second-stage forecast during training (Figure 6a) and validation (Figure 6b), the minimum error was noted for model M10, which is an MLP model, trained with the BP algorithm. The mean, maximum, minimum and standard deviation (SD) of the 10 models' output for the first- and second-stage forecasts were compared by observation (Tables 6 and 7). The mean, maximum, minimum and SD of the observation were about 828.78, 961.40, 661.90 and 78.25 mm respectively. The means from the 10 models for the first-stage forecast were comparable with observed values; however, the maximum, minimum and SD were

**TABLE 6** Descriptive statistics of the observed and model output of Indian summer monsoon rainfall (ISMR) for first-stage forecasting

	Mean	Minimum	Maximum	SD	CC
Actual	828.78	661.90	961.40	78.25	
M1	825.53	717.10	856.72	32.13	0.09
M2	813.65	723.22	899.51	60.13	0.52
M3	813.65	723.22	899.51	60.13	0.52
M4	832.40	803.84	859.31	14.31	0.54
M5	832.46	804.47	858.89	14.03	0.54
M6	832.44	804.29	859.01	14.11	0.54
M7	840.11	810.89	862.11	12.92	0.49
M8	816.80	711.60	930.53	64.08	0.60
M9	819.47	764.86	898.35	45.29	0.54
M10	816.36	702.83	919.08	65.69	0.61

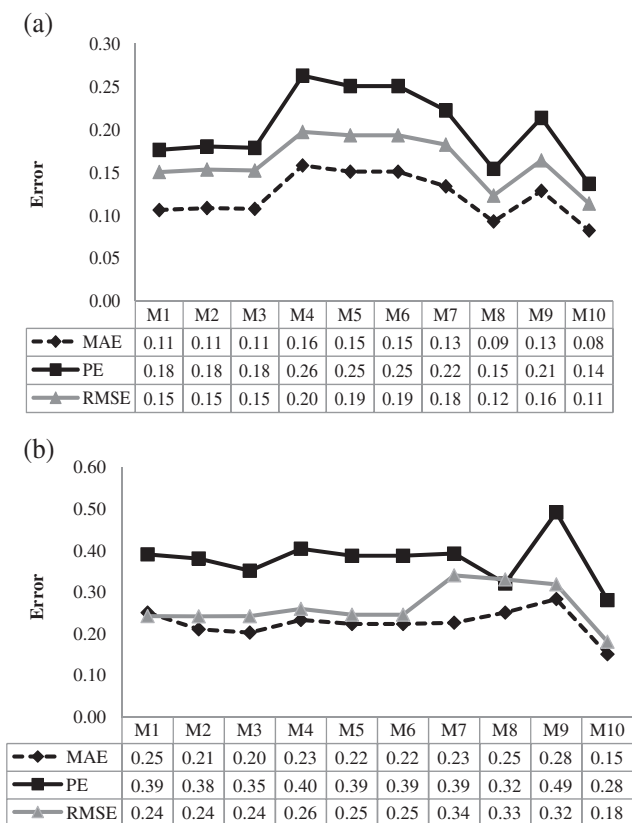
Note: CC: correlation co-efficient; SD: standard deviation.

observed to fluctuate, which shows variability in the data. The deviation was minimum for model M10. The correlation with M10 was observed to be 0.65. A similar result was observed for the second-stage forecast as well. However, the correlation with M10 was observed to be 0.88 in the second-stage forecast. The best model was selected after performance evaluation (Figures 7 and 8). Error evaluation during training and validation shows that model M10 had the lowest error in both stage forecasts of ISMR. Willmott's index (Figure 9) also

**TABLE 7** Descriptive statistics of the observed and model output of Indian summer monsoon rainfall (ISMR) for second-stage forecasting

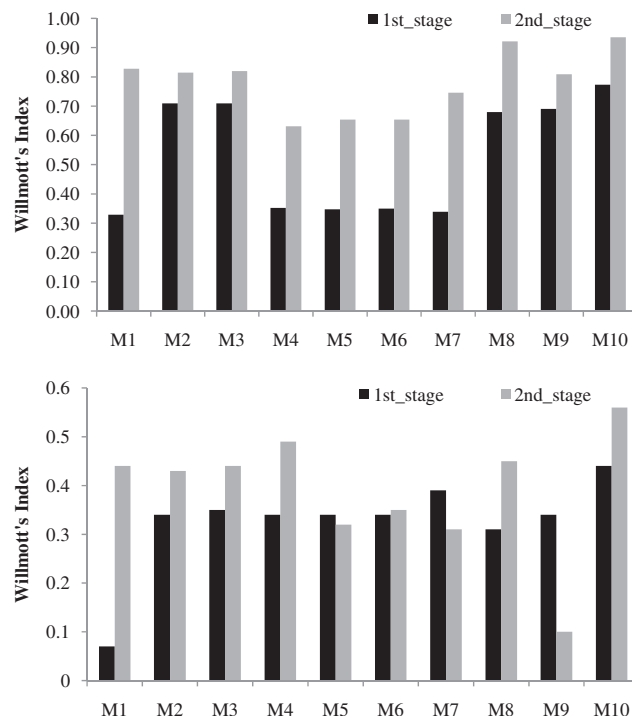
	Mean	Minimum	Maximum	SD	CC
Actual	828.7788	661.9000	961.4000	78.25045	
M1	829.3499	707.4431	909.9450	48.88721	0.78
M2	830.0789	711.1157	906.3796	46.71214	0.77
M3	829.7805	709.5269	907.8180	47.60698	0.77
M4	830.0573	772.4864	899.9267	36.57026	0.53
M5	829.7692	766.7691	911.3792	38.18869	0.56
M6	829.7692	766.7691	911.3792	38.18869	0.56
M7	817.2635	744.0694	902.2057	45.77741	0.64
M8	831.6427	658.7839	964.2943	72.82084	0.85
M9	825.3868	735.0743	940.3594	54.90537	0.71
M10	826.5490	654.6427	950.3877	75.51005	0.88

Note: CC: correlation co-efficient; SD: standard deviation.

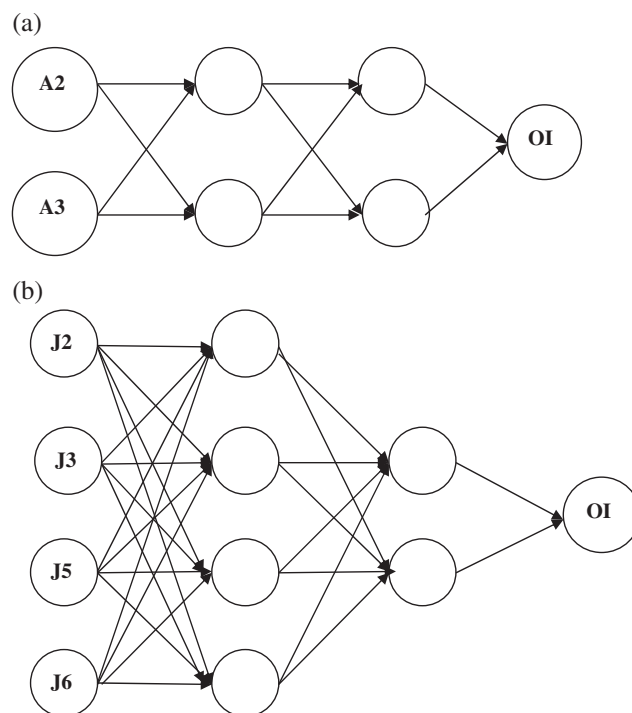


**FIGURE 8** Evaluation of 10 neural networks using mean absolute error (MAE), prediction error (PE) and root mean square error (RMSE) during (a) training with data for the period 1980–2012; and (b) validation with data for the period 2012–2017 for the second-stage long-range forecasting of Indian summer monsoon rainfall (ISMR)

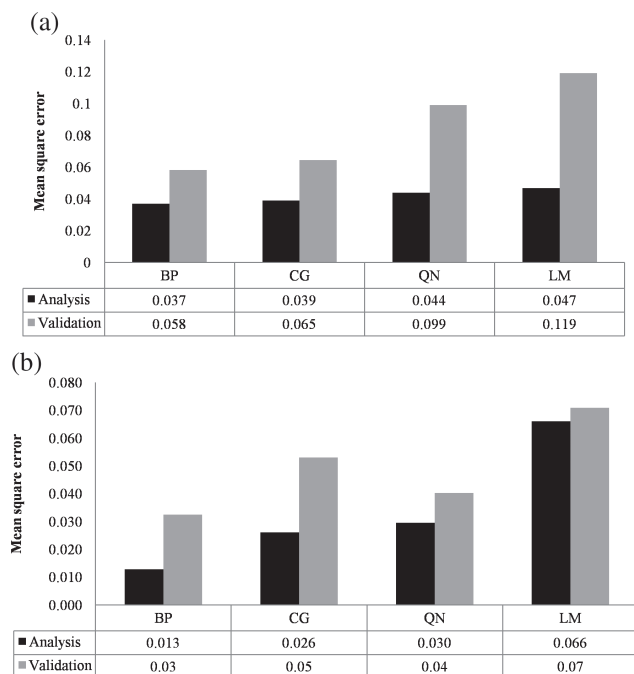
identifies model M10 as the best model in both stages of the forecast. However, the structure of model M10 is different for both cases. Its architecture comprises two inputs and two hidden layers along with two nodes at each hidden layer and one



**FIGURE 9** Comparison of Willmott's index for 10 artificial neural networks (ANN) used to forecast the first- and second-stage forecasts of seasonal rainfall of Indian summer monsoon rainfall (ISMR) during (a) training and (b) validation

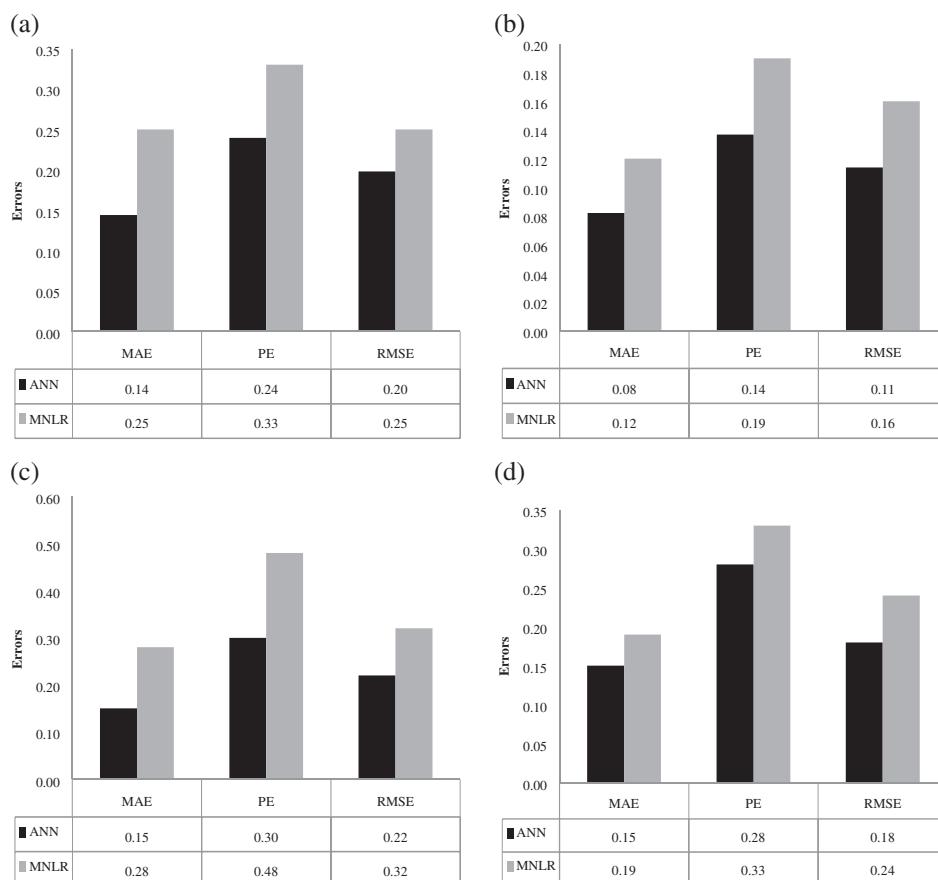


**FIGURE 10** Architecture of a neural network for (a) the first- and (b) the second-stage long-range forecasting (LRF) of Indian summer monsoon rainfall (ISMR)

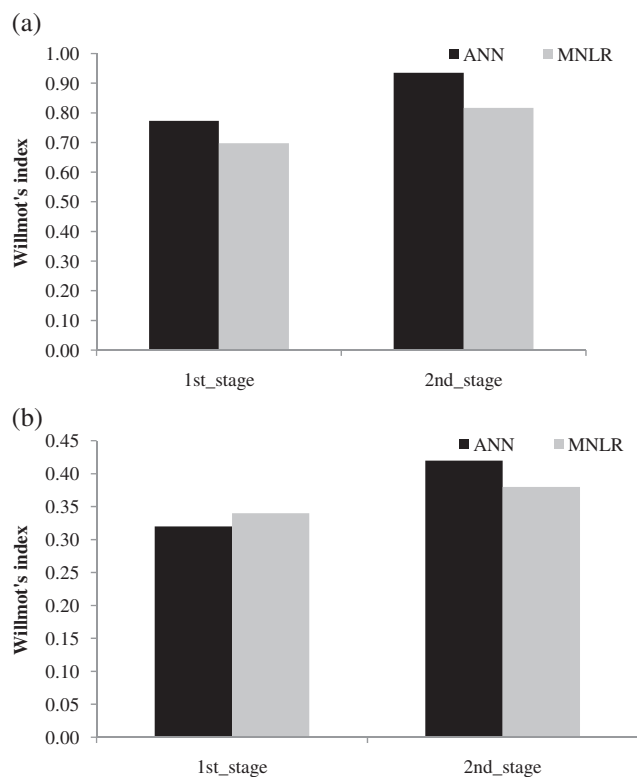


**FIGURE 11** Evaluation of different training algorithms for both (a) the first- and (b) the second-stage long-range forecasting (LRF) of Indian summer monsoon rainfall (ISMR)

output for ISMR for the first-stage forecast (Figure 10a). While the architecture of model M10 comprises four inputs and two hidden layers along with four nodes in the first hidden layer, two nodes in the second hidden layer and one output for ISMR for the second-stage forecast (Figure 10b). Different training algorithms were tested with the best selected MLP model (Figure 11). The BP is observed to be the best algorithm for the present study. To check the skill of the ANN at forecasting ISMR, the best MLP model was compared with an MNLR model. The MNLR model was also analysed with the same input and output and validated for same time period. It was observed that during training, the MAE, PE and RMSE were less in the case of the MLP than the MNLR for both the first- (Figure 12a) and second-stage (Figure 12b) forecasts of seasonal rainfall of the ISM. During validation, in both the first- (Figure 12c) and second-stage (Figure 12d) forecasting of seasonal rainfall, the same results as for training were observed. Error evaluation suggests that the MLP performs better than the MNLR model in forecasting ISMR seasonal rainfall. To ensure the model's forecast accuracy, Willmott's index was computed and compared (Figure 13). Its value was higher in the ANN for both forecast



**FIGURE 12** Comparison of errors obtained through artificial neural networks (ANN) and multiple nonlinear regression (MNLR) during training for (a) the first- and (b) the second-stage forecasts of seasonal Indian summer monsoon rainfall (ISMR) and during validation for (c) the first- and (d) the second-stage forecasts of seasonal ISMR

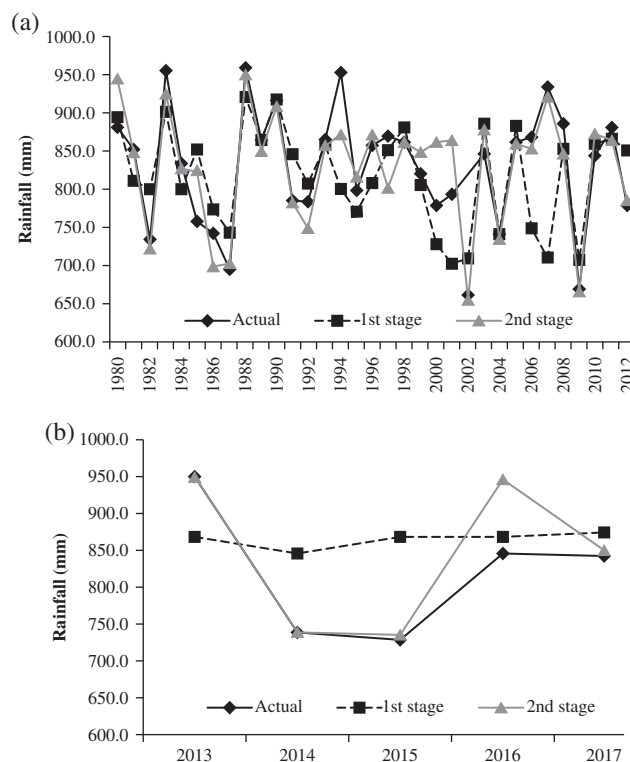


**FIGURE 13** Comparison of Willmott's index for first- and second-stage forecasts of seasonal Indian summer monsoon rainfall (ISMR) during (a) training and (b) validation

stages during training (Figure 13a); however, during validation (Figure 13b), the first-stage forecast showed a lower Willmott's index in the case of the ANN. Model performance was estimated for the LRF of ISMR by comparing it with the observation for both first- and second-stage forecasts (Figure 14). The result shows that the second-stage forecast is well comparable with the observation during both training and validation with observation (Figure 14a, b).

#### 4 | CONCLUSIONS

The long-range forecasting (LRF) of accumulated Indian summer monsoon rainfall (ISMR) was attempted with an artificial neural network (ANN) following the India Meteorological Department's (IMD) two-stage forecast procedure with 38 years (1980–2017) of data. The present paper adopted nine predictors currently in use by the IMD. The Equatorial South East Indian Ocean sea surface temperature anomaly (ESEIOSSTA), the East Asia surface pressure anomaly (EASPA), North Atlantic sea surface temperature anomaly (NASSTA) and the North Central Pacific zonal wind anomaly at 850 hPa (NCPZWA-850) were found to be significant for predicting ISMR. The use of principal component analysis (PCA) with both sets of predictors confirmed the same finding. The first-stage forecast was performed with the ESEIOSSTA and EASPA as inputs for the ANN model, whereas for the second-stage



**FIGURE 14** Comparison of long-range forecasting (LRF) of Indian summer monsoon rainfall (ISMR) between model output and observation (black solid line) for the first stage (solid grey line) and the second stage (black dotted line) during (a) training and (b) validation

forecast, the ESEIOSSTA, EASPA, NASSTA and NCPZWA-850 were considered as inputs for the model. It was observed that the second-stage forecast was better than first-stage forecast when compared with observations. This indicates that the NASSTA and NCPZWA-850 play a significant role during the Indian summer monsoon (ISM). The North Atlantic sea surface pressure relates to the North Atlantic Oscillation, which is a temporal fluctuation of the zonal wind strength across the Atlantic due to pressure variations in the subtropical anticyclonic belt and in the subpolar low near Iceland. The pressure gradient between the two centres of actions, the Icelandic low (IL) and Azores high (AH), is a measure of the mean intensity of the westerly winds in the middle latitudes. This variation in the intensity of the westerlies is implicitly linked to the low and high index phases, which are in turn linked to the blocking high episodes over the Asian region known to affect ISMR on inter-annual time scales (Dugam *et al.*, 1997). The other predictor, the NCPZWA-850, depicts the fact that when the zonal wind at 850 hPa over the North Central Pacific weakens (strengthens) the convection zone over the Central Pacific, the shift eastward pulls down all the moisture over the Western Indian Ocean, which might lead to a weakening of the ISM. The effect of El Niño has, thus, been incorporated in terms of zonal wind. The above physical linkage might play a crucial role in identifying the variability in the rainfall for



second-stage forecasting. The structure of the ANN is more complex for a second-stage forecast; hence, more nonlinearity can be incorporated into second-stage forecast model due to the incorporation of two more predictors. The study also suggests that the MLP model with a back-propagation (BP) algorithm is best among the other ANN models. The MNLR model prepared with the two predictor sets was compared with the best MLP model. The MLP model was found to have better skill in the LRF of ISMR than the MNLR model. Willmott's index reconfirms the skill of the MLP in the LRF of ISMR. The comparison of the MLP model forecast with observations shows that the second-stage forecast is better than the first-stage forecast.

## ACKNOWLEDGEMENT

Sutapa Chaudhuri thanks the Ministry of Science and Technology, Government of India, for providing the opportunity to work for the National Networking of Climate Modelling under the Climate Change Programme. The authors thank the Editor-in-Chief and the anonymous reviewers for providing constructive comments that helped to improve the clarity of the manuscript.

## ORCID

Rajashree Acharya  <https://orcid.org/0000-0002-0907-8133>

## REFERENCES

- Aksoy, H. and Dahamshe, A. (2008) Artificial neural network models for forecasting monthly precipitation in Jordan. *Stochastic Environmental Research and Risk Assessment*, 23, 917–931.
- Bose, N.K. and Liang, P. (1995) *Neural Network Fundamentals with Graphs, Algorithms and Applications*. Cambridge: McGraw Hill International Edition, The MIT Press.
- Bose, N.K. and Liang, P. (1998) *Neural Network Fundamentals with Graphs, Algorithms, and Applications*. New Delhi, India: Tata McGraw-Hill.
- Bryson, A.E. and Ho, Y.C. (1969) *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham, Massachusetts: Blaisdell Pub. Co.
- Chakraverty, S. and Gupta, P. (2007) Comparison of neural network configurations in the long-range forecast of southwest monsoon rainfall over India. *Neural Computing and Applications*, 17, 187–192.
- Chaudhuri, S. (2010) Convective energies in forecasting severe thunderstorms with one hidden layer neural net and variable learning rate back propagation algorithm. *Asia-Pacific Journal of Atmospheric Sciences*, 46, 173–183.
- Chaudhuri, S., Das, D., Goswami, S. and Das, S.K. (2016) Long-range forecast of all India Summer Monsoon Rainfall using adaptive neuro-fuzzy inference system: skill comparison with CFSv2 model simulation and real-time forecast for the year 2015. *Climate Dyn.*, 47(9–10), 3319–3333.
- Chaudhuri, S., Das, D., Sarkar, I. and Goswami, S. (2015) Multilayer perceptron model for nowcasting visibility from surface observations: results and sensitivity to dissimilar station altitudes. *Pure and Applied Geophysics*, 172, 2813–2829.
- Chaudhuri, S., Dutta, D., Goswami, S. and Middey, A. (2013) Intensity forecast of tropical cyclones over North Indian Ocean using multi-layer perceptron model: skill and performance verification. *Natural Hazards*, 65, 97–113.
- Chaudhuri, S., Goswami, S. and Middey, A. (2014) Medium-range forecast of cyclogenesis over North Indian Ocean with multilayer perceptron model using satellite data. *Natural Hazards*, 70, 173–193.
- Comrie, A.C. (1997) Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association*, 47, 653–663.
- Dugam, S.S., Kakade, S.B. and Verma, R.K. (1997) Interannual and long-term variability in the North Atlantic oscillation and Indian summer monsoon rainfall. *Theoretical and Applied Climatology*, 58, 21–29.
- Gadgil, S., Rajeevan, M. and Nanjundiah, R. (2005) Monsoon prediction: why yet another failure? *Current Science*, 88, 1389–1400.
- Gardner, M.W. and Dorling, S.R. (1998) Artificial neural network multilayer perceptron—a review of application in atmospheric sciences. *Atmospheric Environment*, 32, 2627–2636.
- Gowariker, V., Thapliyal, V., Kulsheshta, S.M., Mandal, G.S., Sen Roy, N. and Sikka, D.R. (1991) A power regression model for long range forecast of southwest monsoon rainfall over India. *Mausam*, 42, 125–130.
- Gowariker, V., Thapliyal, V., Sarkar, R.P., Mandal, G.S. and Sikka, D.R. (1989) Parametric and power regression models: new approach for long range forecasting of monsoon rainfall in India. *Mausam*, 40, 115–122.
- Guhathakurta, P. (2006) Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network. *Current Science*, 90, 773–779.
- Guhathakurta, P., Rajeevan, M. and Thapliyal, V. (1999) Long range forecasting Indian summer monsoon rainfall by a hybrid principal component neural network model. *Meteorology and Atmospheric Physics*, 71, 255–266.
- Hagan, M.T., Demuth, H.B. and Beale, M.H. (1996) *Neural Network Design*. Boston: PWS Publishing.
- Haykin, S. (1999) *Neural Networks, A Comprehensive Foundation*, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.
- Hornik, K. (1991) Approximation capabilities of multi-layer neural networks. *Neural Networks*, 4, 251–257.
- Hsieh, W.W. and Tang, B. (1998) Applying neural network models to prediction and data analysis meteorology and oceanography. *Bulletin of the American Meteorological Society*, 79, 1855–1870.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (1996) The NCEP/NCAR reanalysis project. *Bulletin of the American Meteorological Society*, 77, 437–471.
- McPhaden, M.J. (2003) Tropical Pacific Ocean heat content variations and ENSO persistence barriers. *Geophysical Research Letters*, 30, 1480. <https://doi.org/10.1029/2003GL016872>.
- McPhaden, M.J., Busalacchi, A.J., Cheney, R., Donguy, J.R., Gage, K.S., Halpern, D. J., Julian, P., Meyers, G., Mitchum, G.T., Niiler, P.P., Picaud, J., Reynolds, R.W., Smith, N. and Takeuchi, K. (1998) The Tropical Ocean Global Atmosphere (TOGA) observing system: a decade of progress. *Journal of Geophysical Research*, 103, 14169–14240.
- Meinen, C.S. and McPhaden, M.J. (2000) Observations of warm water volume changes in the equatorial Pacific and their relationship to El Niño and La Niña. *Journal of Climate*, 13, 3551–3559.
- Muller, B. and Reinhardt, J. (1991) *Neural Networks: An Introduction*, Vol. 2. Berlin, Heidelberg, New York: Springer.
- Navone, H.D. and Ceccatto, H.A. (1994) Predicting Indian monsoon rainfall: A neural network approach. *Climate Dynamics*, 10, 305–312.
- Pai, D.S. and Rajeevan, M. (2006) Long range prediction models for the Indian summer monsoon rainfall with different lead time periods based on the global SST anomalies. *Meteorology and Atmospheric Physics*, 92, 33–43.
- Patterson, D.W. (1996) *Artificial Neural Networks: Theory and Applications*. Englewood Cliffs: Prentice Hall.
- Rajeevan, M., Guhathakurta, P. and Thapliyal, V. (2000) New models for long range forecasts of summer monsoon rainfall over northwest and peninsular India. *Meteorology and Atmospheric Physics*, 73, 211–225.
- Rajeevan, M. and McPhaden, M.J. (2004) Tropical Pacific upper ocean heat content variations and Indian summer monsoon rainfall. *Geophysical Research Letters*, 31, L18203. <https://doi.org/10.1029/2994GL020631>.
- Rajeevan, M., Pai, D.S., Anil Kumar, R. and Lal, B. (2007) New statistical models for long range forecasting of southwest monsoon rainfall over India. *Climate Dynamics*, 28, 813–828.
- Rumelhart, E.D., Hinton, E.G. and Williams, J.R. (1986) Learning representations by backpropagating errors. *Nature*, 323, 533–536.
- Saha, K., Sanders, F. and Shukla, J. (1981) Westward propagating predecessors of monsoon depressions. *Mon. Wea. Rev.*, 109, 303–343.
- Sahai, A.K., Soman, M.K. and Satyan, V. (2000) All India summer monsoon rainfall prediction using an artificial neural network. *Climate Dynamics*, 16, 291–302.
- Singh, P. (2018) Indian summer monsoon rainfall (ISMR) forecasting using time series data: A fuzzy-entropy-neuro based expert system. *Geoscience Frontiers*, 9(4), 1243–1257.



- Singh, P. and Borah, B. (2013) Indian summer monsoon rainfall prediction using artificial neural network. *Stochastic Environmental Research and Risk Assessment*, 27, 1585–1599.
- Smith, T.M. and Reynolds, R.W. (2004) Improved extended reconstruction of SST (1854–1997). *Journal of Climate*, 17, 2466–2477.
- Swaminathan, M.S. (1998) *Padma Bhusan Prof. P. Koteswaram First Memorial Vayu Mandal*, pp. 3–10.
- Venkatesan, C., Raskar, S.D., Tambe, S.S., Kulkarni, B.D. and Keshavamurty, R.N. (1997) Prediction of all India monsoon rainfall using error-back-propagation neural networks. *Meteorology and Atmospheric Physics*, 62, 225–240.
- Walker, G.T. (1914) Correlation in seasonal variations of weather. III. On the criterion for the reality of relationships or periodicities. *Memoirs of the Indian Meteorological Department*, 21, 13–15.
- Walker, G.T. (1923) Correlation in seasonal variations of weather. VIII. A preliminary study of world-weather. *Memoirs of the Indian Meteorological Department*, 24, 75–131.
- Willmott, C.J. (1982) Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, 63, 1309–1313.
- Wilson, I., Paris, S., Ware, J. and Jenkins, D. (2002) Residential property price time series forecasting with neural networks. *Knowledge-based Systems*, 15, 335–341.

**How to cite this article:** Acharya R, Pal J, Das D, Chaudhuri S. Long-range forecast of Indian summer monsoon rainfall using an artificial neural network model. *Meteorol Appl.* 2019;26:347–361. <https://doi.org/10.1002/met.1766>