Forecasting the winter shower over India through a neurocomputing approach

S. S. De, Goutami Chattopadhyay, Suman Paul, Soumyadip Chattopadhyay and D. K. Haldar

ABSTRACT

The development of a neurocomputing technique to forecast the average winter shower in India has been modeled from 48 years of records (1950–1998). The complexities in the rainfall–sea surface temperature relationships have been statistically analyzed along with the collinearity diagnostics. The presence of multicollinearity has been revealed and a variable selection has been executed accordingly. The absence of persistence has also been revealed. For this reason, an Artificial Neural Net Model as a predictive tool for the said meteorological event in the form of a Multiple Layer Perceptron has been generated with a sea surface temperature anomaly and monthly average winter shower data over India during the above period. After proper training and testing, a Neural Net model with small prediction error is developed and the supremacy of the Artificial Neural Net over conventional statistical predictive procedures has been established statistically.

Key words | artificial neural network, prediction, winter shower

S. S. De (corresponding author)
Goutami Chattopadhyay
Suman Paul
D. K. Haldar
Centre of Advanced Study in Radio Physics and Electronics,
University of Calcutta,
1 Girish Vidyaratna Lane,
Kolkata 700 009,
India
E-mail: *de_syam_sundar@yahoo.co.in*

Soumyadip Chattopadhyay
Department of Economics,
Visva-Bharati University,
Santiniketan 731 235,
West Bengal,
India

INTRODUCTION

It is widely accepted that the climate and its variability are the result of a complex system of air–sea interactions and atmosphere–ocean feedbacks (Latif 1998; Wang *et al.* 2003). Sea surface temperature (SST) and their interaction play a significant role in the phenomenon of rainfall (Maity & Nagesh Kumar 2007). The association between seasonal or annual rainfall and global SST is well documented in the literature for various parts of the world (Reason & Mulenga 1999; Aldrian & Dwi Susanto 2003; Moron *et al.* 2004). Sahai *et al.* (2003) presented a methodology based on correlation analysis for making optimum use of global SST for long lead prediction of the Indian summer-monsoon rainfall. SST anomalies influence the atmosphere by altering the flux of latent heat and sensible heat from the ocean (Holton 1972). Basically, the changes in SST influence the large-scale atmospheric circulation, which in turn influences the rainfall (Maity & Nagesh Kumar 2007).

In the tropics, positive SST anomalies are associated with enhanced convection and the resulting heating is balanced by adiabatic cooling (Reddy & Salvekar 2003). SST anomalies also play an important role in producing rainfall (Clark *et al.* 2000). The El-Niño-Southern Oscillation (ENSO) is a coupled ocean–atmosphere phenomenon that has worldwide impact on climate in general and Indian monsoons in particular (Kumar 2005). The oscillations in wind stress due to the Southern Oscillation are associated with changes in the circulation of the ocean and the SST anomaly that are referred to as El Nino. Warm ENSO episodes are characterized by an increased number and intensity of tropical storms over the Bay of Bengal and hence enhanced winter monsoon rainfall (Kumar 2005).

The importance of the Indian summer monsoon rainfall, on which the country's agriculture, power generation and industrial production heavily depend, is well known (Rao 1999). A number of studies have been made to examine the association of the Indian summer monsoon rainfall with the Southern Oscillation (SO), which designates an oscillation in the sea level pressures between the Pacific

doi: 10.2166/hydro.2011.124

and the Indian Ocean from Africa to Australia (Rao 1999). The studies pertaining to the winter monsoon season are very few in the literature, although this season's rainfall is quite important for southern peninsular India. Literature on the winter-monsoon rainfall over India includes Rao (1963, 1999), Singh (1995) and Raj (1996). Outside the tropical Pacific, significant ENSO-related SST anomalies are found in many places, such as in the tropical North Atlantic, the tropical Indian Ocean, the extratropical North and South Pacific, and the South China Sea (Deser *et al.* 2004). Linkage between the Asian winter monsoon and SST anomalies over the tropical Indian Ocean is examined in detail (Annamalai *et al.* 2005; Yang *et al.* 2010). Wen *et al.* (2000) also discussed the linkage between the Asian winter monsoon and SST anomaly in the tropical Pacific. The winter or north-east monsoon over India gives about 11% of its annual rainfall to the southern peninsular India. The winter monsoon rainfall is of considerable importance for economic interests in this region, which constitutes about 15% of the Indian subcontinent (Rao 1999).

The present paper endeavors to develop an Artificial Neural Network (ANN) model to forecast the average winter shower in India. A handful of literature is available where the summer-monsoon rainfall over India has been predicted using ANN (Venkatesan *et al.* 1997; Sahai *et al.* 2000; Chattopadhyay 2007; Chattopadhyay & Chattopadhyay 2008; Guhathakurata 2008). However, the ANN, which is particularly useful when the underlying physical processes are not fully understood or display chaotic properties (Sivakumar 2000), has not been attempted to forecast the winter-monsoon rainfall in India.

The advantages of ANN over conventional predictive methodologies are the following (Gardner & Dorling 1998):

- Unlike other statistical techniques, the ANN makes no prior assumptions concerning the data distribution.
- It can model highly nonlinear functions and can be trained to accurately generalize when presented with new, unseen data.

Conventional weather forecasting models are highly data-specific and based on complex and expensive-to-maintain mathematical models that are built many months in advance of the event they are attempting to predict. The performance of conventional statistical models very often relies on the availability of accurate real-time data inputs, the quality of the engineering knowledge and the mathematical skills used to specify, build and operate the models, and the ability of the models to respond to the unexpected and rapidly changing environment. The ANN, on the other hand, offer real prospects for an effective, more flexible and less assumption-dependent adaptive methodology well suited for modeling weather forecasting, which by its nature is inherently complex because of nonlinearity and chaotic effects (Maqsood *et al.* 2002). An exhaustive review of issues associated with the suitability of ANN in meteorological forecasting is available (Hsieh & Tang 1998).

In the present paper, the time series pertaining to the winter shower over the study zone has first been viewed for its non-stationarity and associated non-persistence over time. As the Indian winter monsoon is associated with precipitation only over the southern tip of India, extending to about 12° N, the winter convection (or rainfall) is facilitated by the enhanced upward motion in this region. This is consistent with the positive correlation between the Indian winter-monsoon rainfall and the ENSO on an interdecadal timescale. Keeping the relationship between the global SST anomaly and rainfall in mind, the global SST anomalies for the corresponding period have been statistically analyzed with respect to non-stationarity. The presence of any collinearity between the SST anomalies during the winter and the average seasonal rainfall has been analyzed. Subsequently, the ANN model has been generated as a multilayer perceptron (MLP) in a multivariate environment. The outcome of this model work has been compared with conventional multiple linear regression (MLR). It has been observed that an ANN model in the form of an MLP shows lower prediction error than the forecast produced by the MLR.

DATA ANALYSIS

The monthly average winter (November–January) shower data from 1950 to 1998 have been collected from the Indian Institute of Tropical Meteorology (IITM) rainfall data series, available at www.tropmet.res.in. Details of the data are available on the website of IITM, Pune and in Kothawale & Rupa Kumar (2005). The SST anomaly data

pertaining to the same years are collected from http://jisao.washington.edu/data/global_sstanomts. Thus, one set of predictors include the data pertaining to November and December of year x and January of the year $(x+1)$. This combination would predict the average rainfall of the winter-monsoon period comprising November and December of the year $(x+1)$ and January of the year $(x+2)$. Thus, there would be six predictors (i.e. SST anomalies for November, December and January and rainfall amounts for November, December and January) and one predictand (i.e. average winter monsoon). Consequently, the number of time series under consideration would be six. In the study period, the time series for the predictors and the predictand would consist of 48 data points. For example, the time series of the SST anomaly for November would contain successive SSTs in November during the period from 1950 to 1997.

To have a clear pattern of the data under consideration, the auto correlation function (ACF) of the predictors and the predictand have been calculated from the following equation (Wilks 1995):

$$r_k = \frac{\text{Covariance}[(\bar{x}(n-k)),(\underline{x}(n-k))]}{\sqrt{\text{Variance}((\bar{x}(n-k)))}\sqrt{\text{Variance}((\underline{x}(n-k)))}} \quad (1)$$

where $\bar{x}(n-k)$ denotes the first $(n-k)$ data values, $\underline{x}(n-k)$ denotes the last $(n-k)$ data values and r_k denotes the lag-k autocorrelation. The ACFs computed for up to 25 lags are presented in Figure 1. It is apparent from this figure that none of the time series under consideration exhibit significant serial dependence (ACF always lies between ± 0.5). It is further visible that the ACFs do not follow any sinusoidal pattern and do not systematically tend to 0. However, it may be noted that SST anomalies have larger lag-1 autocorrelations than rainfall and it remains at a positive level up to lag-12. This indicates that SST anomaly has some extent of persistence over time. As the ACFs do not tend systematically to 0, it may be concluded that the time series under consideration are not stationary (El-Fandy *et al.* 1994).

In this step, the multiple linear regression is adopted and multicollinearity within the dataset is investigated. A stepwise removal of variables is adopted and consequently six models are generated from the six predictors. The collinearity diagnostics are presented in Table 1. The third column of Table 1 shows the standardized regression coefficients and it is found that positive as well as negative coefficients are appearing. This indicates that there is a positive as well as a negative impact of the predictors on the predictand. Moreover, it is observed that none of the coefficients are of significantly high value. This indicates that none of the predictors are having a major impact on the predictand. However, it is seen that, in model 1, which contains all of the six predictors, the SST of November is almost -0.6. It can, therefore, be said that this predictor has relatively higher influence on the predictand than the remaining five predictors. Subsequently, the values of the t-statistic are calculated and it is revealed from the next column that the significance level is above 0.05 in all cases. This indicates that, in none of the cases, the t-statistic is significant in more than 95% of cases in the long run. This leads us to conclude that the predictors are not having any prominent impact upon the predictand. In the last two columns of Table 1, the tolerance values and the variance inflation factors (VIF) have been displayed. It is observed that for the predictors SST in November and December the tolerance values are less than 0.1 and the VIF are greater than 10. This indicates the presence of multicollinearity, that means, the above two predictors are having a similar impact on the predictand. To remove the multicollinearity, the predictor SST of December is removed and the second model is generated. It is found that none of the tolerance values is less than 0.1 and none of the VIF is greater than 10. Thus, the second model is free from multicollinearity. It is found that the correlation coefficient R is almost equal in models 1 and 2. However, the other models have much smaller correlation coefficients. Thus, we identify model 2 as a relatively acceptable model and in subsequent sections we generate ANN models with the predictors selected after removal of multicollinearity, i.e. model 2. To generate the ANN model all the data are scaled to provide values between 0.1 and 0.9 as follows:

$$z_i = 0.1 + 0.8 \times \left(\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}\right) \quad (2)$$

where z_i denotes the transformed appearance of the raw data x_i. After the modeling is completed, the scaled data

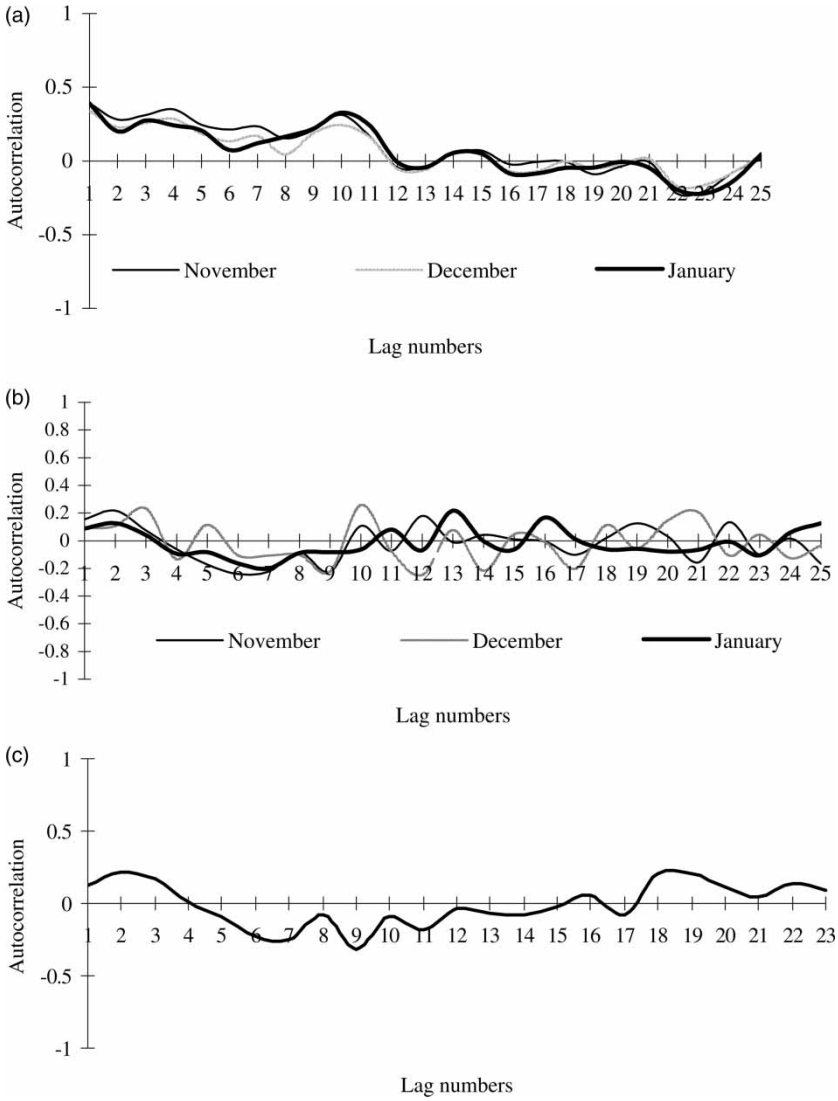

Figure 1 | Autocorrelation functions for SST anomalies (a), monthly rainfall amounts (b) and average rainfall (c) during winter monsoon.

are reverse-scaled according to

$$P_i = x_{\min} + \left(\frac{1}{0.8}\right) \times [(y_i - 0.1) \times (x_{\max} - x_{\min})] \quad (3)$$

where P_i denotes the prediction on the original scale. This transformation is performed to get rid of the asymptotic effect arising from the sigmoid activation function to be used in the ANN model. A thorough discussion on the usefulness of scaling of data prior to ANN model generation is presented (Section 5, Maier & Dandy 2000). In the case of a sigmoid activation function ($f(x) = (1 + e^{-x})^{-1}$), the data are transferred between 0 and 1. However, if the values are scaled to the extreme limits of the transfer function, the size of the weight updates is extremely small and flat spots in training are likely to occur. To avoid this problem, the scaling presented in Equation (2) has been used. The training of the data is done by the back-propagation method. The method may be described as

$$y_d = f_d(x_1, x_2, x_3, x_4, x_5, x_6) \quad (4)$$

where $x_1, x_2, \ldots, x_6$ represent the six predictors mentioned in the first paragraph of this section and y_d

Table 1 | Stepwise multiple regression with the collinearity diagnostics

Model		Standardized coefficients Beta	t	Significance	Collinearity statistics Tolerance	VIF
Model 1	Constant		4.648	0.000		
	R (Nov)	0.151	0.977	0.334	0.911	1.098
	R (Dec)	−0.149	−0.914	0.366	0.812	1.231
	R (Jan)	0.208	1.387	0.173	0.961	1.041
	SST (Nov)	−0.594	−0.996	0.325	0.061	16.434
	SST (Dec)	0.168	0.234	0.816	0.042	23.837
	SST (Jan)	0.402	0.909	0.369	0.111	9.034
Model 2	Constant		4.697	0.000		
	R (Nov)	0.149	0.977	0.334	0.913	1.095
	R (Dec)	−0.149	−0.926	0.360	0.812	1.231
	R (Jan)	0.210	1.415	0.165	0.963	1.039
	SST (Nov)	−0.482	−1.366	0.179	0.170	5.887
	SST (Jan)	0.458	1.242	0.221	0.156	6.425
Model 3	Constant		4.808	0.000		
	R (Nov)	0.172	1.151	0.256	0.940	1.064
	R (Jan)	0.229	1.561	0.126	0.981	1.019
	SST (Nov)	−0.425	−1.225	0.228	0.175	5.707
	SST (Jan)	0.357	1.016	0.315	0.170	5.865
Model 4	Constant		4.986	0.000		
	R (Nov)	0.149	1.009	0.319	0.962	1.040
	R (Jan)	0.244	1.671	0.102	0.991	1.009
	SST (Nov)	−0.106	−0.715	0.478	0.954	1.048
Model 5	Constant		5.009	0.000		
	R (Nov)	0.170	1.177	0.245	1.000	1.000
	R (Jan)	0.234	1.622	0.112	1.000	1.000
Model 6	Constant		7.888	0.000		
	R (Jan)	0.231	1.592	0.118	1.000	1.000

Note: R = rainfall; SST = sea-surface temperature; VIF = variance inflation factor; t = t-statistic.

is the predictand. The form of the function f_d is obtained after adjusting a set of weights by using the training set of the data. Within the training matrix, every row is a sample case. The percent error of the prediction (PE) has been computed (Perez & Reyes 2001) as

$$\mathrm{PE} = \frac{\langle |y_{\mathrm{predicted}} - y_{\mathrm{actual}}| \rangle}{\langle y_{\mathrm{actual}} \rangle} \tag{5}$$

where $\langle\rangle$ implies the average over the whole test set.

METHODOLOGY

In this algorithm, an initial weight vector w_k of a feed-forward neural network is iteratively adopted according to the recursion

$$w_{k+1} = w_k + \eta d_k \tag{6}$$

where w_k denotes the weight vector at the kth step. This recursion relation is used to find an optimal weight vector. Presenting a set of pairs of input and target vectors to the

network, the adaptation is performed sequentially. The quantity η is called the learning rate (Martín del Brío & Serrano Cinca 1993). The direction vector d_k is the negative of the gradient of the output error function E, which is the mean squared error at the kth step. Mathematically d_k is expressed as

$$d_k = -\nabla E(w_k) \tag{7}$$

In the learning scheme for the back-propagation algorithm (Widrow & Lehr 1990), the weight vector w_k contains the weights computed during the kth iteration and the output error function E is the multivariate function of the weights of the network. Mathematically this is expressed as

$$E(w_k) = E_p(w_k) \tag{8}$$

where $E_p(w_k)$ represents the half-sum-of-squares error function of the network outputs for a certain input pattern p. The objective of this supervised learning is to select the set of weights that minimizes E, which is the deviation between the network output and the target pattern over the complete set of training patterns. This is presented to the neural network, called an epoch (Gardner & Dorling 1998). The learning continues until E is less than a preset value at the end of an epoch.

In Equation (1), if the function f_d is nonlinear, then a nonlinear perceptron is achieved. Additional room for a good fitting of data is obtained by introducing a set of hidden nodes z_{dk} $(k = 1, 2, \ldots, n)$ in such a way that

$$z_{dk} = f\left(w_{dk_1}x_1 + \cdots + w_{dk_{24}}x_{24} + w_{dk_0}\right) \tag{9}$$

and

$$y_d = f\left(v_1 z_1 + \cdots + v_{d_n} z_d + v_{d0}\right) \tag{10}$$

The function f is defined as

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{11}$$

This form of transfer function is very popular in the ANN literature because of the form of its derivative. Introduction of this nonlinear transfer function makes the ANN able to deal with the nonlinearity associated with the input variables. Details of usefulness of this transfer function are available in Gardner & Dorling (1998) and Yegnanarayana (2000). In order to find the w and v, the back-propagation method is to be used. To implement the ANN methodology in the present problem, the entire dataset under study is divided into two subsets, namely the training set and the test set. In the ANN literature, an exhaustive variable selection is recommended to get a better performance when the problem is a multivariate prediction. In the previous section, we have already done this variable selection using the method of multiple regression. Two models have been generated using the predictors identified in the previous section. The selected predictors are rainfall amounts of November, December and January and the SST of November and January. The training set has been generated in three different ways:

1. Using a round robin method to select 70% of the dataset as the training set.
2. Using the first 70% as the training set and the remaining 30% as the test set.
3. By continuously withholding 6 years and training through the remaining years.

This ratio of the number of training and test cases (i.e. the ratio 7:3) is similar to that followed in Lundin *et al.* (1999). An adaptive gradient learning (Lundin *et al.* 1999) has been used using a modified cascade method. The sigmoid function given by Equation (11) is used as the activation function in both the hidden and output layers. Minimization of the mean squared error (MSE) has been chosen as the stopping criterion in all models. To examine the impact of variable selection, a third model has been generated using all of the six predictors that were presented to the multiple regression method. Results have been compared in the following section. The third technique is similar to that used by Arulsudar *et al.* (2005).

RESULTS AND COMPARISON

First, we discuss the performance of the multiple linear regression. The results of variable selection have been

presented in Table 1. The table displays the results of six regression models along with their collinearity diagnostics. It has been established that the SST of December should be removed to eradicate the effect of collinearity. The model correlations are displayed in Table 2. It is found that correlations in models 1 and 2 are very close to each other. However, further variable selection does not have a positive impact upon the modeling. Nevertheless, in all cases the correlations are below 0.5, which indicates that the multiple linear regression is not an acceptable tool for the forecasting problem addressed in the present section. In the present analysis, we implemented the ANN model learned through an adaptive gradient. Three ANN models have been generated. In the first ANN model, where 70% of the data have been used as the training cases and the selection has been made based on the round-robin method, the validation has been made over the entire dataset. The results are presented in Table 3. It should be noted that the inputs to the ANN pertain to the variables selected through multiple regression. It has been revealed that the correlation R between actual and predicted values is 0.5983 and the final network architecture (i.e. input nodes–hidden nodes–output node) is 5-13-1. Another ANN model is generated using the first 70% of the data as the training data and the last 30% as the test data. In this case, R is 0.8340 and the network architecture is 5-19-1. To judge the impact of variable selection, another ANN model is generated using all the six predictors, i.e. without any variable selection. In this case R is 0.5274 and the network architecture is 6-5-1. Obviously, inclusion of all predictors is reducing the correlation. It is therefore understood that the model performance is enhanced by variable selection from a given set of predictors. Moreover, it is established that the prediction performance is significantly enhanced by the application of the ANN. The degree of linear association between the actual and the predicted values in the different competitive models is judged through scatterplots in Figure 2. Figure 2(a) shows the scatterplot for the ANN model with five predictors (i.e. rainfall amounts of November, December and January and the SST of November and January) and training set formation through the round robin. The plot shows that a fraction of the set of points is close to a straight line and other fractions are significantly deviated from it. The coefficient of determination R^2, which is 0.358, is not very high. In Figure 2(b), we have used the same set of predictors, but the first 70% of data have been taken as training data. In this figure, a significant linearity is apparent from the scatterplot. Almost all of the data points are very close to a straight line with a significant positive slope. The coefficient of determination R^2 is 0.6955, which is significantly high. In Figure 2(c), we have presented the scatterplot for the ANN model without variable selection, i.e. with six predictors. Here also the first 70% of data have been taken as training data. The coefficient of determination R^2 is 0.2781, which does not indicate a good prediction capacity of the model. This indicates that removal of multicollinearity by exclusion of the collinear predictor enhances the prediction performance of the ANN. Here, the coefficient of determination R^2 further drops to 0.1341. Another ANN model is generated using the third technique mentioned earlier. The method is summarized as follows:

1. The first 6 years (1, 2, 3, 4, 5 and 6) of data are withheld and training is done on the remaining 42 years.

Table 2 | Different multiple linear regression models and the corresponding correlations (R) and coefficients of determination (R^2)

	Multiple linear regression models					
Statistic	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
R	0.3662	0.3646	0.3388	0.3050	0.2867	0.2309
R^2	0.1341	0.1329	0.1148	0.0930	0.0822	0.0533

Note: Predictors corresponding to the models are available in column 2 of Table 1.

Table 3 | The coefficient of determination (R^2) and the prediction error (PE) calculated over the set of validation cases

Model	R^2	**PE**
ANN (model 1)	0.3580	0.2819
ANN (model 2)	0.6955	0.1733
ANN (model 3)	0.2781	0.3134
MLR	0.1341	0.3583

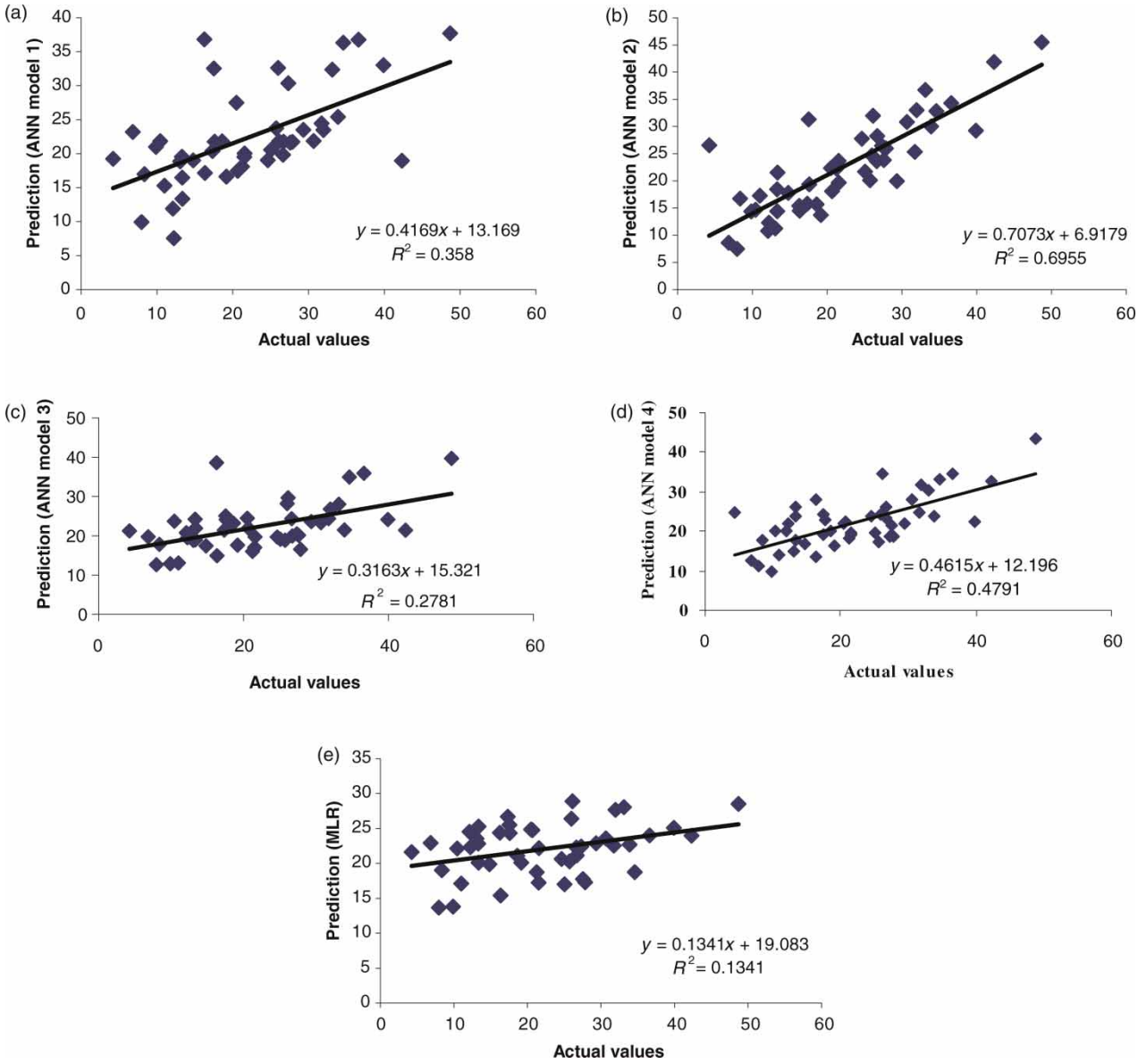

Figure 2 | Scatterplot showing the degree of linear association between the actual rainfall and those predicted by ANN with round-robin training data selection and five predictors (a), with the first 70% of the data as training data and five predictors (b), with the first 70% of the data as input set and six predictors (c), with continuous withholding of 6 years' data and training through the remaining 42 years' data and five predictors (d) and by multiple linear regression (MLR) with five predictors (e). The coefficients of determination R^2 are presented along with the trend equations.

2. Predicted values for these years (1, 2, 3, 4, 5, 6) are calculated
3. Data pertaining to the next 6 years (7, 8, 9, 10, 11 and 12) are withheld and training is done on the remaining 42 years.
4. Predicted values for these years (7, 8, 9, 10, 11 and 12) are calculated.
5. These steps are repeated to get predicted values for all eight pairs of six years' withheld data and thus the predicted values for all 48 years are obtained.

Outcomes of this model are presented in the form of a scatterplot in Figure 2(d), which shows that the coefficient of determination R^2 is 0.4791. It should be stated that, for

this model, the five previously selected variables are used as the predictors. It is observed that the value of R^2 lies above that for the ANN model with five predictors and trained using the round-robin method. However, the R^2 lies below that corresponding to the ANN model with five predictors and the first 70% of the data as the training data. Thus, the above method of selection of the training cases generates a better prediction than the round-robin method. However, selecting the first 70% of the data as the training data proves to be better than this method of selection of the training set. In Figure 2(e) we have presented the scatterplot for the multiple linear regression model with the selected five variables as in Figures 2(a), (b) and (d). The prediction capacity of multiple linear regression is significantly less than the ANN even after removal of multicollinearity. The prediction performance of the multiple linear regression as well as the ANN is further judged by calculating the statistic given in Equation (5). The prediction errors (PE) are presented in Figure 3. It is revealed in this figure that the ANN with selected variables and learned with the first 70% of the data as the training data

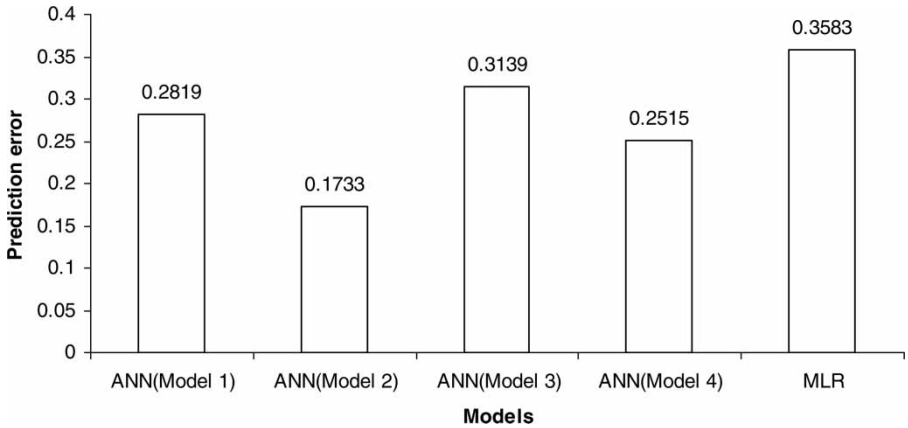

Figure 3 | Comparison of the prediction errors (PE) produced by five competitive models in predicting average winter shower over India.

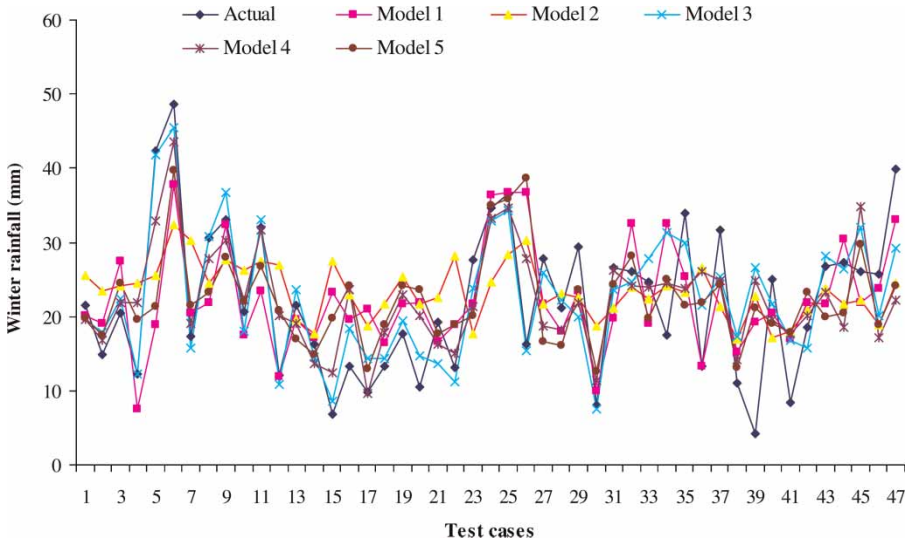

Figure 4 | Actual winter rainfall time series and the predictions from four ANN models and the multiple linear regression model pertaining to the test cases under consideration.

performs the best among all the proposed models. As a further support to the results, a line diagram, showing the actual and predicted time series, is presented in Figure 4. This figure shows that outputs from model 2 (that produced the lowest prediction error) are most closely associated with the observed values of the winter rainfall.

CONCLUSION

In the present paper, we have viewed the autocorrelation structures of monthly rainfall amounts of November, December and January over India and global sea surface temperature anomalies for the same months. Absences of persistence over time have been revealed for all of these variables through the study of autocorrelation functions computed up to several lags. The purpose of the study being the prediction of the average winter shower over India for a given year based on the above variables in the previous year, a literature survey has been done in the relevant field. Subsequently a stepwise multiple linear regression has been executed along with collinearity diagnostics. After studying the tolerance and variance inflation factor the variable sea-surface temperature anomaly for December has been excluded from the list of predictors. The prediction has been made from ANN modelling using the five selected variables. While generating the ANN model, different modes of training data selection have been used and adaptive gradient learning has been executed with the sigmoid activation function and minimization of mean square error as the stopping criteria. Assessing their performance by computation of the prediction error and formation of the scatter plot it is revealed that good prediction is possible if we divide the input set in the ratio 7:3 where the first 70% of the data constitute the data training set and performs better than that with the round-robin method of training data selection. Further, we compare the model performance with that of ANN without variable selection. It is found that the performance of the ANN drops down if no variable selection is followed. Finally, the ANN model has been compared with statistical model multiple linear regression. It is found that, despite extensive variable selection, the multiple linear regression performs very poor in the prediction problem under study. Thus, the best model from the given study is the ANN model, where the first 70% of the data constitute the training set for adaptive gradient learning, and the final architecture contains 5 input nodes, 19 hidden nodes and 1 output node.

ACKNOWLEDGEMENTS

This work is funded by the Indian Space Research Organization (ISRO) through the S K Mitra Centre for Research in Space Environment, University of Calcutta, Kolkata. Sincere thanks are due to the anonymous reviewer who has given constructive comments to enhance the quality of the paper.

REFERENCES

Aldrian, E. & Dwi Susanto, R. 2003 Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. *Int. J. Climatol.* **23**, 1435–1452.

Annamalai, H., Liu, P. & Xie, S.-P. 2005 Southwest Indian Ocean SST variability: its local effect and remote influence on Asian monsoons. *J. Climate* **18**, 4150–4167.

Arulsudar, N., Subramanian, N. & Muthy, R. S. 2005 Comparison of artificial neural network and multiple linear regression in the optimization of formulation parameters of leuprolide acetate loaded liposomes. *Int. J. Pharmacy Pharmaceut. Sci.* **8**, 243–258.

Chattopadhyay, S. 2007 Feed forward artificial neural network model to predict the average summer-monsoon rainfall in India. *Acta Geophys.* **55**, 369–382.

Chattopadhyay, S. & Chattopadhyay, G. 2008 Comparative study among different neural net learning algorithms applied to rainfall time series. *Meteorol. Appl.* **15**, 273–280.

Clark, O. C., Cole, J. E. & Webster, P. J. 2000 Indian Ocean SST and Indian summer monsoon rainfall: predictive relationships and their decadal variability. *J. Climate* **14**, 2503–2519.

Deser, C., Phillips, A. S. & Hurrell, J. W. 2004 Pacific interdecadal climate variability: linkages between the tropics and the North Pacific during boreal winter since 1900. *J. Climate* **17**, 3109–3124.

El-Fandy, M. G., Taiel, S. M. M. & Ashour, Z. H. 1994 Time series models adoptable for forecasting Nile floods and Ethiopian rainfall. *Bull. Am. Meteorol. Soc.* **75**, 83–94.

Gardner, M. W. & Dorling, S. R. 1998 Artificial neural network (multilayer perceptron) – a review of applications in atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636.

Guhathakurata, P. 2008 Long lead monsoon rainfall prediction for meteorological sub-divisions of India using deterministic artificial neural network. *Meteorol. Atmos. Phys.* **101**, 93–108.

Holton, J. R. 1972 *An Introduction to Dynamic Meteorology*. Academic, San Diego.

Hsieh, W. W. & Tang, B. 1998 Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Am. Meteorol. Soc.* **79**, 1855–1870.

Kothawale, D. R. & Rupa Kumar, K. 2005 On the recent changes in surface temperature trends over India. *Geophys. Res. Lett.* **32** (L18), 714.

Kumar, B. 2005 Impact of ENSO on winter monsoon rainfall over South India. *Geophys. Res.* Abstract 7, 03869, SRef-ID: 1607-7962/gra/EGU05-A-03869.

Latif, M. 1998 Dynamics of interdecadal variability in coupled ocean–atmosphere models. *J. Climate* **11**, 602–624.

Lundin, M., Lundin, J., Burke, H. B., Toikkanen, S., Pylkkanen, L. & Joensuu, H. 1999 Artificial neural networks applied to survival prediction in breast cancer. *Oncology* **57**, 281–286.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ. Modell. Software* **15**, 101–124.

Maity, R. & Nagesh Kumar, D. 2007 Hydroclimatic teleconnection between global sea surface temperature and rainfall over India at sub divisional monthly scale. *Hydrol. Process.* **21**, 1802–1813.

Maqsood, I., Khan, M. R. & Abraham, A. 2002 Neurocomputing based Canadian weather analysis. In *Proc. 2nd International Workshop on Intelligent Systems Design and Application, Atlanta, GA*. Dynamic Publishers Inc., Atlanta, GA, pp. 39–44.

Martín del Brío, B. & Serrano Cinca, C. 1993 Self-organizing neural networks for analysis and representations of data: some financial cases. *Neural Comput. Appl.* **1**, 193–206.

Moron, V., Philippon, N. & Fontaine, B. 2004 Simulation of West African monsoon circulation in four atmospheric general circulation models forced by prescribed sea surface temperature. *J. Geophys. Res.* **109** (D24), 105.

Perez, P. & Reyes, J. 2001 Prediction of particulate air pollution using neural techniques. *Neural Comput. Appl.* **10**, 165–171.

Raj, Y. E. A. 1996 Inter and intra-seasonal variation of thermodynamic parameters of the atmosphere over coastal Tamilnadu during northeast monsoon. *Mausam* **47** (3), 259–268.

Rao, G. N. 1999 Variations of the SO relationship with summer and winter monsoon rainfall over India: 1872–1993. *J. Climate* **12**, 3486–3495.

Rao, K. V. 1963 A study of the Indian northeast monsoon season. *Ind. J. Meteorol. Geophys.* **14**, 143–155.

Reason, C. J. C. & Mulenga, H. 1999 Relationships between South African rainfall and SST anomalies in the Southwest Indian Ocean. *Int. J. Climatol.* **19**, 1651–1673.

Reddy, P. R. C. & Salvekar, P. S. 2003 Equatorial East Indian Ocean sea surface temperature: a new predictor for seasonal and annual rainfall. *Curr. Sci.* **85**, 1600–1604.

Sahai, A. K., Grimm, A. V., Satyan, V. & Pant, G. B. 2003 Long-lead prediction of Indian summer monsoon rainfall from global SST evolution. *Climate Dyn.* **20**, 855–863.

Sahai, A. K., Soman, M. K. & Satyan, V. 2000 All India summer monsoon rainfall prediction using an artificial neural network. *Climate Dyn.* **16**, 291–302.

Singh, O. P. 1995 Influence of Bay of Bengal on winter monsoon rainfall. *Mausam* **46** (3), 307–312.

Sivakumar, B. 2000 Chaos theory in hydrology: important issues and interpretations. *J. Hydrol.* **227**, 1–20.

Venkatesan, C., Raskar, S. D., Tambe, S. S., Kulkarni, B. D. & Keshavamurty, R. N. 1997 Prediction of all India summer monsoon rainfall using error-back-propagation neural networks. *Meteorol. Atmos. Phys.* **62**, 225–240.

Wang, B., Wu, R. & Li, T. 2003 Atmosphere–warm ocean interaction and its impacts on Asian–Australian monsoon variation. *J. Climate* **16**, 1195–1211.

Wen, C., Graf, H. F. & Ronghui, H. 2000 The interannual variability of East Asian winter monsoon and its relation to the summer monsoon. *Adv. Atmos. Sci.* **17**, 48–60.

Widrow, B. & Lehr, M. A. 1990 30 years of adaptive neural networks: perceptron, madaline and backpropagation. *Proc. IEEE* **78**, 1415–1442.

Wilks, D. S. 1995 *Statistical Methods in Atmospheric Sciences*. Academic, New York.

Yang, J., Liu, Q. & Liu, Z. 2010 Linking Asian monsoon to Indian Ocean SST in the observation: possible roles of Indian Ocean basin mode and dipole mode. *J. Climate* **23**, 5889–5902.

Yegnanarayana, B. 2000 *Artificial Neural Networks*. Prentice-Hall of India Pvt Ltd, New Delhi.

First received 2 October 2010; accepted in revised form 16 April 2011. Available online 2 August 2011