

Artificial Neural Network – a Tool for Prediction of Monsoon Rainfall over North and South Assam in India

S.S. De, G. Chattopadhyay, A. Debnath

Centre of Advanced Study in Radio Physics and Electronics 1, Girish Vid-yaratna Lane, Kolkata 700 009, India

Received 1 March 2009

Abstract. The possibility of predicting average summer-monsoon rainfall over North and South Assam in India has been analyzed in this paper through Artificial Neural Network models. In formulating the Artificial Neural Network based predictive model, three layered networks have been constructed with sigmoid non-linearity. The models under study are different in the number of input layers. After a thorough training and test procedure, neural net with two inputs is found to be the best predictive model both for North and South Assam. Finally, the results are compared with outputs of regression approach. After comparison it is found that for South Assam, Neural Network completely outperforms the regression. But in North Assam regression can be used as an alternative approach to Neural Network.

PACS number: 92.60.Ry, 92.60.Wc, 92.60.Nv

1 Introduction

The word “monsoon” is used to designate seasonally reversing circulation system and is usually associated with the tropical regions of Asia. This monsoon governs the climate of Indian subcontinent. The principal attention is usually focussed on the surface winds of the summer monsoon of Southern Asia. Prominent circulation patterns also exist aloft. Near 150 mb a relatively strong easterly jet is common. It extends from Southeast Asia to Eastern Africa that generates very strong winds over Southern India. The existence of this tropical easterly jet implies the existence of a deep layer of relatively warm air to its North and colder air to its South over the Indian Ocean. The heating of the jet at its North arises from the strong solar heating of the elevated surface of the Tibetan plateau which contributes large amounts of latent heat in the process. Here humid southwest monsoon air encounters the highlands of Northeastern India.

Air entering the eastern end, faces the stronger southward-directed pressure gradient that exits within the region of the jet. By this, air temporarily gains a slight northerly component, which in turn produces an area of upper level divergence to the North of the eastern end of the jet and an area of upper level convergence to the South. The area of upper level divergence coincides with the region of ascending air and heavy monsoon rainfall over Northeastern India.

Northeast India (east of 88° E and north of 21° N) has distinct precipitation and drainage patterns due to its unique location and orography [1-3]. From different analysis, it has been established that seasonal rainfall patterns over the northeastern region contrasts to that over the rest of the country. During the months of June to September, southwesterly monsoons supply about 70% of the annual precipitation. The monsoon period over Assam is from the first week of June up to the second week of October. The rainfall is highest in June and then it gradually decreases thereafter.

The maximum annual rainfall of India is due to southwest monsoon. In the case of Assam, the normal monsoon rainfall (June to September) is about 164 cm and forms about 66 per cent of the annual rainfall; but for the individual months of the monsoon season, the variability is 19.3 in June, 18.4 in July, 18.2 in August and 24.4 per cent in September. The rainfall during the rest of the months is made up of a few spells of above or even below normal rainfall. This type of monsoon rainfall causes floods in certain periods of the month even when the total monthly rainfall itself is just normal or even below normal.

The atmosphere is very much chaotic by nature and no prior assumptions can be made while developing any model for chaotic atmospheric processes. Unlike the stochastic modelling techniques, the Artificial Neural Networks (ANN) are capable of modelling highly non-linear relationships without any prior assumption and can be trained to generalize accurately when presented with a new data set. The ANN acts as parallel computational models, comprised of densely organized adaptive processing units. The vital characteristic of neural networks is their adaptive nature that makes the ANN techniques very alluring in application domains for solving problems where the internal physical processes are highly complex and non-linear [4].

This paper endeavors to develop an ANN model to forecast average rainfall during summer-monsoon period in North and South Assam in India. There is developed statistical model using regression method to predict the anomaly in Indian summer monsoon rainfall. The problem has been studied extensively by different workers. Some works revealed that Indian summer monsoon predictability exhibits epochal variation. Kishitawal *et al.* [5] evoked the feasibility of a nonlinear technique based on genetic algorithm (Artificial Intelligence) for the prediction of summer rainfall over India. Guhathakurata [6] introduced ANN to forecast the summer-monsoon rainfall over the Kerala state in India. Instead of choosing a particular state, the present authors implemented Backpropagation

ANN to forecast the average summer-monsoon rainfall over the whole country. The aroma of newness further lies in the fact that here various multilayer ANN models have been attempted to find out the best fit.

Hu [7] initiated the implementation of ANN methodology in weather forecasting. Various workers utilized ANN as a forecasting tool involving atmosphere related phenomena [8,9]. Michaelides *et al.* [10] compared the performance of ANN with multiple linear regressions in estimating missing rainfall data over Cyprus. By implementing ANN, reconstruction of the rainfall time series over Cyprus has been carried out [11]. Moreover, the method is used in rainfall prediction by splitting available data into homogeneous subpopulation [12]. Wong *et al.* [13] constructed fuzzy rule bases with the aid of SOM and Backpropagation Neural Networks which are used to develop predictive model for rainfall over Switzerland by spatial interpolation.

2 Methodology

In the recent years, the ANN has been applied to model large data with large dimensionality [14]. This paper introduces ANN model step-by-step to predict the average rainfall over North and South Assam during summer- monsoon between the periods 1871 to 2000 by exploring the data available at the website <http://www.tropmet.res.in> of Indian Institute of Tropical Meteorology.

The ANN approach has several advantages over conventional phenomenological or semi-empirical models. It requires known input data set without any assumptions [1, 15]. It exhibits rapid information processing and is able to develop a mapping of the input and output variables. Such a mapping can subsequently be used to predict desired outputs as a function of suitable inputs. A multilayer neural network can approximate any smooth, measurable function between input and output vectors by selecting a suitable set of connecting weights and transfer functions or activation functions [16].

The model building process consists of four sequential steps:

- Selection of the input and output for the supervised Backpropagation learning
- Selection of the activation function
- Training and testing of the model
- Testing the goodness of fit of the model

The advent of Backpropagation algorithm (BP) and the adaptation of steepest descent method opened up application of Multilayered ANN for many problems of practical interest [17-21]. A multilayered ANN contains three basic types of

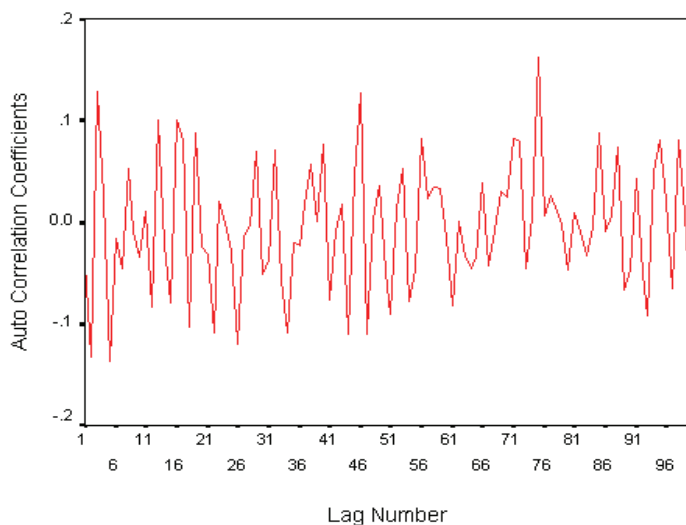


Figure 1. North Assam rainfall autocorrelation function.

layer: input layer, hidden layer(s), and output layer. Basically, the Backpropagation learning involves propagation of error backwards from the output layers to the hidden layers in order to determine the update for the weights leading to the units in the hidden layer(s). The non-linear relationship between input and output parameters in any network requires a function, which can appropriately connect and/or relate the corresponding parameters.

The BP algorithm can be mathematically written as

$$\Delta w_{ij}(t+1) = \eta(t+1)\delta_i(t+1)o_j(t+1) + \rho\Delta w_{ij}(t). \quad (1)$$

Equation (1) is used to compute the entity of weight change at step $(t+1)$. Here $\eta(t)$ is the learning rate at time t , $\delta_i(t)$ is the usual delta factor for unit i as obtained from direct application of the delta rule at time t , and $o_j(t)$ is the output from the unit j at the same time.

The present study explores the data of June, July, August and September corresponding to the years 1871-2000. From these 130 years data, the last year data is deleted as it would not lead to any prediction. Prior to developing ANN based predictive model, it is necessary to discern whether there is any necessity of adopting neural net approach instead of conventional forecasting techniques. To do the same, the autocorrelation function has been computed up to several lags and in all the cases it is apparent that the autocorrelation coefficient values are significantly small (Figures 1 and 2). This indicates that there is almost no persistence in the time series under consideration. This lack of persistence implies that no pattern is maintained within the time series over a considerable

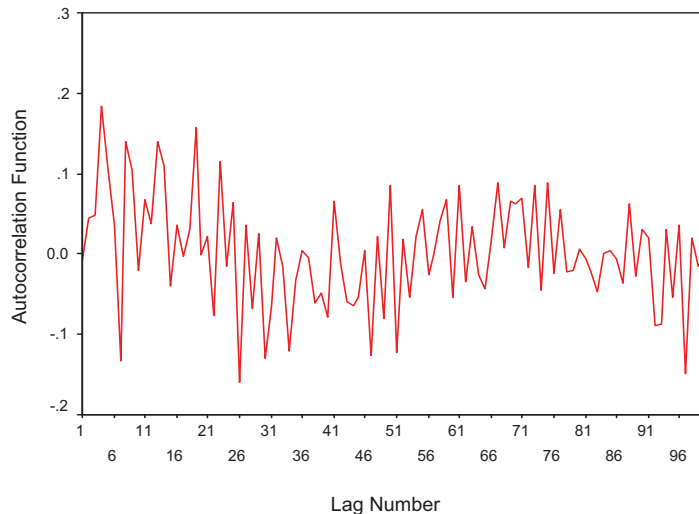


Figure 2. South Assam rainfall autocorrelation function.

period of time and finally it can be concluded that there is significant degree of chaos within this time series. Beauty of ANN in predicting problem lies in the fact that it can forecast even in the face of severe uncertainty and chaos. Thus, necessity for adopting ANN as our predictive methodology is established.

Now, aim of this paper is to develop a multilayer feed forward ANN model so that the average summer-monsoon rainfall of a given year can be predicted using the rainfall data of the summer-monsoon months of the immediately previous year over North and South Assam. Thus, the input matrix would consist of three, four, and five columns of which the first two, three, and four columns would correspond to the summer-monsoon months' rainfall of year ' n ' and the next column would correspond to the average summer-monsoon rainfall of the year $(n+1)$. Basically, the last column would correspond to the 'desired output' in the supervised Backpropagation learning procedure. The first 50% data (*i.e.*, 65 rows out of 129 rows) are taken as the training set and the remaining 50% data (*i.e.*, 64 rows out of 129 rows) are taken as the test set or validation set. To avoid the asymptotic effect, the raw data are scaled according to

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

where, z_i denotes the transformed appearance of the raw data x_i .

After the modelling is completed, the scaled data are reverse scaled according to

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

where, P_i denotes the prediction in original scale, and the corresponding scaled prediction is y_i .

There are two standard learning schemes for the BP algorithm: on-line learning and batch learning. In on-line learning, the weights of the network are updated immediately after the presentation of each pair of input and target patterns. In batch learning, all the pairs of patterns in the training sets are treated as a batch and the network is updated after processing of all training patterns in the batch. In either case, the vector w_k contains the weights computed during k -th iteration, and the output error function E is a multivariate function of the weights in the network

$$E(w_k) = \begin{cases} E_p(w_k) & \text{on-line} \\ \sum_p E_p(w_k) & \text{batch} \end{cases}, \quad (4)$$

where, $E_p(w_k)$ denotes the half-sum of squares error functions of the network output for a certain input pattern p . The purpose of the supervised learning (or training) is to find out a set of weights that can minimize the error E over the complete set of training pair. Every cycle in which each one of the training patterns is presented once to the neural network is called an epoch.

The direction vector d_k , expressed in terms of error gradient depends upon the choice of activation function. When the sigmoid function is adopted, the BP algorithm becomes ‘Back propagation for the Sigmoid Adaline’ [17-21]. In this method, the input matrix is multiplied by the weight matrix and the product is used as the variable for the sigmoid activation function. For example, at epoch k , the sigmoid non-linearity is produced as

$$f(W_k X_k) = \frac{1}{1 + e^{-\left(\sum_i w_i x_i\right)}}, \quad (5)$$

where $W_k = [w_1 \ w_2 \ \dots \ w_n]$ and $X_k = [x_1 \ x_2 \ \dots \ x_n]^T$ are the weight matrix

Table 1. Basic network components of the ANN model

Network architecture			
Number of inputs	2	Number of outputs	1
Number of hidden layers	1	Hidden layer sizes	3
Learning parameter	0.2	Initial wt range (0 +/- w)	0.5
Momentum	0.9		
Training options			
Total number of rows in data	129	Number of training cycles	500
	Training mode	On line	
Save network weights	With least training error		
Training/validation set	Partition data into training/validation set		

and the transpose of the input matrix respectively at epoch k .

After training or learning the ANN with BP algorithm with sigmoid non-linearity, a ultimate weight matrix is obtained. This weight matrix is applied to another set of independent inputs to examine the efficiency of the model. This phase is called the testing or the validation phase.

After developing the model through training and testing, goodness of fit of the model is examined statistically. Over-all prediction error (PE) is measured as

$$PE = \frac{\langle |y_{\text{predicted}} - y_{\text{actual}}| \rangle}{\langle y_{\text{actual}} \rangle}, \quad (6)$$

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

The predictive model is identified as a good one if the PE is sufficiently small, *i.e.*, close to 0. The model with minimum PE is identified as the best prediction model.

3 Results and Discussion

Details of the input and output variables are presented in the previous Section. The learning rate η is taken to be 0.9. A three-layered feed forward neural net is now designed. The network components are presented in Table 1. The problem is to find out the number of hidden nodes producing the best model. Since the number of adjustable parameters in a one hidden-layer feed forward neural network with n_i input units, n_o output units, and n_h hidden units is $[n_o + n_h(n_i + n_o + 1)]$ for $n_i = 3$, three models are generated for both North

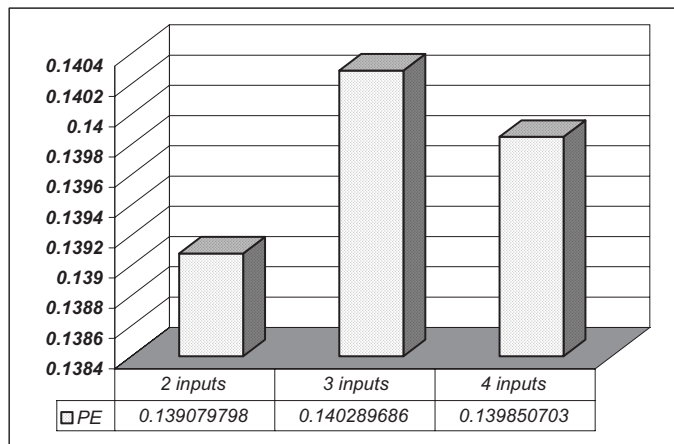


Figure 3. Prediction errors corresponding to 2-inputs, 3-inputs, and 4-inputs based neural net predictive models for the North Assam rainfall data.

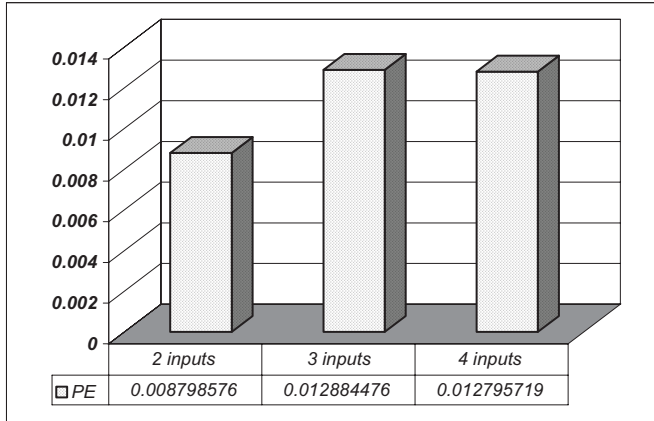


Figure 4. Prediction errors corresponding to 2-inputs, 3-inputs, and 4-inputs based neural net predictive models for the South Assam rainfall data.

and South Assam. In both models, the initial weights are chosen randomly from -0.5 to +0.5. After each training iterations/epochs the network is tested for its performance on validation data set. The training process is stopped when the performance would reach the maximum on validation data set.

After training and testing, the *PE* [Equation (6)] values are computed for each model. The results are schematically presented for North Assam (Figure 3) and South Assam (Figure 4).

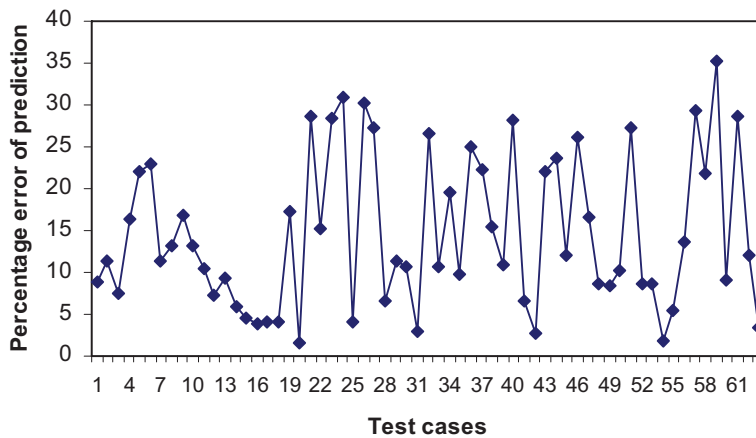


Figure 5. Percentage errors of prediction in the test cases with 2-inputs ANN model applied to North Assam rainfall data.

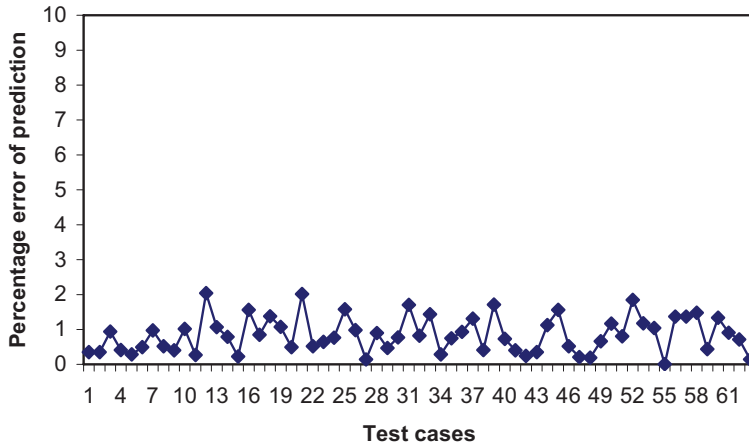


Figure 6. Percentage errors of prediction in the test cases with 2-inputs ANN model applied to South Assam rainfall data.

The result shows that the two inputs model produces the lowest prediction error among the three possible predictive models. Figures 3 and 4 depict the performance of two inputs model of North and South Assam. It is found that, the prediction error of two inputs model is minimum in both cases.

Now, the percentage errors of prediction from the two inputs ANN models are computed for both North Assam and South Assam rainfall data to have a thorough look into the performance of the two input ANN models (Figures 5 and 6). It is found that in the case of North Assam, in 37 out of 64 test cases the percentage errors of prediction lie below 15%.

Thus, it can be said that if we make 15% error in yearly monsoon rainfall forecast, then in 58% cases a forecast with maximum 15% error is possible. Thus, forecast yield is 0.58 if 15% error is allowed. It is quite interesting to see that, in the case of South Assam, the actual and predicted monsoon rainfall amounts almost coincide. In this case, in all 64 test cases, the percentage errors of prediction are much lower than 15%. Thus, if 15% error is allowed, the forecast yield is 1. It can, therefore, be concluded that in both North and South Assam, ANN produces significant forecast yield. But, in the case of South Assam, the performance of two-input ANN is overwhelmingly adroit. To have a complete look into the prediction capability of two-input models, the actual North and South Assam rainfall amounts are plotted against predicted rainfall amounts in Figures 7 and 8. It is apparent from the said figures that in both cases ANN can generate a good forecast for monsoon rainfall. But, in South Assam, the actual rainfall amount almost coincides with the predicted amount.

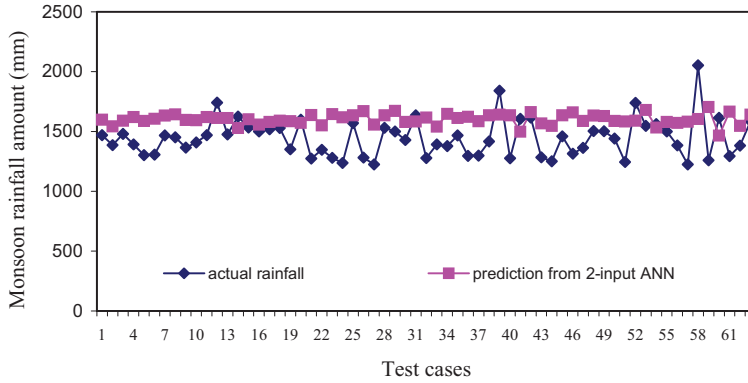


Figure 7. Schematic showing the actual and predicted monsoon rainfall over North Assam.

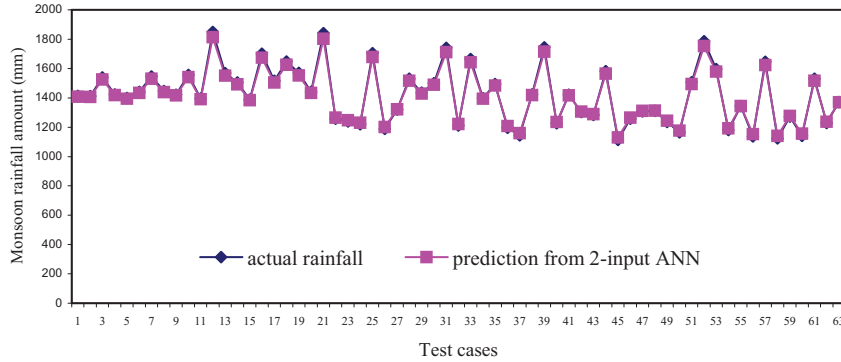


Figure 8. Schematic showing the actual and predicted monsoon rainfall over South Assam.

4 Comparison of ANN with Regression

From the painstaking study explained in the earlier Sections, it is established that 2-input ANN can be considered as a good predictive methodology for monsoon rainfall time series derived from North and South Assam. In this Section, the prediction from the said neural net model would be placed against multiple linear regression to have a comparative study. A multiple linear regression is of the form

$$\hat{y} = a_0 + \sum_{i=1}^n a_i x_i. \quad (7)$$

In the two-input models, the two inputs are taken as predictor, and the third value is the predictand. The absolute prediction errors from the multiple linear regression models are plotted against those from ANN (Figures 9 and 10). The Figure

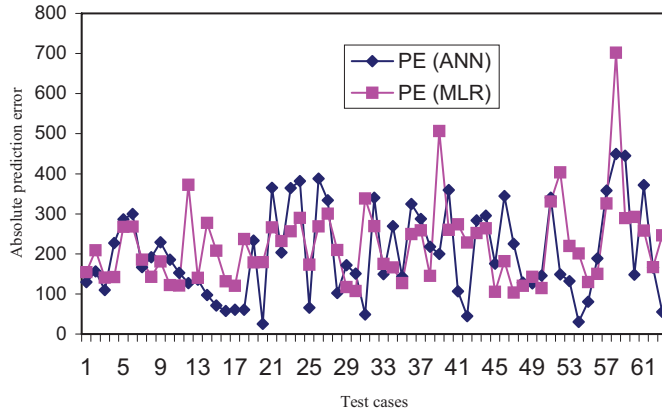


Figure 9. Schematic showing the absolute prediction errors from ANN and MLR in the case of North Assam.

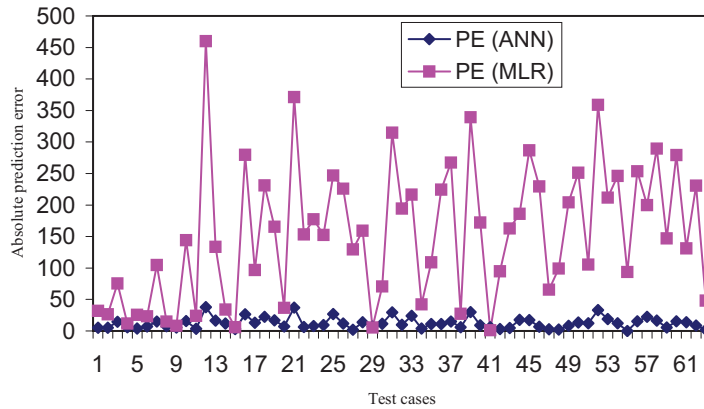


Figure 10. Schematic showing the absolute prediction errors from ANN and MLR in the case of South Assam.

9 shows that in the case of North Assam, the ANN is not significantly better than multiple linear regressions. In this case, sometimes ANN and sometimes regression is producing lower prediction error. But in the case of South Assam (Figure 10), the multiple linear regression is absolutely outperformed by ANN. The PE (Eq. 6) is 0.14 for North Assam and 0.26 for South Assam. Thus, it is found that for both of the regions the ANN produces better prediction than multiple linear regression. But, in the case of South Assam, the PE is much higher for regression than ANN.

5 Conclusion

After comparing the performance of three ANN models with sigmoid non-linearity, the ANN model with two inputs is found to be able for prediction of mean summer-monsoon rainfall over North and South Assam on the basis of previous years' rainfall data of the summer-monsoon months. After comparing two-inputs ANN with multiple linear regression, it is observed that in the case of North Assam ANN produces better forecast to some extent than multiple linear regression. But for South Assam, the two-inputs ANN is found to be the best predictive model for monsoon rainfall.

Acknowledgements

This work is funded by Indian Space Research Organization (ISRO) through S. K. Mitra Centre for Research in Space Environment, University of Calcutta, Kolkata 700 009, India. The authors are thankful to Indian Institute of Tropical Meteorology for the data of the work made available at their website <http://www.tropmet.res.in>

References

- [1] K. Parthasarathy (1960) *Indian Meteor. Dept.* 185.
- [2] WMO (1965) Guide to Hydro Meteorological Practices, WMO.
- [3] P.K. Raman, O.N. Dhar (1966) *Indian J. Met. Geophys.* 87.
- [4] S.M.S. Nagendra, M. Khare (2006) *Ecol. Modell.* **190** 99.
- [5] C.M. Kishtawal, S. Basu, F. Patadia, P.K. Thapliyal (2003) *Geophys. Res. Lett.* **30** doi: 10.1029/2003GL018504.
- [6] P. Guhathakurta (2006) *Current Science* **90** 773.
- [7] M.J.C. Hu (1964) *Application of ADALINE System to Weather Forecasting*, Technical Report, Stanford Electron.
- [8] M.W. Gardner, S.R. Dorling (1998) *Atmospheric Environment* **32** 2627.
- [9] W.W. Hsieh, T. Tang (1998) *Bull. Am. Meteorol. Soc.* **79** 1855.
- [10] S.C. Michaelides, C.C. Neocleous, C.N. Schizas (1995) *Proceedings of the DSP95 International Conference on Digital Signal Processing, Limassol, Cyprus.* 668.
- [11] S.A. Kalogirou, C.N. Constantinou, S.C. Michaelides, C.N. Schizas (1997) *EUFIT '97*, September 8-11, 2409.
- [12] S. Lee, S. Cho, P.M. Wong (1998) *J. Geogr. Inform. Decis. Anal.* **2** 233.
- [13] K.W. Wong, P.M. Wong, T.D. Gedeon, C.C. Fung (1999) www.it.murdoch.edu.au/~wong/publications/SIC97.pdf 213.
- [14] M. Gevrey, I. Dimopoulos, S. Lek (2003) *Ecol. Model.* **160** 249.
- [15] S.V. Kartalopoulos (1996) *Understanding Neural Networks and Fuzzy Logic – Basic Concepts and Applications*, Prentice Hall, New-Delhi.
- [16] P. Perez, J. Reyes (2001) *Neural Comput. Appl.* **10** 165.

- [17] S.V. Kamarthi, S. Pittner (1999) *Neural Networks* **12** 1285.
- [18] T.J. Sejnowski, C.R. Rosenberg (1987) *Complex Systems* **1** 145.
- [19] A.K. Sahai, M.K. Soman, V. Satyan (2000) *Climate Dynamics* **16** 291.
- [20] A.K. Sahai, D.R. Patanik, V. Satyan, A.M. Grimm (2003) *Meteorol. Atmos. Phys.* **84** 217.
- [21] B. Widrow, M.A. Lehr (1990) *Proc. IEEE* **78** 1415.