COMPARATIVE STUDY OF RAINFALL FORECASTING MODELS

Mohita Anand Sharma[1] and Jai Bhagwan Singh[2]

1. Research Scholar,
2. Seinor Professor Statistics.

Department of Mathematics and Statistics, School of Basic and Applied Science,
Shobhit University, Meerut, Uttar Pradesh, 250110, India.
Email: mohita_anand@rediffmail.com

Abstract: The weekly average of seven weather variables viz. rainfall, maximum and maximum temperature, relative humidity at 7.00 am and 2.00 PM., bright sunshine hours and pan evaporation of 39 years for the month of June were collected from the IMD approved Metrological Observatory situated at GB Pant University of Agriculture and Technology, Pantnagar, India. Comparisons of the forecasting models were made to identify the appropriate model for the prediction of rainfall. Stepwise regression analysis was used to identify the significantly contributed variables. These selected variables were used for Artificial Neural Network analysis. The Neural network work tool of Matlab is used for prediction of weekly rainfall. The data were analyzed to find the mean, standard deviation and minimum and maximum value for all the seven variables. The inter correlation between these variables were also observed. The result shows that the variability in rainfall during the month of June was highest among the all variables and it was lowest for minimum temperature and the correlation between rainfall and relative humidity at 2PM was maximum (0.5529) and it was minimum between rainfall and minimum temperature (-0.0224). The result shows that the value of the coefficient of determination (R^2) = 50.91% for multiple regression model while it was 76.68% using neural network model, secondly, the prediction error for multiple regression model is 0.368 which is higher than that of the prediction error obtained through neural network model as 0.131 and thirdly, the absolute mean value is 13.393 for regression model and 4.796 for neural network model which further indicate that the mean absolute difference from actual rainfall is less obtained through the neural network method. Finally on the basis of maximum value of coefficient of determination(R^2), minimum prediction error and minimum mean difference it was concluded that neural network approach is better than multiple regression approach for predicting weather parameters.
[Sharma, M.A. and Singh, J.B. Comparative study of Rainfall forecasting models. New York Science Journal 2011;4(7):115-120]. (ISSN: 1554-0200). http://www.sciencepub.net/newyork.

Key words: Rainfall Data, Multiple Linear Regression (MLR), Artificial Neural Network (ANN).

1. INTRODUCTION:

Weather forecasting for the future is one of the most important attributes to forecast because agriculture sectors as well as many industries are largely dependent on the weather conditions. It is often used to predict and warn about natural disasters that are caused by abrupt change in climate conditions (Paras et al., 2007). Several researchers have utilized statistical principles for studying the weather forecasting models. In general, weather prediction models use as a combination of empirical and dynamic techniques.

Among the different approaches, the multiple regression approach and artificial neural network approach are most commonly used methods for forecasting weather variable. Forecasting models based on time series data are being developed for prediction of the different variables. Linear and non-linear multiple regression models of different orders are also being used for predicting purpose based on the time series data. There are some limitations of multiple regression approach such as multiple collinearly, inter relation, extreme observation and non-linear relationship between dependent and independent variables.

The prediction in an artificial Neural Network Method (ANN) always takes place according to any data situation (without limitation) based on initial training. However, networks can be different depending on type and number of layers and having control feedback etc. The ANN method will be more efficient when the non-linear relationship between dependent and independent variables exists.

Chattopadhyay (2007) analyzed that neural net with three nodes in the hidden layer is found to be the best predictive model for possibility of predicting average summer-monsoon rainfall over India. Paras *et al.* (2007) concluded that neural networks are capable of modeling a weather forecast system. Statistical indicators chosen are capable of extracting the trends, which can be considered as features for developing the models.

Fallah-Ghalhary et al. (2009) studied the relationship between climate large-scale synoptic patterns and rainfall in Khorasan region. The research attempted to train Fuzzy Inference System (FIS) based prediction models. The model predicted outputs were compared with the actual rainfall data. Simulation results reveal that soft computing techniques are promising and efficient. The root mean square error by Fuzzy Inference System was obtained 52 millimeter.

Kulshrestha et al. (2009) examined that ANN gives more accurate results to predict the probability of extreme rainfall than the probability by Fisher-Tippet Type II distribution.

Chang'a et al. (2010) described how farmers in south-western highland of Tanzania predict rainfall using local environmental indicators and astronomical factors.

Systematic documentation and subsequent integration of indigenous knowledge into conventional weather forecasting system was recommended as one of the strategy that could help to improve the accuracy of seasonal rainfall forecasts under a changing climate.

Summary of descriptive statistics for maximum daily rainfall is presented by Sharma and Singh (2010) and it was concluded that the wet season in this area starts in the month of june every year.

Wang and Sheng (2010) proposed the application of generalized regression neural network (GRNN) model to predict annual rainfall in Zhengzhou. The simulation results of GRNN showed more advantageous in fitting and prediction as compared with back propagation (BP) neural network and stepwise regression analysis methods because GRNN has smaller prediction error.

Wu et al. (2010) proposed a hybrid rainfall-forecasting approach which is based on support vector regression, particle swarm optimization and projection pursuit technology. The computing results show that the present model yields better forecasting performance in this case study, compared to other rainfall-forecasting models.

Wu et al. (2010) made an attempt to seek a relatively optimal data driven model for rainfall forecasting from three aspects: model inputs, modeling methods, and data preprocessing techniques. Prediction was performed in two modes in which normal mode indicate that Modular artificial neural network (MANN) performs the best among all four models and for data processing mode the improvement of model performance generated by singular spectrum analysis (SSA) is considerable whereas those of moving average (MA) or principle component analusis (PCA) are slight. He analysed that the proposed optimal rainfall forecasting model can be derived from MANN coupled with SSA.

El-Shafie et al. (2011) developed feed forward neural network FFNN model and implemented to predict the rainfall on yearly and monthly basis. The analysis was made on the basis of four parameters viz. root mean square error (RMSE), mean absolute error (MAE), coefficient of correlation (CC) and mean error (BIAS) which suggested that ANN provide better results than the MLR model.

Recently, Zaefizadeh *et al.* (2011) recommended ANN approach as better predictor of yield in Barley than in multiple linear regression.

After going through the work related to our topic, it was observed that the selection of variables (which are significant to the dependent variable) was an important aspect. In the present work stepwise regression analysis was used to identify the significantly contributing variables and a comparative analysis was done using three different analytical methods. Multiple regression approach was compared with artificial neural network approach using weekly average time series data to identify the appropriate model.

2. MATERIALS AND METHODS:

The present study is based on time series rainfall data of 39 years collected from the IMD approved meteorology observatory situated at GB Pant University of Agriculture and Technology Pantnagar, India. It is located at 29°N latitude and 79°3' E longitudes. The wet season in this area starts in the month of June every year. Based on the previous time series analysis weekly data of seven variables viz. rainfall, maximum and maximum temperature, relative humidity at 7.00 am and 2.00 pm. bright sunshine hours and pan evaporation were collected for the month of June.

The descriptive statistics of the weekly average 156 data set was computed resulting the mean, standard deviation and sum for all the seven variables. Minimum and maximum weekly average value was also found for each variable. The Inter correlation between these seven weather variables was also computed. The description of Multiple regression model and neural network model used in this study are given below in brief.

2.1. Multiple regression models: Multiple regression analysis is to incorporate a number of independent variables simultaneously for predicting the value of a dependent variable. In the present study rainfall was treated as dependent variable and maximum temperature, minimum temperature, Relative humidity at 7 am and 2 pm, Pan evaporation and Bright sunshine as independent variables. The form of the multiple linear regression equation fitted to the weekly average weather parameters is given below:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots\ldots + \beta_6 X_6 + \varepsilon \qquad (1)$$

Where:
Y = rainfall (mm)
X_1= Maximum temperature (°C)
X_2 = Minimum temperature (°C)
X_3= Relative humidity at 7 am (%)
X_4= Relative humidity at 2 pm (%)
X_5= Pan evaporation (mm/day)
X_6= Bright sunshine (hours)
α = Intercept
β_i = regression coefficient of i^{th} independent variables (i=1,2,…6)
ε = error term

Stepwise regression analysis was used to identify the significant variables for predicting the dependent variable based on the six independent variables. The best fit multiple regression equation was fitted by stepwise regression analysis through SAS. Stepwise procedure starts with a simple regression model in which most highly correlated one independent variable was only included at first with a dependent variable. Correlation coefficient is further examined in the procedure to find an additional independent variable that explains the largest portion of the error remaining from the initial regression model. Until the model includes all the significant variables, the procedure keeps on repeating. A potential bias in the stepwise procedure results from the consideration of only one variable at a time.

2.2. Neural network model: An Artificial Neural Network (ANN) is a flexible mathematical structure that can discover patterns adaptively between input and output data

set. ANN models have been found useful and efficient. Feed Forward Neural Network (FFNN) is the most popular Neural Network model for time series forecasting applications. ANN provides input-output simulation and forecasting models in situations that do not require modeling of the internal structure of the parameters. The model building process consists of the following sequential steps:

- The significant variables selected through SAS in fitting the multiple regression equation were used as input variables to train the neural network model
- Select the best suited architecture of Feed Forward Neural Network Model (FFNN) for weekly rainfall data. In the proposed neural network model three layers are considered, the input environment with the significant variables, two hidden layers and an output layer. The (8×10×1) configuration is considered with 8 neurons in the first hidden layer and 10 neurons in the second hidden layer and only one neuron in the output layer (Fig. 1) below.

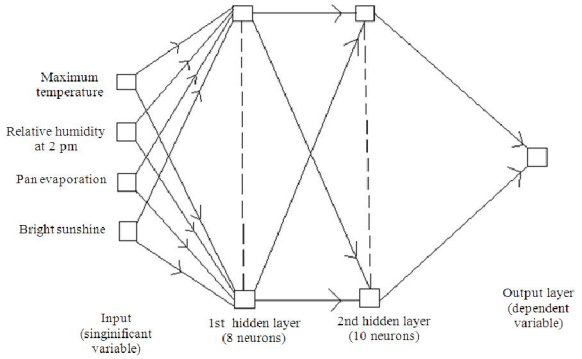

Figure 1. Artificial neural network for weekly average rainfall prediction

- The input enrollments of 92 data sets of weekly average of selected variables were used to train the proposed neural network
- Scaled Conjugate Gradient Algorithm was used for training Multilayer Perceptron (MLP)
- Appropriate activation function for each layer was decided. In the first and second hidden layer tan sigmoid and in the output layer log sigmoid activation function is used
- 9432 epochs were used to train the neural network model with 0.00064 goals, shown in (Fig. 2)
- The test enrollment of 64 data sets of significant weather variables were used to check the performance of the proposed method
- The set of random values distributed uniformly between -1 to +1 are used to initialize the weight of the neural network model
- The Neural Network tool of Mat lab is used for prediction of weekly rainfall
- The predicted value of weekly average rainfall was obtained from the tested data set

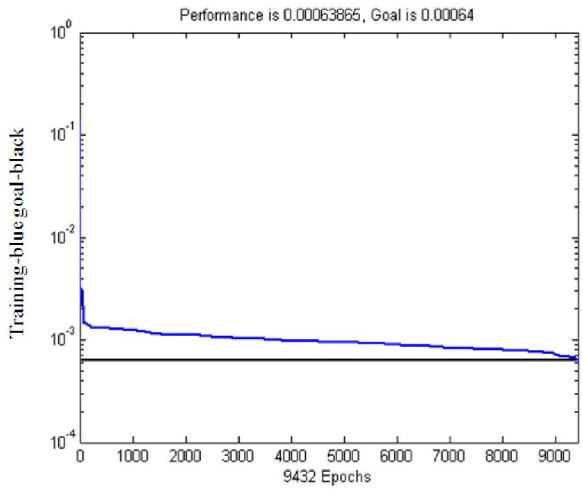

Figure 2. Mapping of the number of epochs obtained for desired goal

2.3. Comparative study: Comparison among the multiple regression model and artificial neural network was made to identify the adequacy of the fitted models using the following three analytical methods:

2.3.1. Coefficient of Determination (R^2): A data set has values y_i, each of which has an associated modeled value $\hat{y}_i$. Here, the values y_i are called the observed values and the modeled values $\hat{y}_i$ are called the predicted values. The "variability" of the data set is measured through different sum of squares:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \text{ , the total sum of squares} \quad (2a)$$

$$SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2 \text{ , the regression sum of squares} \quad (2b)$$

$$SS_{err} = \sum_i (y_i - \hat{y}_i)^2 \text{ , the sum of squares of residuals} \quad (2c)$$

In the above $(\bar{y})$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_i^n y_i \quad (2d)$$

Where, n is the number of observations. The coefficient of determination is the ratio of the regression sum of squares to the total sum of squares. R^2 is a statistic that will give information about the adequacy of the model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. Values of R^2 outside the range 0-1 can occur where it is used to measure the agreement between observed and modeled values. The maximum value of R^2 decides the best fit model.

2.3.2. Prediction Error (PE): After developing the model through training and then testing, adequacy of the model is examined statistically. Over all Prediction Error (PE) is measured as:

$$PE = \frac{\left\langle \left| y_{predicted} - y_{actual} \right| \right\rangle}{\left\langle y_{actual} \right\rangle} \quad (3)$$

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.

2.3.3. Absolute mean difference: The absolute value of the deviation of actual rainfall to the predicted rainfall from regression and neural network model were calculated and their comparative result using paired T-test is presented in results. The paired T-test was used to test the null hypothesis that the equality of two population mean for the estimated value of rainfall using multiple regression analysis and neural network approach are equal against the alternative hypothesis that these two estimated values are not equal at 5% level of significance. T-test statistic was calculated using this formula:

$$t = \frac{\sum d}{\sqrt{\frac{n\left(\sum d^2\right) - \left(\sum d\right)^2}{n-1}}} \quad (4)$$

And the decision was made about the more affective approach, having the less absolute mean deviation.

3. RESULTS:

The methodology presented above was applied to the 39 years weather data for the month of June was taken from meteorological observatory pant agar. The weather data was further classified into weekly averages for further analysis.

The multiple regression model was fitted to predict the weekly average rainfall as dependent variable taking the other weekly average independent variables as maximum temperature, minimum temperature, Relative humidity at 7 am and 2 pm, Pan evaporation and Bright sunshine. The data set which was not included in developing the model was used for comparing the actual and estimated values.

The Descriptive statistics of each variable is given below in Table 1. Where, the numbers of observations are 156 for all the variables. The mean for weekly average maximum temperature was 36.04°C ranging from 24.40-43.20°C and for weekly average minimum temperature the mean was 24.75°C ranging 18.80-34°C. Similarly, 76.18% was the mean for weekly average relative humidity at 7 am and ranging from 38- 95%, for relative humidity at 2 pm the mean was 52.11% varying between 16 and 82%. Also for pan evaporation we observe the variation from 2.10-18.50 mm day ^{1} with a mean of 8.67 mm day ^{1}. Bright sunshine variable with a mean of 7.57 h observed the range of variation between the minimum 1.7 h and maximum 11.6 h weekly average value. Whereas, rainfall with weekly average mean of 42.24 mm observes the widest range from 0.00-291.80 mm among all the other variables.

Table 1. Descriptive statistics of weekly average weather variables for the month of June

Variable	Number of weeks	Mean	Standard Deviation	Sum	Minimum	Maximum
Temperature (°C)	156	36.0372	3.3044	5622	24.40	43.20
Minimum Temperature (°C)	156	24.7526	1.8689	3861	18.80	34.00
Relative Humidity (7 AM) (%)	156	76.1821	12.3689	11884	38.00	95.00
Relative Humidity (2 PM)(%)	156	52.1122	16.4573	8130	16.00	82.00
Pan Evaporation (mm/day)	156	8.6647	2.9420	1352	2.10	18.50
Bright Sunshine (hours)	156	7.5731	1.9320	1181	1.70	11.60
Rainfall (mm)	156	42.2423	54.0589	6590	0.00	291.80

The standard deviation of rainfall is the highest 54.0589 which shows a lot of variability in rainfall during the month of June where as it is the lowest for minimum temperature.

Further the Inter correlation coefficient between different variable was computed and presented in Table 2. From Table 2 it was observe that maximum temperature is positively correlated with minimum temperature but highly positively correlated with pan evaporation and bright sunshine. Relative humidity at 7 pm is also highly positively correlated with relative humidity at 2 pm. Rest all other inter correlation are negatively correlated. There is negative correlation between rainfall and maximum temperature, minimum temperature, pan evaporation and bright sunshine hours while it is positively correlated with relative humidity at 7 am and 2 pm. It can be observed that there is highest correlation between rainfall and relative humidity at 2 pm and lowest correlation among rainfall and minimum temperature.

The most significantly contributed variables were selected using stepwise regression analysis based on 92 data set of 23 years and the best fit multiple regression model is given below as equation 5:

$$Y = 18.19598 - 5.1012X_1 + 2.28600X_4 + 5.48338X_5 + 6.17105X_6 \quad (5)$$

where, Maximum temperature (X_1), Relative humidity at 2 pm (X_4), Pan Evaporation (X_5), Bright sunshine (X_6) were observed as most contributing variable. Further, this model was used to find the predicted value of weekly average rainfall for different data sets used for testing purpose.

The same four significant variables observed in stepwise regression model were used from the training data sets for development of neural network model. Then using this neural network program, predicted value of weekly average rainfall was estimated using the testing data set.

Table 2. Inter correlation coefficient between weather variables

Variables	Maximum Temperature	Minimum Temperature	Relative humidity (7 am)	Relative humidity (2 pm)	Pan Evaporation	Bright Sunshine	Rainfall
Maximum temperature	1.00000	0.01164	-0.84095	-0.82149	0.75499	0.54579	-0.49670
Minimum Temperature	0.01164	1.00000	0.04085	0.08632	-0.09945	-0.07181	-0.02237
Relative Humidity (7 am)	-0.84095	0.04085	1.00000	0.87818	-0.76055	-0.57829	0.47226
Relative Humidity (2 pm)	-0.82149	0.08632	0.87818	1.00000	-0.74103	-0.63545	0.55286
Pan Evaporation	0.75499	-0.09945	-0.76055	-0.74103	1.00000	0.54456	-0.27748
Bright Sunshine	0.54579	-0.07181	-0.57829	-0.63545	0.54456	1.00000	-0.26515
Rainfall	-0.49670	-0.02237	0.47226	0.55286	-0.27748	-0.26515	1.00000

Table 3. Comparative results

Models	Mean	Mean difference	T statistic	P value
$\lvert y - \hat{y}_{regression} \rvert$	13.393	8.5962	2.1905	0.0322
$\lvert y - \hat{y}_{NN} \rvert$	4.796			

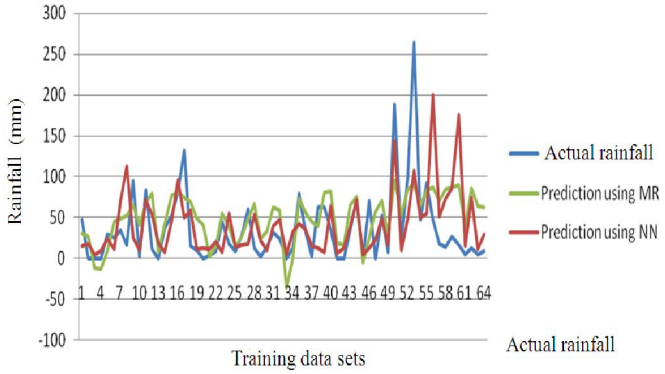

Figure 3. Actual and Predicted weekly average rainfall using multiple regression and neural network model

Figure 3 clearly demonstrates the comparison of the actual rainfall with predicted rainfall using multiple regression and neural network models, which shows that neural network model, is approaching to actual rainfall in comparison to multiple regression model.

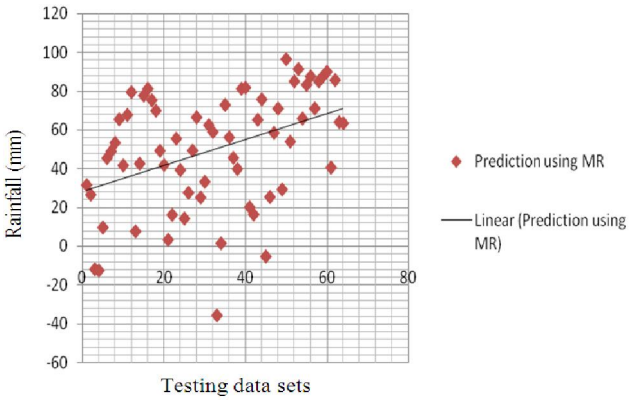

Figure 4. Trend analysis using multiple regression model

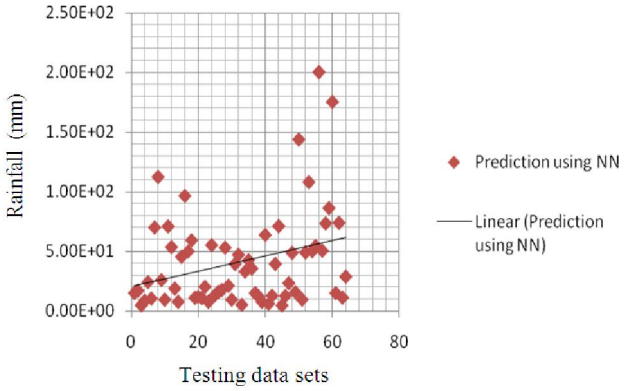

Figure 5. Trend analysis using neural network model

The similar trend is also presented in Fig. 4 and 5 for multiple regression model and neural network model respectively.

3.1. Comparison of two methods: The comparative analysis of these two models based on three approach viz. coefficient of determination, prediction error and absolute mean difference is analyzed below to finally identify the adequacy of the model.

3.1.1 Coefficient of determination(R^2): The values of coefficient of determination R^2 = 50.91% was observed using multiple regression model and (R^2) = 76.68% using neural network model, which clearly indicates that Neural Network model provides better estimation of rainfall in comparison to the regression analysis in the case of unknown test data sets.

3.1.2 Prediction Error (PE): After training and testing, the predicted error values were computed for each model. Prediction error obtained from multiple regression models was 0.368 and through neural network model was 0.131 which was sufficiently small indicating that neural network model is best prediction model.

3.1.3 Absolute mean difference: The mean absolute difference from actual rainfall in the regression method was significantly more than the neural network method which explains that the error in the estimation by the regression method was more than the error in neural network method,

given in Table 3. Hence, neural network model give more effective result than regression approach.

4. DISCUSSION:

The result shows that the variability in rainfall during the month of June was highest among the all variables and it was lowest for minimum temperature and the correlation between rainfall and relative humidity at 2pm was maximum (0.5529) and it was minimum between rainfall and minimum temperature (-0.0224). Further, comparisons of the forecasting models were made to identify the appropriate model for the prediction of rainfall. The result shows that the value of the coefficient of determination (R^2) =50.91% for multiple regression model while it was 76.68% using neural network model, secondly, the prediction error for multiple regression model is 0.368 which is higher than that of the prediction error obtained through neural network model as 0.131 and thirdly, the absolute mean value is 13.393 for regression model and 4.796 for neural network model which further indicate that the mean absolute difference from actual rainfall is less obtained through the neural network method. The graphs further indicate that neural network model is approaching to actual rainfall in comparison to multiple regression models. Finally it can be concluded that neural network approach is better than multiple regression approach for estimating weather parameters.

Corresponding Author:
Mohita Anand Sharma, Research Scholar,
School of Basic and Applied Science,
Shobhit University, Meerut, Uttar Pradesh, 250110, India
Email: mohita_anand@rediffmail.com

REFERENCES

1. Chang'a, L.B., Yanda, P.Z. and Ngana J. Indigenous knowledge in seasonal rainfall prediction in Tanzania: A case of the south-western highland of Tanzania. Journal of Geography and Regional planning 2010; 3(4): 66-72.
2. Chattopadhay, S. Multiplayer feed forward artificial neural network model to predict the average summer monsoon rainfall in India. Acta Geophysica 2007; 55(3): 369-382.
3. El-Shafie, A.H., El-Shafie, A., El-Mazoghi, H.G., Shehata, A. and Taha, Mohd. R. Artificial neural network technique for rainfall forecasting applied to Alexandria, Egypt. International Journal of the Physical Science 2011; 6(6): 1306-1316.
4. Fallah-Ghalhary, G.A., M. Mousavi-Baygi and H.N. Majid. Annual rainfall forecasting by using mamdani fuzzy inference system. Res. J. Env. Sci. 2009; 3: 400-413.
5. Kulshrestha, M., R.K. George and A.M. Shekh. Application of artificial neural networks to predict the probability of Extreme rainfall and comparison with the probability by Fisher-Tippet Type II distributions. Int. J. Applied Math. Computations 2009; 1: 118-131.
6. Paras, M.S., A. Kumar and M. Chandra. A feature based neural network model for weather forecasting. Int. J. Comput. Intel. 2007; 4: 209-216.
7. Sharma, M.A. and Singh, J.B. Use of probability distribution in rainfall analysis. New York Science Journal 2010; 3(9):40-49.
8. Wang, Zhi-L. and Sheng, Hui-h. Rainfall prediction using generalized regression neural network: case study Zhengzhou. Computational and Information Science (ICCIS) 2010; 1265-1268.
9. Wu, C.L., Chau, K.W. and Fan, C. Prediction of rainfall time series using modular artificial neural networks coupled with data preprocessing techniques. Journal of Hydrology 2010; 389(1-2): 146-167.
10. Wu, J., Liu, M. and Jin, L. A hybrid support vector regression approach for rainfall forecasting using particle swarm optimization and projection pursuit technology. International Journal of Computational Intelligence and Application (IJCIA) 2010; 9(2): 87-104.
11. Zaefizadeh, M., M. Khayatnezhad and R. Gholamin. Comparision of Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) in predicting the yield using its components in the Hulless Barley. American-Eurasian J. Agric. Environ. Sci. 2011; 10: 60-64.

27th June 2011.