Performance of Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) for the Prediction of Monthly Maximum Rainfall in Benin City, Nigeria

Ilaboya, I.R and Igbinedion O. E*

* Department of Civil Engineering, Faculty of Engineering, University of Benin, P.M.B 1154, Benin City, Edo State, Nigeria.

(rudolph.ilaboya@uniben.edu; Mobile: +2348038027260)

Received: 28.11.2018 Accepted:13.03.2019

Abstract- The focus of the study was to investigate the capability of linear and non-linear regression techniques for long-term rainfall prediction. Of the linear regression techniques, multiple linear regression method was employed. One of the non-linear regression techniques being widely used in time series prediction is Artificial Neural Networks (ANN) approach which has the ability of mapping between input and output patterns. ANNs are among the numerous empirical models available and have proven to be especially good in modelling time-dependent systems. The study area was restricted to Benin City. Monthly rainfall data, wind speed, evaporation, temperature, and relative humidity for the period of 1981 to 2015 spanning to about thirty-four (34) years was collected processed and used for the analysis. Data analysis tools, namely; EViews, SPSS and MATLAB were employed to conduct the analysis. Results of the descriptive statistics show a marked variation in the mean and standard deviation of the data used. Rainfall for example had a mean value of 459.643 and standard deviation of 1.0655E2. The bell shaped configuration observed in the histogram plot of the variables revealed that the climatic variables used in the study are statistically normally distributed. On the performance of multiple linear regression (MLR) and artificial neural network (ANN), it was observed that artificial neural network performed better than multiple linear regressions. This conclusion was based on the calculated value of the coefficient of determination (R2) for which ANN was 0.9999 and MLR was 0.1755. The performance of ANN compared to MLR was based on the non-linear dependence of rainfall on other climatic variables such as temperature, wind speed, relative humidity, and vapour pressure.

Keywords- Artificial Neural Network (ANN), Multiple Linear Regressions (MLR), Coefficient of Determination (R2), Test of Normality, Skewness and Kurtosis

1. Introduction

Accurate forecasting of rainfall events is an important issue in hydrological research. In addition, early warnings of severe weather made possible by timely and accurate forecasting can help prevent damages caused by natural disasters. Models aimed at prediction of long-term rainfall have not been very satisfactory in terms of accuracies when compared with short-term rainfall prediction models (Mekanika et al., 2011). One of the probable reasons that account for the difficulties in conducting long-term rainfall prediction is the complexity of the atmospheric processes and the uncertainty of relationships between rainfall and hydro-meteorological variables. Secondly, most available time series analysis models consider linear relationships between rainfall and other meteorological variables. One of such model is multiple linear regressions (MLR). The model is mostly used in various research areas especially in extreme and daily rainfall prediction and is widely accepted in hydrological analysis. In spite of their obvious success in many applications, the acute variability associated with climatic variables can limit the reliability of multiple linear regression results especially for long term prediction (Leahy, 2001).

However, in the real world, temporal variations in data are difficult to analyze and predict since they do not always show simple regularities (Wong et al., 2003). The use of non-linear models such as Artificial Neural Networks (ANN), which are capable of modelling complex non-linear problems, can be suitable for real world temporal data (Hung et al., 2008). Neural networks procedure is considered data driven as opposed to model driven procedures. This is due to its dependence on the available data (Shaymaa, 2014). In addition, neural networks possessed the capacity to work with data without having prior knowledge of the process from which the data were generated.
In this study, attempt was made to compare the performance of multiple linear regression and artificial neural network in the prediction of monthly maximum rainfall.

2. Research Methodology

2.1. E-signature at Browser

The study area is Benin City. Benin City is a humid tropical urban settlement which comprises of three Local Government Areas namely Egor, Ikpoba Okha and Oredo. It is located within latitudes 6020'N and 6058'N and longitudes 5035'E and 5041'E. It broadly occupies an area of approximately 112.552 sq km. This extensive coverage suggests spatial variability of weather and climatic elements. Benin City lies visibly in the southern most corner of a dissected margin: a prominent topographical unit which lies north of the Niger Delta, west of the lower Niger Valley, and south of the Western Plains and Ranges (Okhakhu, 2010). Figure 2.1 shows the base map of Benin City

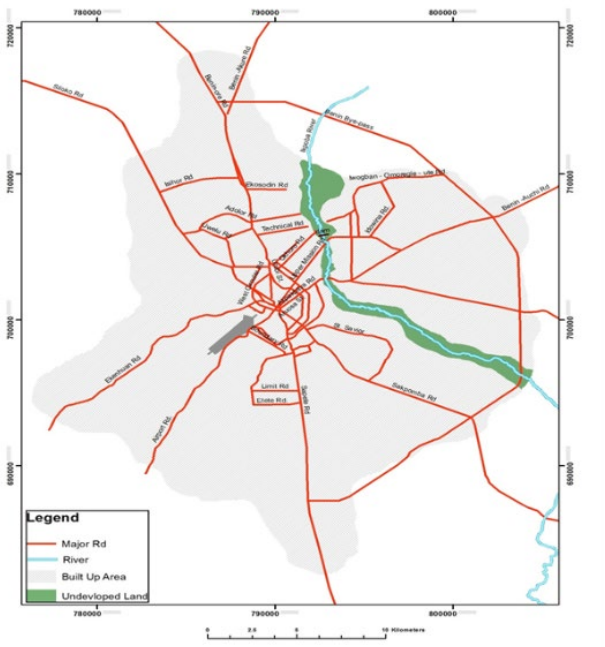


Fig 1. Base Map of Benin City

2.2 Data Collection and Pre-processing

The data used for this study were collected from the Nigerian Meteorological Agency (NIMET), Warri Delta State Nigeria. The data include monthly temperature, vapour pressure, relative humidity, wind speed and precipitation data for thirty four (34) years spanning between; 1981 to 2015. Pre-processing of the data was aimed at:

i. Missing data analysis

ii. Test of Homogeneity;

iii. Detection of outliers; and

2.1.1 Descriptive Statistics of the Data

The data used for this study were collected from the Nigerian Meteorological Agency (NIMET), Warri Delta State Nigeria. The data include monthly temperature, vapour pressure, relative humidity, wind speed and precipitation data for thirty four (34) years spanning between; 1981 to 2015. Pre-processing of the data was aimed at:

i. Missing data analysis

ii. Test of Homogeneity;

iii. Detection of outliers; and

2.1.2 Analysis of Missing Data

The missing data was filled using Expectation Maximization Algorithm (EMA) which is arguably one of the best missing data analysis methods second to Multiple Imputation method which is the best missing data analysis method. Expectation Maximization Algorithm was selected for use in this study owing to its high degree of flexibility and ease of execution. It is important to note that before you perform missing data analysis using EMA, you need to first run the little MCAR (missing completely at random) test which include; chi-square, DF and Sig. The null hypothesis for the little MCAR test was formulated as;

H0; the data are missing completely at random

H1; data are not missing completely at random

The analysis was done at 99.5% confidence interval that is 0.05 degree of freedom. Since p-value > 0.05 then H0

was accepted and it was conclude that the data are missing completely at random

2.1.3 Detection of Outliers

Outlier detection was done using the labeling rule method. The labeling rule is the statistical method of detecting the presence of outliers in the data using the 25th percentile (lower bound) and the 75th percentile (upper bound). The underlying mathematical equation based on the lower and the upper bound is presented as follows:

$$Lower\ Bound\ \mathrm{Q}_1 - (2.2 \times (Q_3 - Q_1) \qquad (2.1)$$

$$Upper\ \text{Bound}\ \mathrm{Q}_3 + (2.2 \times (Q_3 - Q_1) \qquad (2.2)$$

At 0.05 degree of freedom, any data lower than Q1 or greater than Q3 was considered an outlier and was removed before the analysis (Levi et al., 2009).

2.1.4 Test of Homogeneity

Frequency analysis of time series data requires that the data be approximately homogeneous and independent. Homogeneity test was carried out to establish the fact that the data used for the analysis are from the same population distribution. Homogeneity test is based on the cumulative deviation from the mean as expressed using the mathematical equation below (Raes et al., 2006).

$$S_k = \sum_{i=1}^{k}\left(X_i - \bar{X}\right) \quad \text{k = 1, ---------- n} \qquad (2.3)$$

Where;

X_i = The record for the series X1 X2 ---------- Xn

$\bar{X}$ = The mean

Sks = the residual mass curve

The initial value of Sk = 0 and last value of Sk = n are equal to zero. For a homogeneous record, one may expect that the Sks fluctuate around zero in the residual mass curve since there is no systematic pattern in the deviation Xi's from the average values . To perform the homogeneity test, a software package (Rainbow) for analyzing hydrological data was employed (Raes et al., 2006).

3.

3.1 Prediction of Monthly Maximum Rainfall using Multiple Regression Approach

To apply multiple linear regression models, the meteorological variables that influenced rainfall must first be known. From available literatures, it was observed that temperature, wind speed, vapour pressure and relative humidity are the most important climatic variables affecting rainfall events (Lee and Wong, 1998; Kin et al 2001). In developing the multiple linear regression equation, temperature, wind speed, vapour pressure and relative humidity were taken as the independent variables while rainfall events was the dependent variable. To ascertain the dependence of the dependent variable on the selected independent variables, multiple linear regressions were applied to generate a regression equation of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + - - - - - - - - - \beta_5 x_5 + \varepsilon \qquad (3.1)$$

Where;

X1, X2 ---------- Xn = the selected independent variables

Y = the dependent variable (Rainfall events), β_0 , β_1 are the regression constant; ξ is the deviation. To execute the multiple linear regression modelling and generate the regression equation, statistical software (EViews 9.0) was employed.

3.2 Prediction of Monthly Rainfall Events using Artificial Neural Network

To train a network, large volume of data is required. In this study, thirty four (34) years monthly data of rainfall, temperature, wind speed, relative humidity and vapour pressure resulting to about four hundred and eight (408) data was used. To apply neural network, 60% of the data was employed to train a network, 25% of the data was used to validate the network while the remaining 15% was used to test the performance of the network. The neural network modelling and prediction was done with the aid of a neural network modelling tool (MATLAB 10.1). The basic steps involved in the application of the network are as follows:

i. Normalization of the data

ii. Selection of optimum training algorithm or learning rule

iii. Selection of optimum number of hidden neurons

iv. Training of the network

v. Network validation, testing and prediction

3.2.1 Normalization of Input and Output Data

To avoid the problem of weight variation which can subsequently affect the efficiency of the training process, the input and output data were first normalized to obtain a value of between 0.1 and 1.0 using the visual gene developer software.

3.2.2 Selection of Training Algorithm and Hidden Neurons

Input and output data training resulting in the design of network architecture is of paramount importance in the application of neural network to data modeling and prediction. To obtain the optimal network architecture that possesses the most accurate understanding of the input and output data, two major factors were considered. First is the selection of the most accurate training algorithm and secondly, the number of hidden neurons. Based on this consideration, different training algorithm and hidden neurons were tested to determine the best training algorithm and accurate number of hidden neurons that will produce the most accurate network architecture. Selectivity was based on the coefficient of determination (r2) and the mean square error value (MSE) (Kin et al 2001).

3.2.3 Network Training/Performance of ANN

To train the network, 3 runs of 1000 epochs, each with a precision rate of 0.00001 and a learning rate of 0.05 was used. In addition, cross validation data representing about 25% of the total input data was introduced to monitor the progress of training and prevent the network from memorizing the input data instead of leaning which is a common problem associated with overtraining (Kin et al 2001). The progress of the training was monitored using the mean square error (MSE) graph for training and cross validation.

3.2.4 Network Testing

To test the efficiency of the trained network, 15% of the input data was introduced to the network to generate the predicted monthly rainfall event.

3.2.5 Reliability of Trained Network

To test the reliability of the network and acertain the prediction accuracy, a reliability plot of output using the predicted rainfall as the vertical axis and the observed rainfall as the horizontal axis was obtained with the aid of Microsoft Excel Spreadsheet. Reliability of the network was then evaluated using the value of the coefficient of determination (r2) between the predicted and the observed rainfall event.

To test the reliability of the network and acertain the prediction accuracy, a reliability plot of output using the predicted rainfall as the vertical axis and the observed rainfall as the horizontal axis was obtained with the aid of Microsoft Excel Spreadsheet. Reliability of the network was then evaluated using the value of the coefficient of determination (r2) between the predicted and the observed rainfall event.To compare the performance of artficial neural network (ANN) and multiple linear regression (MLR), the following steps were employed;

i. Prediction of rainfall using selected input variable combinations was done using ANN and MLR

ii. A regression plot of output between the observed rainfall and the predicted rainfall was generated using ANN and MLR

iii. Coefficient of determination (r2) was calculated for ANN predicted values of rainfall and MLR predicted values of rainfall

iv. The rule of higher the better was employed to select the best model for predicting rainfall

4. Results and Discussion

4.1 Descriptive Statistics of Data

The descriptive statistics of the data employed in this study is presented in Tables' 4.1a to 4.1d.

			Statistic	Std. Error
Rainfall	Mean		459.643	18.0102
	95% Confidence Interval for Mean	Lower Bound	423.042	
		Upper Bound	496.244	
	5% Trimmed Mean		459.497	
	Median		458.800	
	Variance		1.135E4	
	Std. Deviation		1.0655E2	
	Minimum		222.3	
	Maximum		722.5	
	Range		500.2	
	Interquartile Range		121.1	
	Skewness		.111	.398
	Kurtosis		.700	.778

Table 4.1a: Descriptive statistics of monthly maximum rainfall

Temperature	Mean		33.86	.362
	95% Confidence Interval for Mean	Lower Bound	33.12	
		Upper Bound	34.59	
	5% Trimmed Mean		33.93	
	Median		34.00	
	Variance		4.597	
	Std. Deviation		2.144	
	Minimum		29	
	Maximum		37	
	Range		8	
	Interquartile Range		4	
	Skewness		-.335	.398
	Kurtosis		-.644	.778

Table 4.1b: Descriptive statistics of monthly maximum temperature

Wind Speed	Mean		12.177	.3519
	95% Confidence Interval for Mean	Lower Bound	11.462	
		Upper Bound	12.892	
	5% Trimmed Mean		12.135	
	Median		12.300	
	Variance		4.334	
	Std. Deviation		2.0817	
	Minimum		8.5	
	Maximum		16.7	
	Range		8.2	
	Interquartile Range		2.9	
	Skewness		.247	.398
	Kurtosis		-.533	.778

Table 4.1c: Descriptive statistics of monthly maximum wind speed

Relative Humidity	Mean		83.86	.552
	95% Confidence Interval for Mean	Lower Bound	82.74	
		Upper Bound	84.98	
	5% Trimmed Mean		83.83	
	Median		84.00	
	Variance		10.655	
	Std. Deviation		3.264	
	Minimum		77	
	Maximum		93	
	Range		16	
	Interquartile Range		4	
	Skewness		.076	.398
	Kurtosis		.868	.778

Table 4.1d: Descriptive statistics of monthly maximum relative humidity

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Rainfall	.117	35	.200*	.976	35	.621
Temperature	.132	35	.131	.953	35	.145
Wind Speed	.075	35	.200*	.978	35	.689
Relative Humidity	.111	35	.200*	.961	35	.239
Vapour Pressure	.182	35	.005	.836	35	.000

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Table 4.1e: Descriptive statistics of monthly maximum vapour pressure

For normality; the computed probability p-value based on Shapiro-Wilk and Kolmogorov- Smirnov must be greater than 0.05 that is; ($p > 0.05$). Therefore, for $p>0.05$, the null hypothesis of normality was accepted and it was concluded that the data are normally distributed otherwise; the null hypothesis would be rejected. From the results of Table 4.2, the following observations were made

1. For rainfall, the p-value based on Shapiro-Wilk and Kolmogorov- Smirnov was observed to be 0.621 and 0.200 respectively. Since p-value is greater than 0.05, the null hypothesis was accepted and it was concluded that the rainfall data used in this study are normally distributed.

2. For temperature, the p-value based on Shapiro-Wilk and Kolmogorov- Smirnov was observed to be 0.145 and 0.131 respectively. Since p-value is greater than 0.05, the null hypothesis was accepted and it was concluded that the temperature data used in this study are normally distributed.

3. For wind speed, the p-value based on Shapiro-Wilk and Kolmogorov- Smirnov was observed to be 0.689 and 0.200 respectively. Since p-value is greater than 0.05, the null hypothesis was accepted and it was concluded that the wind speed data used in this study are normally distributed.

4. For relative humidity, the p-value based on Shapiro-Wilk and Kolmogorov- Smirnov was observed to be 0.239 and 0.200 respectively. Since p-value is greater than 0.05, the null hypothesis was accepted and it was concluded that the relative humidity data used in this study are normally distributed.

5. For vapour pressure, the p-value based on Shapiro-Wilk and Kolmogorov- Smirnov was observed to be 0.000 and 0.005 respectively. Since p-value is less than 0.05, the null hypothesis was rejected and it was concluded that the vapour pressure data used in this study are not normally distributed.

4.3 Test of Homogeneity

Homogeneity test was done to ascertain the singular fact that the rainfall, temperature, wind speed, relative humidity and vapour pressure data employed in this analysis are from the same population distribution. The detail of the homogeneity test is presented as follows.

4.3.1 Homogeneity Test of Benin City Rainfall Data

For homogeneity, it is expected that the rainfall data points fluctuates around the zero horizontal center line of the mass curve. Result of the homogeneity test conducted for Benin City monthly maximum rainfall data is presented in Figure 4.2a.

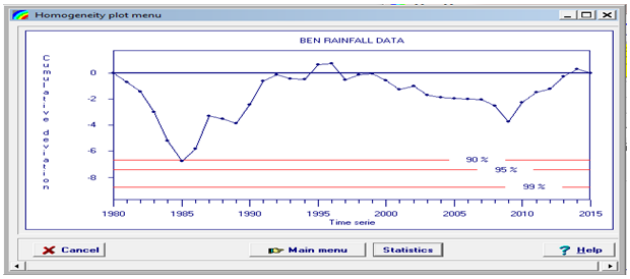


Figure 4.2a: Homogeneity test of Benin City monthly maximum rainfall data

Since the data points fluctuate around the center line as observed in figure 4.2a, it was concluded that the rainfall data from Benin City are statistically homogeneous.

4.3.2 Homogeneity test of Benin City Temperature Data

Result of the homogeneity test conducted for Benin City monthly maximum temperature data is presented in Figure 4.2b.

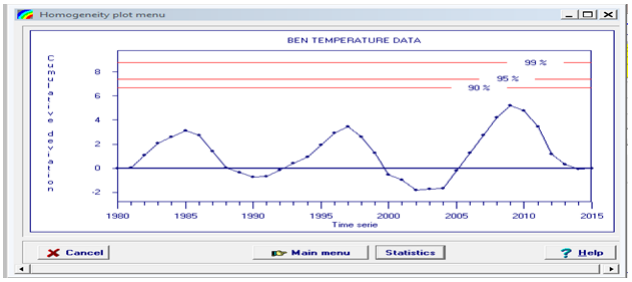


Figure 4.2b: Homogeneity test of Benin City monthly maximum temperature data

Since the data points fluctuate around the center line as observed in Figure 4.2b, it was concluded that the temperature data from Benin City are statistically homogeneous. The same results were obtained for wind speed, relative humidity and vapour pressure. Based on the results, it was concluded that the data employed for the analysis are statistically homogenous.

4.4 Outlier Detection Using Labeling Rule Method

The labeling rule method was employed to determine the presence of outliers in the monthly maximum rainfall, temperature, wind speed, relative humidity and vapour pressure. The outlier statistics of the data is presented in Table 4.3a

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Rainfall	232.620	325.120	394.800	458.800	515.900	614.740	669.460
	Temperature	29.80	31.00	32.00	34.00	36.00	37.00	37.00
	Wind Speed	8.900	9.360	10.400	12.300	13.300	15.480	16.140
	Relative Humidity	77.80	78.60	82.00	84.00	86.00	87.00	89.00
	Vapour Pressure	19.340	19.580	20.700	21.200	21.800	22.440	24.260
Tukey's Hinges	Rainfall			404.200	458.800	511.000		
	Temperature			32.50	34.00	35.50		
	Wind Speed			10.500	12.300	13.300		
	Relative Humidity			82.00	84.00	86.00		
	Vapour Pressure			20.750	21.200	21.650		

Table 4.3a: Computed percentiles for Benin City climatic data

Using the weighted average definition, the 25th percentile (Q1) for rainfall was observed to be 394.800 while the 75th percentile (Q3) was observed to be 515.900. Adopting the labeling rule equation of the form:

$$Lower\ Bound\ Q_1 - (2.2 \times (Q_3 - Q_1) \quad (4.1)$$

$$Lower\ Bound\ Q_1 - (2.2 \times (Q_3 - Q_1) \quad (4.2)$$

The lower and upper bound statistics were calculated as follows:

Lower bound = 394.800 – (2.2(515.900 – 394.800)) = 128.38

Upper bound = 515.900 + (2.2(515.900 – 394.800)) = 782.32

The extreme value statistics of Benin City monthly maximum rainfall data which shows the highest and lowest case numbers is presented in Table 4.3b

			Case Number	Value
Rainfall	Highest	1	7	722.5
		2	11	656.2
		3	30	615.1
		4	10	614.5
		5	15	580.8
	Lowest	1	6	222.3
		2	4	235.2
		3	5	305.2
		4	17	338.4
		5	29	342.9

Table 4.3b: Extreme value statistics of Benin City monthly maximum rainfall data

From the result of Table 4.3b, it was observed that the highest rainfall value is 722.5mm which is less than the calculated upper bound of 782.32mm. The lowest rainfall value was observed to be 222.3mm which is greater than the calculated lower bound of 128.38mm. Since no rainfall value is greater than the calculated upper bound or lower than the calculated lower bound, it was concluded that the monthly maximum rainfall data from Benin City is devoid of possible outliers. The same procedure also applied for wind speed, temperature, relative humidity and vapour pressure.

4.5 Rainfall Prediction Using Multiple Linear Regressions

Results of the multiple linear regression analysis are presented as follows;

4.5.1 Pearson correlation matrix of regression

The computed Pearson correlation matrix of regression is presented in Table 4.4a

Correlations

		Rainfall	Temperature	Wind Speed	Relative Humidity	Vapour Pressure
Pearson Correlation	Rainfall	1.000	-.282	.101	.190	-.270
	Temperature	-.282	1.000	.148	-.100	-.021
	Wind Speed	.101	.148	1.000	.027	-.339
	Relative Humidity	.190	-.100	.027	1.000	.055
	Vapour Pressure	-.270	-.021	-.339	.055	1.000
Sig. (1-tailed)	Rainfall	.	.050	.281	.137	.059
	Temperature	.050	.	.198	.284	.453
	Wind Speed	.281	.198	.	.438	.023
	Relative Humidity	.137	.284	.438	.	.376
	Vapour Pressure	.059	.453	.023	.376	.
N	Rainfall	35	35	35	35	35
	Temperature	35	35	35	35	35
	Wind Speed	35	35	35	35	35
	Relative Humidity	35	35	35	35	35
	Vapour Pressure	35	35	35	35	35

Table 4.4a: Computed correlation coefficient

The Pearson correlation matrix of regression was employed to give an insight into the relationship between the independent variables. From the result of Table 4.4a, it was observed that;

1. The correlation between rainfall and other independent variables, namely; temperature, wind speed, relative humidity and vapour pressure is very poor

2. The correlation between temperature and other independent variables, namely; rainfall, wind speed, relative humidity and vapour pressure is very poor

3. The correlation between wind speed and other independent variables, namely; temperature, rainfall, relative humidity and vapour pressure is very poor

4. The correlation between relative humidity and other independent variables, namely; temperature, wind speed, rainfall and vapour pressure is very poor

5. The correlation between vapour pressure and other independent variables, namely; temperature, wind speed, relative humidity and rainfall is very poor

The conclusion of poor correlation between the independent variables is based on the poor correlation coefficient between each of the independent variables. If the correlation coefficient between the independent variables is poor, then the dependence of the dependent variable (rainfall) on the independent variable will also be poor.

4.5.2: Validation of the Multiple Linear Regression Models

Relative humidity and vapour pressure) and one dependent variable (rainfall). The analysis was conducted at 95% confidence interval which is 0.05 degree of freedom and result obtained is presented in Table 4.4c

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	72920.528	4	18230.132	1.747	.166[a]
	Residual	313077.998	30	10435.933		
	Total	385998.526	34			

a. Predictors: (Constant), Vapour Pressure, Temperature, Relative Humidity, Wind Speed

b. Dependent Variable: Rainfall

Table 4.4c: ANOVA table for assessing the significance of the model

From the result of Table 4.4c, it was observed that the p-value was greater than 0.05. For ($p>0.05$), the null hypothesis was rejected and it was concluded that the multiple linear regression model is not significant in explaining the underlying relationship between rainfall and other independent variables, namely temperature, wind speed, relative humidity and vapour pressure.

4.5.4: Generation of Multiple Linear Regression Equation

To generate the regression equation, the coefficient estimate statistics pr esented in Table 4.4d was employed.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	862.401	625.231		1.379	.178			
	Temperature	-13.754	8.315	-.277	-1.654	.109	-.282	-.289	-.272
	Wind Speed	2.371	9.066	.046	.262	.795	.101	.048	.043
	Relative Humidity	5.760	5.413	.176	1.064	.296	.190	.191	.175
	Vapour Pressure	-21.090	13.728	-.269	-1.536	.135	-.270	-.270	-.253

a. Dependent Variable: Rainfall

Table 4.4d: Unstandardized coefficient statistics

Based on the unstandardized coefficient statistics, the multiple linear regression equation was derived as follows.

Rainfall = 862.401–3.754(temperature) + 2.371(wind speed) + 5.760(relative humidity) – 21.090(vapour pressure)

(4.1)

4.5.5: Rainfall prediction using multiple linear regression equation

Using the model equation 4.1, the observed and predicted rainfall including the prediction errors were generated and presented in Table 4.4e.

Temperature	Wind Speed	Relative Humidity	Vapour Pressure	Observed Rainfall (mm)	Predicted Rainfall (mm)
34	9.6	85	21.5	394.8	793.6916
36	11.3	83	20.7	391.0	795.5663
36	9.4	83	21.9	305.7	765.7534
35	16	87	21.8	235.2	810.305
35	10.1	78	26.9	305.2	636.9171
33	9.3	83	21.1	567.2	793.6503
31	10.7	86	20.2	722.5	840.7387
31	10.4	79	20.8	445.4	787.0534
33	10.2	85	20.8	427.8	813.6312
33	11.5	85	21	614.5	812.4955
34	14.2	83	21.4	656.2	795.1872
35	11.2	86	22.3	515.9	782.6192
35	12.7	87	21.5	433.9	808.8077
35	14	81	20.5	461.4	798.42
36	11.6	88	20.8	580.8	822.9686
36	10.6	93	21.4	476.0	836.7436
35	13.3	83	21.2	338.4	793.5173
32	10.1	83	21.3	506.1	795.0831
31	12.2	87	23.6	472.5	778.3492
30	13.1	84	20.1	413.7	840.7721
33	15.4	84	20.3	394.3	830.7454
32	12.8	87	21.4	491.8	822.4158
34	8.5	85	22.1	393.3	778.4295
34	14.4	85	21.4	448.9	807.1814
37	13	87	22.5	458.8	780.921
37	16.7	84	19.1	462.4	844.1197
37	15.6	80	19.4	462.5	812.1446
37	12.9	77	19.7	413.6	782.1359
36	9	78	21.4	342.9	746.55
33	13.7	82	20.8	615.1	804.6497
31	11.6	82	19.4	550.8	836.7046
29	12.3	87	22.4	490.4	811.4023
32	13.3	81	21.1	564.0	795.3683
33	13	85	21.2	523.4	811.834
34	12.5	82	22.2	436.5	768.5245

Table 4.4e: Observed and predicted rainfall based on multiple linear regressions (MLR)

4.6: Rainfall prediction using ANN

For rainfall prediction using artificial neural network (ANN), monthly rainfall, temperature, wind speed, relative humidity and vapour pressure data were employed. The data range was between 1981 to 2015 resulting to four hundred and eight (408) data.

4.6.1 Normalization of data

Normalization of the input and output data was done to reduce the effect of weight variation that may consequently result to overtraining. The whole idea was to reduce the weight of the input and output data to between zeros to one (0 to 1). Table 4.5a and 4.5b shows a section of the raw training data and the normalized form of the data.

Neural Network - Training Set [35 sets]

Open pattern | Save pattern | Paste all data | Copy | include titles

	In 1	In 2	In 3	In 4	Out 1	P
1	34.0000000	9.6000000	85.0000000	21.5000000	394.8000000	
2	36.0000000	11.3000000	83.0000000	20.7000000	391.0000000	
3	36.0000000	9.4000000	83.0000000	21.9000000	305.7000000	
4	35.0000000	16.0000000	87.0000000	21.8000000	235.2000000	
5	35.0000000	10.1000000	78.0000000	26.9000000	305.2000000	
6	33.0000000	9.3000000	83.0000000	21.1000000	567.2000000	
7	31.0000000	10.7000000	86.0000000	20.2000000	722.5000000	
8	31.0000000	10.4000000	79.0000000	20.8000000	445.4000000	
9	33.0000000	10.2000000	85.0000000	20.8000000	427.8000000	
10	33.0000000	11.5000000	85.0000000	21.0000000	614.5000000	
11	34.0000000	14.2000000	83.0000000	21.4000000	656.2000000	
12	35.0000000	11.2000000	86.0000000	22.3000000	515.9000000	
13	35.0000000	12.7000000	87.0000000	21.5000000	433.9000000	
14	35.0000000	14.0000000	81.0000000	20.5000000	461.4000000	
15	36.0000000	11.6000000	88.0000000	20.8000000	580.8000000	
16	36.0000000	10.6000000	93.0000000	21.4000000	476.0000000	
17	35.0000000	13.3000000	83.0000000	21.2000000	338.4000000	
18	32.0000000	10.1000000	83.0000000	21.3000000	506.1000000	
19	31.0000000	12.2000000	87.0000000	23.6000000	472.5000000	
20	30.0000000	13.1000000	84.0000000	20.1000000	413.7000000	
21	33.0000000	15.4000000	84.0000000	20.3000000	394.3000000	
22	32.0000000	12.8000000	87.0000000	21.4000000	491.8000000	
23	34.0000000	8.5000000	85.0000000	22.1000000	393.3000000	

Table 4.5a: Section of the raw data used for ANN modelling

Neural Network - Training Set [35 sets]

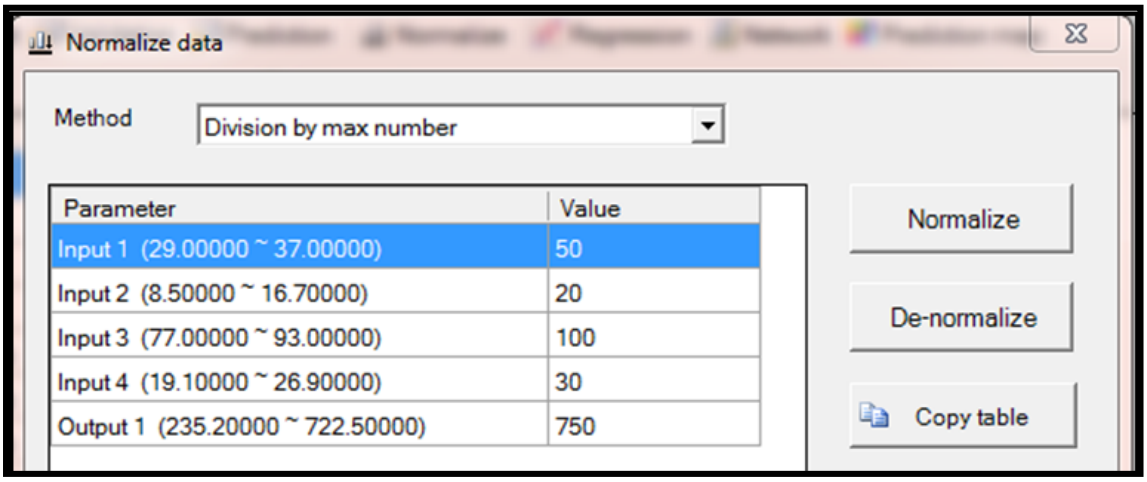
Open pattern | Save pattern | Paste all data | Copy | include titles

	In 1	In 2	In 3	In 4	Out 1
1	0.6800000	0.4800000	0.8500000	0.7166666	0.5264000
2	0.7200000	0.5650000	0.8300000	0.6900000	0.5213333
3	0.7200000	0.4700000	0.8300000	0.7300000	0.4076000
4	0.7000000	0.8000000	0.8700000	0.7266666	0.3136000
5	0.7000000	0.5050000	0.7800000	0.8966666	0.4069333
6	0.6600000	0.4650000	0.8300000	0.7033333	0.7562667
7	0.6200000	0.5350000	0.8600000	0.6733333	0.9633333
8	0.6200000	0.5200000	0.7900000	0.6933333	0.5938666
9	0.6600000	0.5100000	0.8500000	0.6933333	0.5704000
10	0.6600000	0.5750000	0.8500000	0.7000000	0.8193333
11	0.6800000	0.7100000	0.8300000	0.7133333	0.8749334
12	0.7000000	0.5600000	0.8600000	0.7433333	0.6878667
13	0.7000000	0.6350000	0.8700000	0.7166666	0.5785334
14	0.7000000	0.7000000	0.8100000	0.6833333	0.6152000
15	0.7200000	0.5800000	0.8800000	0.6933333	0.7744000
16	0.7200000	0.5300000	0.9300000	0.7133333	0.6346667
17	0.7000000	0.6650000	0.8300000	0.7066667	0.4512000
18	0.6400000	0.5050000	0.8300000	0.7100000	0.6748000
19	0.6200000	0.6100000	0.8700000	0.7866667	0.6300000
20	0.6000000	0.6550000	0.8400000	0.6700000	0.5516000
21	0.6600000	0.7700000	0.8400000	0.6766666	0.5257333
22	0.6400000	0.6400000	0.8700000	0.7133333	0.6557333
23	0.6800000	0.4250000	0.8500000	0.7366667	0.5244000

Table 4.5b: Normalized form of the raw data used for ANN modelling

As indicated in Table 4.5a and 4.5b; In1, In2, In3, In4 represent the input variables, namely; temperature, wind speed, relative humidity and vapour pressure while (Out 1) represent the output variable (rainfall). The parameters of normalization is presented in Table 4.5c

Table 4.5c: Parameters used in normalizing the training data



4.6.2 Selection of training algorithm and hidden neurons

Table 4.5d shows the performance of the different training algorithm tested.

Table 4.5d Selection of optimum training algorithm for ANN

S/No	Training Algorithm (Learning Rule)	Training MSE	Cross Validation MSE	R-Square (r^2)
1	Gradient information (Step)	0.03578	0.06734	0.6592
2	Gradient and weight change (Momentum)	0.07645	0.08923	0.8802
3	Gradient and rate of change of gradient (Quick prop)	0.03874	0.02712	0.7456
4	Adaptive step sizes for gradient plus momentum (Delta Bar Delta)	0.04424	0.003452	0.8023
5	Second order method for gradient (Conjugate gradient)	0.06747	0.065743	0.7023
6	Improved second order method for gradient (Levenberg Marquardt)	0.00025*	0.00045*	0.992*

Result of Table 4.5d revealed that improved second order method of gradient also known as Levenberg Marquardt back propagation training algorithm was the best learning rule and was therefore adopted in designing the network architecture. To determine the exact numbers of hidden neuron, different numbers of hidden neurons were selected to create a trained network using Levenberg Marquardt Back Propagation training algorithm. Performance of the trained network was assessed using mean square error (MSE) and coefficient of determination r2. The number of hidden neuron corresponding to the lowest MSE and the highest r2 as presented in Table 4.5e was selected to design the network architecture.

S/No	Number of Hidden Neurons	Training MSE	Cross Validation MSE	R-Square (r^2)
1	2	0.002100	0.006455	0.888
2	3	0.004305	0.019234	0.795
3	5	0.004700	0.077239	0.864
4	8	0.003305	0.134096	0.828
5	10	0.000412	0.000214	0.9768

Based on the results of Table 4.5d and 4.5e, Levenberg Marquardt Back Propagation training algorithm having 10 hidden neurons in the input layer and output layer was used to train a network of 4 input processing elements (PEs) and 1 output processing elements. The input layer of the network uses the hyparbolic targent (tan-sigmoid) transfer function to calculate the layer output from the network input while the output layer uses the linear (purelin) transfer function. The number of hidden neuron was set at 10 neurons per layer and the network performance was monitored using the mean square error of regression (MSEREG). A learning rate of 0.01, momentum coefficient of 0.1, target error of 0.01, analysis update interval of 500 and a maximum training cycle of 1000 epochs was used. The network generation process divides the input data into training data sets, validation and testing. For this study, 60% of the data was employed to perform the network training, 25% for validating the network while the remaining 15% was used to test the performance of the network. Using these parameters, an optimum neural network architecture was generated as presented in Figure 4.3

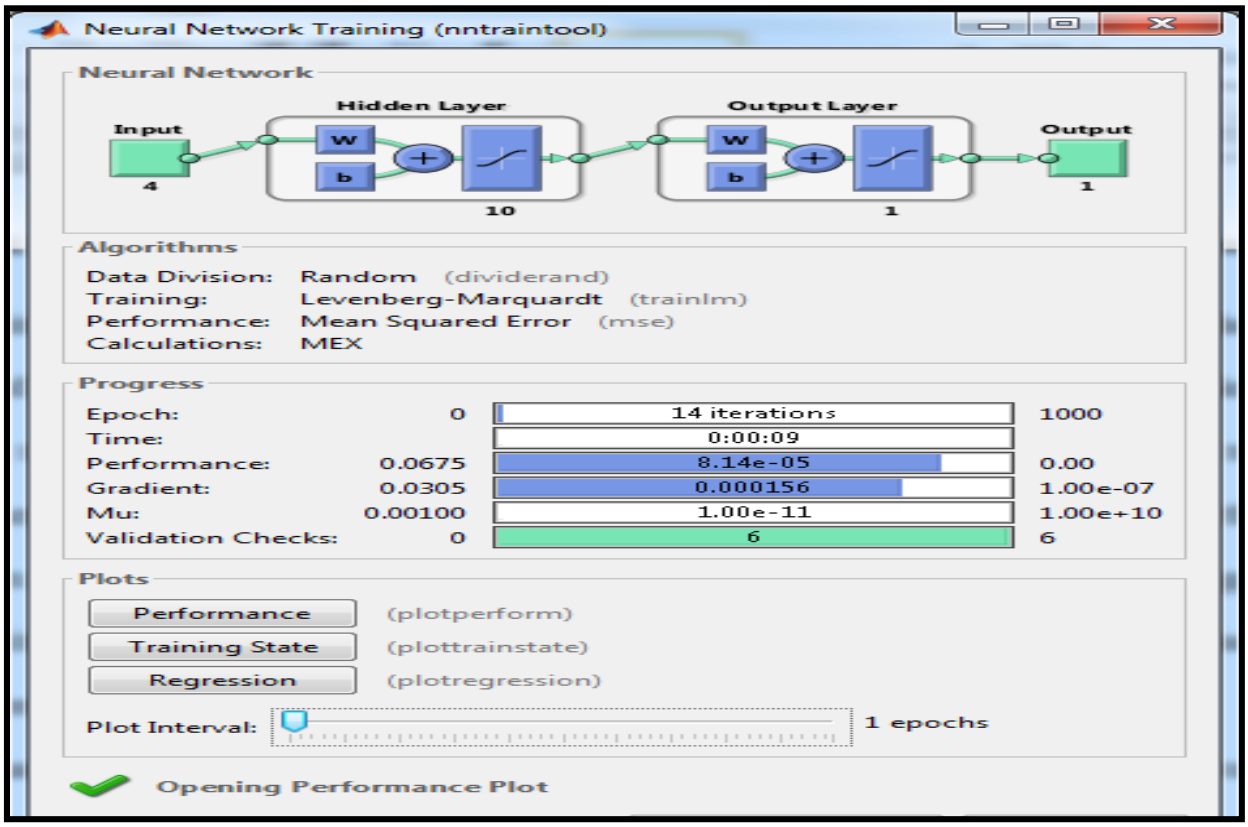


Figure 4.3: Network training diagram for predicting rainfall

From the network training diagram of Figure 4.3b, it was observed that the network performance was significantly good with a performace error of 8.14e-05 which is far lesser than the set target error of 0.01. The maximum number of iteration needed for the network to reach this performance was observed to be 14 iterations which is also lesser than the initial 1000 epochs. The gradient function was calculated to be 0.000156 with a training gain (Mu) of 1.00e-11. Validation check of six (6) was recorded which is expected since the issue of wieght biased had been addressed via normalization of the raw data. A performance evaluation plot which shows the progress of training, validation and testing is presented in Figure 4.4

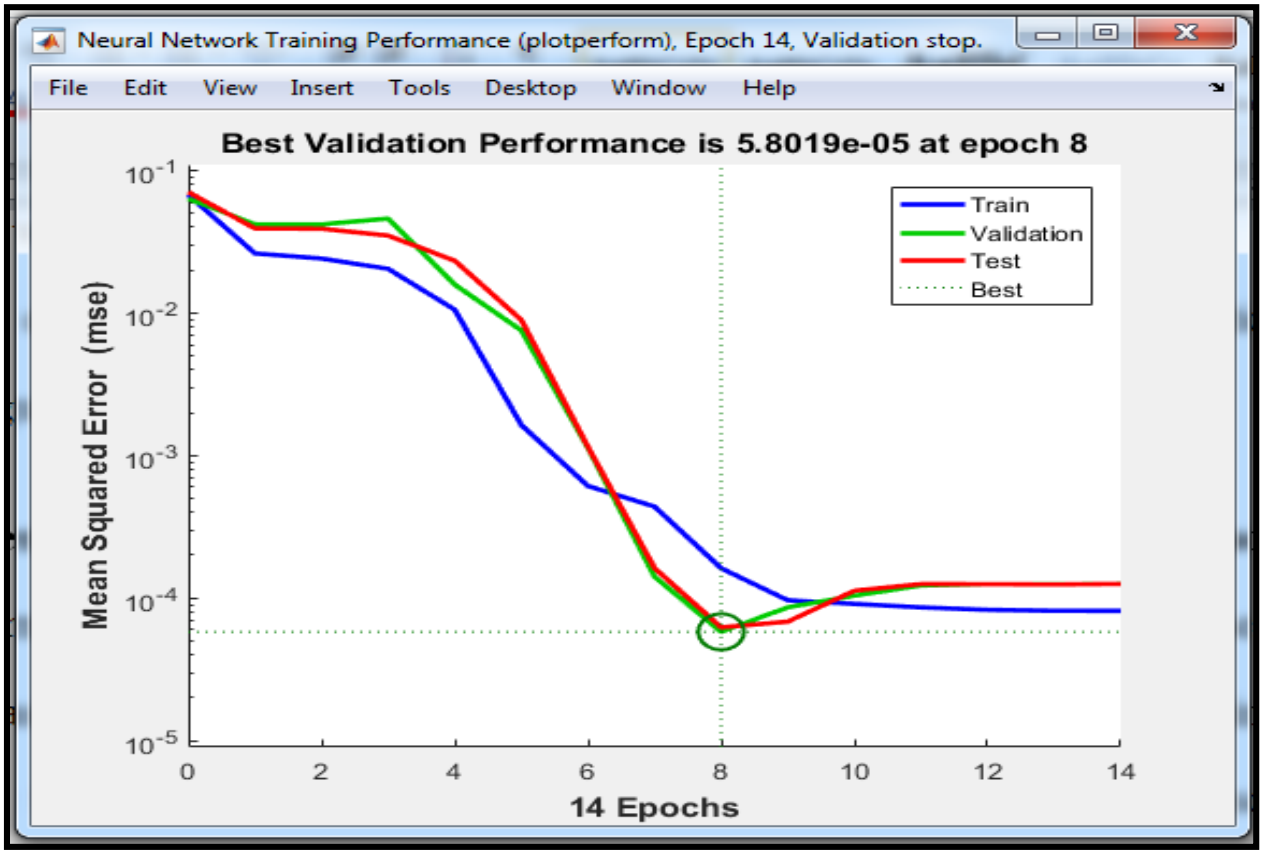


Figure 4.4: Performance curve of trained network for predicting rainfall

From the performance plot of Figure 4.4, no evidence of over fitting was observed. In addition similar trend was observed in the behaviour of the training, validation and testing curve which is expected since the raw data were normalized before use. Lower mean square error is a fundamental criteria used to determine the training accuracy of a network. An error value of 5.8019e-05 at epoch 8 is an evidence of a network with strong capacity to predict rainfall. The training state, which shows the gradient function, the training gain (Mu) and the validation check, is presented in Figure 4.5

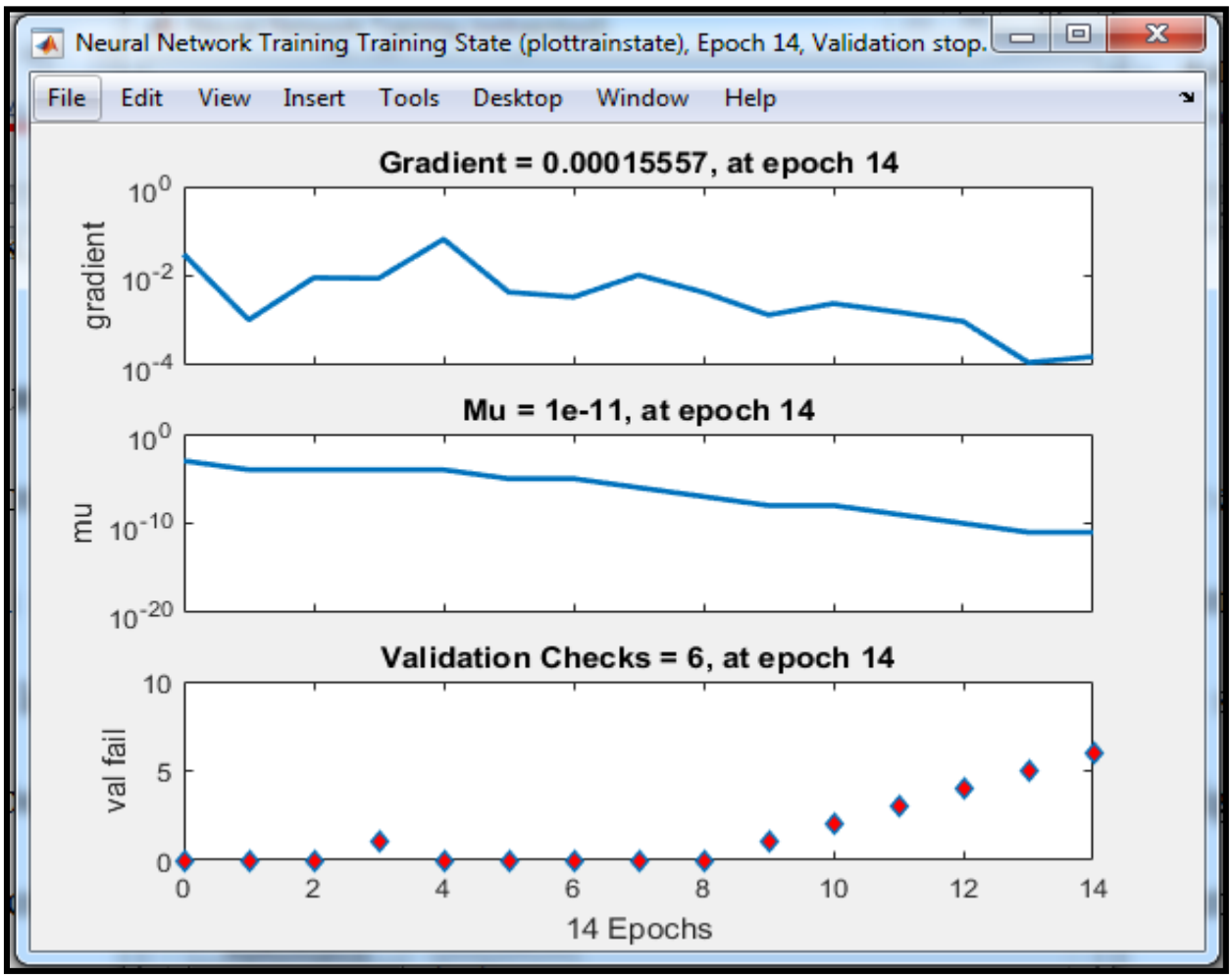


Figure 4.5: Neural network training state for predicting rainfall

Back propagation is a method used in artificial neural networks to calculate the error contribution of each neuron after a batch of data training. Technically, the neural network calculates the gradient of the loss function to explain the error contributions of each of the selected neurons. Lower error is better. Computed gradient value of 0.00015557 as observed in Figure 4.5 indicates that the error contributions of each selected neurons is very minimal. Momentum gain (Mu) is the control parameter for the algorithm used to train the neural network. It is the training gains and its value must be less than one. Momentum gain of 1.0e-11 shows a network with high capacity to predict rainfall. The regression plot which shows the correlation between the input variables (temperature, wind speed, relative humidity and vapour pressure) and the target variable (rainfall) coupled with the progress of training, validation and testing is presented in Figure 4.6

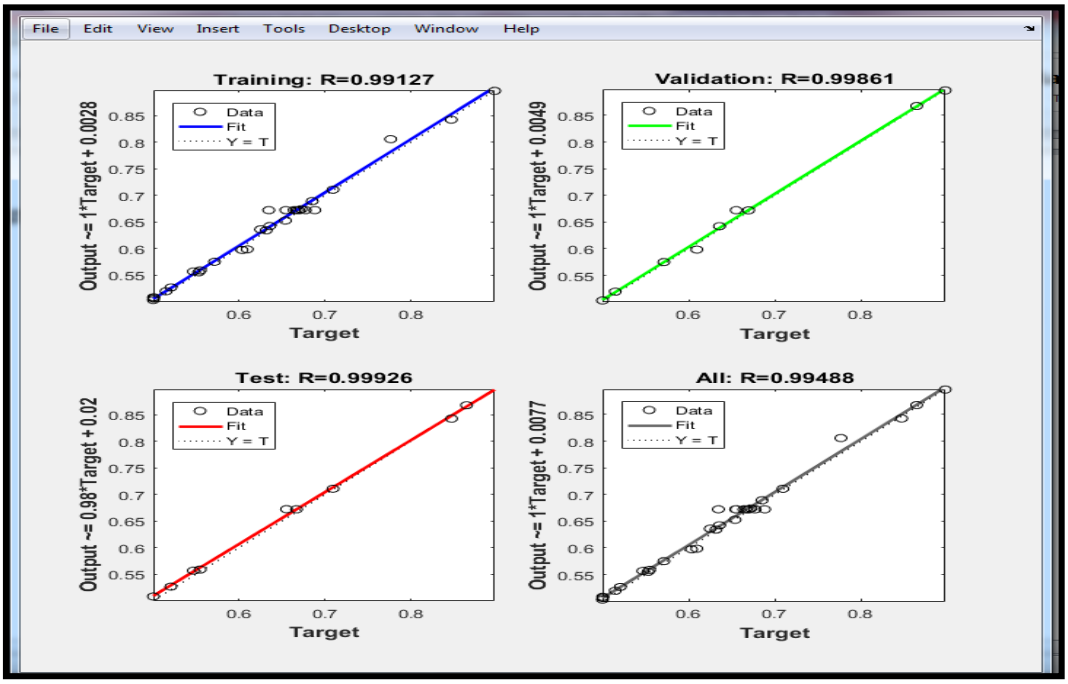


Figure 4.6: Regression plot showing the progress of training, validation and testing

Based on the computed values of the correlation coefficient (R) as observed in Figure 4.6, it was concluded thet the network has been accurately trained and can be employed to predict rainfall.

To test the reliability of the trained network, the network was thereafter employed to predict its own values of rainfall using the same sets of input parameters (temperature, wind speed, relative humidity and vapour pressure). Based on the observed and the ANN predicted values of rainfall, a regression plot of outputs was thereafter generated as presented in Figure 4.7

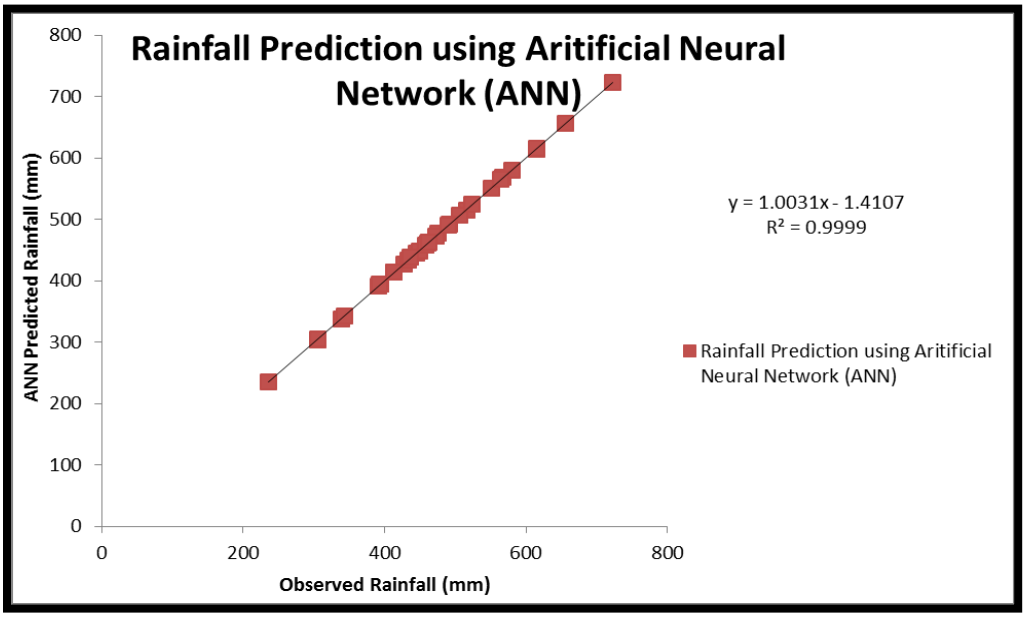


Figure 4.7: Regression plot of observed versus ANN predicted rainfall

Coefficient of determination (r^2) values of 0.9999 as observed in Figures 4.7 was employed to draw a conclusion that the trained network can be used to predict rainfall. To evaluate the performance of ANN in predicting rainfall, a comparative analysis between ANN and multiple linear regressions was done. Table 4.6a and 4.6b shows the result of using ANN and MLR to predict rainfall

Table 4.6a: Observed and predicted rainfall based on multiple linear regressions (MLR) and Neural Network (ANN)

Temperature	Wind Speed	Relative Humidity	Vapour Pressure	Observed Rainfall (mm)	MLR Predicted Rainfall (mm)	ANN Predicted Rainfall (mm)
34	9.6	85	21.5	394.8	793.6916	394.3
36	11.3	83	20.7	391.0	795.5663	391.5
36	9.4	83	21.9	305.7	765.7534	303.2
35	16	87	21.8	235.2	810.305	235.1
35	10.1	78	26.9	305.2	636.9171	304.6
33	9.3	83	21.1	567.2	793.6503	568.9
31	10.7	86	20.2	722.5	840.7387	723.4
31	10.4	79	20.8	445.4	787.0534	445.6
33	10.2	85	20.8	427.8	813.6312	427.1
33	11.5	85	21	614.5	812.4955	614.3
34	14.2	83	21.4	656.2	795.1872	656.5
35	11.2	86	22.3	515.9	782.6192	515.1
35	12.7	87	21.5	433.9	808.8077	433.4
35	14	81	20.5	461.4	798.42	461.4
36	11.6	88	20.8	580.8	822.9686	580.5
36	10.6	93	21.4	476.0	836.7436	476.9
35	13.3	83	21.2	338.4	793.5173	337.5
32	10.1	83	21.3	506.1	795.0831	506.5
31	12.2	87	23.6	472.5	778.3492	472.2
30	13.1	84	20.1	413.7	840.7721	414.4
33	15.4	84	20.3	394.3	830.7454	394.7
32	12.8	87	21.4	491.8	822.4158	491.2
34	8.5	85	22.1	393.3	778.4295	393.7
34	14.4	85	21.4	448.9	807.1814	447.8

Table 4.6b: Observed and predicted rainfall based on multiple linear regressions (MLR) and Neural Network (ANN) cont.

37	13	87	22.5	458.8	780.921	458.6
37	16.7	84	19.1	462.4	844.1197	462.7
37	15.6	80	19.4	462.5	812.1446	462.9
37	12.9	77	19.7	413.6	782.1359	413.3
36	9	78	21.4	342.9	746.55	342.2
33	13.7	82	20.8	615.1	804.6497	615.5
31	11.6	82	19.4	550.8	836.7046	550.3
29	12.3	87	22.4	490.4	811.4023	490.6
32	13.3	81	21.1	564.0	795.3683	564.4
33	13	85	21.2	523.4	811.834	524.5
34	12.5	82	22.2	436.5	768.5245	438.9

Based on the result of Tables 4.6a and 4.6b, a regression plot of output was obtained as presented in Figure 4.8

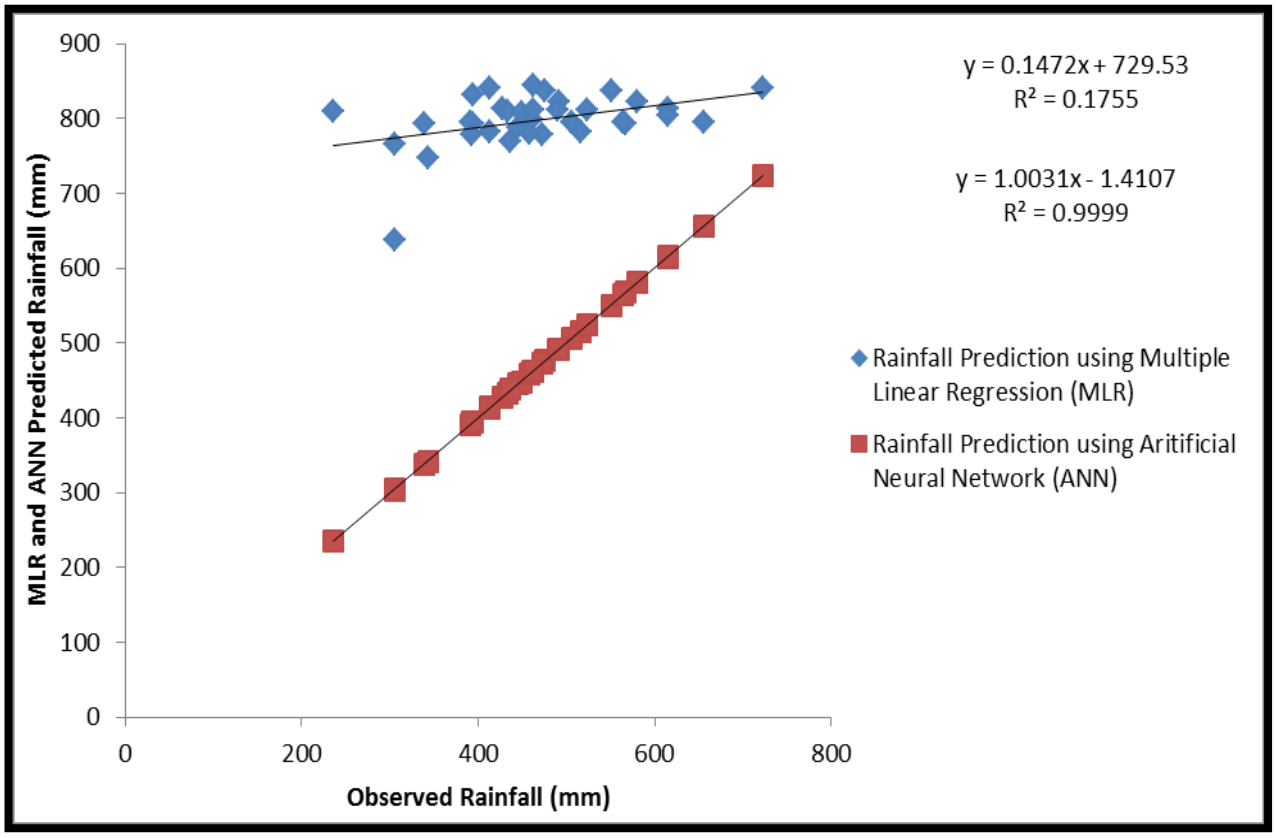


Figure 4.8: Performance of ANN and MLR in predicting rainfall

With a coefficient of determination of 0.9999, ANN was acclaimed the best model for predicting rainfall ahead of MLR having coefficient of determination of 0.1755. When compared with previous work by French et al., (1992), it was observed that the results of this study are similar. ANN was acclaimed by French et al., 1992 as one the best non-linear regression model for the prediction of rainfall and other hydrological variables. Hsu et al (2015) also had similar results, thus showing the adequate of the results obtained from this study.

5. Conclusion

The focus of the present study was to employ multiple linear regression (MLR) and artificial neural network (ANN) for the prediction of rainfall in Benin City. The overall target was to compare the performance of MLR and ANN for the prediction of extreme rainfall using selected input parameters namely; temperature, wind speed, relative humidity and vapour pressure. Based on the overall analysis of the data and the results obtained, the following conclusions were made:

i. The labeling rule method for outlier detection was very effective. Based on the outcome of the outlier detection analysis, it was concluded that the climatic data used for this analysis are devoid of possible outliers
ii. It was also concluded based on the homogeneity test that the data used are from the same population distribution.
iii. Finally, it was concluded that artificial neural network is a better prediction tool than multiple linear regression in the prediction of extreme rainfall in Benin City

6. Reference

French, M. N., Krajewski, W. F., and Cuykendall, R. R (1992): Rainfall forecasting in space and time using neural network, Journal of Hydrology, vol. 137, pp: 1–31.

Hsu, K; Gupta, H. V and Sorooshian, S (2015): Artificial neural network modeling of the rainfall-runoff process, Water Resources Research Journal; vol. 31(10), pp: 2517–2530

Hung, N. Q; Babel, M. S; Weesakul, S and Tripathi, N. K (2008): An artificial neural network model for rainfall forecasting in Bangkok, Thailand; Hydrology Earth System Sciences Discussion 5, pp: 183–218

Kin, C; Luk, J; Ball, E and Sharma, A (2001): An Application of Artificial Neural Networks for Rainfall

Forecasting; Meteorological and Atmospheric Physics Journal, vol. 33, pp: 883-699.

Leahy, K (2001): Multicollinearity: When the solution is the problem, in: Data mining cookbook: Modelling data for marketing, risk and customer relationship management, edited by: Rud, O. P., John Wiley and Sons, New York, pp: 106–108

Lee, S; Cho, S and Wong, P.M (1998): Rainfall prediction using artificial neural network", J. Geog. Inf. Decision Anal, vol. 2, pp: 233–242

Levi, D. B.; Julie, E. K.; Olsen, J. R.; Pulwarty, R. S.; Raff, D. A.; Turnipseed, D. P.; Webb, R. S and Kathleen D. W (2009); Climate Change and Water Resources Management: A Federal Perspective, circular 1331, pp: 1 – 72

Mekanika, F; Leeb, T. S and Imteaza, M A (2011): Rainfall modeling using Artificial Neural Network for a mountainous region in West Iran;19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011; http://mssanz.org.au/modsim2011, PP: 23-45

Okhakhu, P.A. (2014): Environmental and Human Challenges in the Niger-Delta Region of Nigeria, Journal of Environment and Earth Science, Vol. 4(23), pp: 112 - 134

Raes, D; Willens, P and Gbaguidi (2006), Rainbow – A software package for analyzing data and testing the homogeneity of historical data sets, vol. 1, pp: 1-15

Shaymaa, A.A (2014): Prediction of Monthly Rainfall In Kirkuk Using Artificial Neural Network And Time Series Models; Journal of Engineering and Development, Vol. 18, No.1, ISSN 1813- 7822; pp: 129-143

Wong, K.W; Wong, P.M; Gedeon, T. D. and Fung, C. C (2003): Rainfall Prediction Using Soft Computing Technique, Soft Computing Journal, vol.7, pp: 434 – 438.