International Journal of Soft Computing 12 (1): 13-19, 2017
ISSN: 1816-9503
© Medwell Journals, 2017

Skill Scores Verification for all India Rainfall Data Using Artificial Neural Network

[1]Neeta Verma, [2]Y.D.S Arya and [3]K.C. Tripathi
[1, 3]Department of Computer Science and Engineering,
Inderprastha Engineering College, Ghaziabad, 201004 Uttar Pradesh, India
[2]Invertis University, Bareilly, 201004 Uttar Pradesh, India

Abstract: There is great regional and temporal variation in the distribution of rainfall. The great variation in the amount of rainfall both spatially and temporally, the high degree of uncertainty related to the date of arrival, etc., are unexplained. A neural network model for analyzing the all India rainfall has been developed by using the 142 years of data. In formulating artificial neural network based predictive models three layered network has been constructed. The models under study are different in the number of hidden neurons. The main objective of this study is to evaluate the applicability of ANN. The performance of different networks have been evaluated and tested. The reason of using the ANN (Artificial Neural Network) model is based on prediction by smartly analyzing the trend from the previously existing data set. In the present research, the last 142 years data of all Indian rainfall has been analyzed through artificial neural network models. The Artificial Neural Network (ANN) technique with back-propagation algorithm for the predictability of AIR with 1 lag by analysing the historical time series of 142 years of AIR data. The ANN model used to forecast rainfall that is validated using the correlation coefficient and Root Mean Square Error (RMSE). The statistical parameters are not sufficient for model accuracy, so for accuracy skill scores for verification areobtained in this study.

Key words: Artificial Neural Network (ANN), statistical parameters, backpropagation, skill scores parameters, India

INTRODUCTION

The monsoons effect most part of India, the amount of rainfall varies from heavy to scanty on different parts, so the prediction of rainfall is very demanding and challenging problem. Meteorologists have been trying to explain these phenomena from different angles relating to wide variety of generalisation. Rainfall is one of the most important climate variable that can affect the environments by two destructive ways droughts and floods. Rainfall plays a destructive role f or agriculture. Heavy rainfall at the end of the crop cycle causes damages of crops. The positive affect of the rainfall prediction is the planning of agricultural strategies. The agriculture production highly depends on the rainfall. The decision of crop selection and the output of agricultural production are highly determined by rainfall and water availability. Crops are affected through rainfall in two different ways-high and low rainfalls. So, the prediction of rainfall will play an important role for agriculture. The technological advancement will play the most crucial role to solve the problems of agriculture.

The forecasting of the rainfall is one of the challenging problem. An Artificial Neural Networks (ANN) based on feed-forward back-propagation architecture is an innovative technique for rainfall forecasting. ANN is significantly improved approach for rainfall forecasting over a long period. An artificial neural network is a capable model for very irregular rainfall. By using the ANN for a catchment located in a semiarid climate in Morocco, it is observed that the ANN approach is more suitable to predict river runoff than classical regression (Kumarasiri and Sonnadara, 2006; Riad *et al*., 2004). The multilayer perceptron network with back propagation and linear regression is used for rainfall time series prediction. By the comparison of results for tested data it is predicted that ANN is more accurate technique as compared to linear regression for rainfall prediction (Kumar *et al*., 2012; Gupta *et al*., 2014). ANNs is also compared with linear regression tested for predictability of sea surface temperature anomalies of small area of indian ocean region, the better results are obtained by AAN (Tripathi *et al*., 2006). It is observed that the Back Propagation Algorithm (BPA) is the best algorithm among BPA, LRN CBP in multi-layer architecture of ANN. LEARNGDM is the best learning function to train the data, TRAINLM is the best training function for rainfall prediction (Kumar *et al*., 2012). The ANN is one of the

Corresponding Author: Neeta Verma, Department of Computer Science and Engineering, Inderprastha Engineering College, Ghaziabad, 201004 Uttar Pradesh, India best approach for cyclone intensity prediction also. By the analysis of dataset for western north Pacific Ocean for 1997-2004 from Joint Typhoon Warning Centre, it is observed that using the Multiple Linear Regression (MLR) and the ANN-based, the system performance improves upon MLR as the lead hour increases from 12-120 h (Sharma *et al.*, 2013). A Neural Network Approach to Estimate Tropical Cyclone Heat Potential (TCHP) in the Indian Ocean shows the superiority with other approaches. The approach is justified by evaluating the root-mean-square error and the scatter index. The utility of the ANN technique in estimating TCHP with better accuracy for the North Indian Ocean helps in improving the cyclone track and intensity predictions (Ali *et al.*, 2012).

The ANN is suitable for the monthly time series data. By using the ANN, prediction of monthly precipitation data in Mashhad synoptic station is done and good results are found (Khodashenas *et al.*, 2010). The ANN can be used for prediction of Indian summer monsoon rainfall using Niño indices. The seasonal forecast skills of the Indian Summer Monsoon Rainfall Index (ISMRI) are improved by ANN. The ANN is tested by correlation analysis to see the effect of SST indices of Nino-1+2.

Nino-3, Nino-3.4 and Nino-4 regions on ISMRI with a lag period of 1-8 seasons models. A comparison of multiple linear regression and Artificial Neural Networks (ANNs) states that the ANN model has better predictive skills than all the linear regression models (Shukla *et al.*, 2011). The large scale climate indices like El-Nino Southern Oscillation (ENSO), EQUitorial Indian Ocean Oscillation (EQUINOO) and a local climate index of Ocean-Land Temperature Contrast (OLTC) are used for rainfall forecasting. To handle the highly non-linear and complex behaviour of the climatic variables for forecasting of rainfall, the Artificial Neural Networks (ANNs) can be used (Kumar *et al.*, 2007).

In addition to ANN alone, the wavelet technique with Artificial Neural Network (ANN) is more effective approach. By combination of ANN and wavelet technique the monthly rainfall is forecasted at Darjeeling rain gauge station using 74 years data. The result shows that combination is better approach than ANN alone (Ramana *et al.*, 2013). The ANN can be combined with other approaches to improve the performance of the system. The ANN is trained using RGA will improve the performance for forecasting. By using Average Absolute Relative Error (AARE), Normalized Mean Bias Error (NMBE) and mean error in estimating peak flow (MF%), it is observed that RGA is a good technique for training of ANN. The combination of SOM and BPA model obtained the best for Pearson's correlation coefficient (R), Nash-Sutcliff Efficiency (E), Normalized Root Mean Square Error (NRMSE) and persistence coefficient (Eper) statistics. So, RGA is one of the suitable approach as compared to BPA for train the ANN (Srinivasulu and Jain, 2006). Many parameters affects the rainfall forecasting. Sea surface temperature, sea level pressure, humidity, zonal (u), meridional (v) winds. The ANN is capable for using the multiple parameters. Self Organizing Maps (SOM) via a clustering based ANN technique is one of reasonably good technique for accuracy of rainfall forecasting. The forecasting of the rainfall based of division of datasets like monthly, biannually, quarterly. Focused Time Delay Neural Networks (FTDNN) for rainfall forecasting is the good approach for variation of datasets. The forecast accuracies decrease for the biannual, quarterly and monthly datasets (Htike and Khalifa, 2010). Neural networks and support vector regression are the approaches that can be used for forecasts of tropical pacific sea surface temperatures. SVR has two structural advantages over neural network models that are no multiple minima in the optimization process and an error norm robust to outliers in the data but it did not give better overall forecasts than ANN (Martinez and Hsieh, 2009). The relationship between rainfall and lagged indices can be observed by the ANN. That relationship will affect the rainfall of the region. The ANN is able to provide the higher correlations using lagged indices. The correlations can be improved upto 99% (Mekanik and Imteaz, 2012). The feed forward multi-layered artificial neural network is useful technique for the estimation of the maximum surface temperature and relative humidity needed for the genesis of severe thunderstorms. The ANN is an efficient forecasting tool to forecasting the occurrence of high frequency small-scale weather systems like severe local storms (Chaudhuri and Chattopadhyay, 2005). The rainfall data and temperature were analysed to find the seasonal trend. The statistical significance of the trend in the the time series was analysed by Mann-Kendall test. The significance was tested at 95 and 99% levels of confidence (Laskar *et al.*, 2014). The statistical parameters like correlation coefficient, root mean square error and standard deviation are used to investigate the simulation of Antartic sea ice. The quality of fore cast can not be proved by onle these parameters, so other skill attiribute of the forecasting are also considered. The results shows that simulation have fairly good accuracy (Tripathi and Das, 2008).

MATERIALS AND METHODS

Data: Selection of dataset for any research work is very important to meet the objectives. The 140 years monthly data set of (1871-2010) of ALL-INDIA RAINFALL of 30 meteorological subdivisions encompassing 2,880,324 SQ.

KM with a resolution of up to 0.1 mm/month obtained from the Indian Institute of Tropical Meteorology website.

Preprocessing and partitioning: To protect the output of a neuron from being driven to saturation, the input to it, i.e., the activation, must not be too small or too large. In the present study, the data is normalized to the range (0.2, 0.8) using the following normalization equation. With lag value 1 the predictor and predictand series has been normalized using the following equation:

$$X_n = \left[(X\text{-}X)/(X_{max}\text{-}X_{min})\right]\times 0.6+0.2 \quad (1)$$

Where:

X = The rainfall value
X_{max} = The maximum value
X_{min} = The min value
X_n = The normalized value of the rainfall

After obtaining the normalized data, the next step is to train the input data using back-propagation algorithm. The training data is all India rainfall data for the years (1871-2010). The (1871-2000) data is used for training and validation with 1 month lag value. Total 1560 data points are used for training. The correlation coefficient between input and target is calculated and it is observed that CC is >5. The remaining (2001-2010) 120 data points are used for testing purpose:

- Training and validation: 92% (130 years data)
- Testing: 8% (10 years data)

The training is performed for 1000 epochs. The training data is divided into 70% for training, 15% for validation and 15% for testing purpose. Gradient descent backpropagation is used for training purpose. The momentum 75 is used for this purpose.

Introduction of ANN

Artificial neural network: An ANN is a highly interconnected network of many simple processing units called neurons which are analogous to the biological neurons in the human brain. The strength of connection between the two neurons in adjacent layers is represented a 'connection strength' or 'weight'. An ANN normally consists of three layers, an input layer, a hidden layer and an output layer.

Figure 1 represent the model of simple neuron. It will represent the calculation of summing junction. The net input at the summing junction can be written as:

$$U_k = \sum_{j=1}^{m} W_{kj}X_{kj}$$

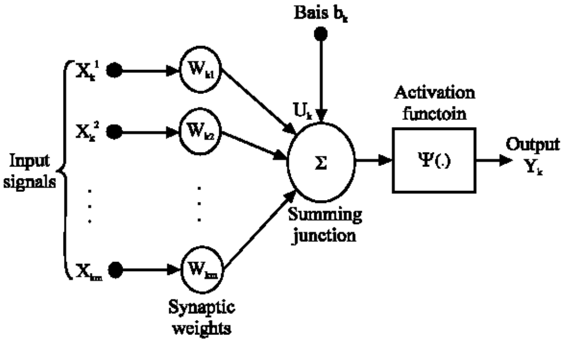

Fig. 1: Model of a simple neuron

The output U_k is worked upon an activation function, whose sole purpose is to limit the output of the neuron to a desired value, increasing the net performance of the network. Then the output from the kth Neuron is: $Y_{k=\Psi}(U_k+b_k)$. ANNs are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found. The main parameters for ANN modelling are network topology, neurons characteristics, training and learning rules. The number of neurons is based on the number of input and output data. For preparing the suitable network, following are the criteria to be considered:

- Suitable input data
- Determination of the appropriate number of hidden layers and neurons
- Proper training and testing of the network

The selection of a suitable ANN architecture will play a significant role in performance of the model. If the architecture is too small, the network may not have sufficient degree of freedom to learn the process correctly. On the other hand if the network is too large, it may not converge during training, or it may overfit the data. In this study the ANN model consists one hidden layer and an output layer. The hyperbolic tangent sigmoid transfer function:

$$f(n) = 2/\left(1+\exp(-2\times n)\right)-1 \quad (2)$$

With bias has been used as the activation function for the neurons in hidden layer. For the neurons in the output layer the identity function has been used as the activation function:

$$f(n) = n \quad (3)$$

One neuron for input and one for the output. In this model, we have tested with different number of neurons

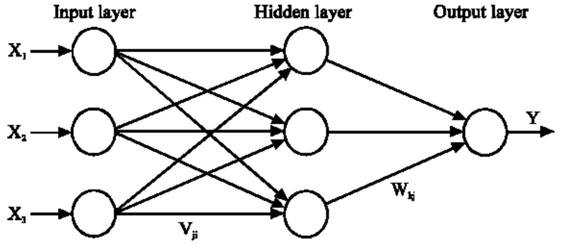

Fig. 2: Structure of feed forward ANN

and find out that five neurons for the hidden layer with a learning rate of 0.2 are providing some valuable results.

Backpropagation: Backpropagation is a common method of training artificial neural networks so as to minimize the objective function. It is a supervised learning method with generalization of the delta rule. BPN repeatedly adjust the parameters (weights and biases) to minimize the error between target output and estimated output according to generalized delta rule (Fig. 2).

The term feed-forward refers to the inputs being swept forward through the network, getting multiplied by each synaptic weight and being summed at each node until the output node is met. Then, back-propagation occurs where the desired output is compared with the produced output and errors are backwardly propagated through the network. The synaptic weights for each layer are adjusted in proportion to minimize this error, the adjustment of weights is limited by a factor defined by the user because the large adjustments may cease to occur for stray values in the data set. The ANNs learning is the process of finding the optimal weight matrices in a systematic manner in order to achieve the desired value of target outputs. The performance measure used to train ANNs is minimizing the network error function which is given as:

$$E = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{i,j} - y^{t}_{i,j})^2 \qquad (4)$$

Where:

E = The error
$y^t_{i,j}$ = Desired (target) value for jth output node and ith pattern
$y_{i,j}$ = Computed value for jth output node and ith pattern
q = Number of output nodes
p = Number of training patterns

RESULTS AND DISCUSSION

The optimal structure of developed artificial neural network for obtaining the minimum prediction error is shows in Table 1. Table 2 shows various statistics to evaluate the results. The correlation between (input and target) and (target and output) is >0.5 in both the cases. The ANN is trained by using the Back propogation algorithm. The performance of various networks are analised on the basis of statistical parameters. The standard deviation and root mean square error is calculated for different networks. The networks are designed with the variation of no of neurons. It is observed that the network with 5 neurons have minimum error for training and testing data. In both the cases (training and testing) root mean square error is less than the standard deviation. With these results, it is concluded that designed system is capable for prediction. The RMSE is close for training and testing data. Similarly, SD for training and testing data is also closer.

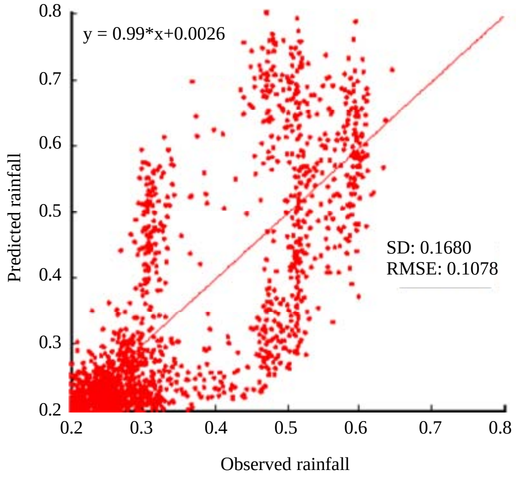

Fig. 3: Scattered plot of observed and predicted rainfall values for training data (1871-2000)

Table 1: Parameters of ANN

Parameters	Feature
Epochs	1000
Learning rate	0.2
No of hidden layers	1
Activation function (hidden layer)	Hyperbolic tangent sigmoid transfer function
Activation function (outer layer)	Identity function
Training algorithm	Back propagation

Table 2: Training and testing results on the basis of statistical parameters

CC B/W input and target	CC B/W output and target	RMSE	SD
Training results			
0.7487	0.7711	0.1078	0.168
Testing results			
-	0.7228	0.1152	0.164

No. of neurons = 5; CC: correlation coefficient; RMSE: Root Mean Square Error, SD: Standard Deviation

Figure 3 shows the scattered plot of predicted and observed rainfall values for training data during the years (1871-2000). Figure 3 and 4 shows the flow of predicted and observed training data during the (1857-2000) and testing data during the year (2001-2010).

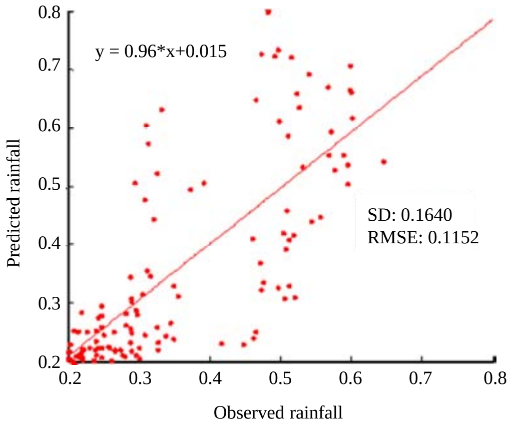

Fig. 4: Scattered plot of observed and predicted rainfall values for testing data (2001-2010)

Skill scores for verification: The network is designed with different no of neurons and different activation functions. By the analysis, it is observed that the root mean square error is less than standard deviation in both the cases (training and testing). Such network can be used to provide the suitable results for forecasting. The correlation between output and target is also suitable for the network utility. The Indian rainfall parameters are simulated using ANN on the basis of commonly used statistical parameters such as correlation coefficient, root mean square value and standard deviation but these parameters are not sufficient to define the accuracy of the forecast, therefore other skill scores are to be used to define the accuracy of the forecast system.

Evaluation of forecast skills: On the basis of the analysis of these skill scores we can conclude that ANN is capable for rainfall forecasting with good accuracy. Verification is one aspect of measuring forecast goodness. Skill Score (SS) measures forecast accuracy relative to some set of control or reference forecast. With the dichotomous variables the precipitation/reflectivity above or below threshold on a grid is represented. A dichotomous forecast says, “yes, an event will happen” or “no, the event will not happen”. Verification measures the quality of forecasts:

- Hit-event forecast to occur and did occur
- Miss-event forecast not to occur but did occur
- False alarm-event forecast to occur but did not occur
- Correct negative-event forecast not to occur and did not occur

To verify this type of forecast we start with a contingency table that shows the frequency of “yes” and “no” forecasts and occurrences. The four combinations of forecasts (yes or no) and observations (yes or no), called the conditional distribution. Table 3 is a contingency table that illustrating the counts used in verification statistics for dichotomous (e.g., Yes/No) forecasts and observations. The counts in this table can be used to compute a variety of traditional verification measures. For evaluating the quality of forecast based on Binary of dichotomous (yes/no) forecasts, the test data has been classified into two classess, normal events and rare events. A normal event is one when the observed anomaly is under the standard deviation of the observed data. A rare event is defined as the case when the observed anomaly exceeds the standard deviation of the observed data. A perfect forecast system would produce only hits and correct negatives and no misses or false alarms.

Table 3: The 2×2 contingency table

Observed	Yes	No	Total
Forecast			
Yes	Hit	False alarm	Forecast Yes
No	Miss	Correct negative	Forecast No
Total	Obs. yes	Obs. No	Total

Table 4: Measured parameters

Observed	Yes	No	Total
Forecast			
Yes	15	5	20
No	13	87	100
Total	28	92	120

Table 4 shows the results of counts used in the verification statistics for dichotomous forecast. There are several categorical statistics that can be computed from the yes/no contingency table.

Accuracy: The accuracy is the measure of level of agreement between the forecast an the truth. The difference between the forecast and the observation is the error.

Bias: This is the fraction of the forecast frequency of the yes events and the observed frequency of the yes events.

Probabilty of dection: The Probability of Detection (POD) indicates the fraction of the observed “yes” events that were correctly forecast. It is the fraction of the total number of times the rare event is detected to the total number of times the rare event actually occurred.

False alarm ratio: The false alarm ratio is a measure of the fraction of the predicted “yes” events that actually did not occur (false alarm).

Probability of false detection: Probability of False Detection (POFD) is another attribute that tells us what fraction of the observed “no” events were incorrectly forecast as “yes”.

Table 5: Forecast verification scores

ACC	BIAS	POD	FAR	POFD	TS	ETS
0.85	0.714	0.535	0.25	0.05	0.45	0.3641

ACC: Accuracy of forecast; BIAS: Bias of forecast, POD: Probability of Detection; FAR-False Alarm Ratio; POFD: Probability of False Detection; TS: Threat Score; ETS: Equitable Threat Score

Threat Score (TS): Threat Score (TS) measures the degree of correspondence between the forecast "yes" events and the observed "yes" events.

Equitable Threat Score (ETS): Equitable Threat Score (ETS) Measures the fraction of observed and/or forecast events that were correctly predicted after adjusted for hits associated with random chance.

Verification results: Table 5 shows the forecast verification scores. Table 5 shows the various skill scores of the dichotomous forecasts for test cases. The accuracy is observed 85%, means forecast is correct by 85%. Bias that is a fraction of forecast frequency of yes events and observed frequency of yes events is 0.714.The standard value of this should be 1. In the worst case the ratio can be 0 (if there is no "yes" in the forecast set or infinity (if there is no "yes" in the observed case). If this bias value is >1 then there is overforecast and if <1 there is underforecast. In our study, there is underforecast. The "Probability of Detection (POD)", ideally should be 1 and worst case (if there is no hits) it is 0. In our case its value is 0.535. The false alarm ratio that is measure of fraction of predicted "yes" events that actually did not occur (false alarm). The ideal value should be 0. In our case it is observed 0.25. The Probability Of False Detection (POFD) is ideally should be 0, in our case it is observed 0.05 that is very close to zero. Threat score that is measure of degree between forecast "yes" events and the observed "yes" events. Ideal value to be 1 and worst case is 0, means no correspondence. In our study, TS is 0.45. TS is used in conjunction with Equitable Threat Score. ETS is always less than the TS. In our study, its value is 0.3641.

We have analysed the ANN for rainfall prediction on the basis of statistical parameters such as correlation coefficient, root mean square error, standard deviation. All parameters are supporting to our system but for rare events various other attributes are also calculated. On the basis above calculated parameters and attributes, we can suggest that ANN is capable for rainfall prediction.

CONCLUSION

The weather prediction using different parameters is an innovative problem. Different parameters are used for prediction but only forecasting is not a solution for the problem, it should be analysis for goodness of system. The verification of the forecasting system accuracy can be done on the basis of skill scores. This model is tested to forecast and verified the accuracy for all India rainfall. It is observed that the results obtained for forecasting are suitable. The accuracy is 85% for the tested data which is a satisfactory result for rainfall prediction. The other skill score parameters like bias of forecast, probability of dection, false alarm ration, probability of false dection, threat score and equitable threat score are producing the suitable and satisfactory results. The model was statically evaluated and strong correlation coefficient are obtained between target and output values.

REFERENCES

Ali, M.M., P.S.V. Jagadeesh, I.I. Lin and J.Y. Hsu, 2012. A neural network approach to estimate tropical cyclone heat potential in the Indian Ocean. IEEE. Geosci. Remote Sens. Lett., 9: 1114-1117.

Chaudhuri, S. and S. Chattopadhyay, 2005. Neuro-computing based short range prediction of some meteorological parameters during the pre-monsoon season. Soft Comput., 9: 349-354.

Gupta, P., S. Mishra and S.K. Pandey, 2014. Time series data mining in rainfall forecasting using artificial neural network. Intl. J. Sci. Eng. Technol., 3: 1060-1065.

Htike, K.K. and O.O. Khalifa, 2010. Rainfall forecasting models using focused time-delay neural networks. Proceedings of the 2010 International Conference on Computer and Communication Engineering (ICCCE), May 11-12, 2010, IEEE, New York, USA., ISBN:978-1-4244-6233-9, pp: 1-6.

Khodashenas, S.R., N. Khalili, K. Davari and M.M. Bayagi, 2010. Monthly precipitation prediction by artificial neural networks (Case study: Mashhad synoptic station). Novatech, Portsmouth, England.

Kumar, A., D.S. Pai, J.V. Singh, R. Singh and D.R. Sikka, 2012. Statistical models for long-range forecasting of southwest monsoon rainfall over India using step wise regression and neural network. Atmos. Clim. Sci., 2: 1-15.

Kumar, D.N., M.J. Reddy and R. Maity, 2007. Regional rainfall forecasting using large scale climate teleconnections and artificial intelligence techniques. J. Intell. Syst., 16: 307-322.

Kumarasiri, A.D. and D.U.J. Sonnadara, 2006. Rainfall forecasting: An artificial neural network approach. Proc. Techn. Sessions, 22: 1-13.

Laskar, S.I., S.D. Kotal and S.K.R. Bhowmik, 2014. Analysis of rainfall and temperature of selection stations over North East India during last century. Mausam, 1: 497-508.

Martinez, S.A. and W.W. Hsieh, 2009. Forecasts of tropical pacific sea surface temperatures by neural networks and support vector regression. Intl. J. Oceanogr., 2009: 1-13.

Mekanik, F. and M.A. Imteaz, 2012. A multivariate artificial neural network approach for rainfall forecasting: Case study of Victoria, Australia. Proceedings of the World Congress on Engineering and Computer Science, Vol. I, October 24-26, 2012, IAENG, San Francisco,California,USA.,ISBN:978-988-19251-6-9, pp: 557-561.

Ramana, R.V., B. Krishna, S.R. Kumar and N.G. Pandey, 2013. Monthly rainfall prediction using wavelet neural network analysis. Water Resour. Manage., 27: 3697-3711.

Riad, S., J. Mania, L. Bouchaou and Y. Najjar, 2004. Rainfall-runoff model usingan artificial neural network approach. Math. Comput. Modell., 40: 839-846.

Sharma, N., M.M. Ali, J.A. Knaff and P. Chand, 2013. A soft-computing cyclone intensity prediction scheme for the Western North Pacific Ocean. Atmos. Sci. Lett., 14: 187-192.

Shukla, R.P., K.C. Tripathi, A.C. Pandey and I.M.L. Das, 2011. Prediction of Indian summer monsoon rainfall using Nino indices: A neural network approach. Atmosp. Res., 102: 99-109.

Srinivasulu, S. and A. Jain, 2006. A comparative analysis of training methods for artificial neural network rainfall-runoff models. Applied Soft Comput., 6: 295-306.

Tripathi, K.C. and I.M.L. Das, 2008. Simulation of Antarctic sea ice area with artificial neural network. J. Marine Sci., 37: 77-85.

Tripathi, K.C., I.M.L. Das and A.K. Sahai, 2006. Predictability of sea surface temperature anomalies in the Indian Ocean using artificial neural networks. J. Marine Sci., 35: 210-220.