J. Ind. Geophys. Union (April 2006)
Vol.10, No.2, pp.141-151

Modelling and prediction of rainfall using artificial neural network and ARIMA techniques

V.K.Somvanshi, O.P.Pandey, P.K.Agrawal, N.V.Kalanker[1], M.Ravi Prakash and Ramesh Chand
National Geophysical Research Institute, Hyderabad -500 007
[2]Swami Ramanand Teerth Marathwada University Nanded – 431 602

ABSTRACT
Climate and rainfall are highly non-linear and complicated phenomena, which require sophisticated computer modelling and simulation for accurate prediction. An artificial intelligence technology allows knowledge processing and can be used .as forecasting tool. For example, the application of Artificial Neural Networks (ANN), to predict the behaviors of nonlinear systems has become an attractive alternative to traditional statistical methods. In this paper, we present tools for modeling and predicting the behavioral pattern in rainfall phenomena based on past observations. The paper introduces two fundamentally different approaches for designing a model, the statistical method based on autoregressive integrated moving average (ARIMA) and the emerging computationally powerful techniques based on ANN. In order to evaluate the prediction efficiency, we made use of 104 years of mean annual rainfall data from year 1901 to 2003 of Hyderabad region (India). The models were trained with 93 years of mean annual rainfall data. The ANN and the ARIMA approaches are applied to the data to derive the weights and the regression coefficients respectively. The performance of the model was evaluated by using remaining 10 years of data. The study reveals that ANN model can be used as an appropriate forecasting tool to predict the rainfall, which out performs the ARIMA model.

INTRODUCTION

Rainfall is natural climatic phenomena whose prediction is challenging and demanding. Its forecast is of particular relevance to agriculture sector, which contributes significantly to the economy of the nation. On a worldwide scale, numerous attempts have been made to predict its behavioral pattern using various techniques. In the present work, we make a comparative study of rainfall behavior as obtained by autoregressive integrated moving average (ARIMA) and the artificial neural network (ANN) techniques. The former is basically a linear statistical technique and has been quite popular for modeling the time series and rainfall forecasting due to ease in its development and implemention. In contrast, the application of the ANN in time series for forecasting is relatively (Mirko & Christian 2000). It is primarily based on the ability of neural networks to approximate nonlinear functions. This technique corresponds to human neurological system, which consists of a series of basic computing elements, called as neurons interconnected together to form a network, [Rummelhart & McClelland 1996]. The parallel-distributed processing architecture of ANN has proved to be a very powerful computational tool which is now being used in several fields to model the dynamic processes successfully [Mirko & Christian 2000; Mary 2002] including the rainfall [Singh & Chowdhury 1986; Cigizoglu 2002]. This technique has the ability to learn and generalise from examples to produce meaningful solutions. The present work convincingly demonstrates the advantages of using ANN over that of ARIMA technique to model the rainfall behavior.

DATA

The data for mean annual rainfall over Hyderabad region, of Andhra Pradesh, India, which is bounded by latitude 17°-18° N, and longitude 78°-79° E is being used for the present study. The database was provided by the India Meteorological Department, Hyderabad airport. It consists of the mean annual rainfall from year 1901 to 2003 (104 years) of Hyderabad region as shown in Fig.1. The series is regarded as Nonlinear and Non-Gaussian and is used to evaluate the effectiveness of the nonlinear model. We use first 93 years of mean annual rainfall time series data for model training while the prediction is carried out using the rest of the 10 years data using both ARIMA and ANN models, the details of which are given below.

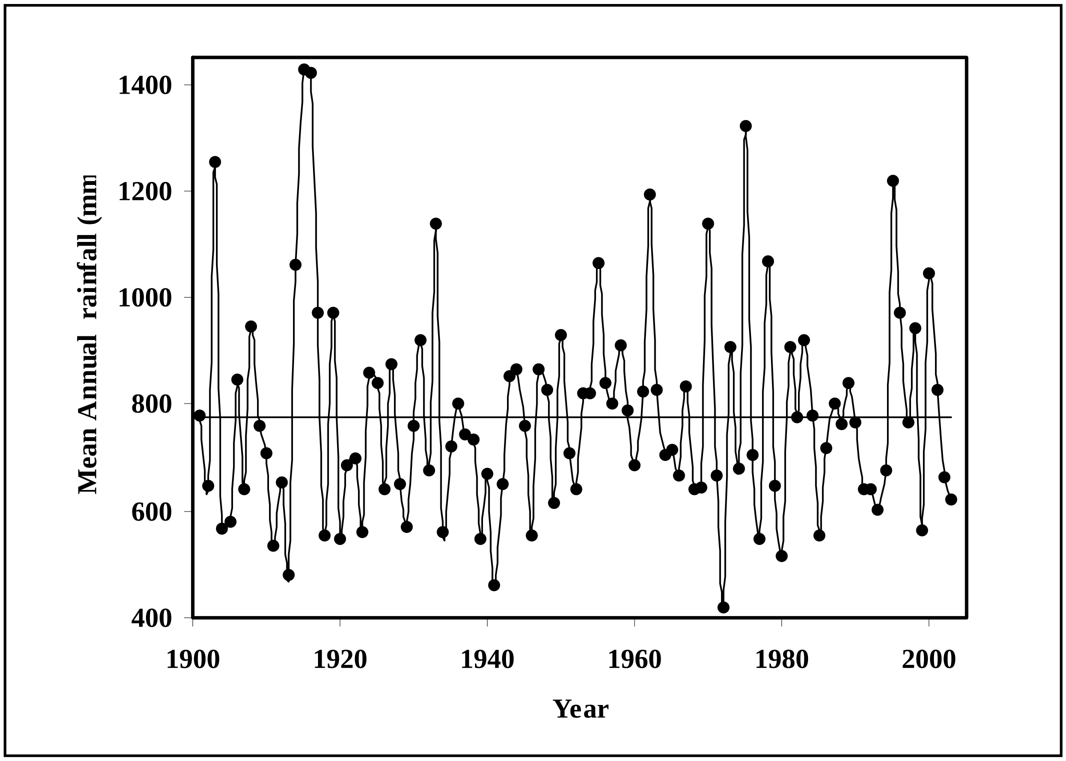


Figure 1. Pictorial representation of mean annual rainfall data bounded by latitude 17°-18° N and longitude 78°-79° E. The horizontal line represent the mean rainfall.

ARIMA Model

Box & Jenkins (1970) developed this forecasting technique which is still very popular among hydrologists. The autoregressive integrated moving average ARIMA(p,d,q) model of the time series $\{r_{1,} r_2 .....\}$ is defined as

$\phi(B)\Delta^d r_t = \theta(B) e_t$ (1)

where r_t and e_t respectively represent mean annual rainfall time series and random error terms at time t. B is the backward shift operator defined by $Br_y = r_{y-1}$ and related to Δ by Δ = 1- B; d is the order of difference. The $\phi(B)$ and $\theta(B)$ of order p and q are defined as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots\cdots\cdots \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots\cdots\cdots \theta_q B^q$$

where $\phi_1, \phi_2, ...\phi_p$ are the autoregressive coefficients and $\theta_1, \theta_2, ...\theta_q$ are the moving averages coefficients

In this ARIMA(p,d,q) modelling, the first step is to determine whether the time series is stationary or non stationary. If it is non stationary it is transformed into a stationary time series by applying suitable degree of differencing by selecting proper value of d. The appropriate values of p and q are chosen by examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series.

ANN Model

An ANN is a massively parallel-distributed processor that has a natural propensity for storing the experimental knowledge and making it available for further use. It resembles the human brain whose speed and efficiency has been always fascinating to researchers for quite a long time. The quest to understand these processes and to solve the associated problems has led to the development of ANN technique. Neural networks essentially involve a nonlinear modelling approach that provides a fairly accurate universal approximation to any function. Its power comes from the parallel processing of the information from data. No prior assumption of the model form is required in the model building process. Instead, the network model is largely determined by the characteristics of the data. Single hidden layer feed-forward network is the most widely used model form for time series modeling and forecasting. The back-propagation network (BPN) is one of the neural network algorithm which is formalized by Parker, (1986), Lippmann (1987) and Rummelhart & McClelland (1986) etc. It has been extensively used for inversion, prediction that consist of two passes: a forward pass and a backward pass. In the forward pass the input is applied to input layer and its effect is propagated through network, layer by layer. The net effect is computed as the weighted sum of the output of the neurons of the previous layer. The sum of squared deviation of the output from the target value at the nodes of the output layer defines the error signal that is to be propagated back to previous layers such that the parameters are adjusted to minimize the error in further computations.

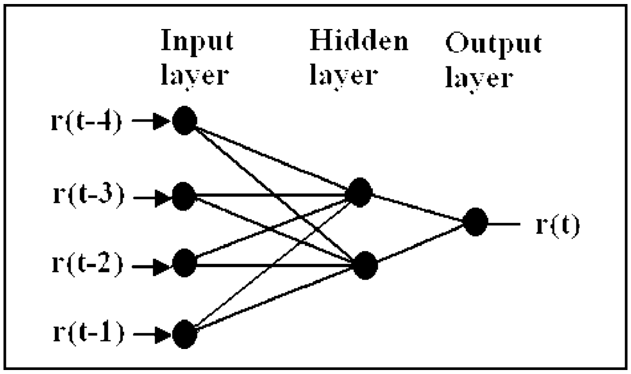


Figure 2. ANN architecture with four input and one output for mean annual rainfall data.

As shown in Fig.2, an ANN consists of layers of neurons. The model is characterized by a network of three layers of simple processing units, which are connected to each other. The first layer, which receives input information, is called an input layer. The last layer, which produces output information, is called an output layer. Between output and input layers are hidden layers. There can be one or more hidden layers. Information is transmitted through the connections between nodes in different layers.

The relationship between the output (y_t) and the inputs ($r_{t''1}$; $r_{t''2}$ . . . $r_{t''i}$) can be represented by the following mathematical equation:

$$y(t) = s_1\left(\sum_{j=1}^{J} w_j s_2\left(\sum_{i=1}^{I} w_i r_{(t-i)}\right)\right) \qquad (2)$$

where y(t) is an output from the network, $r_{(t-i)}$ is the inputs to network. W_j and W_i are the connection weights. S_1 and S_2 are activation function, the most commonly used function is a logistic sigmoid function given by equation:

$$s(y) = \frac{1}{1 + e^{-x}}$$

The main control parameters of any ANN are the weights. The processes of estimating these parameters are known as training where optimal connection weights are determined by minimizing an objective function.

Data analysis and model selection

One of the most common problems that a modern data analyst encounters is the extraction of meaningful conclusions about a complicated system using data from a single measured parameter. The most popular treatment of mean annual rainfall data is to feed the neural networks with either the data at each observation, or the data from several successive observations. The treatment can be described as

$$r_{k+1} = NNF(r_k, r_{k-1}, \cdots\cdots r_{k-l})$$

where NNF() stands for the neural network forecaster and *l* is the number of successive observations. Fig 3a and 3b show plot of autocorrelation and partial autocorrelation coefficient for various lags (in year) of mean annual rainfall data with 95% confidence level. These figures exhibit significant correlations at lags 4 and 10. Thus, the above analysis shows that the neural network forecaster should use four or ten past observations as inputs to neural network (Sudheer, Gosain & Ramasastri). The present analysis uses four past observations as inputs to neural network model.

The use of one hidden layer is generally recommended at least in preliminary studies. As the use of more than one hidden layer substantially increases the number of parameters to be estimated. Such an increase in the number of the parameters may slow down the training process without substantially improving the efficiency of the network. A single hidden layer was adopted in the present study. The determination of the appropriate number of neurons in the hidden layer is important for the successful application, since it greatly enhances the performance of the neural network. If the hidden layer has too few neurons then the performance of the neural network may deteriorate. On the other hand, if the hidden layer has too many neurons, then there are too many parameters and there is a danger of over-fitting the training data set. The best strategy for selecting the appropriate number of neurons in the hidden layer is to experiment, i.e. a trial and error procedure

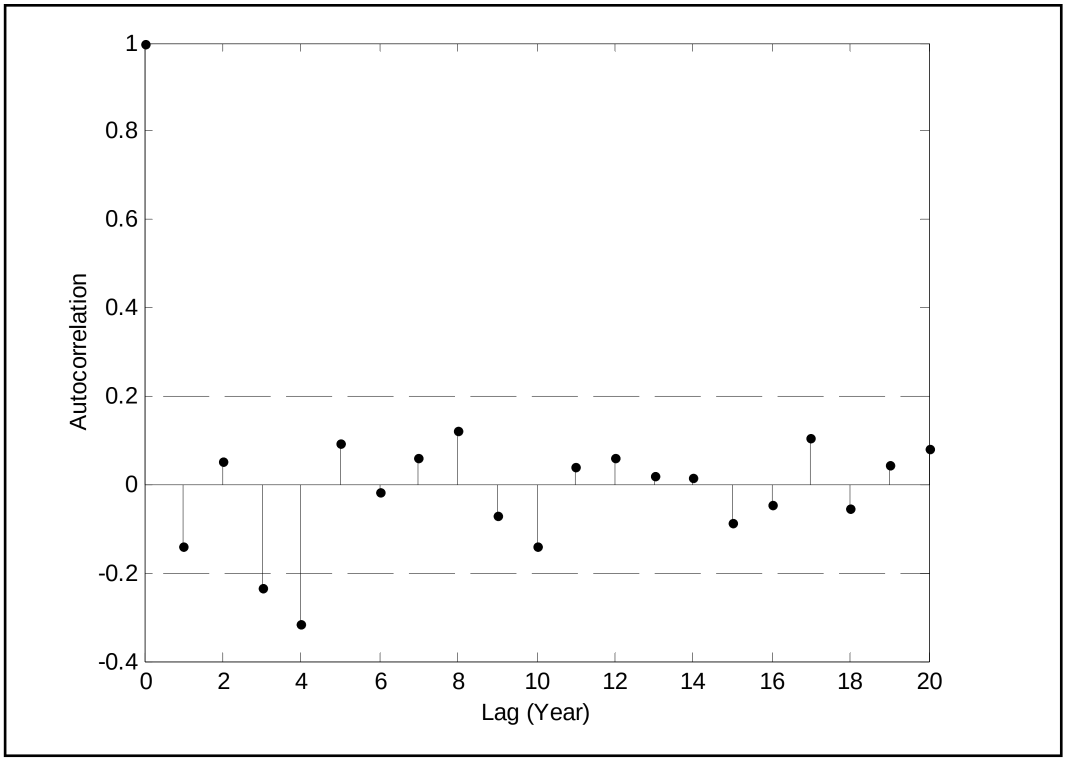


Figure 3a. Plot of autocorrelation coefficient and time lags of mean annual rainfall data.

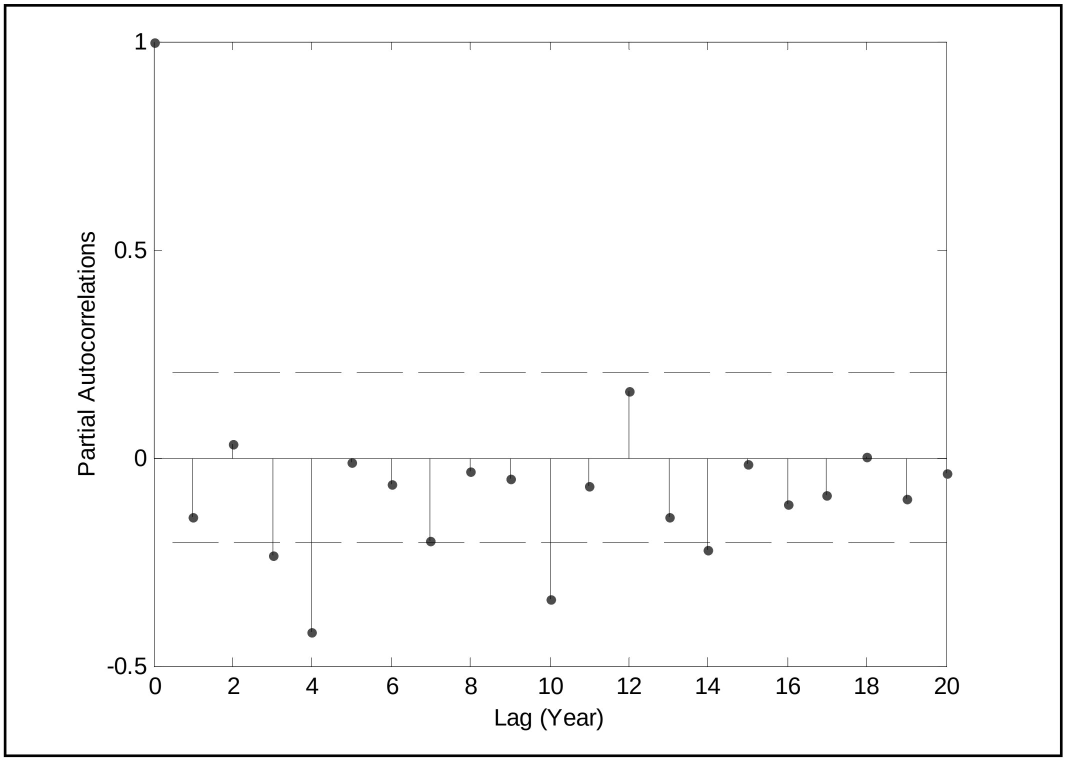


Figure 3b. Plot of partial autocorrelation coefficient and time lags of mean annual rainfall data

(Hammerstrom 1993) is adopted. Training the network (with a small data set) and changing the hidden neurons from 2 to 10 decide the number of hidden neurons in the hidden layer. The root mean square error (RMSE) is considered as criteria for selecting the number of hidden neurons. With 2 hidden neurons RMS error was found to be minimum. The ANN architecture selected for the present analysis is (4,2,1).

To assess the forecasting performance of models, data set is divided into two samples of modeling and testing. The modeling data set is used exclusively for model development and the test sample is used to evaluate the established model. For the present analysis values of learning rate, target error and momentum are 0.8, 0.7 and 0.001 respectively. The network converged within 3000 iterations on Pentium-4 PC with a speed of 2.6 GHz

The ACF and PACF of mean annual rainfall time series are displayed in Fig. 3(a) and 3(b), were used for estimating the parameters of ARIMA model. Both the ACF and PACF have two significant terms at lag 4 and 10, the second term at lag 10, indicates that if moving average or autoregressive models are used, they should be of order 10. Following the principle of parsimony, we choose autoregressive model of order 4 for fitting the data. Thus, the ARIMA(4 0 0) model is used for present study.

Performance evaluation criteria

The World Meteorological Organization (WMO) and other investigators (WMO 1975; Aitken 1973; Kachroo 1992) have proposed the evaluation and intercomparison of different models, which can be evaluated in terms of graphical representation and numerical computations. The graphical performance criteria involves:

- A linear scale plot of the predicted and observed for both the calibration and the verification periods.
- A scatter plot of the predicted versus observed rainfall for the both calibration and the verification periods.

In comparison, numerical performance criteria relate to root mean square error and Mean Absolute Error (MAE), given by

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}\left(t(k) - a(k)\right)^2}{N}} \qquad (3)$$

$$MAE = \frac{1}{N}\sum_{k=1}^{N}\left|t(k) - a(k)\right| \qquad (4)$$

Where t(k) is the target mean annual rainfall and a(k) the observed mean annual rainfall.

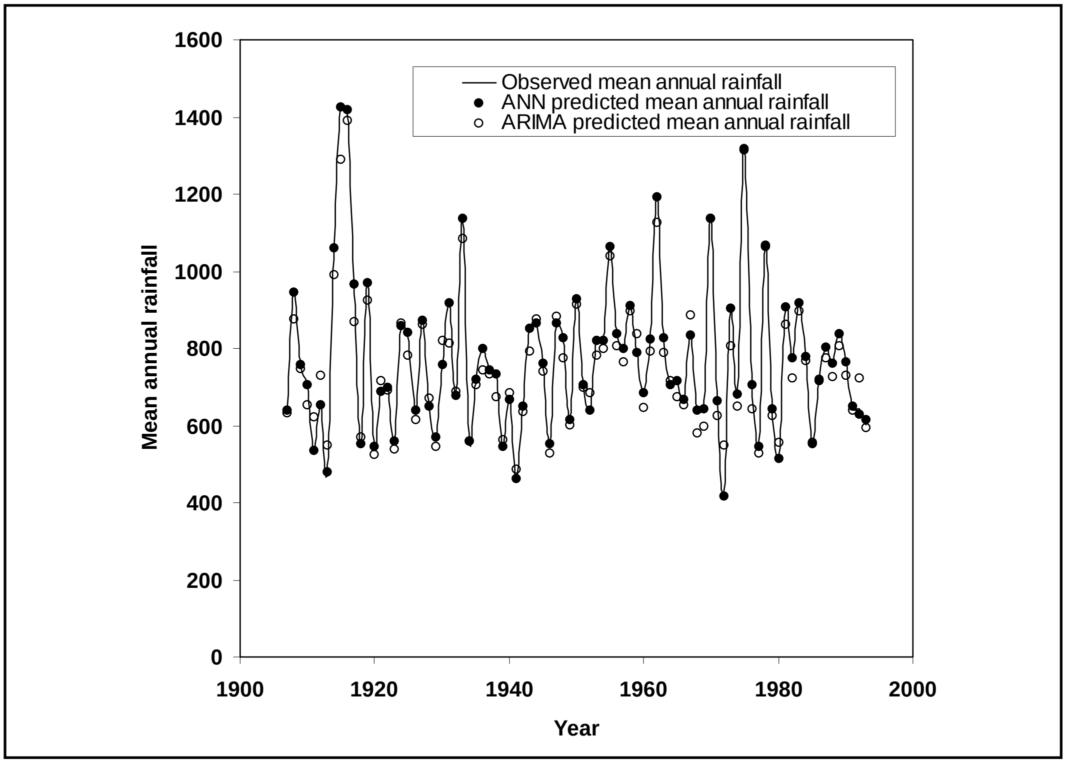


Figure 4. A linear scale plot of the predicted and observed mean annual rainfall for model data set using ANN and ARIMA model.

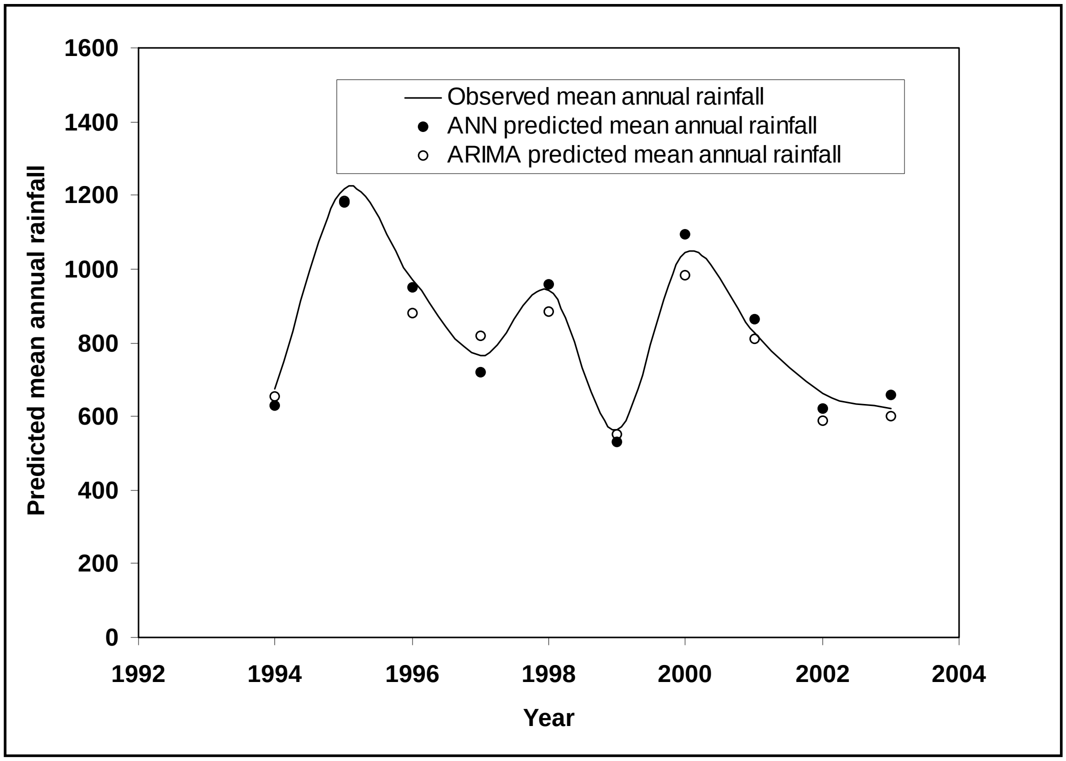


Figure 5. A linear scale plot of the predicted and observed mean annual rainfall for test data (1994 -2003) set using ANN and ARIMA model.

Another A Information Criterion (AIC of order two) (Akaike, 1974) and the B information criteria (BIC) (Rissanen 1978) are also computed for both ANN and ARIMA models by using the equations:

$$AIC = m \ln(RMSE) + 2\ npar + [2*npar\ (npar+1]/[m-npar-1]$$

$$BIC = m \ln(RMSE) + npar*\ln(m)$$

Where m is the number of input-output patterns and npar is the number of parameters to be identified. Notice, that while the RMSE statistics are expected to progressively improve as more parameters are added to the model, the AIC and BIC statistic penalize the model for having more parameters and therefore tend to result in more parsimonious models. In the present case the ratio of (m/npar) is less than 40, hence second order AIC is evaluated for measuring the model performance.

RESULTS AND DISCUSSION

Figure 4 shows the plot of predicted models vs. observed values of the mean annual rainfall data from year 1904 to 1994 by both the ANN and ARIMA. The ANN model fits extremely well with the actual data values as compared to the ARIMA model. Both the models were tested using the test data set for the period 1994 to 2003, which is shown in Fig.5. From this figure it can be observed that the mean annual rainfall values predicted by the ANN model are quite closer to observed mean annual rainfall as compared to the ARIMA model.

The scatter plot of observed and predicted mean annual rainfall for calibration data set and test data set using both ANN and the ARIMA techniques are shown in Figs 6 and 7 and Figs 8 and 9 respectively. These figures reveal that scattering along the regression line for ARIMA model (Fig. 6) is larger compared to ANN model (Fig.8). A similar inference can be drawn from figs (7&9) for test data sets also. The coefficient of determination (R^2) for model and test data set for ARIMA is 0.9535 and 0.9404, while it is considerably higher at 0.9841 and 0.9695 for the ANN. For the considered mean annual rainfall data the fitted ARIMA(4 0 0) model can be expressed by following empirical relation:

$r(t) = 16.311+0.1743*r(t-1)-0.0985*r(t-2)+0.0159*r(t-3)+0.8872*r(t-4)$

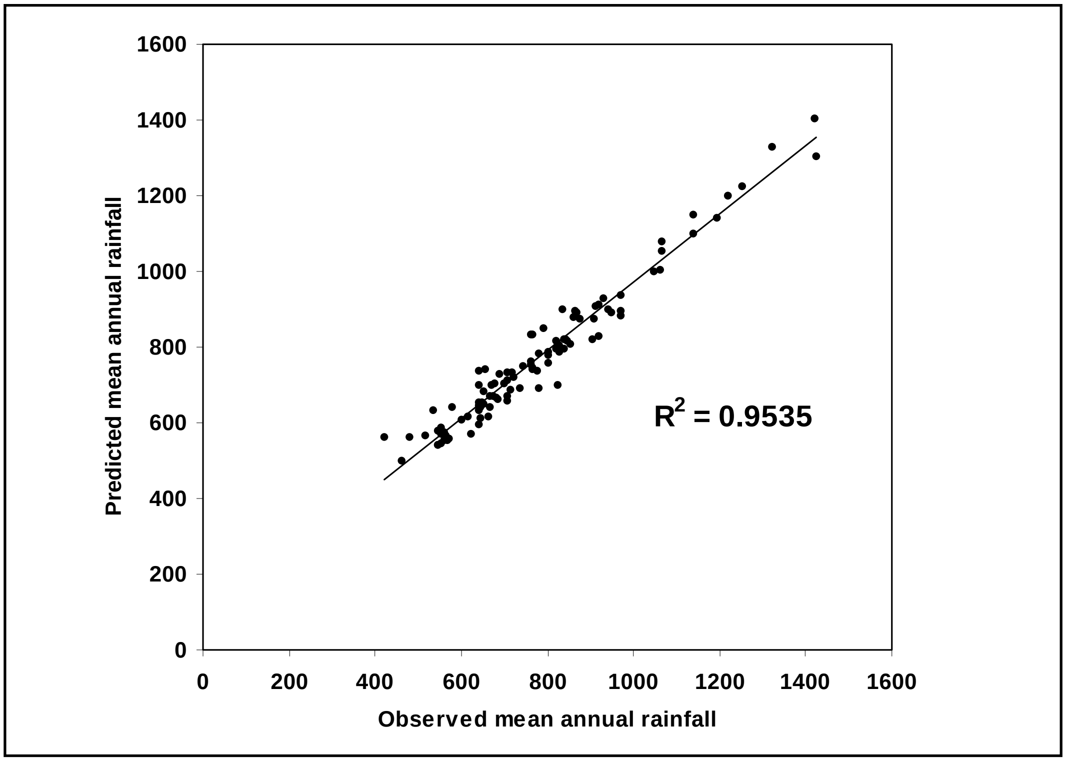


Figure 6. Scatter plot of observed and predicted mean annual rainfall for model data set (1904 – 1994) using ARIMA model.

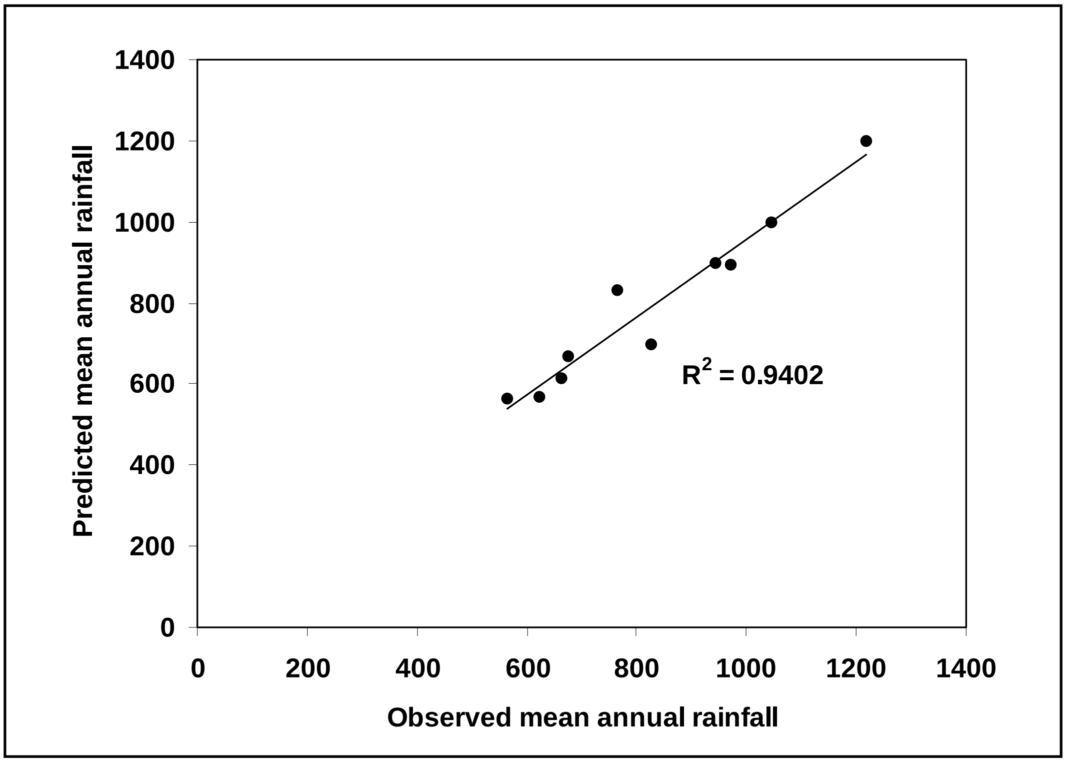


Figure 7. Scatter plot of observed and predicted mean annual rainfall for test data set (1994 -2003) using ARIMA model.

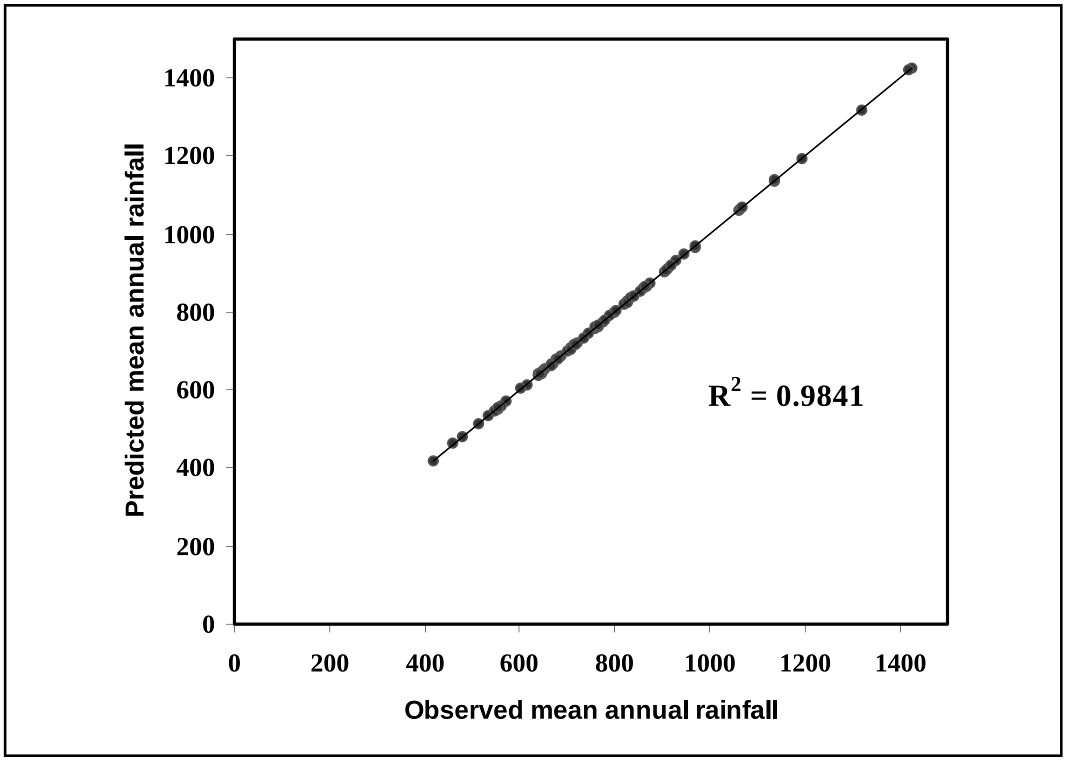


Figure 8. Scatter plot of observed and predicted mean annual rainfall for model data set (1904 – 1994) using ANN model.

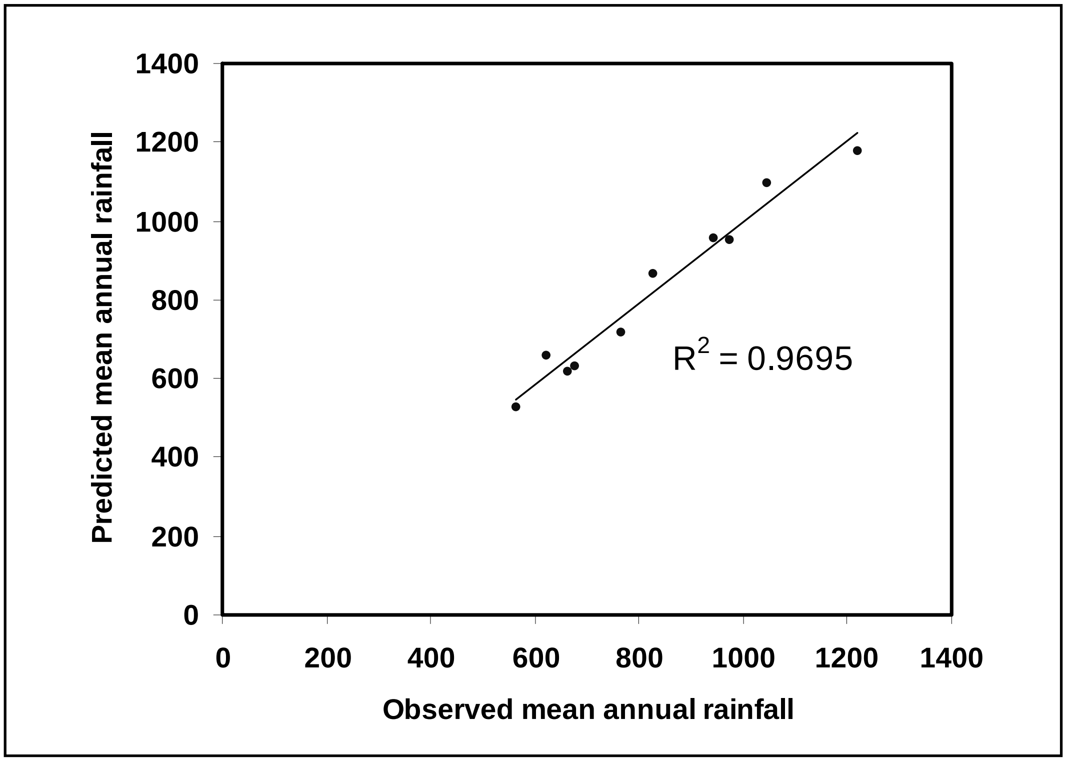


Figure 9. Scatter plot of observed and predicted mean annual rainfall for test data set(1994 -2003) using ANN model.

Table 1.

Techniques	Error measures for mean annual rainfall model data set					Error measures for mean annual rainfall test data set		
	RMSE	MAE	R^2	AIC	BIC	RMSE	MAE	R^2
ANN(4,2,1)	5.5186	0.9198	0.9841	186.5	209.6	145.14	36.773	0.9695
ARIMA(4 0 0)	35.882	24.388	0.9535	352.1	361.9	262.57	44.132	0.9402

which can be used to predict the mean annual rainfall by providing past four values.

The performance measures of ANN and ARIMA models in terms of numerical computations are shown in Table 1. The table indicates that the ANN model outperforms the ARIMA model. The MAE error for model data set and test data set for ARIMA model is 24.388 and 44.132 respectively. While the same error measure is considerably lower at 0.9198 and 36.773 for the ANN model. The other performance measures such as RMSE and R^2 depict that the ANN forecast is superior to ARIMA forecast. The number of parameters for ANN and ARIMA models are 11 and 4. The RMSE error progressively improve as more parameters are added to the model, the AIC and BIC statistic penalize the model for having more parameters and therefore tend to result in more parsimonious models. The AIC and BIC for ANN model are 186.5 and 209.6 which are lower than 352.1 and 361.9 of ARIMA model. On the basis of AIC, BIC and RMSE the ANN model is more appropriate than the ARIMA model. Therefore, our study establishes that ANN method should be favored as an appropriate forecasting tool to model and predict annual rainfall than the ARIMA model.

CONCLUSIONS

Complexity of the nature of annual rainfall record has been studied using the ANN and ARIMA techniques. An annual rainfall data spanning over a period of 1901-2003 of Hyderabad region was used to develop and test the models. Autocorrelation and partial autocorrelation coefficient for various lags (in year) of rainfall data was used to find out number of past observations as inputs to neural network. The present analysis uses four past observations as inputs to neural network model. The study reveals that ANN model can be used as an appropriate forecasting tool to predict the rainfall, which out performs the ARIMA model. Further refinement of the model using the data separately from the different zones of the country may be useful for the long-term prediction.

ACKNOWLEDGEMENTS

The authors are thankful to the Director, National Geophysical Research Institute, Hyderabad for giving the permission to publish this study. We also thank to Indian Meteorological Department, Hyderabad Airport, Hyderabad for providing the monthly total rainfall data set used in the study.

REFERENCES

Aitken, A.P. , 1973. Assessing systematic errors in rainfall-runoff models, J. of Hydrology. 20,121-136.

Akaike, H., 1974. A new look at the statistical model identification , IEEE Trans. Autom. Control, AC-19,716-723.

Box, G.E.P. & Jenkins, G.M., 1970. Time Series Analysis, Forecasting and Control, Holden-Day, CA, San Francisco.

Cigizoglu, H.K., 2002a. Forecasting of meteorologic data by Artificial Neural Networks, In: *Advances in Soft Computing* (Proc. Sixth Int. Conf. On Neural Networks and Soft Computing, Zakopane, Poland, 11-15 June 2002).

Hammerstrom D., 1993 Working with neural networks, IEEE Spectrum 46-53.

Kachroo, R.K., 1992. River flow forecasting part1. A discussion of the principles, J. of Hydrology, 133, 1-15.

Lippmann R.P., 1997. An Introduction to computing with neural nets. IEEE ASSP Magazine April 1987, pp 4-20.

Mirko van der Baan and Christian Jutten, 2000. Neural network in geophysical Applications, Geophysics, 65 (4),1032-1047.

Mary M. Poulton, 2002. Neural network as an intelligence amplification tool: A review of applications. Geophysics, 67 (3), 979-993.

Parker D.B., 1986. A comparison of algorithm for neuron-like cell in J. S. Denker (e.d.) AIP Conference Proceedings. 151 Neural networks for Computing, Snowbird Utah, AIP

Rissanen, J., 1978. Modeling by short data description , Automation , 14, 467-471.

Rummelhart , D. E. & McClelland J. L., 1996. Parallel Distributed Processing , MIT Press, Cambridge, Massachusetts, pp 318-362.

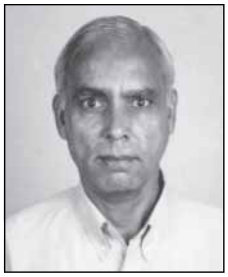
Singh, V.P. and Chowdhury, P.K., 1986. Comparing some methods of estimating mean area rainfall. Water Resources Bulletin 22(2),275-282.

Sudheer K. P., Gosain A.K. and Ramasastri K.S., 2002. A data-driven algorithm for constructing artificial neural network rainfall-runoff models. Hydrological Processes 16 (6), 1325-1330.

World Meteorological Organization, 1975. Inter-comparison of conceptual models used in operational hydrological forecasting, World Meteorological Organization, Technical report No. 429, Geneva, Switzerland.

(Accepted 21st March 2006. Received 18th March 2006; in original form 14th February 2006)

Somvanshi V.K. presently working as a scientist at National Geophysical Research Institute, Hyderabad. He obtained B.Sc. Physics (hons) and M.Sc. Physics (electronics special) from Marathawada University Aurangabad and M.Tech. (Computer Science) from Jawaharlal Neharu Technological University, Hyderabad. He has registered for Ph.D at Swami Ramanand teerth Marathawada University Nanded. His research interest include modeling and prediction using artificial neural network, fuzzy logic and genetic algorithm techniques.



Dr. O. P. Pandey is working as a senior scientist at National Geophysical Research Institute, Hyderabad He joined this institute in 1971. He obtained B.Sc. from Lucknow university, B.Sc. (hons) and M.Sc. in applied geophysics from Indian School of Mines (Dhanbad) and Ph. D. in Geophysics from Victoria University of Wellington (New Zealand). His current interest is Heat flow and Geodynamics

Dr. P.K. Agrawal obtained his master degree in geophysics in 1966 from Banaras Hindu University, Varanasi. He retired as head of theoretical geophysics group with a position of scientist "G" in National Geophysical Research Institute ,Hyderabad. He has immensely contributed in the understanding of the evolution, geodynamics and nature of Indian lithosphere by applying gravity, magnetic , Magsat , heat flow and other geophysical techniques. He has a large number of publication in the international and national journal of geophysics and geology.

Dr. N.V. Kalyankar joined Yeshwant Mahavidyalaya, Nanded Maharashtra India as Lecturer and presently working as Principal.. He completed graduation (B.Sc.) in 1978 and obtained Masters Degree (M.Sc.) Physics in 1980 and his doctorate in 1995 in Physics. He is recognized research guide and 12 students are working under his guidance for Ph.D. He has many publications in International Journals to his credit. He is actively engaged in Coordinating research activities in the faculty of Arts, Social Science, Commerce and Science and responsible for research work in fourteen different disciplines and subjects

Dr. M. Ravi Prakash Graduated in 1975 and obtained Ph.D in Statistics (1998) at Osmania University, Hyderabad, India. He is a senior Scientist at National Geophysical Research Institute, Hyderabad. His research interest include Fractals, GeoStatistics, NonParametric Estimation and Stochastic Modeling of Earth Systems

Ramesh Chand is presently working as Deputy Director at the National Geophysical Research Institute (NGRI), Hyderabad. He obtained his M.Sc. Physics (Meerut University) and M.Tech. Applied Geophysics (IIT Roorkee) in year 1871 and 1974 respectively. Since 1974 he is working with NGRI and pursuing his research in the field of isotope Hydrology. He was awarded German Fellowship (CDG) during 1977-79 and worked at Institute of Radio-hygrometry (GSF) Neuherberg, Munich Germany. He also worked at National Institute of Hydrology Roorkee during 1985-87. He published over 50 papers in international/national journals.