

Original Research Article

Regional time series forecasting of chickpea using ARIMA and neural network models in central plains of Uttar Pradesh

ABSTRACT:

Climate and yield prediction are the most important and challenging tasks in modern agriculture during the climate change era. In general, climate and yield are highly non-linear and complicated phenomena. India is an agricultural country and most of its economy depends upon agriculture therefore early prediction of climate and yield is necessary for the planned economic growth of our country. This research identifies superior forecasting models of Autoregressive Integrated Moving Average (ARIMA) as well as Artificial Neural Network (ANN) for predicting future climate and chickpea yield. Historical data for the climate and crop were used (1996-2020) and forecasting was done for the next 5 years (2020-2025). By using, RMSE and R^2 statistical tools simultaneously, the predictive accuracy of ARIMA and ANN models was compared. By comparing the R^2 values of ARIMA (0.591) and ANN (0.96), this study revealed that ANN models can be used as more accurate forecasting tools to predict the climate as well as yield, enabling timely agricultural management.

Keywords: ARIMA, ANN, MLP, RMSE, BIC, Nodes, Hidden Layer, Learning rate.

1. INTRODUCTION:

The yield of a crop is dependent on various technological, biological, and environmental aspects. Among various environmental aspects soil fertility, topography, water quality, and climate play a major role in deciding the yield and growth of the plant. Climate changes are majorly influenced by changes in temperature and rainfall. For assessing the impact of climate on crop yield various statistical models and methods are being used such as Auto Regressive Integrated Moving Averages (ARIMA), Artificial Neural Network (ANN), Seasonal Autoregressive Integrated Moving Averages (SARIMA) for interpreting the relation between climate and crop yield. ARIMA modeling was developed by Box-Jenkins which is used for predicting future values based on past trends. It is used in analyzing and forecasting weather and climate-based parameters in meteorological studies. ARIMA makes use of lagged moving averages to smooth time series data. They are widely used in technical analysis to forecast future security prices. ANN modeling utilizes the processing of the brain for developing algorithms that will be used to model complex patterns and predict problems. It allows the non-linear complex relations between the response variable and its predictors. It also compares the efficiency of models in fitting and future prediction. In the present study, time series data on climate and yield parameters of chickpeas was analyzed and forecasted by using ARIMA and ANN models.

Chickpea (*Cicer arietinum*) is a member of the legume, pea, or pulse. It is an ancient cool season food legume crop, cultivated mainly in semiarid environments in almost all parts of the world covering Asia, Africa, Europe, Australia, North America, and South America continents

(Saxena1990). Chickpea is the second most important food legume crop after the common bean (FAOSTAT 2011). Estimating the yield parameters of chickpeas is very important, so in this connection, the present study was used for estimating the climate and yield parameters by using ARIMA and ANN. A large amount of research has been done using time series models such as Multi Linear Regression (MLR), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), and Seasonal Autoregressive Integrated Moving Average (SARIMA)(Cornillon et al., 2008; Ibrahim et al., 2010; Samsuri et al., 2017; Muhamad Safiih et al., 2017). However, the major weakness of these models is generated from a linear component which has difficulties in capturing the nonlinear component. On the other hand, Artificial Neural Network (ANN) is nonlinear in nature and is influenced by the behavior of neurons in them. It can approximate the function to a satisfying level of accuracy. Interest in the use of Artificial Neural Networks (ANNs) for developing climate change prediction models has increased in recent years due to ever-changing climate patterns in the world (Acharya et al., 2014; Belayneh et al., 2014; Hashim et al., 2017; Mishra et al., 2018). ANNs are computer systems inspired by biological neural networks to model relationships between independent and dependent variables. Hence, ARIMA and ANN are the most utilized technologies for data mining. The use of these models has the advantage of capturing patterns of data sets as well as improving prediction accuracy (Qiu & Song, 2016). Hence, we conducted our study with the following objectives:

- Forecasting of climatic parameters (maximum temperature, minimum temperature, and rainfall) using ARIMA.
- Time series (ARIMA) forecasting for a yield of chickpea in Prayagraj.
- To compare ARIMA and ANN models for chickpea yield prediction, based on climatic parameters as independent variables.

2. MATERIALS AND METHODS:

In this study, the climate prediction of the Autoregressive Integrated Moving Average (ARIMA) model for time series data and Artificial neural network (ANN) were used for forecasting algorithms based on the information in the past values of the time series to predict the future values.

2.1 Study area:

Prayagraj is a district in the central plains of southeastern Uttar Pradesh. The city is situated at the confluence of the two rivers – Ganga and Yamuna. It lies between the parallels of 24° 47' North latitude and 81° 19' East longitudes. The exact location of Prayagraj in India is shown in fig.1. Due to climate change and rapidly changing rainfall patterns, the city faces frequent floods.



Figure.1 Location of Prayagraj in India

Data collection:

The historical data for rainfall, minimum and maximum temperature were obtained from Indian Meteorological Department (IMD) for the years (1996-2020). The yield data was taken from ICRISAT.

Software used:

In this study, we used two software programs, SPSS for ARIMA and MATLAB for ANN. MATLAB is an integrated software program that numerically computes a high-level statistical language as well as provides visualization in the form of graphics and simulations which can be used for data exploration. The application of this software program can help the user develop models and algorithms in a system interface. SPSS analyses data to solve research problems through an interface that is easy to use. It has the capabilities of advanced statistical procedures. These procedures can use extensions such as R, and Python which ensure accurate data analysis and progressive decision-making.

Time series forecasting:

Time series forecasting is the technique of data science to predict based on historical data. Time series forecasting is also an important area of machine learning (ML) and can be cast as a supervised learning problem. ML methods such as Regression, Neural Networks, and Support Vector Machines, be applied to it. Forecasting involves taking models to fit historical data and using them to predict future observations.

Process of time series forecasting:

The Box-Jenkins Model forecasts data using three principles: autoregression, differencing and moving average. These three principles are known as p, d, and q, respectively. Each principle is used in the Box-Jenkins analysis; together, they are collectively shown as ARIMA (p, d, q).

- a) Ensuring stationarity (d)
- b) Identification of AR(p) and MR(q)
- c) Estimations using appropriate p, d, q values
- d) Remove seasonality
- e) Selecting the best seasonal ARIMA model

In time series analysis, to better understand the data and for future forecasting, the auto-regressive (p) integrated (d) moving average (q) (ARIMA) model was used. The basic idea of using the ARIMA model was to remove the trend of the series by differencing so that a stationary series is obtained by transforming a non-stationary series (Bahadir 2012, Wang et al. 2014; Afrifa-Yamoah 2015). This ARIMA model is based on Box-Jenkins approach. The AR part of the ARIMA model shows that the variable under concern is regressed on its prior values. The MA part of the ARIMA model shows that the regression error is a linear combination of error values occurring at various time intervals in the past. The 'I' part shows the number of times differencing has been performed. The entire objective of finding adequate AR, I, and MR terms is to make the model fit the data in the best possible way. The model assumes data to be a non-seasonal series which needs to depersonalize before modeling. A non-seasonal ARIMA model is generally denoted as ARIMA(p, d, q), where p is the lag order, d is the order of differencing, and q is the order of moving average. A seasonal ARIMA model is denoted as ARIMA(p, d, q) (P, D, Q)_m, where m is the number of periods in each season, and the uppercase P, D, Q refer to the autoregressive (AR), differencing (I), and moving average (MR) terms respectively, for the seasonal part of the ARIMA model. ARIMA methodology has its limitations of relying on past values; however, it works best for long and stable time series. It does not explain the structure of the underlying data mechanism but simply approximates the historical patterns (Bari et al. 2015; Naz 2015; YoosefDoost et al. 2017).

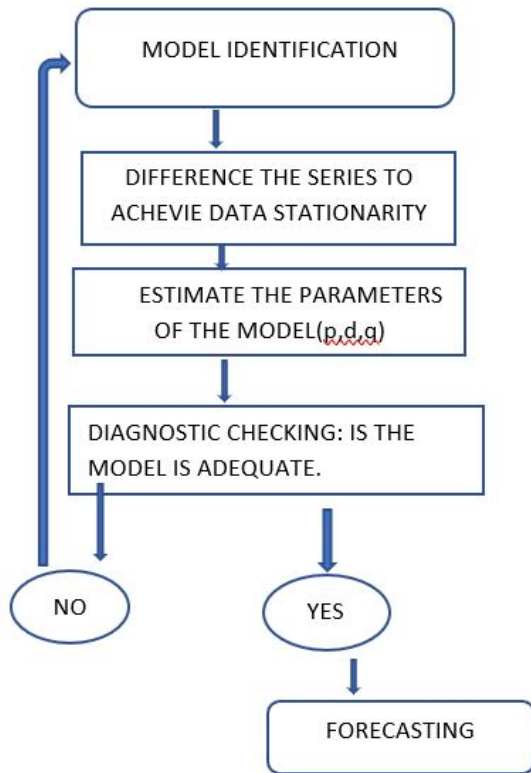


Figure2.Arima model flow chart.

ANN (Artificial Neural network)

The Artificial Neural Network (ANN) is a deep learning method that arises from the concept of the human brain's Biological Neural Networks. MATLAB(R2021a) was used to develop to predict the monthly minimum and maximum temperature and rainfall. The input variables of the algorithms were used monthly rainfall and temperature and the target variable of the algorithm was used yearly yield data.

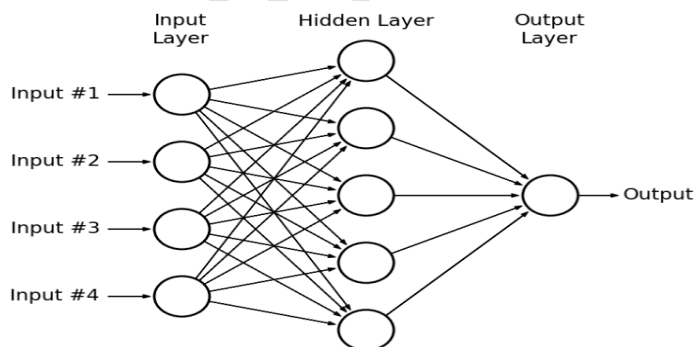


Figure.3 Artificial neural network (ANN) architecture with input, hidden layer, and output.

As shown in Fig.3, 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 (yt) and the inputs ($rt^1 ; rt^2 . . . rt^i$) can be represented by the following mathematical equation:

$$Z = W_0 + W_1X_1 + W_2X_2 + \dots + W_nX_n \quad (1)$$

The creation of the ANN predictive model (in MATLAB software) involves the following aspects:

- (i) Creating the network topology which involves the selection of the number of input neurons, the number of hidden layers, the number of hidden neurons in the hidden layer, and the number of output neurons.
- (ii) Training the network that involves selecting the network type/ training algorithm (feed-forward backpropagation algorithm in the present case. The network is devoid of links that would allow the information exiting the output node to be sent back into the network. The purpose of feedforward neural networks is to approximate functions. There is a classifier using the formula $y = f^*(x)$. This assigns the value of input x to the category y), feeding the training and target data, selecting the training function, selecting the adaptation learning function, performance function (MSE), and the transfer function. Simulation experiments were conducted on different ANN model configurations to ascertain the best-performing model. The sound level time-series data set for 25 years was divided into training data (70%), testing data (15%), and validation data (15%). Training set, Validation set, and Test set are three main aspects of ANN.
- (iii) The training set is the one that has to use for the training of the algorithm. A validation set is used to find out how accurate the Algorithm is, and to calculate the efficiency of the algorithm in terms of Root mean squared error (RMSE).

3.RESULTS AND DISCUSSION

3.1 Autoregressive Integrated Moving Average (ARIMA):

The rainfall and temperature data were taken from 1996-2020. The Rainfall and temperature (maximum and minimum) data were separately analyzed using different ARIMA models in SPSS statistical package. Then we start with the initial preprocessing of the data to make it stationary, and then we choose possible values of p and q which we can of course adjust as model fitting progresses. And ACF and PACF graphs are shown in Figure 4, 5, 6.

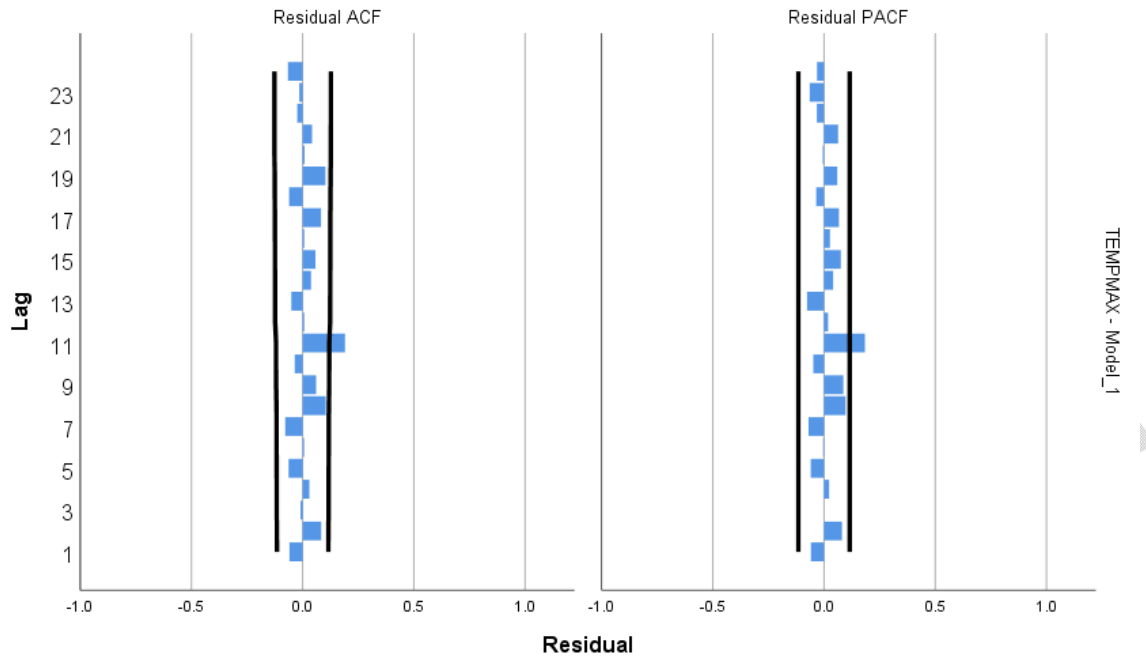


Figure 4: ACF and PACF values of maximum temperature.

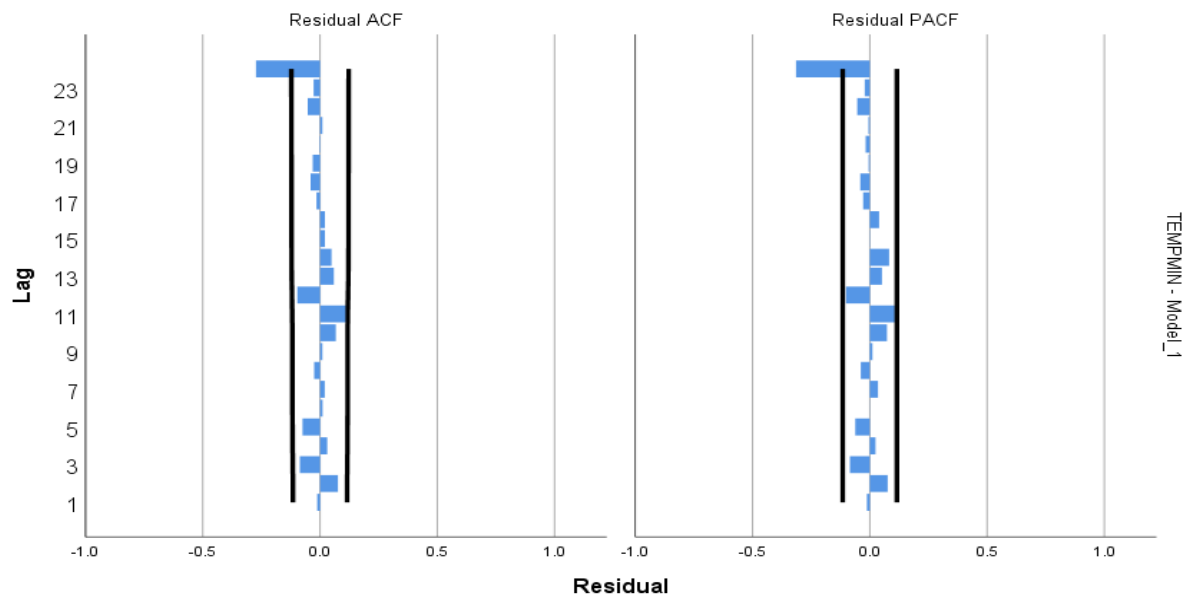


Figure 5: ACF and PACF values of minimum temperature.

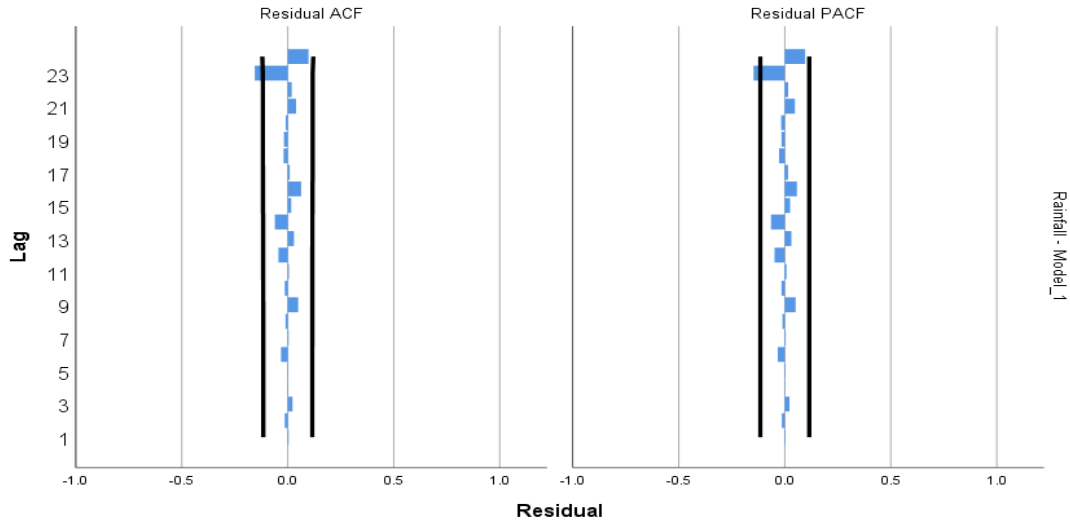


Figure 6. ACF and PACF values of Rainfall.

Model Identification

We experimented with different parameters of autoregressive (p) and moving average (q) to determine the best model that will give the best forecast as indicated in Table 1.

Table.1 Best fitted models of ARIMA.

ARIMA	BEST FITTED MODELS
Rainfall	(0,0,0) (1,1,1)
Maximum Temperature	(1,0,0) (0,1,1)
Minimum Temperature	(1,0,1) (1,1,0)

After determining the three parameters p , d , and q evaluating the model with fit statistics are required to quantify the performance of the forecast within its acceptable limits. Some of the statistical measures are RMSE (Root Mean Square Error), MAPE (Mean Absolute percentage errors), and MAE (Mean Absolute Error). The values of these errors should be minimum for better performance of the model.

Model validation and Forecasting

To test the adequacy and predictive ability of the chosen models, the actual data sets, and predicted values were plotted and displayed in figure 7 below. The graphs show that the predicted values are well-fitted through the original data. There is a slight over-prediction in March 2000, May 2004, and November 2016 as compared with the original data. The rest of the predicted values are well-fitted through the original data with the lower and upper limits containing the majority of the original data. This indicates that the models were chosen for rainfall best-fitted ones for the data sets. The ARIMA model is used to forecast rainfall in Prayagraj's upcoming 5 years (2021-2025).

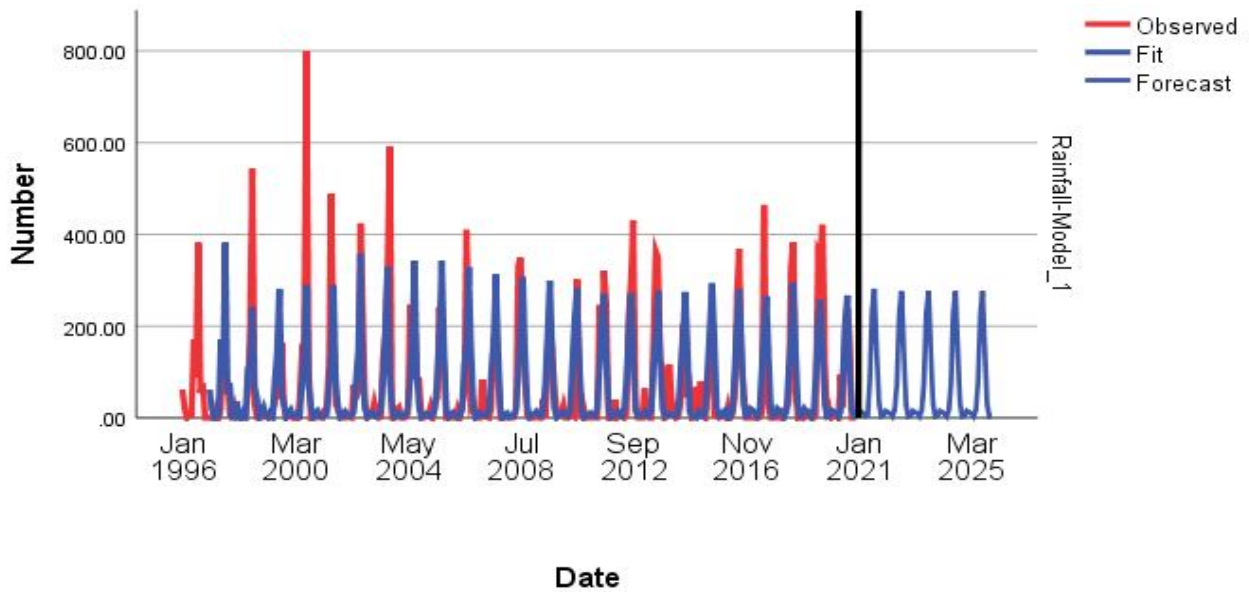


Figure 7. Observed and fitted values of rainfall series.

To test the adequacy and predictive ability of the chosen models, the actual data sets, predicted values, and fit values were plotted and displayed in figure 8. And the predicted values are well-fitted through the original data with the lower and upper limits containing the majority of the original data. This indicates that the models chosen for the maximum temperature series are the best-fitted ones for the Prayagraj. The ARIMA model is used to forecast rainfall in Prayagraj for the upcoming 5 years (2021-2025).

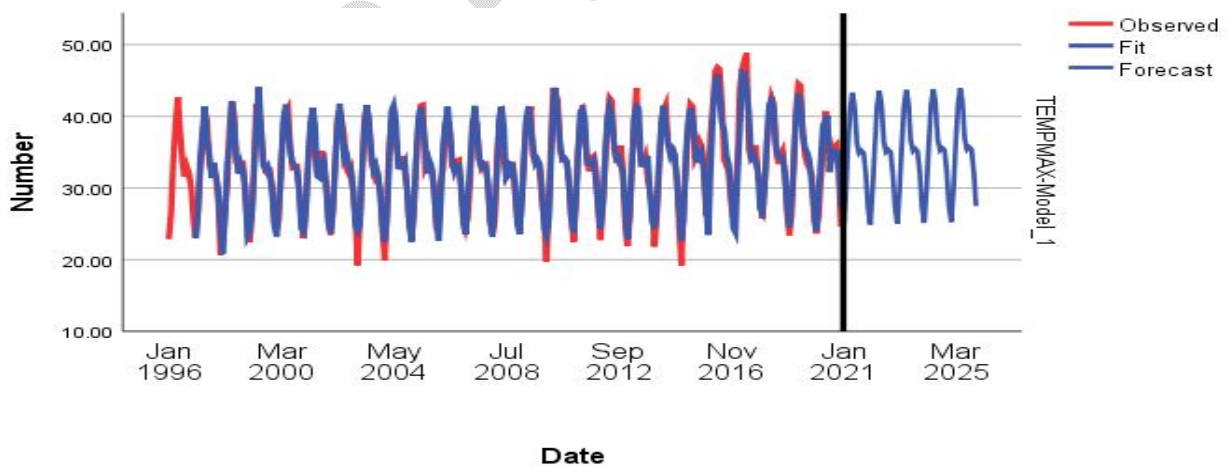


Figure 8. Observed and fitted values of maximum temperature series.

To test the adequacy and predictive ability of the chosen models, the actual data sets, predicted values, and fit values were plotted and displayed in figure 9. Slight over can be seen in January 2017 as with the original data. Rests of the predicted values are well-fitted through the original data with the upper limits containing majorities of the original data. This indicates that the

models chosen for the minimum temperature series are the best-fitted ones for Prayagraj. The ARIMA model is used to forecast the minimum temperature in Prayagraj for the upcoming 5 years (2021-2025).

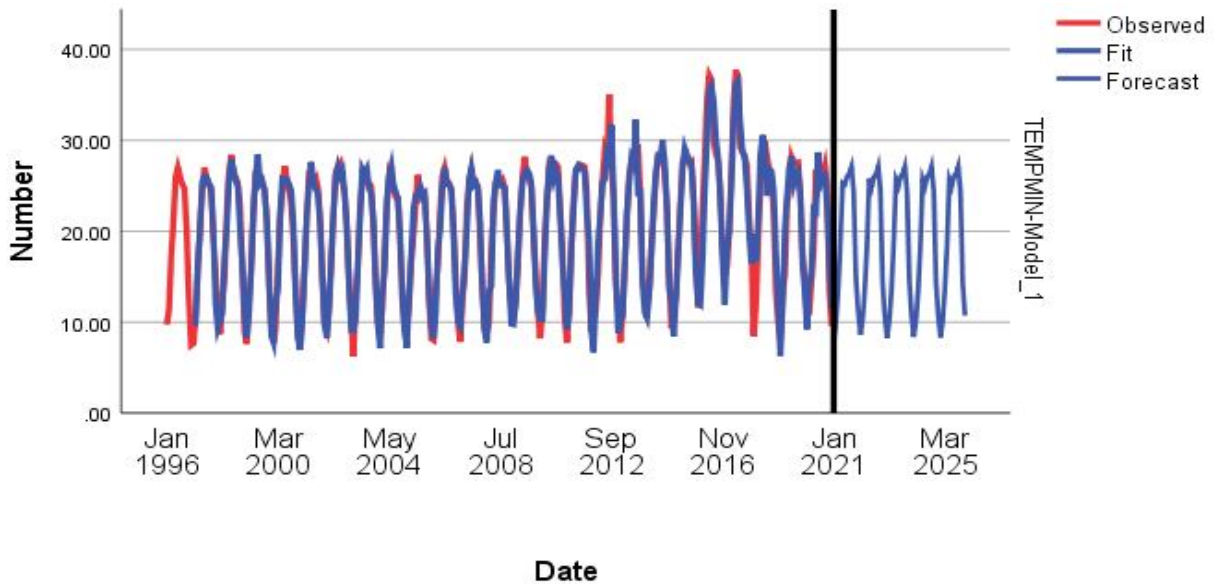


Figure9. observed and fitted values of minimum temperature series.

Diagnostic analysis

After determining the three parameters p , d , and q evaluating the model with fit statistics are required to quantify the performance of the forecast within its acceptable limits. Some of the statistical measures are RMSE (Root Mean Square Error), MAPE (Mean Absolute percentage errors), and MAE (Mean Absolute Error). The values of these errors should be minimum for better performance of the model.

Table2. Forecasting accuracy statistics.

Parameters	Rainfall model Arima (0,0,0)(1,1,1)	Arima model Maximum temperature (1,0,0) (0,1,1)	Arima model Minimum temperature (1,0,1) (1,1,0)
StationaryR-squared	0.501	0.551	0.464
R- squared	0.497	0.908	0.927
RMSE	85.508	1.812	1.929
MAPE	337.466704	4.134	8.069
MAE	44.989	1.324	1.314

3.2 Forecasting of the annual yield of chickpea by using the Arima model

Arima model was used to forecast the yield of chickpeas. The data used for model building is from the year 1996-2020. The data from 2021-2025 is used for cross-validation of the selected model and the forecasting is done by the years 2021-2025 by using the selected best fit model. The yield data were separately analyzed using different ARIMA models in SPSS statistical package when the time series is not stationary. To make the data stationary, differencing was done. After the time series has been rationalized by differences, the next step is to determine whether any AR or MA terms are required to correct the autocorrelation that remains in the differenced series to fit an ARIMA model. The ACF and PACF plot suggest the tentative values of q and p that would be suitable for a yield of chickpea are $q=0$ and $p=0$. Thus the Arima model that was found to be best fitted for a yield of chickpea is $(0,1,0)$.

After determining the three parameters p, d, and q evaluating the model with fit statistics are required to quantify the performance of the forecast within its acceptable limits.

RESIDUAL ACF AND RESIDUAL PACF

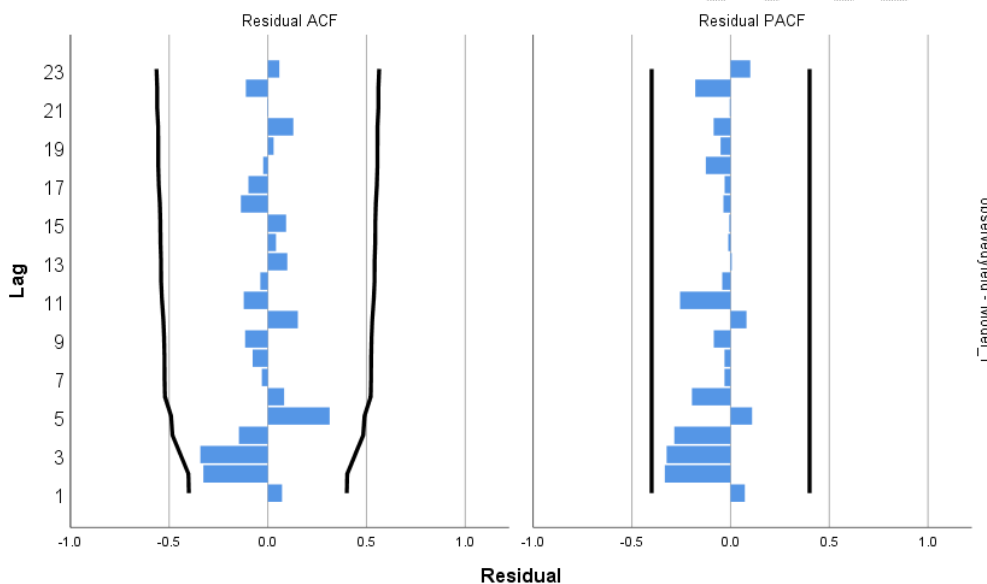


Figure 10. RESIDUAL ACF and PACF of yield.

The above Figure 10 are showing the residual ACF and PACF of yield. From the time plot of the residuals against time, we can see that there is no obvious pattern in the plot except for a possible outlier, and it looks like an independently and identically distributed sequence of mean zero with a constant variance. The plots of the ACF and PACF (yield) of the residuals lack enough evidence of significant spikes which clearly shows that the residuals are white noise. The results showed that the residuals are non-significant with the Box-Ljung test. From the above tests, it is clear that the fitted model is adequate since the residuals are white noise. That is, ARIMA $(0, 1, 0)$ is adequate for modeling the log-transformed yield data in PRAYAGRAJ.

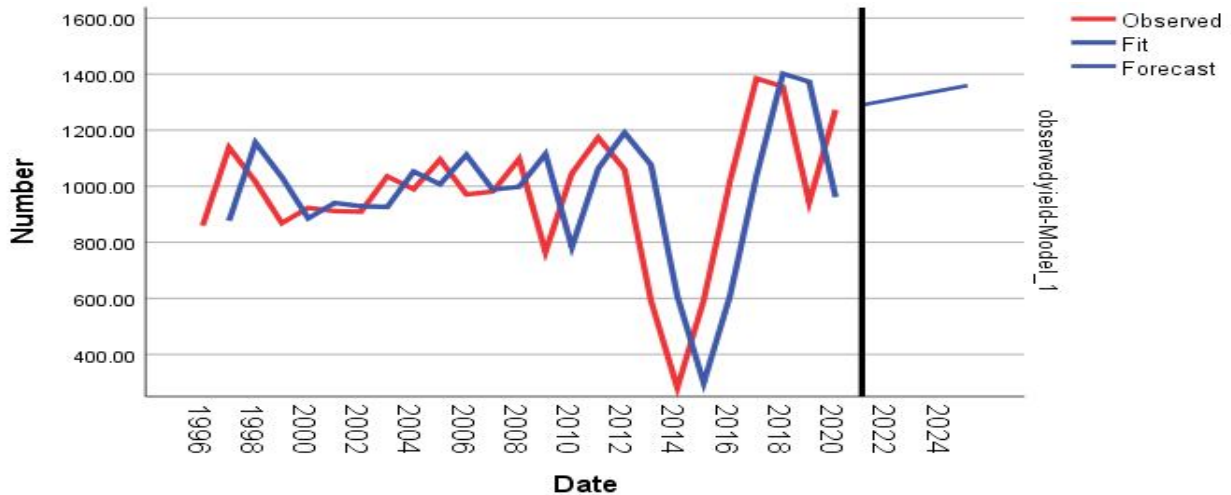


Figure11. Observed and fitted values of yield data.

To test the adequacy and predictive ability of the chosen models, the actual data sets, predicted values, and fit values were plotted and displayed in figure 11. Slight over can be seen in January 2014 as with original data. Rests of the predicted values are well-fitted through the original data with the upper limits containing majorities of the original data. This indicates that the models chosen for yield data are the best-fitted ones for Prayagraj. The ARIMA model is used to forecast yield in Prayagraj for the upcoming 5 years (2021-2025).

Table 3. FORECASTING OF YIELD PARAMETERS

Parameters	Forecasting of Yield Arima model (0,1,0)
Stationary R-squared	-2.220E-16
R- squared	-.025
RMSE	246.079
MAPE	24.748
MAE	194.370

3.3 COMPARISON OF ARIMA AND ANN

ARIMA (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE):

Based on data yearly production and monthly climate parameters were calculated for measuring the quantitative relationship between these variables. Here we take independent variables such as temperature (maximum, minimum), rainfall, and dependent variable as a yield. And the best fitted ARIMA model is (0,0,1) among all other ARIMA models. This has the R-Square value which is

0.591 with an RMSE value of 159.632 which is the lowest among all other models. Among all models, ARIMA(0,0,1) is the best-fitted model for forecasting production with the highest R – Square value of 0.59.

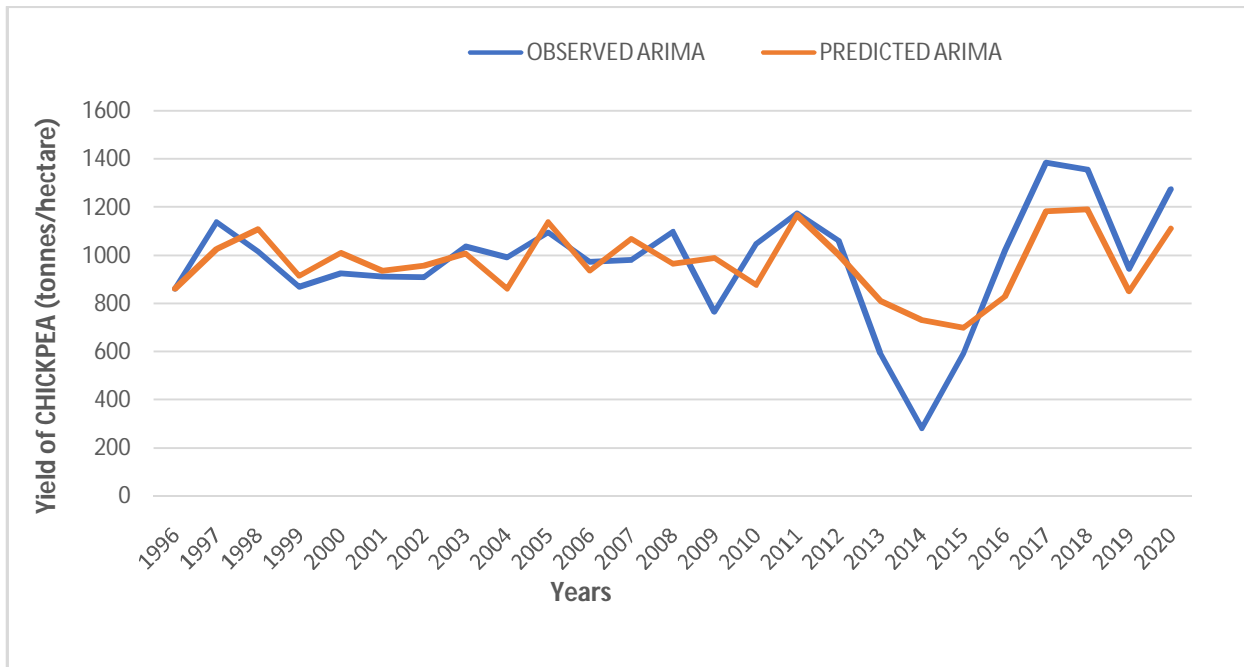


Figure12. Observed and fitted values of ARIMA.

ANN (Artificial Neural Network):

In Prayagraj, timely sowing of chickpea crops was done in September month, hence monthly climate data (maximum temperature, minimum temperature, and rainfall) from September-November was considered for the study. And yearly wise 25 years(1996-2020) data was taken. Here independent variable (input) was taken as (maximum temperature, minimum temperature, rainfall), and the dependent variable(target) was taken as yield.

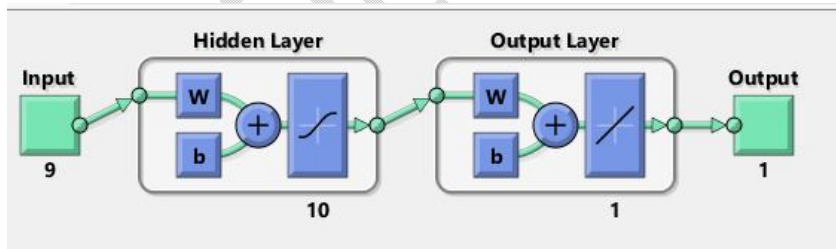


Figure 13. Observed and fitted values of ANN.

The ANN was trained with 70% of target data and 15% of data were used to validate and 15% of data was used for testing. In addition, 10 hidden neurons and 2 delays were used in the network. The details are shown in Fig.10. Performance of each training algorithm in predicting the atmospheric temperatures was evaluated using the Mean Squared Error (MSE) and the correlation coefficient (R).

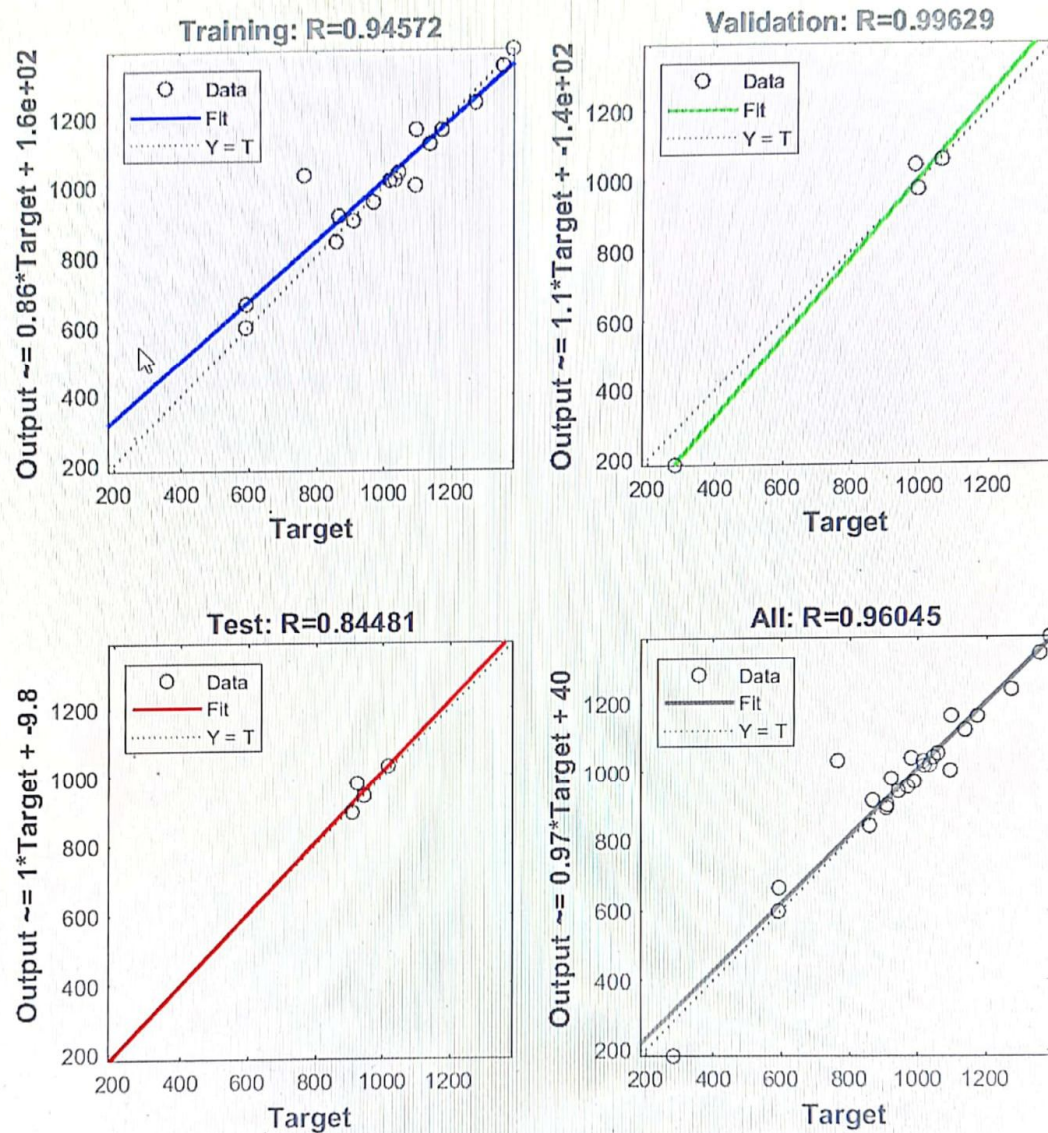


Figure 14. Neural network training regression graph.

On the same data set, the ANN model was used. After a lot of training the ANN best-fit Model is chosen. The best fit model of the ANN is with the R - Square value of 0.960 and RMSE value of 66.716.

3.3 COMPARISON OF ARIMA AND ANN

For comparison purposes, the training and testing performance of the ANN model were compared with the ARIMA model. The ARIMA and ANN forecasts are close to actual values. It shows that both approaches work well for chickpea forecasting. The comparison of training and testing precision among the two approaches is based on RMSE, MAPE, and MSE statistical measures.

Empirical results on the chickpea forecasting data set using two different models clearly reveal the efficiency of the ANN model. It shows ANN models are the best fit when compared to the ARIMA model.

The reason could be the nonlinear machine learning techniques can capture the heterogeneous trend (dissimilar) in the data set and perform well as compared to the ARIMA model.

Based on the results obtained in this work one can conclude that ARIMA models are not always adequate for the time series that contains non-linear structures. In this context, nonlinear artificial intelligence technologies like neural networks can be an effective way to improve forecasting performance. It is seen from the comparison of the R – Square value and RMSE of the models. In the study, it is clear that ANN is the best fit model for forecasting the impact of climate change in the Prayagraj district. ANN models always have the highest R-Square value and low RMSE than all other models so ANN is the best model to fit. From the graph, it could be observed that the yield parameter was more accurate with the ARIMA model when compared with the ANN model.

Table 4. Forecasting parameters for ARIMA and ANN.

Model	RMSE	R ² value
ARIMA	159.632	0.591
ANN	66.716	0.96

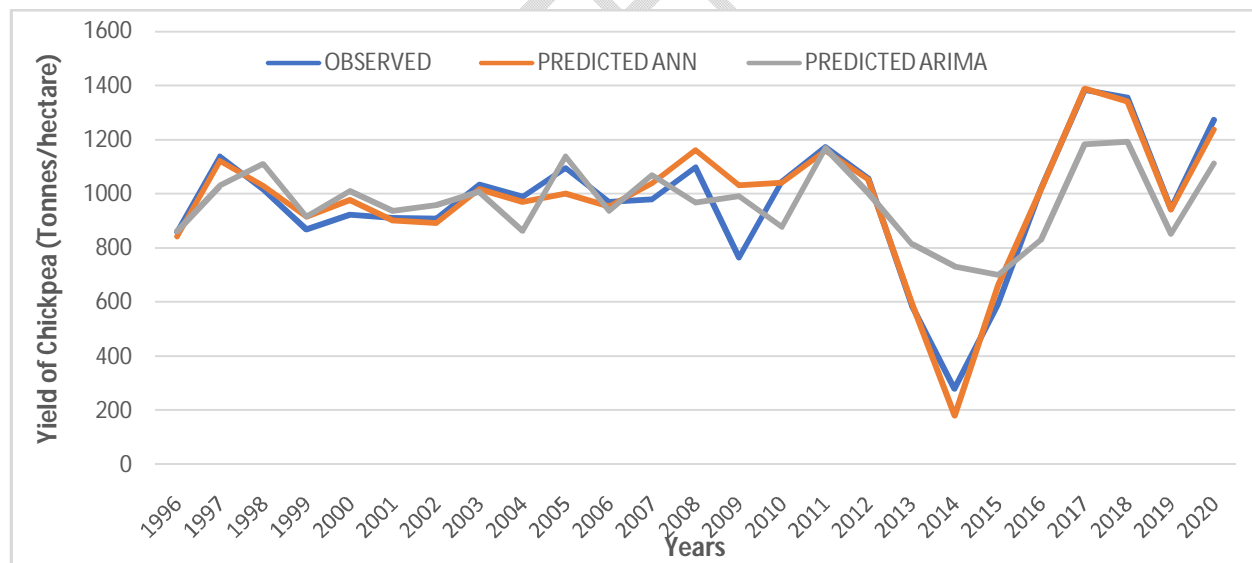


Figure 15. Observed and forecast values of Chickpea yield by ARIMA and ANN.

4. CONCLUSION:

Rainfall and temperature are the main factors governing the dynamic structure of climate resulting in climate change. In the present study, rainfall and temperature data time series were

studied and the best fitted ARIMA model was found after the removal of seasonality, and forecasting was done using the same model. The forecast results for rainfall were found to be overpredicting the values for extreme rainfall events, while it matches in the case of other rainfall events. However, the forecast results for temperature (minimum and maximum) are matching well and are showing an increasing trend. And we also used annual yield data to predict the future yield. The ARIMA and ANN forecasts are close to the actual value. It shows that both of the approaches work well for chickpea forecasting.

Based on the results obtained in this work one can conclude that ARIMA models are not always adequate for the time series that contains non-linear structures. In this context, nonlinear artificial intelligence technologies like neural networks can be an effective way to improve forecasting performance. It is seen from the comparison of the R – Square value and RMSE of the models. In the study, it is clear that ANN is the best fit model for forecasting the impact of climate change in the Prayagraj district. ANN models always have the highest R-Square value and low RMSE than all other models so ANN is the best model to fit. From the graph, it could be observed that the yield parameter was more accurate with the ARIMA model when compared with the ANN model.

REFERENCE

- **Acharya N (2014)** Development of an artificial neural network based multimodel ensemble to estimate the northeast monsoon rainfall over south peninsular India: An application of extreme learning machine. *Clim. Dynam.*, 43, 5-6, 303-310, doi: 10.1007/s00382-013-1942-2.
- **Belayneh A (2014)** Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression. 508, 418-429, doi: 10.1016/j.jhydrol.2013.10.052.
- **Box, G.E.P and Jenkins, G.M (1976)** Time Series Analysis: Forecasting and Control. Revised Edition, Holden-Day, San Francisco.
- **Cornillon P (2008)** Forecasting time series using principal component analysis with respect to instrumental variables. *Computational Statistics and Data Analysis*, 52: 1269– 1280.
- **Hashim, F. R (2017)** Prediction of rainfall based on weather parameter using artificial neural network, *J. Fundam. Appl. Sci.*, 9, 3S, 493-502.
- **Haviluddin (2015)** Rainfall monthly prediction based on artificial neural network: A case study in tenggarong station east kalimantan-indonesia. *Procedia Computer Science*, 59, 142-151.
- **Ibrahim M.Z (2010)** Time-series Analysis of Pollutants in East Coast Peninsular Malaysia. *Journal of Sustainability Science and Management*, 5(1): 57-65.
- **Mishra N (2018)** Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data 16-23, doi: 10.5815/ijisa.2018.01.03.
- **Moghims (2017)** Bias Correction of Climate Modeled Temperature and Precipitation Using Artificial Neural Networks, *J. Hydrometeorology.*, 18, 1867-1884, doi:10.1175/JHM-D-16-0247.1.

- **Mohapatra(2017)** Rainfall prediction based on 100 years of meteorological data, Computing and Communication Technologies for Smart Nation (IC3TSN), 2017 International Conference on, 162-166, IEEE.
- **Qiu M&SongY(2016)**Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model. PLoS ONE, 11(3): 1-11.
- **Sachan and Abhishek (2014)**Forecasting of rainfall using ANN, GPS and meteorological data, Convergence of Technology (I2CT), 2014 International Conference for 1-4IEEE.
- **SamsuriA (2017)** Multiple Linear Regression (MLR) models for long term Pm10 concentration forecasting during different monsoon seasons. Journal of Sustainability Science and Management, 12(1): 60-69.
- **Saxena MC (1990)** Problems and potential of chickpea production in nineties. In: Chickpea in the nineties: proceedings of the second international workshop on chickpea improvement, 4–8 Dec 1989, ICRISAT Center Patancheru India.
- **Varshney RK (2005)**Genic microsatellite markers in plants features and applications Trends Biotechnology 23:48–55.

UNDER PEER REVIEW