Nigerian Journal of Theoretical and Environmental Physics (NJTEP)
ISSN Print: 3026-9601
ISSN Online: 1597-9849
DOI: https://doi.org/10.62292/njtep.v4i1.2026.104
Volume 4(1), March 2026

Using Machine Learning To Predict Rainfall in Abeokuta Nigeria

Ayanda, Oluwatosin Emmanuel

University of Plymouth, United Kingdom.

*Corresponding author's email: ayandaoluwatosin@gmail.com

ABSTRACT

Rainfall is extremely important in Abeokuta as it supports agriculture, influences daily life, and the local economy. Abeokuta receives substantial seasonal precipitation; however, variability poses significant risks. This work focuses on Abeokuta and examines rainfall variability from 2001 to 2010. Using machine learning models is one way to extract information, patterns, and trends from historical data, enabling stakeholders to make informed decisions. It is essential to measure the climatic variables that correlate with rainfall, as this will help identify ways to mitigate their effects in pursuit of a balanced nature. In this study, the Seasonal Autoregressive Integrated Moving Average (SARIMA), Random Forest (RF), and Artificial Neural Network (ANN) were used to examine seasonality and patterns in our dataset for predictive purposes. The results show that the ANN model's MSE (7767.4691) and MAE (64.7500) are the lowest among the models used in this work. This suggests that the ANN model's predictions are closer to the actual rainfall values than those of other models. Furthermore, a correlation analysis revealed that the ANN correlated with evaporation and relative humidity. This suggests that during wet seasons or prolonged rainy periods, when rainfall is high, evaporation is low, and less irrigation may be needed, but soils stay wet longer. This might result in flooding. In contrast, in dry seasons with little rain, evaporation is high, and soils dry out quickly, increasing irrigation demand.

Keywords:
Rainfall,
Machine learning models,
Environmental impact.

INTRODUCTION

It is well known that rainfall depends on some climatic variables such as temperature, evaporation, relative humidity, and atmospheric pressure. Evaporation is the process by which water changes from liquid to vapor and enters the atmosphere. As a key process in the water cycle, evaporation affects both climate and water availability. Temperature is also crucial for rainfall and measures the average kinetic energy of a substance's molecules. It indicates how hot or cold a body is. According to Lindsey and Dahlman (2025), Earth's temperature has risen by an average of 0.11° Fahrenheit (0.06° Celsius) per decade since 1850. This totals about 2° F overall. The increase may likely influence rainfall globally, especially in the study area.

In Nigeria, global warming causes prolonged drought by altering temperature patterns and increasing evaporation rates. This intensifies water stress in regions already prone to aridity. The IPCC highlights West Africa, including Nigeria, as vulnerable to increased drought risk due to rising temperatures and changing precipitation patterns (IPCC, 2014). Global warming also increases the risk of extreme rainfall events, leading to more frequent and severe floods in Nigeria. These changing climate conditions increase the likelihood of heavy precipitation. Such events can overwhelm drainage systems, leading to both riverine and urban flooding. Berlie (2018) noted that both human and natural forces contribute to global warming. These forces result in changes in rainfall and temperature.

Climate is a complex system, and analysing its data can be challenging. Time series models are often used to analyse climate datasets. A time series model is a mathematical or statistical method used to analyze and predict future values. Thus, climate time-series analysis uses statistical methods to understand the temporal evolution of the climate. The main advantage of time series analysis is its simplicity in the context of linear models, such as ARIMA and SARIMA. Time series analysis offers a framework for studying temporal variation in climate variables (Fowler et al., 2007). This approach involves examining sequential data points

collected over time. It enables scientists to detect patterns, trends, and fluctuations that are crucial to understanding climate dynamics.

The SARIMA (Seasonal Autoregressive Integrated Moving Average) model is a widely used extension of the ARIMA (Autoregressive Integrated Moving Average) model in climate research. ARIMA models handle non-seasonal trends, while SARIMA incorporates additional seasonal parameters, enabling accurate forecasting of both seasonal and non-seasonal patterns in climate time series.

Minor climate deterioration can cause disastrous socioeconomic consequences. Temperature increases can alter biophysical relationships for crops, livestock, and fisheries. In Nigeria, rainfall is the main form of precipitation. The country gets heavier rainfall in the south than in the north, due to its geography. The agricultural sector accounts for almost 19% of Nigeria's GDP (Statista, 2023). Thus, any change in rainfall pattern can reduce crop yields (Nwosu et al., 2021). EPA (2009) notes that global warming intensifies the water cycle, enhancing water availability. A warmer atmosphere evaporates more water, allowing it to hold more vapour.

Descriptive statistics present data using measures such as mean, median, and mode. Inferential statistics use sample data to make predictions or draw conclusions about a larger population. The five steps in statistical analysis are data collection, data organization, data presentation, data analysis, and data interpretation.

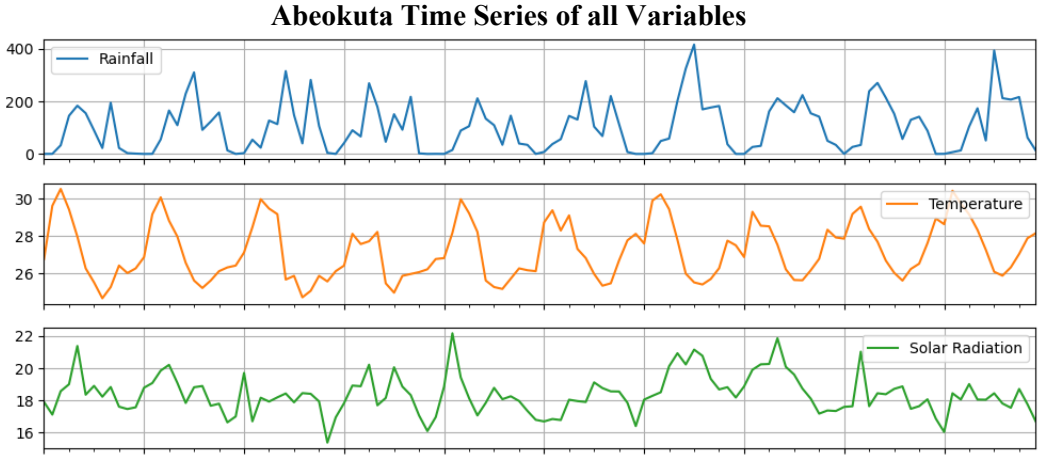
Helfer et al. (2012) examined the impact of climate change on temperature and evaporation in Australia. They found that 40% of water storage capacity evaporated annually, primarily due to rising surface air temperatures. Cifuentes et al. (2020) analyzed the influence of historical climate change over the past decade and demonstrated that machine learning techniques enable accurate temperature prediction. Ray et al. (2021) employed the SARIMA model to forecast monthly rainfall and temperature in South Asian countries, observing changes during the study period. These researchers indicated that SARIMA is highly relevant and outperforms other models.

This study aims to develop and apply machine learning to predict rainfall in this location, as it can help to address environmental challenges, water resources management, and climate adaptation, by integrating meteorological data such as temperature, solar radiation, wind speed, relative humidity, atmospheric pressure, and evaporation into Seasonal Autoregressive Integrated Moving Average, Random Forest, and Artificial Neural Network. The use of SARIMA, RF, and ANN models underscores this research's modern approach to addressing real-world challenges in a data-driven manner and is tailored to the location's unique climatic and environmental characteristics.

MATERIALS AND METHODS

The location, Abeokuta, that is used for this study is known for agricultural produce like ofada rice, vegetables, cassava, and cocoa. It is in the southwestern part of Nigeria, north of Lagos and near the Ogun River. Abeokuta is vulnerable to flooding, and this is the major reason why this location is suitable for this study. The dataset was obtained from the Nigerian Meteorological Agency (NiMet), and it span from 2001 to 2010. The 2001 to 2010 dataset was used because it is the most complete set accessible for the location considered in this work. It includes atmospheric pressure (mb), rainfall (mm), air temperature (°C), solar radiation (MJ/m2/day), relative humidity (%), wind speed (m/s), and pan evaporation (mL).

Data was processed in a Jupyter notebook using Anaconda Python libraries, including Pandas for data manipulation and Matplotlib and Seaborn for visualizations.



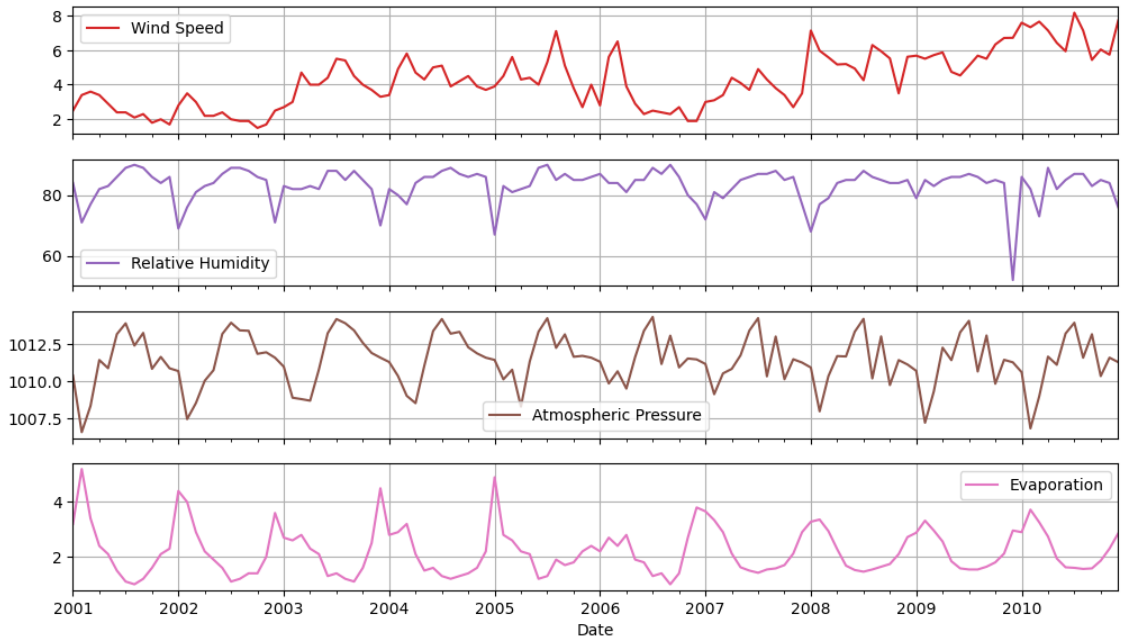


Figure 1: Graphical Representation of Dataset from 2001-2010 for Abeokuta

In this study, rainfall is used as the dependent variable while temperature, relative humidity, solar radiation, atmospheric pressure, and evaporation are used as independent variable. This is in line with physical observation of the relationship between these variables.

The machine models that were used include Seasonal Autoregressive Integrated Moving Average (SARIMA), Artificial Neural Network (ANN), and Random Forest (RF). SARIMA is a time series and linear model which was used with other general and non-linear models (ANN and RF) to identify which among the models performs better.

SARIMA is used in this work because of its ability to model seasonal patterns. There are four components in SARIMA (Vanlaar et al., 2014; Box et al., 2013).

The non-seasonal and seasonal Auto Regressive (AR) polynomial term of order p and P

$$\emptyset_p(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p \quad (1)$$

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \quad (2)$$

The non-seasonal and seasonal Moving Average (MA) part of order q and Q

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \quad (3)$$

$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs} \quad (4)$$

Non-seasonal differencing operator is the of order d used to eliminate polynomial trends,

$$(1 - B)^d \quad (5)$$

Seasonal differencing operator is the order of D used to eliminate seasonal patterns

$$(1 - B^s)^D \quad (6)$$

The generalised form of SARIMA model has a general multiplicative form, SARIMA (p,d,q)x(P,D,Q)s (Hipel and McLeod, 1994), which can be written as

$$\emptyset_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (7)$$

During the hyperparameter search, the SARIMA specification ((0, 0, 0), (0, 1, 1, 12)) achieved the lowest validation MSE of 3619.88. This result indicates that it minimized forecast error on unseen data more effectively than other configurations. Using validation MSE as the optimization criterion prioritizes generalization over in-sample fit and reduces the risk of overfitting. The model's AIC value of 804.58, among the lowest of tested candidates, further supports the optimization outcome. The low AIC confirms that the improvement in validation MSE was achieved without unnecessary model complexity, reflecting an efficient balance between fit and simplicity.

Correlation analysis was used in this study to assess the relevance of each feature to rainfall modeling. A correlation analysis is a statistical method used to measure the direct relationship between two variables. This analysis is fundamental in identifying associations that can inform further investigation or decision-making. The correlation coefficient, typically denoted as Pearson's r, ranges from -1 to +1. A value of +1 signifies a perfect positive direct relationship, -1 signifies a perfect direct linear relationship, and 0 suggests no direct relationship. It is essential to know that correlation is not causation. While it highlights relationships, further analysis is needed to understand the underlying

mechanisms driving these associations. Anscombe (1973) noted that coefficient analysis should accompany the data plot for a visual inspection of the relationship.

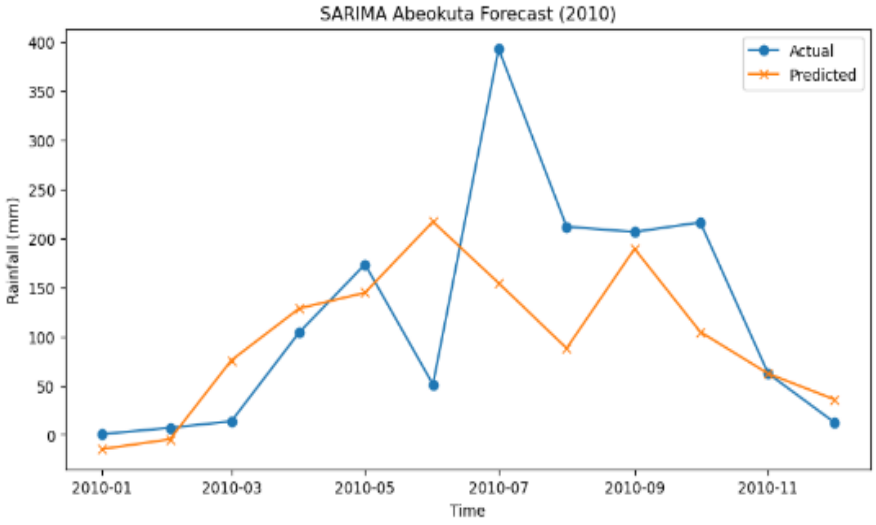
The dataset is split into 80% for training, 10% for validation, and 10% for test, i.e., train: 2001–2008; validation: 2009; test: 2010. This is to ensure that only the past dataset is used to predict future values. To optimize results, the models employed a time-series cross-validation strategy. Before training, input features were normalized using the MinMaxScaler to ensure comparable ranges across predictors, which is particularly important for ANN performance. Subsequently, hyperparameter tuning was performed for both the Random Forest model (number of estimators, maximum depth, and minimum samples per split) and the Artificial Neural Network (ANN), including parameters such as the number of hidden layers, number of neurons, learning rate, activation functions, and regularization terms. Finally, to ensure reproducibility, all random processes were controlled with a fixed random seed (random_state = 42).

Table 1: Parameters for the Models

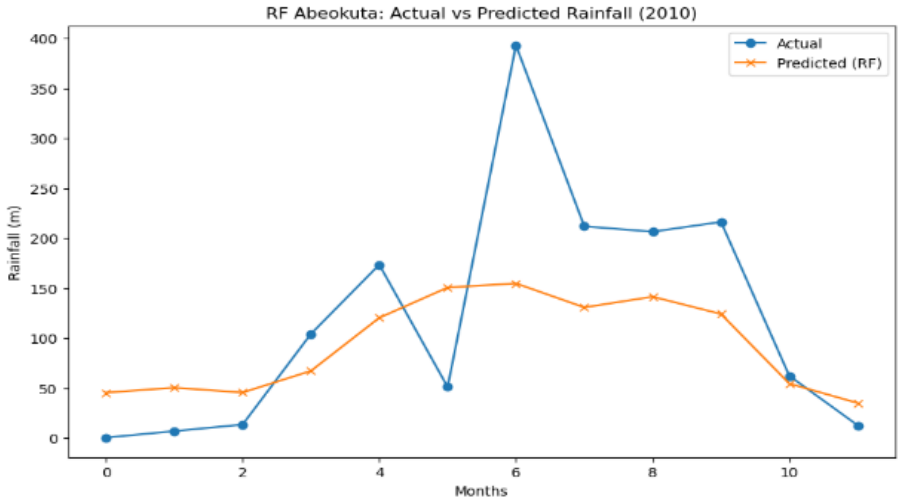
Location	Random Forest	SARIMA
Abeokuta	Best parameters (RF): {'rf__max_depth': 3, 'rf__min_samples_split': 5, 'rf__n_estimators': 50}	BEST by validation MSE: params: ((0, 0, 0), (0, 1, 1, 12)) validation MSE: 3619.878123394447 AIC of best candidate: 804.5779240992412

RESULTS AND DISCUSSION

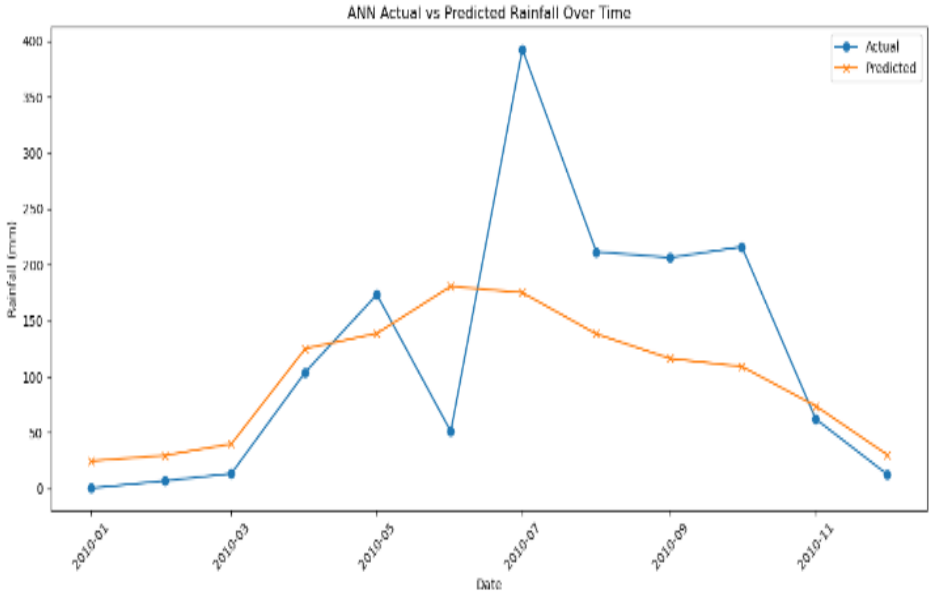
The performance of both SARIMA, RF, and ANN models were evaluated by comparing the predicted vs actual evaporation from Nigerian Meteorological Agency (NiMet). The results are shown in Fig. 2.



(a)



(b)



(c)

Figure 2: Predicted Rainfall using the (a) SARIMA, (b) RF, and (c) ANN Models vs the Actual Rainfall for Abeokuta

After training the models, the figures above show the predicted rainfall for Abeokuta. Notably, the closer the actual and predicted dots are, the more accurate the model is. The results indicate that the ANN performs best in this location.

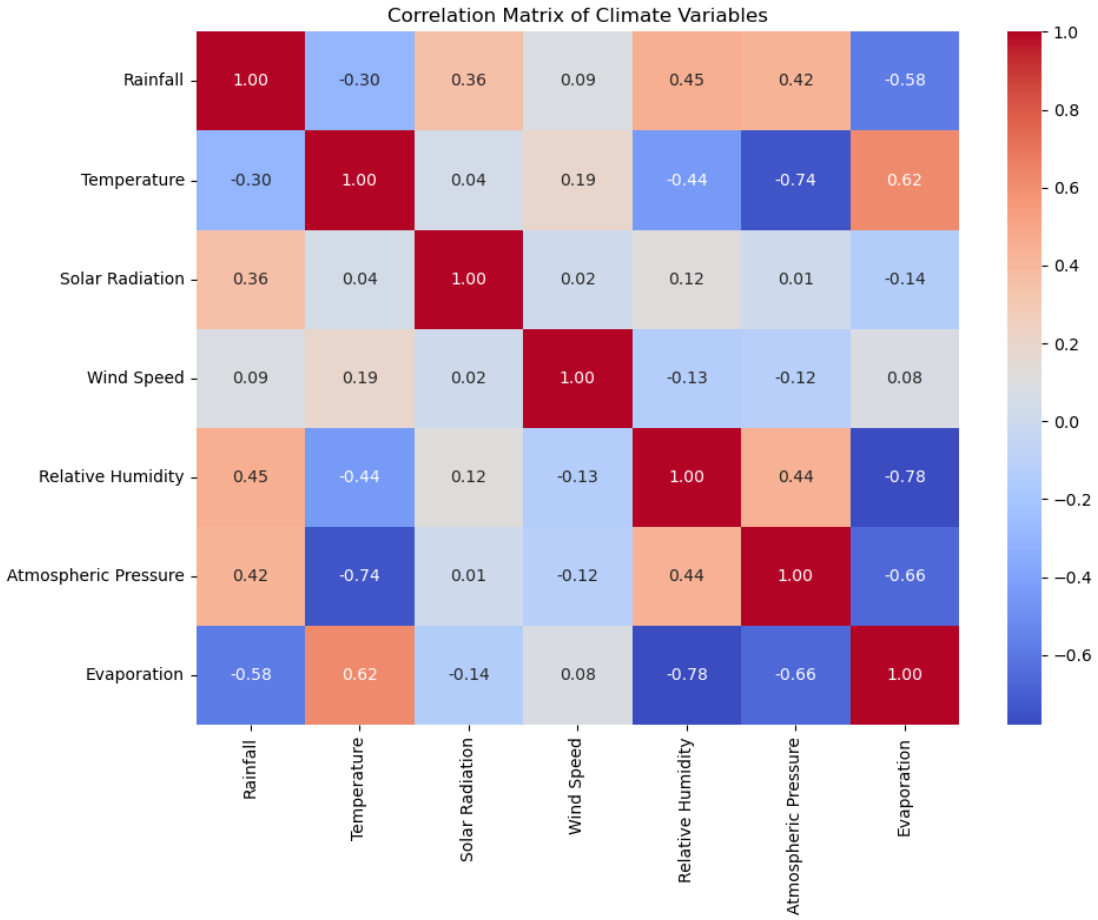


Figure 3: The Result of Correlation Analysis Conducted using the ANN Model's Built-in Feature Importance Metric

Building on these results, Figure 3 shows that rainfall and evaporation have a strong negative correlation (-0.58), meaning that as rainfall increases, evaporation tends to decrease. Additionally, rainfall moderately correlates with relative humidity (+0.45) and atmospheric pressure (+0.42), indicating that higher rainfall is typically associated with higher humidity and pressure in Abeokuta. These correlations suggest that during wet seasons or prolonged rainy periods, when rainfall is high, evaporation is low, and less irrigation may be needed, but soils stay wet longer. In contrast, in dry seasons with little rain, evaporation is high and soils dry out quickly, increasing irrigation demand.

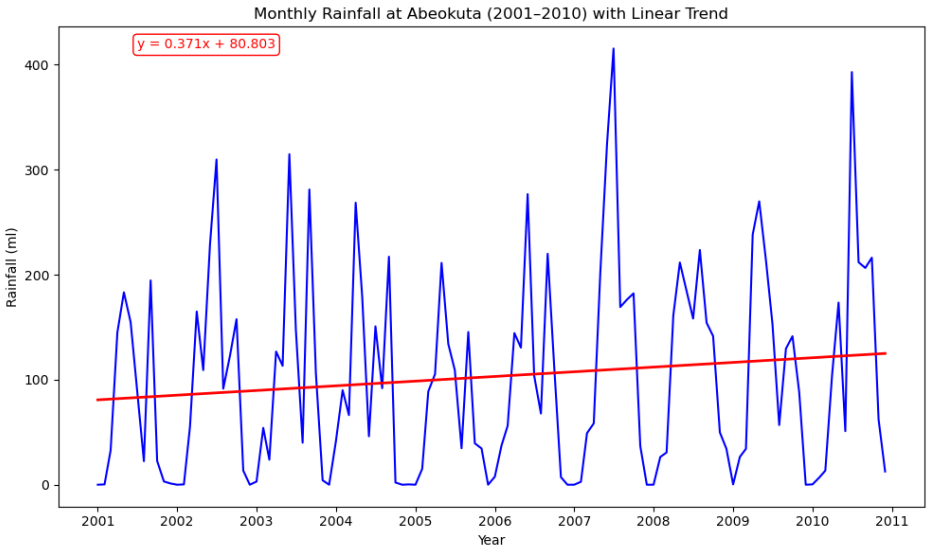


Figure 4: Trend of Rainfall in Abeokuta from 2001 – 2010

The trend analysis in Figure 4 shows an increasing trend, suggesting that rainfall in Abeokuta has increased during this period, likely accompanied by decreasing evaporation and increasing relative humidity, as observed in the correlation analysis.

Model performance and Comparison

This is the stage where we assess the models' performance and compare them based on their error margins. It helps us to determine how well SARIMA, RF, and ANN predict future values based on past data. This enables us to evaluate the reliability and trustworthiness of the learning models. The evaluation metrics that were used in this study are Residual Plot, Mean Squared Error (MSE), Mean Absolute Error (MAE), and the R^2 Score.

Table 2: Model Evaluation for Abeokuta for a Period of Ten Years (I.E. 2001 - 2010)

Locations	SARIMA	Random Forest	ANN
Abeokuta	Mean Squared Error (MSE): 9898.4578 Mean Absolute Error (MAE): 68.5877 R^2 Score: 0.2616	Mean Squared Error (MSE): 7973.5237 Mean Absolute Error (MAE): 68.0811 R^2 Score: 0.4052	Mean Squared Error (MSE): 7767.4691 Mean Absolute Error (MAE): 64.7500 R^2 Score: 0.4206

Table 1 shows that the ANN outperforms the SARIMA and RF models in Abeokuta. The MSE and MAE for the ANN model are the lowest among the models used in this work. This suggests that the ANN model's predictions are closer to the actual rainfall values than those of other models. Also, ANN captures the seasonal and temporal patterns well.

The R^2score for ANN is the highest (0.4206), which implies that ANN can forecast more than 40% of the rainfall values in Abeokuta.

The research work on rainfall prediction in Nigeria done by Dada et al. (2021) using ANN, Aiyelokun et al. (2023) using RF, and Adams and Bamanga (2020) using SARIMA suggests that these models have the capacity to predict rainfall in Nigeria. Also, their work suggests an increasing trend of rainfall, which aligns with this work.

CONCLUSION

The consistent occurrence of extreme weather events, such as flooding and drought, in Nigeria, especially in Abeokuta, has affected lives and property. This rise threatens all forms of life on earth if nothing is done to abate it. In this study, I apply machine learning to predict rainfall at this location, which can help address environmental challenges, support water resource management, and support climate adaptation. Monthly average rainfall from 2001 to 2010 was used to understand the seasonal patterns of these variables with our model. Our study also provided insights into how relative humidity and evaporation moderately correlated with increases and decreases in rainfall during this period. The correlation suggests that during wet seasons or prolonged rainy periods, when rainfall is high,

evaporation is low, and less irrigation may be needed, but soils stay wet longer. In contrast, in dry seasons with little rain, evaporation is high and soils dry out quickly, increasing irrigation demand. In addition, the modes were used to forecast rainfall trends using other climatic variables. This way, we can detect changes in parameters and their effects on rainfall, helping us assess the likelihood of flooding and drought in Abeokuta.

REFERENCES

Adams, S. O., & Bamanga, M. A. (2020). Modelling and forecasting seasonal behavior of rainfall in Abuja, Nigeria; A SARIMA approach. Am. J. Math. Stat, 10(1), 10-19.

Aiyelokun, O. O., Aiyelokun, O. D., & Agbede, O. A. (2023). Application of random forest (RF) for flood levels prediction in Lower Ogun Basin, Nigeria. Natural Hazards, 119(3), 2179-2195.

Anscombe, F. J. (1973). Graphs in statistical analysis. The american statistician, 27(1), 17-21.

Box, G.E., Jenkins, G.M., Reinsel, G.C., (2013). Time Series Analysis: Forecasting and Control; John Wiley & Sons, Inc.: Hoboken, NJ, USA.

Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using machine learning techniques: a review. Energies, 13(16), 4215.

Dada, E. G., Yakubu, H. J., & Oyewola, D. O. (2021). Artificial neural network models for rainfall prediction. European Journal of Electrical Engineering and Computer Science, 5(2), 30-35.

EPA., (2009). Climate change, United States Environmental Protection Agency. www.epa.gov

Fowler, H.J., Blenkinsop, S. and Tebaldi, C., (2007). Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. International Journal of Climatology: A Journal of the Royal Meteorological Society, 27(12), pp.1547-1578.

Helfer, F., Lemckert, C. and Zhang, H., (2012). Impacts of climate change on temperature and evaporation from a large reservoir in Australia. Journal of hydrology, 475, pp.365-378.

IPCC. (2014). Climate Change: Synthesis Report. Contribution of Working Groups I, II, and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland: IPCC.

Nwosu, C.A., Echeta, D.O., Ukwunna, J.C., (2021). Impact of Climate Shocks on Agricultural Productivity in Nigeria .14th NAEE/IAEE Annual International Conference, Abuja, and July 26th – 28th, 2021.

Ray, S., Das, S. S., Mishra, P., & Al Khatib, A. M. G. (2021). Time series SARIMA modelling and forecasting of monthly rainfall and temperature in the South Asian countries. Earth Systems and Environment, 5(3), 531-546.

Statista, (2023). Share of Gross Domestic Production (GDP) generated by the agricultural sector in Nigeria as of 2023. https://www.statista.com/statistics/1207940/share-of-gdp-by-agricultural-sector-in-nigeria/

Vanlaar, W., Robertson, R. and Marcoux, K., (2014). An evaluation of Winnipeg's photo enforcement safety program: Results of time series analyses and an intersection camera experiment. Accident Analysis & Prevention, 62, pp.238-247.