Forecasting Rainfall using SARIMA comparing with LSTM and Stacked LSTM

1. Dr. S. Selva Arul Pandian, Assistant Professor, Department of Statistics, Loyola College
2. Joeal Astile L, M.sc Student, Department of Statistics, Loyola College, Chennai
3. Nishanthini S, M.sc Student, Department of Statistics, Loyola College, Chennai

Abstract

India is an agricultural country mainly reckon on well-timed rainfall. It makes impact not only on agricultural production as well as agricultural product's prices and normal routine life of human being. The rainfall is unmanageable by farmers and which is also naturally fluctuating everywhere. Studying the significance of rainfall and its forecast using yearly rainfall data is ready lend a hand for formers in great anticipate of agricultural production and prices. This study has attempted to relate regression approach to recognize the best model to fit the yearly rainfall data in Tamil Nadu. Finely different techniques of regression which are SARIMA, supervised machine learning algorithm, Artificial Neural Network (ANN) specially LSTM and Stacked LSTM are compared to recognize the more suitable in forecasting. To foresee the most suited model which has lower of MAE, MSE and RMSE might be treated as the diagnostics checking parameters. The study tries to identify the most suitable model for rainfall prediction in Tamil Nadu to manage water successfully for agricultural usages.

Key words - SARIMA, LSTM, Stacked LSTM, Forecasting, machine learning, linear regression

1. Introduction and Historical Background

Agriculture- the backbone of India. Nearly 53% of Indian agriculture irrigation system depends on rainfall. So, it becomes mandatory to forecast the pattern of rainfall and to harvest it. Rainfall prediction is important not only just for irrigation, since rainwater is the purest form of water and root cause for all available fresh water in the world. Since the pattern of rainfall data is a time series data with seasonal changes Seasonal Autoregressive Integrated Moving Average (SARIMA) suits the best. ARIMA model uses the past information to predict the future information whereas SARIMA uses the past information same as ARIMA but also consider the seasonality patterns. Though the main purpose of ARIMA model is to predict the upcoming

conditions or satiations of time series dataset, it has a limitation that it cannot perform best with dataset which holds seasonality with it.

Long short-term memory network (LSTM) is recurrent neural network (RNN) which is used to treat the sequential data since, these networks can learn long-term dependencies between time series dataset. LSTM neural networks overcome the issue of Vanishing Gradient in RNNs by adding a special memory cell that can store information for long periods of time. LSTM uses gates to define which output should be used or forgotten. It uses three gates: Input gate, Output gate and Forget gate. The Input gate controls what all data should be kept in memory. The Output gate controls the data given to the next layer and the forget gate controls when to dump/forget the data not required

The stacked LSTM is an extension model that consist of many layers of LSTM where every layer has multiple memory cell. The stacked LSTM, also known as deep LSTM, was first formulated by A. Graves and was applied to speech recognition problems. Similar to the framework underpinning the DRNN model, the stacked LSTM model uses multiple LSTM layers that are stacked before the forwarding to a dropout layer and output layer at the final output. In a stacked LSTM, the first LSTM layer produces sequence vectors used as the input of the subsequent LSTM layer. Moreover, the LSTM layer receives feedback from its previous time step, thus allowing for the capturing of data patterns. The dropout layer also excludes 10% of the neurons to avoid over fitting.

India's monthly rainfall data from 1951 to 2014 has been examined and forecast by Subbaiah Naidu K.CH.V (2016). According to this study, SARIMA $(0,0,0)x(0,1,1)_4$ is adequate, using the Seasonal ARIMA (SARIMA) approach. The findings of this study can be applied in developing appropriate drinking water and agricultural sector planning strategies.

Time series rainfall data in Maiduguri, Northeastern Nigeria, from 1981 to 2011 were examined by Emmanuel Sambo Uba and H.R. Bakari (2015). The Autoregressive Integrated Moving Average (ARIMA) approach was used to estimate monthly rainfall, and it was discovered that ARIMA(1,1,0) provides a decent fit and is suitable for short-term forecasts. Priorities for managing water demand and agriculture are set using the findings.

Convolutional LSTM model for sentiment analysis in social big data, Ranjan Kumar Behera,Department of Computer Engineering, Atilim University, Ankara Turkey, dealt with the data of consumer reviews posted on social media is found to be essential for several business applications. To handle the big data they have used Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) (RNN with memory). The paper resulted machine learning approaches in terms of accuracy and other parameters.

Yong Yu,Department of Automation, Xi'an Institute of High-Technology, China . Made a review on Recurrent Neural Networks: LSTM Cells and Network Architectures. The long term dependancies in the data can be easily handled by using LSTM model. Moreover LSTM has become the Main focus of deep learning. The LSTM networks are divided into two broad categories: LSTM-dominated networks and integrated LSTM networks.

Dastan Hussen Maulud, Ministry of Higher Education and Scientific Research
Made a Review on Linear Regression Comprehensive in Machine Learning,December 2020,Journal of Applied Science and Technology Trends, Linear regression is one of the most commonly used comprehensive statistic and machine learning algorithms. Linear regression can optimizes to best predictions and precision. And it can determine a models efficiency by analysing the correlation of the actual values to the axplanatory variables.

In a recent study by TriptiDimri, Shamshad Ahmad, and Mohammad Sharif (2020), the climate variables of the Bhagirathi river basin, which is located in the Indian state of Uttarakhand, were examined across time periods from 1901 to 2000. For temperature and precipitation forecasts, they used the Seasonal ARIMA (SARIMA) approach and established that $SARIMA(0,1,1)x(0,1,1)_{12}$ is the appropriate one to model precipitation data while $SARIMA(0,1,0)x(0,1,1)_{12}$ is the suitable one to model temperature data. The outcomes of the SARIMA modelling has good scope of application in the development of policies for effective prediction of floods, urban planning, and environmental planning.

2. Description of Data for the Current Study

The rainfall data used in this study was obtained from the Open Government Data (OGD) Platform India. It is a univariate time series data that provides statistics on India's yearly and monthly rainfall from 1901 to 2015. This analysis uses Python-written software and specifically considers the data from Tamil Nadu.

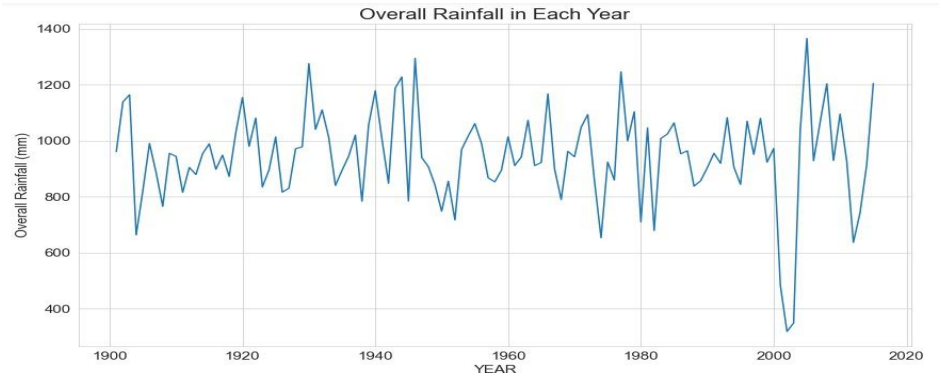


Fig 2.1

The basic descriptive statistics of the data are provided in Table 2.1 below:

No. of observations	1380	0.25	27
Mean	78.643986	0.5	59.2
Standard Deviation	70.412739	0.75	113.65
Minimum	0	Maximum	436.1

Table 2.1

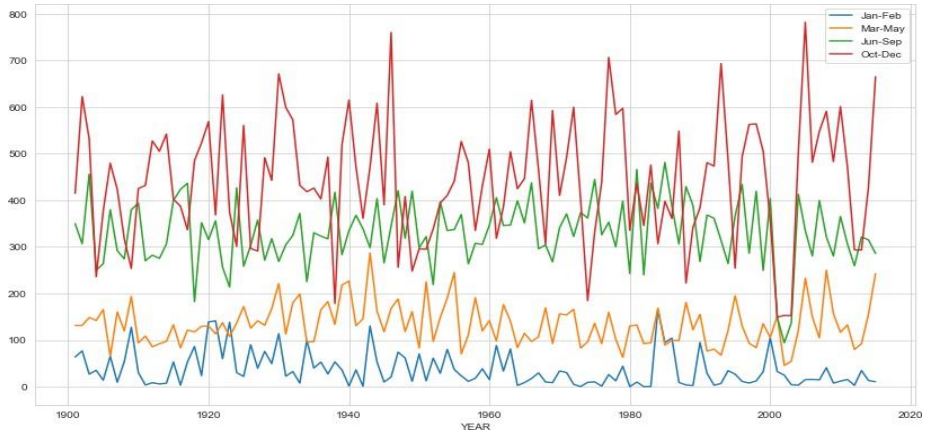


Fig 2.2

Fig2.1 a graph of Tamil Nadu's annual rainfall totals from 1901 to 2015

Fig 2.2 shows the total seasonal precipitation that fell throughout the four seasons of January through February, March through May, June through September, and October through December from 1901 to 2015.

3. Modeling Methodologies

3.1 Seasonal Auto-Regressive Integrated Moving Average (SARIMA) Model:

To produce the seasonal auto-regressive integrated moving average model, ARIMA is expanded with a seasonal component. A seasonal ARIMA model is produced by adding seasonal components to the ARIMA. But doing so necessitates seasonal backshifts. The terms that make up the model's seasonal element are very comparable to those that make up its non-seasonal parts. A typical form of the model is SARIMA (p, d, q) x (P, D, Q)s, where p, d, and q are the three ARIMA model parameters and P, D, and Q are the orders for seasonal auto regression, seasonal difference order, and seasonal moving average, respectively. The number of time steps in a seasonal cycle is shown here by the symbol.

3.2 Linear Regression

It is a method for showing how one or more independent variables are related to a dependent variable. A time-series dataset with associated residuals will prevent linear regression from correctly identifying the trend. This can be prevented by utilizing machine learning techniques to build a linear regression model using lag values (time series regression analysis). The supervised machine learning technique can be used to solve a time series problem by taking measurement lags into account as inputs

3.2 LONG SHORT-TERM MEMORY NETWORKS

By including a unique memory cell that can store data for extended periods of time, LSTM neural networks get around the problem of Vanishing Gradient in RNNs. LSTM uses gates to define which output should be used or forgotten. It makes use of the input, output, and forget gates. The input gate regulates what data is stored in memory. The forget gate regulates when to dump/forget the data that is not needed, while the output gate regulates the data that is passed to the next layer.

3.4 STACKED LONG SHORT TERM MEMORY NETWORKS

An LSTM model that consists of several LSTM layers is known as a stacked LSTM architecture. Graves developed the stacked LSTM, also known as deep LSTM, and used it to solve speech recognition issues. The stacked LSTM model uses multiple LSTM layers that are stacked prior to forwarding to a dropout layer and output layer at the final output, which is similar to the architecture supporting the DRNN model. Sequence vectors generated by the first LSTM layer of a stacked LSTM feed into the second layer. Additionally, the LSTM layer gets input from its prior time step, enabling the detection of data patterns. In order to prevent over fitting, the dropout layer additionally rejects 10% of the neurons.

4. Model identification and construction

4.1 SARIMA model construction

An essential phase in this research study is identifying and developing a model based on the facts at hand. Therefore, the purpose of the study is to determine the values of the parameters p, d, q, P, D, and Q. Due to the use of monthly data in this study, s is set to be equal to 12, indicating there are 12 time steps in a yearly seasonal cycle.

The research study first creates a time series plot of the entire data and thoroughly examines it. The decomposition plot provides the study with a clear image of the existence of residuals, seasonality, and trends in the series. The Augmented Dickey-Fuller test is then used to check for stationarity. The values of the parameters are mostly determined by the autocorrelation function (ACF), partial autocorrelation function (PACF), and inverse autocorrelation function (IACF). The ACF shows how closely the current data points are related to the previous data points and aids in calculating the value of q. The PACF is used to calculate the value of p and provides the partial correlation between the series and its own lagged values controlling for the effect of the intervening lags. The IACF assists in identifying a parsimonious model and provides rough initial estimates of the parameters. The values of the autoregression order p, the moving average order q, the differencing order d for the trend component of the time series, as well as the related seasonal parameters P, D, and Q, are then determined.

The study creates two loops, the first for the three ARIMA parameters and the second for the four seasonal component parameters, arriving at a set of parametric combinations in the range of 0 to 2. By applying the minimal value of the Akaike Information Criterion, it selects the parameter values. Here, the research investigation employs a fitting process to ascertain the regression model's coefficients. The model is constructed after the parametric values have been determined.In this stage the suitability of model and the model’s assumptions are examined whether they are true by comparing different models and then plots the residuals for further analysis. And then it will show whether the residuals follow normality or not. Q-Q plot is used to check the fitting level of model and check for normality

Fitting and Prediction

As a next step to model identification and parameter estimation future prediction is done.

ADF Statistic	-5.661516
p-value	0.000001
Critical Values	
1%	-3.436
5%	-2.864
10%	-2.568

Table 4.1

It is evident from Table 4.1 that the p-value is below the threshold value, which suggests that the data is stationary. And no differencing is necessary for the data.

	Co-efficient	Std. Error	Z	P>|z|	0.025	0.975
ar.L1	0.905	0.088	10.251	0	0.732	1.078
ma.L1	-0.8819	0.097	-9.079	0	-1.072	-0.692
ma.S.L12	-1	10.789	-0.093	0.926	-22.145	20.145
sigma2	2049.092	2.21e+04	0.093	0.926	-4.13e+04	4.54e+04

Table 4.2

Data for the study is divided into a train and test set. Using the provided parameters, a seasonal ARIMA model is constructed on the train set. SARIMA (1, 0, 1) x (0, 1, 1)12 is found to be the suitable model from the model summary statistics given in Table 4.2.

The following Figure 4.4 depicts the future forecasts for the following months starting from year 1901 to 2015.

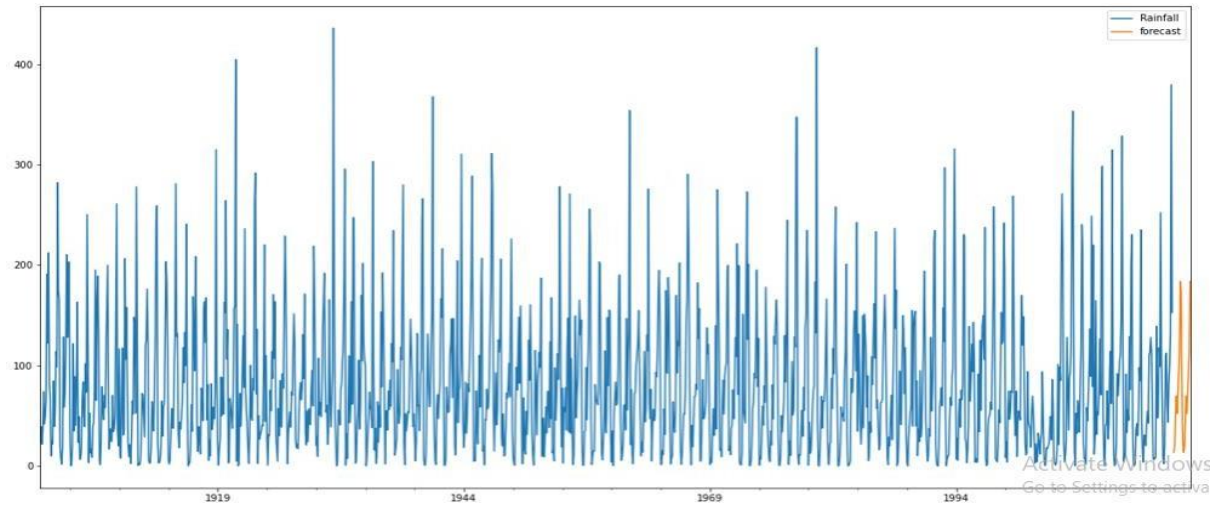


Fig 4.1

The mean absolute percentage error is 3.12. The forecast is 3.12% off, while the model is 96.88% correct. The Root Mean Square Error (RMSE) of the SARIMA model fitted is found to be 33.29.

4.2 Long Short-Term Memory networks (LSTM)

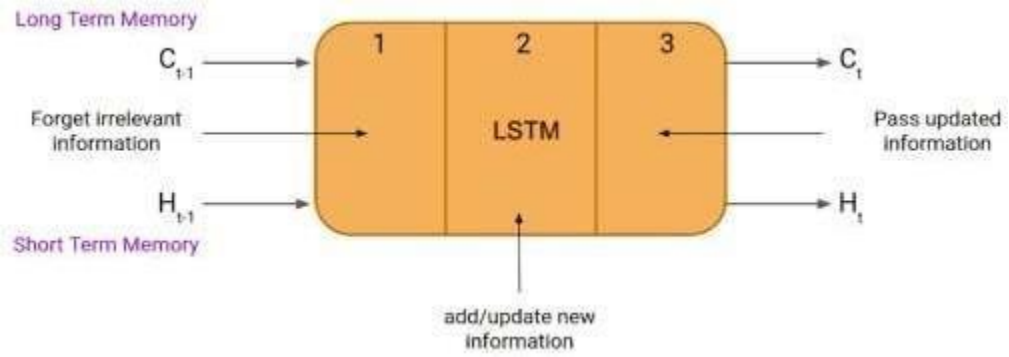
LSTM is nothing but a high level RNN cell. LSTM has three parts which are capable of performing an individual function.

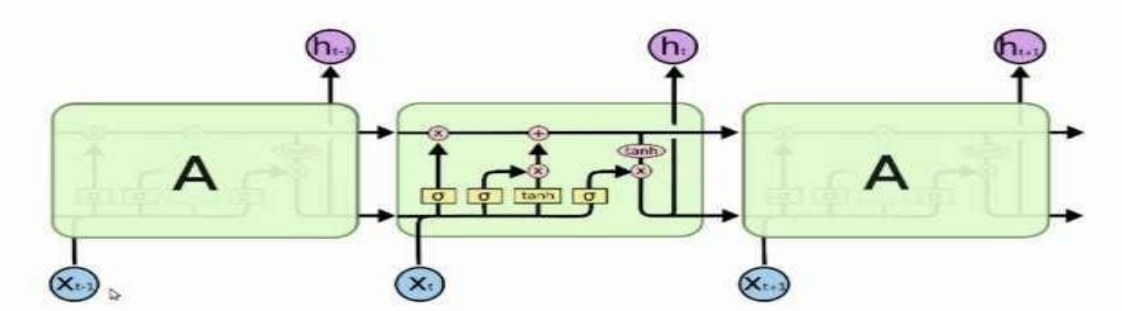
1) **First part(Forgot gate):** it will select the previous timestamp information is to be remembered or to be forgotten since it is irrelevant.

2) **Second part(Input gate):** it will get more information from the input cell.

3) **Third part(Output gate):** the collected information the present timestamp is then passed to the next the timestamp by the cell.

The hidden states are termed as short term memory and the cell states are termed as long state memory.





4.3 Stacked Long Short-Term Memory networks (stacked LSTM)

The stacked LSTM is an extension model that consist of many layers of LSTM where every layer has multiple memory cell. It actually makes the model deeper and more accurate.

The Perceptron Neural Network(PNN) is made deeper by adding multiple hidden layers with it. Additional hidden layers can be added to a Multilayer Perceptron neural network to make it deeper. The added hidden layers are used in recombining the prior layers and create new layers at high levels. Increasing the depth of the network provides an alternate solution that requires fewer neurons and trains faster. Ultimately, adding depth it is a type of representational optimization.

4.4 Linear Regression

The essential step is to rebuild the time series dataset as a supervised learning problem by predicting the value of the study variable at the previous time step. The research study takes into account 12 time steps, thus the input component will consist of 12 columns of earlier data, and the output component will consist of the rainfall for the next month. A model for making twelve-step predictions is being developed in this study. The procedure involves performing a shift down operation to generate the 12 input columns, maintaining the order of the observations while performing the shift operation. In essence, the study predicts the following time step by using the previous time steps. After dividing the data into train set (from 1901 to 2015) and test set (from 1951 to 2000), a regression model is built to make predictions. To assess the model's accuracy, the predictions are obtained from the model and the root mean square error is computed.

Using a supervised machine learning technique, the linear regression operator predicts the rainfall and the actual and predicted rainfall are depicted in Fig. 4.5. The developed linear regression model has a 57.48 RMSE (Root Mean Square Error).

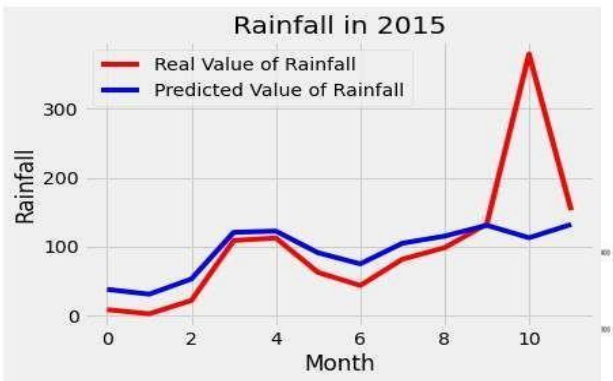


Fig-4.1

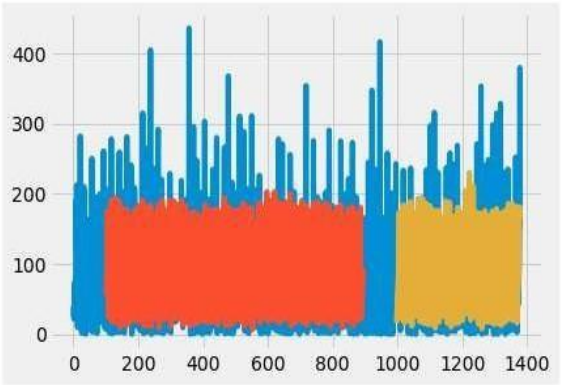


Fig-4.2

Fig-4.3 **Fig-4.4**

Fig-4.1 shows the predicted output by LSTM.

The Root Mean Square Error (RMSE) value of the LSTM model built is found to be 59.62.

Fig-4.2 shows the predicted output by Stacked LSTM.

The Root Mean Square Error (RMSE) value of the Stacked LSTM model built is 61.52.

Fig-4.3 shows the predicted output by Linear regression.

The Root Mean Square Error (RMSE) value of the Linear Regression model built is 57.48.

Fig-4.4 shows the predicted output by SARIMA model.

The Root Mean Square Error (RMSE) value of the SARIMA model fitted is found to be 33.29.

5. Conclusion

By considering the diagrams above (Fig-4.1, Fig-4.2, Fig-4.3,Fig-4.4) we can infer that the SARIMA model can give a clear interpretation.

The SARIMA (1, 0, 1) x $(0, 1, 1)_{12}$ model is constructed using the time series of seasonal rainfall data, and the model fits well. The graph shows that, with the exception of a few high values, the prediction was correct.

Linear regression is used to create a Supervised Machine Learning model, and the plot shows that, with the exception of a few values, the predictions are fairly accurate.

In addition, a Stacked LSTM model and a Long Short-Term Memory (LSTM) model are constructed. With the exception of a few extreme values, all forecasts are accurate.

We analyze the four distinct ways to estimating the rainfall using Root Mean Square Error (RMSE) as the accuracy metric. In comparison to the Linear Regression model, the LSTM model, and the Stacked LSTM model, the RMSE Value for the SARIMA model is lower. Thus, based on comparisons between SARIMA (1, 0, 1) x $(0, 1, 1)_{12}$ and the other three models, we infer that SARIMA (1, 0, 1) x $(0, 1, 1)_{12}$ is more accurate at predicting future rainfall for this particular dataset. We can implement techniques to increase crop productivity and the efficient use of water resources using this prediction about the trend and seasonality of the rainfall.

6. REFERENCES

[1] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung.

Time series analysis - Forecasting and control. Fifth edition, Wiley, 2015.

[2] Rob J. Hyndman and George Athanasopoulos. Forecasting: Principles and Practice. Second edition, OTexts, 2018.

[3] Brockwell, P.J. and Davis, R. A. Introduction to Time Series Analysis. Springer, 2003.

[4] Chatfield, C. Time Series Forecasting. Chapman & Hall, 2001.

[5] Joos Kostranje. Advanced Forecasting with Python: With State of the Art Models Including LSTMs, First Edition, 2021.

[6] James D. Hamilton. Time Series Analysis. Princeton University Press, 1994.

[7] Kolla Bhanu Prakash and Kanagachidambaresan G.R. Programming with TensorFlow: Solution for Edge Computing Applications, First Edition, Springer, 2021.

[8] Dongkyum Kim and Seokkoo Kang, ain-fall runoff LSTM model predicting daily runoff, Hongik University, 2009-20022

[9] Shuli Wang, North China University of Water Resources Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation, Version of Record 19 February 2022

[10] SARIMA Modeling and Forecasting of Seasonal Rainfall Patterns in India, Subbaiah Naidu K.CH.V Department of Mathematics, BT College, Madanapalle, India.

[11] Time Series Analysis and Forecasting of Rainfall for Agricultural Crops in India: An Application of Artificial Neural Network, Debasis Mithiya, Kumarjit Mandal & Simanti Bandyopadhyay, November 6, 2020

[12] Time series analysis of climate variables using seasonal ARIMA approach, TRIPTI DIMRI* , SHAMSHAD AHMAD and MOHAMMAD SHARIF Department of Civil Engineering, Jamia Millia Islamia, New Delhi, India.(20 March 2020)

[13] An Application of Time Series Analysis in Modeling Monthly Rainfall Data for Maiduguri, North Eastern Nigeria Emmanuel Sambo Uba Department of Statistics, Ramat Polytechnic, Maiduguri H R Bakari Department of Mathematics & Statistics, University of Maiduguri.

[14] Time Series Analysis of Nyala Rainfall Using ARIMA Method Tariq Mahgoub Mohamed1 and AbbasAbd Allah Ibrahim2 1Khartoum College of Technology, Sudan 2College of Water and Environmental Engineering, Sudan University of Science and Technology (SUST). (2nd,March-2014)

[15] Forecasting daily meteorological time series using ARIMA and regression models** Małgorzata Murat1 , Iwona Malinowska1 , Magdalena Gos2 , and Jaromir Krzyszczak2 * 1 Department of Mathematics, Lublin University of Technology, Nadbystrzycka 38a, 20-618 Lublin, Poland 2 Institute of Agrophysics, Polish Academy of Sciences, Doświadczalna 4, 20-290 Lublin, Poland January 31, 2018

[16] Development and evaluation of seasonal rainfall forecasting (SARIMA) model for Kumaon region of Uttarakhand, Utkarsh Kuma, November, 2022

[17] SARIMA Approach to Generating Synthetic Monthly Rainfall in the Sinú River Watershed in Colombia, Luisa Martinez-Acosta, Universidad Autónoma de Baja California, 8 June 2020

[18] SARIMA Modelling and Forecasting of Monthly Rainfall Patterns for Coimbatore, Tamil Nadu, India, Kokilavani S, Tamil Nadu Agriculture University, May 2020,

[19] Forecasting monthly rainfall using autoregressive integrated moving average model (ARIMA) and artificial neural network (ANN) model: A case study of Junagadh, Gujarat, India, D. K. Dwivedi* College of Forestry, ACHF, Navsari Agricultural University, Navsari (Gujarat), January 27, 2019.

[20] Forecasting and Modeling Monthly Rainfall in Bengaluru, India: An Application of Time Series Models, Hemlata Josh, Department of Statistics, CHRIST(Deemed to be University), Bengaluru, International Journal of Scientific Research in Mathematical and Statistical Sciences Volume-8, Issue-1, pp.39-46, February (2021)

[21] Application of ARIMA Models in Forecasting Average Monthly Rainfall in Birzeit, Palestine, Deeb Aborass, Birzeit University, Palestine, International Journal of Water Resources and Arid Environments 11(1): 62-80, 2022