Techno-computing Journal 2(1) 41-49(2026)
https://doi.org/10.71170/tecoj.2025.2.1.41-49

Research Paper

Ensemble Machine Learning Approaches for Statistical Downscaling and Future Precipitation Projection in Maiduguri, Nigeria

Jazuli Abdullahi[1,2*], Bashir Abba Kabir[1]

1 Department of Civil Engineering, Faculty of Engineering, Baze University, Abuja, Nigeria
2 Center for Clean Energy and Climate Change, Baze University, Abuja
*Corresponding Author Email: jazuli.abdullahi@bazeuniversity.edu.ng

Abstract

Climate change impact studies are carried out using the general circulation models (GCMs) to mitigate the greenhouse gas effect on the environment. This study aimed to determine the long-term future changes of precipitation in Maiduguri Nigeria. To achieve this, the GCM variables were downloaded in coarse resolution and converted into local scale using machine learning (ML) and conventional downscaling models including booted regression trees (BOT), support vector machine (SVM) and multiple linear regression (MLR) with prior selection of dominant inputs by Pearson correlation analysis. To improve the downscaling performance of the standalone models, ensemble approach was applied. The ensemble approach has the ability to combine the strength and weaknesses of the single models thereby improving performance. Thereafter, BOT and SVM models were applied to forecast the future precipitation changes. The results showed that with appropriate selection of the GCM variables, BOT, SVM and MLR models could be employed for the statistical downscaling in Maiduguri station. Moreover, the applied ensemble model has improved performance up to 8%, 29% and 30% over BOT, SVM and MLR respectively. The forecast results indicated a decrease in precipitation amount most especially in rainy season months from June to September towards the year 2092.

Keywords: Statistical Downscaling, Ensemble Learning, Precipitation Projection, Climate Change, Maiduguri Nigeria

1. Introduction

The changing global climate alters the hydrological cycle, which in response is causing variability in the frequency of the extreme events, availability of water, irrigation water use, and quality of freshwater resources [1]. Thus, the varying nature of the climate, due to the perturbations induced by human activities, draws significant attention to water resources and hydrology. Available evidence showed that Nigeria is already being plagued with diverse ecological problems, which have been directly linked to the ongoing climate change [2]. The southern ecological zone of Nigeria, largely known for high rainfall, is currently confronted by irregularity in the rainfall pattern, while the Guinea savannah is experiencing a gradually increasing temperature. The Northern zone faces the threat of desert encroachment at a very fast rate per year, occasioned by a rapid reduction in the amount of surface water, flora, and fauna resources on land [3][4]. This makes people exploit more previously undisturbed lands, leading to depletion of the forest cover and an increase in sand dunes/Aeolian deposits in the Northern axis of Nigeria.

General circulation models (GCMs) are used to study the future fluctuations of temperature and precipitation in such a way that up to the end of the century, large-scale climate data are simulated under the greenhouse gas changes effect. In the local climate studies, the developed GCM outputs in coarse spatial resolution cannot directly be used. As a result, suitable techniques should be used to downscale the coarse spatial resolution of GCM outputs into finer local climate data [5]. The

(Received 08 Feb 2026; revised 28 March 2026; accepted 30 March 2026; first published online 10 April 2026);

downscaling models are generally classified into: (a) dynamical downscaling method that uses regional climate models (RCMs) based on boundary conditions set by GCM data to extract local-scale information, and (b) statistical downscaling method, which is based on creating a statistical relationship between the predictand (local-scale weather data) and predictors (large-scale climate variables). The methods used for statistical downscaling are simple to use and require little effort for computations, and can be applied in different regions for different GCM outputs [6]. Many methods of statistical downscaling can be found in the literature, including statistical downscaling model (SDSM) [7], correlation analysis [8], multiple linear regression (MLR) [9], the nearest neighborhood (Zorita and Von Storch 1999), adaptive neuro fuzzy inference system (ANFIS) [10], and artificial neural network (ANN) [11] among others, which have already been used to downscale statistically the GCM outputs.

Focusing on the survey of literature on the consequence results achieved for statistical downscaling modeling of climatic parameters based on the application of AI-based methods, e.g., ANNs, it was found that there are contradicting issues regarding their performance; while some studies show efficiency and superiority of ANNs, others demonstrate inferiority and drawback of ANNs over MLR-based models [12]. The quantity and quality of the used data may contribute to the inconsistency of these results. AI-based modeling challenges issues for huge amounts of data arise from the presence of redundant information in the data set. While using the nonlinear models (such as ANN and ANFIS), over the simulation time, the noise present in the data set can be nonlinearly magnified. According to a study by [13], in ANN-based hydro-climatic processes modeling, utilization of several input variables may lead to inefficient results owing to: (i) irrelevant input variables, which lead to difficulty in training process; (ii) lack of convergence and low precision can be caused by the irrelevant input variables; (iii) being more time-consuming and an increase in computational memory; and (iv) it is more difficult to understand the complex models formed using huge inputs in comparison to simple models, more especially when the results of the models are compared. As such, the AI-based downscaling models' efficiency can be largely enhanced by input feature extraction methods as a preprocessing technique.

Some studies, such as [14], utilized data preprocessing techniques in addressing the complexity in statistical downscaling modeling of GCM data through the extraction of dominant predictors. However, feature extraction methods are the techniques employed for the selection of the dominant predictors for downscaling modeling with maximum impact. [15] Applied feature extraction methods for hydro-climatic variables' modeling using a fuzzy model and a coupled genetic algorithm. [16] employed feature extraction method via mutual information (MI) for predictors screening for downscaling modeling using SDSM. [17] applied data preprocessing methods of MI and correlation coefficient (CC) for variables' input selection. Despite an increase in variation threats caused by climate change and the efficient performance of ML models in predicting future changes in hydro-climatological variables, previous studies show that no efforts have been made to ascertain and evaluate the impact that climate change may have on precipitation in Maiduguri through statistical downscaling. Therefore, to fill this gap, ML models including boosted regression trees (BOT) and support vector machine (SVM), as well as traditional multiple linear regression (MLR) were employed as statistical downscaling techniques for precipitation projections over Maiduguri station.

2. Materials and Methods

2.1 Study Area

Maiduguri, a major city and metropolis situated in northeastern Nigeria, is situated between latitudes 11°04'N and 11°44'N and longitudes 13°04'E and 13°44'E. Maiduguri's climate is characterized by a long dry season with a high evaporation rate from October to May and a brief wet season for the remainder of the year. With a total land area of 543 km2, Maiduguri is the largest metropolis in the northeastern region of Nigeria. There are four distinct seasons in the region: the Harmattan or Cool Season (December to February), the Rainy Season (June to September), the Harvest Season (September to November), and the Hot Season (March to May). According to the 2006 census, its population is predicted to be 1.275 million [18]. Figure 1 shows the location of the study.

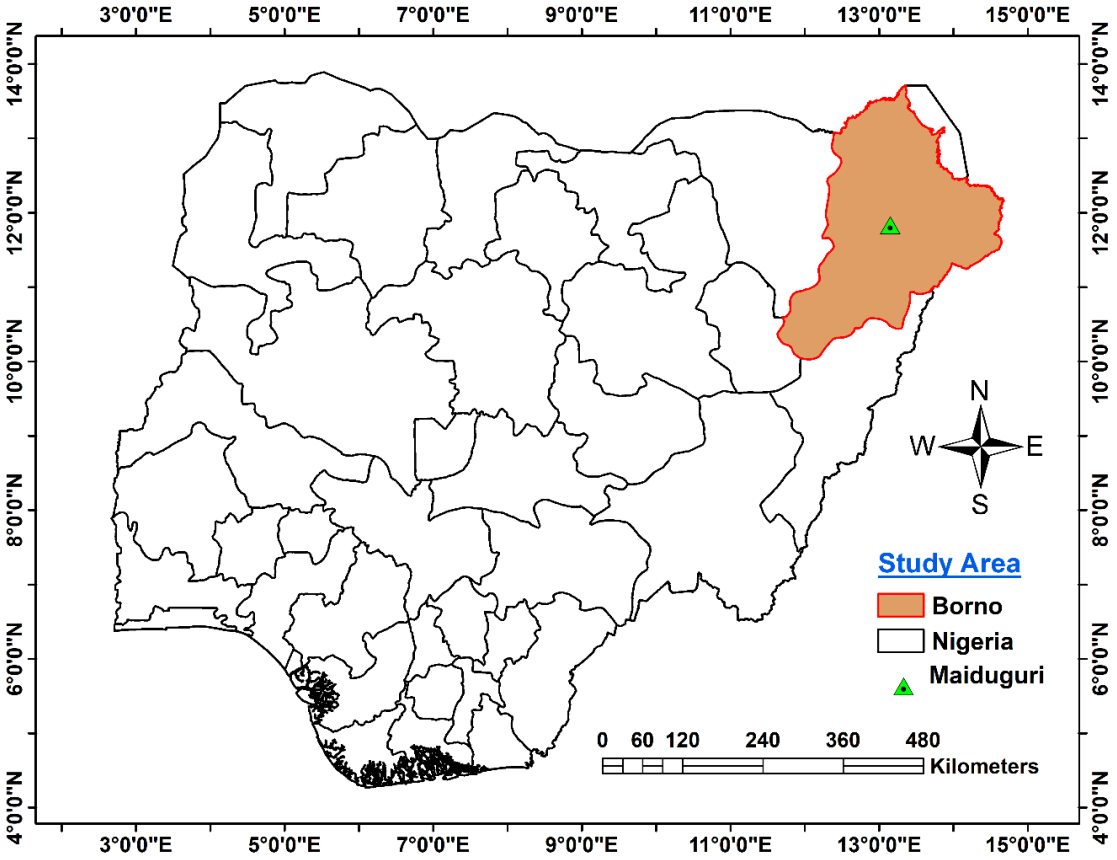

Figure 1: *Map of the study area*

2.2 Performance Criteria and Data Normalization

To determine the accuracy and performance of the applied models for the precipitation downscaling in Maiduguri, 4 global statistical indices were used, including mean absolute deviation (MAD), mean square error (MSE), root mean square error (RMSE), and determination coefficient (DC) [11] [19] given by;

$$MAD = \frac{1}{N}\sum_{i=1}^{n}|p_i - a_i| \quad (1)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(p_i - a_i)^2 \quad (2)$$

$$DC = 1 - \frac{\sum_{i=1}^{N}(a_i - p_i)^2}{\sum_{i=1}^{N}(a_i - \bar{a})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(a_i - p_i)^2}{N}} \quad (4)$$

Where a_i, p_i, $\bar{a}$ and Nare the actual values, predicted values, mean of the actual values, and number of observations, respectively. To ensure all variables have equal attention and to eliminate their dimensional discrepancy, data normalization is usually applied for AI-based modeling. For the normalization purpose in this study, the observations were scaled between 0 and 1. The normalization procedure is given by [11] [20];

$$DC_n = \frac{DC_i - DC_{min}}{DC_{max} - DC_{min}} \quad (5)$$

Where DC_n, DC_{max}, DC_{min} and DC_i represent the normalized value, maximum value, minimum value, and ith values, respectively.

2.3 Boosted Regression Trees (BOT)

The BRT method combines regression trees and a boosting technique to improve the predictive performance of multiple single models (Yang et al. 2016). Boosting is a forward and stage-wise procedure in which a subset of the data is randomly selected to iteratively fit new tree models to minimize the loss function (Elith et al., 2008). This process introduces a stochastic gradient boosting procedure that can improve model performance and reduce the risk of over-fitting (Friedman, 2002). The BRT algorithm is an iterative process in which tree-based models are fitted iteratively using recursive binary splits to identify poorly modeled observations in existing trees until a minimum model deviance was reached. The final fitted model is a linear function of the sum of all trees multiplied by the learning rate (LR) based on all data (Elith et al., 2008).

2.4 Support Vector Machine

The SVM algorithm developed by Vapnik (2013) is a supervised machine learning model for pattern recognition and data analysis. It has been widely employed for regression and forecasting in the fields of agriculture, hydrology, meteorology, and environmental studies. The SVM model estimates the regression function using a series of kernel functions that implicitly map the original, lower-dimensional input data to a higher-dimensional feature space (Fan et al. 2017).

2.5 Multiple Linear Regression

MLR is a conventional technique that models the linear relevance between two or more predictors (independent variables) and a predictand (dependent variable). In general, the n and y representing predictors and dependent variables might have a relation, given by (Elkiran et al. 2021):

$$y = b0 + b1\,x1 + b2\,x2 + b3\,x3 + \cdots + bi\,xi + \xi \quad (6)$$

Where the i^{th} predictor value is xi, the constant of regression is b0, the i^{th} predictor coefficient is bi, and the error term is ξ.

2.6 Ensemble Techniques

For a particular set of information or data, it is observable that the performance of one bright technology could outshine another; at the same time, if dissimilar sets of information are applied, the outcomes may totally be contrary (Nourani et al., 2019). In order not to lose simplification and also to benefit from the significances of all procedures, an ensemble model is formed, which makes use of the individual output of every technique with a definite precedence level assigned to every one of them with the aid of a mediator to proffer the output (Kiran and Ravi, 2008). Therefore, in this study, an ensemble technique was applied to improve the performance of the downscaling modeling.

2.7 Proposed methodology

In this study, the projection of precipitation in the future was performed based on (i) Data collection and feature extraction to select the best inputs, (ii) statistical downscaling of historical data from 1990 to 2022 using BOT, SVM, and MLR, and (iii) long-term precipitation forecast from 2060 to 2092. The proposed method employed in this study is shown in Figure 2.

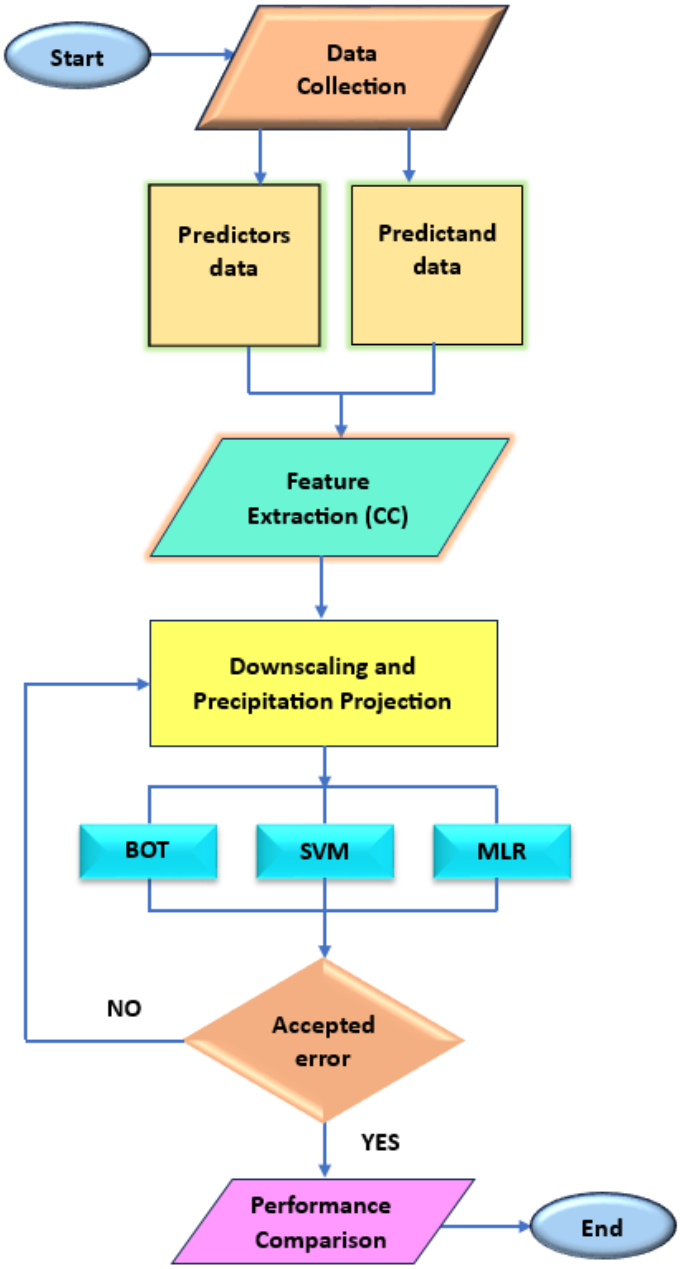

Figure 2: *General methodology applied*

3.0 Application of Results and Discussion

The results and discussion in this study are presented according to the proposed method.

3.1 Dominant Input Selection Results

To determine the most appropriate GCM variables to be used for enhanced downscaling performance, correlation analysis was performed. The results of the correlation analysis is given in Figure 3.

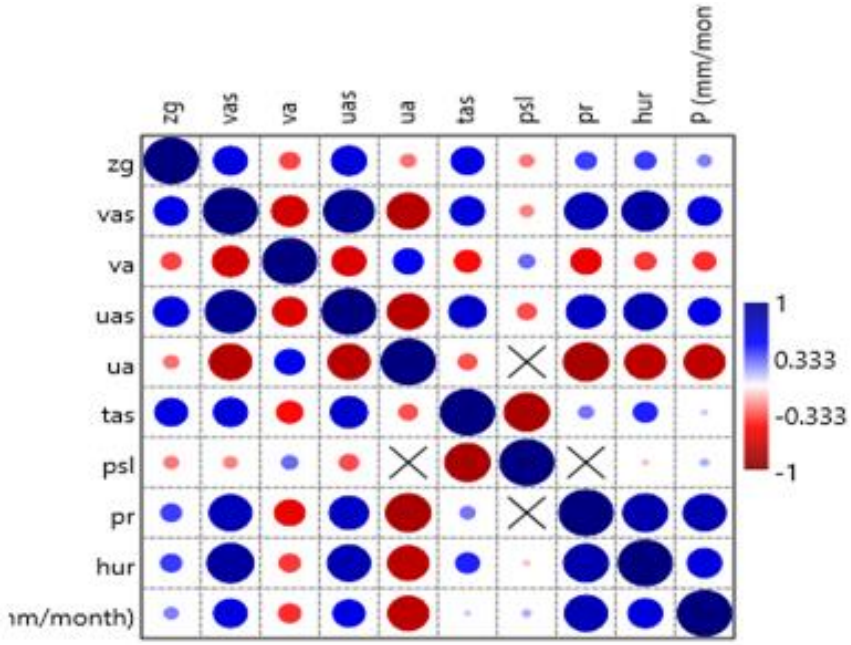

Figure 3: *Results of the correlation analysis performed*

As seen in Figure 3, irrespective of color, the bigger the circle the better the correlation and hence, the higher possible downscaling performance. Pr which is the coarse regional precipitation GCM variable, has the highest correlation with the local precipitation variable, followed closely by ua and tas, which is the least dominant variable based on the applied correlation analysis. Owing to this, two models were developed based on 3 and 4 most correlated GCM variables.

3.2 Results of the Statistical Downscaling Modeling

This section is divided into (i) single model statistical downscaling and (ii) ensemble model statistical downscaling. Therefore, the results are presented accordingly. Single model statistical downscaling results. The standalone models of BOT, SVM, and MLR were used for the statistical downscaling, and performance was assessed using MAD, MSE, RMSE, and DC statistical indicators. The best 3 and 4 GCM variables were used as inputs and classified as model 1 (M1) and model 2 (M2). The results of the downscaling models are given in Table 1.

Table 1: *Single model statistical downscaling results*

		Training				Validation			
Combination	**Model**	**MAD**	**MSE**	**RMSE**	**DC**	**MAD**	**MSE**	**RMSE**	**DC**
M1	BOT	0.047	0.007	0.084	0.846	0.043	0.006	0.076	0.836
	SVM	0.083	0.016	0.127	0.644	0.076	0.013	0.115	0.624
	MLR	0.080	0.017	0.129	0.631	0.073	0.014	0.117	0.616
M2	BOT	0.044	0.006	0.079	0.861	0.041	0.005	0.073	0.851
	SVM	0.083	0.016	0.127	0.645	0.077	0.013	0.115	0.626
	MLR	0.080	0.017	0.129	0.634	0.073	0.014	0.116	0.618

As seen in Table 1, the applied models can be employed to statistically downscale the coarse GCMs values into a finer or local scale value. However, with different methodological approaches of the models, different results were achieved. As demonstrated by Table 1, the 3 or 4 GCM input variables may not have a significant difference in results. For instance, based on the model combinations, it can be observed that in the validation step in terms of DC, the results are 0.836, 0.624, and 0.618 for M1

and 0.851, 0.626, and 0.618 for M2 based on BOT, SVM, and MLR respectively. By thoroughly looking at the Figures, it can be seen that the only differences in performance between M1 and are 0.015 (1.5%), 0.002 (0.2%), and 0.002 (0.2%) based on BOT, SVM, and MLR, respectively. This means that a successful downscaling performance in the study region can be achieved by both 3-input and 4-input models. Moreover, by careful observation of the results, it can be explained that the ML models, being nonlinear models capable of simulating the nonlinear aspect of the GCM variables, led to better results. On the other hand, MLR can only determine the linear aspect of the variables, thereby generating errors for nonlinear phenomena. Figure 4 shows the graphical performance comparison of the applied models.

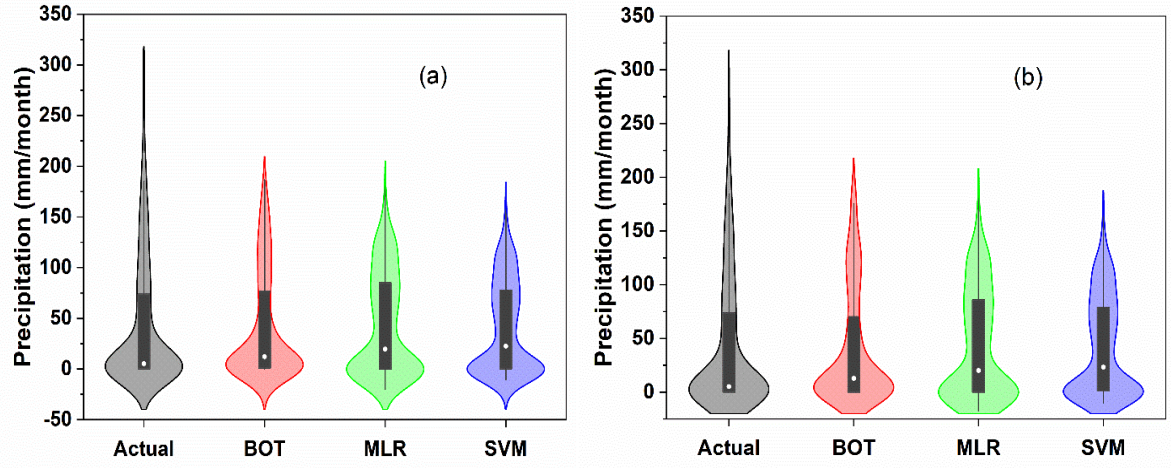

Figure 4: *Performance comparison between the actual and downscaled precipitation for (a) M1 and (b) M2*

It can be seen in Figure 4 that the BOT model is more fitted to the actual precipitation value, thereby leading to the best results in both M1 and M2 downscaling scenarios. Although SVM and MLR models do not match the outstanding performance demonstrated by the BOT model, they clearly indicate their capability of downscaling the GCM variables resolution to a finer scale, with similarity to the trend of the violin box of the actual values.

3.3 Ensemble modeling statistical downscaling results

Despite the outstanding performances of the applied models in the precipitation performance, there is still room to enhance the results. To bridge the gap in the performance, ensemble modeling was applied. Ensemble modeling is the method that combines the strengths of the individual models to enhance modeling capability [21]. Table 2 shows the results of the applied ensemble model.

Table 2: *Ensemble model statistical downscaling results*

		Training				Validation			
Combination	**Model**	**MAD**	**MSE**	**RMSE**	**DC**	**MAD**	**MSE**	**RMSE**	**DC**
M1	BOT	0.034	0.004	0.062	0.914	0.033	0.003	0.056	0.913
M2	BOT	0.034	0.004	0.060	0.920	0.032	0.003	0.055	0.914

As can be seen Table 2, the ensemble modeling has increased the efficiency of the precipitation downscaling to a greater extent. For example, the model with highest performance in Table 1 is BOT which has DC values in the training and validation steps of 0.846 and 0.836 for M1 and 0.861 and 0.851 for M2, whereas DC values in the training and validation by ensemble model up to 0.914 and 0.913 for M1 and 0.920 and 0.914 for M2 were achieved. This signifies an improved performance.

3.4 Results of the Precipitation Forecast

In this section of the study, the precipitation for the future was forecasted from 2060 to 2092 using BOT and SVM models. The results of the future precipitation projection are given in Figure 5. The results of the future forecast showed that there would be a decrease in precipitation in Maiduguri, Nigeria, to a certain extent towards the end of the 21st century. Both BOT and SVM results showed that there would be a decrease in

precipitation amount, most especially during periods with the highest amount of precipitation from May through October.

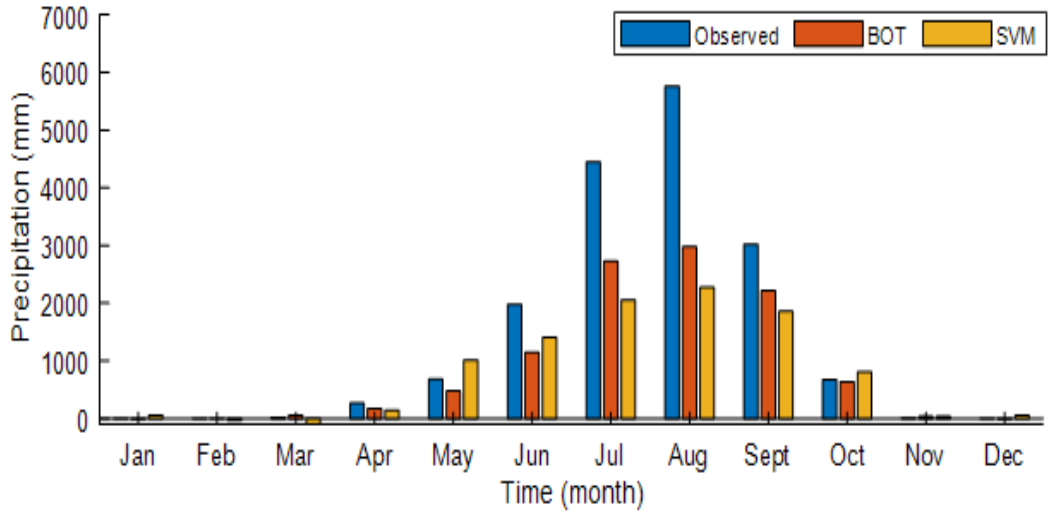

Figure 5: *Comparison between previous and future precipitation*

4.0 Conclusion

This study was performed to determine the possibility of employing machine learning (ML) approaches for the future forecast of precipitation in Maiduguri, Nigeria. In this way, coarse resolution GCM data from coupled model intercontinental project phase 6 (CIMIP6) were downloaded from 1990 to 2092 and subjected to correlation analysis to determine the dominant inputs. Thereafter, statistical downscaling using BOT, SVM, and MLR was conducted to convert the coarse resolution into finer/local scale variables. However, ensemble modeling was employed to improve the accuracy of the applied models. Finally, the impact of climate change on precipitation was assessed through future forecasts. The results indicated that the statistical downscaling could be performed efficiently using the employed models and BOT and SVM, with the most outstanding performance. The future forecast showed that towards the end of the 21st century, the precipitation amount would decrease in the study region, especially in the months of June, July, August, and September.

Competing Interests: The authors declare that they have no competing interests.

Data Availability Statement: The supported data associated with this researcher is available upon request from the corresponding author.

Acknowledgements: The authors gratefully acknowledge the Department of Civil Engineering, Baze University, Abuja, Nigeria, for its valuable support and for providing a conducive environment for this research.

References

[1] S. P. Simonovic, “Bringing future climatic change into water resources management practice today,” *Water Resour. Manag.*, vol. 31, no. 10, pp. 2933–2950, 2017.

[2] D. Adefolalu, “Climate change and economic sustainability in Nigeria,” presented at the International Conference on Climate Change and Economic Sustainability held at Nnamdi Azikiwe University, Enugu, Nigeria, 2007, pp. 12–14.

[3] O. Bello *et al.*, “Evidence of climate change impacts on agriculture and food security in Nigeria,” *Int. J. Agric. For.*, vol. 2, no. 2, pp. 49–55, 2012.

[4] E. E. Obioha, “Climate change, population drift and violent conflict over land resources in northeastern Nigeria,” *J. Hum. Ecol.*, vol. 23, no. 4, pp. 311–324, 2008.

[5] D. E. Mora, L. Campozano, F. Cisneros, G. Wyseure, and P. Willems, “Climate changes of hydrometeorological and hydrological extremes in the Paute basin, Ecuadorean Andes,” *Hydrol. Earth Syst. Sci.*, vol. 18, no. 2, pp. 631–648, 2014.

[6] B. Timbal, A. Dufour, and B. McAvaney, "An estimate of future climate change for western France using a statistical downscaling technique," *Clim. Dyn.*, vol. 20, no. 7, pp. 807–823, 2003.
[7] S. Samadi, G. J. Carbone, M. Mahdavi, F. Sharifi, and M. Bihamta, "Statistical downscaling of river runoff in a semi arid catchment," *Water Resour. Manag.*, vol. 27, no. 1, pp. 117–136, 2013.
[8] W. A. Landman, S. J. Mason, P. D. Tyson, and W. J. Tennant, "Statistical downscaling of GCM simulations to streamflow," *J. Hydrol.*, vol. 252, no. 1–4, pp. 221–236, 2001.
[9] W. H. Klein, "Objective specification of monthly mean surface temperature from mean 700 mb heights in winter," *Mon. Weather Rev.*, vol. 111, no. 4, pp. 674–691, 1983.
[10] K. Alotaibi, A. R. Ghumman, H. Haider, Y. M. Ghazaw, and M. Shafiquzzaman, "Future predictions of rainfall and temperature using GCM and ANN for arid regions: a case study for the Qassim Region, Saudi Arabia," *Water*, vol. 10, no. 9, p. 1260, 2018.
[11] G. Elkiran, V. Nourani, O. Elvis, and J. Abdullahi, "Impact of climate change on hydro-climatological parameters in North Cyprus: Application of artificial intelligence-based statistical downscaling models," *J. Hydroinformatics*, vol. 23, no. 6, pp. 1395–1415, 2021.
[12] M. Abdellatif, W. Atherton, and R. Alkhaddar, "A hybrid generalised linear and Levenberg–Marquardt artificial neural network approach for downscaling future rainfall in North Western England," *Hydrol. Res.*, vol. 44, no. 6, pp. 1084–1101, 2013.
[13] G. J. Bowden, G. C. Dandy, and H. R. Maier, "Input determination for neural network models in water resources applications. Part 1—background and methodology," *J. Hydrol.*, vol. 301, no. 1–4, pp. 75–92, 2005.
[14] M. Babel, T. Sirisena, and N. Singhrattna, "Incorporating large-scale atmospheric variables in long-term seasonal rainfall forecasting using artificial neural networks: an application to the Ping Basin in Thailand," *Hydrol. Res.*, vol. 48, no. 3, pp. 867–882, 2017.
[15] C.-T. Cheng, X.-Y. Wu, and K. W. Chau, "Multiple criteria rainfall–runoff model calibration using a parallel genetic algorithm in a cluster of computers/Calage multi-critères en modélisation pluie–débit par un algorithme génétique parallèle mis en œuvre par une grappe d'ordinateurs," *Hydrol. Sci. J.*, vol. 50, no. 6, 2005.
[16] H. Pahlavan, B. Zahraie, M. Nasseri, and A. Mahdipour Varnousfaderani, "Improvement of multiple linear regression method for statistical downscaling of monthly precipitation," *Int. J. Environ. Sci. Technol.*, vol. 15, no. 9, pp. 1897–1912, 2018.
[17] U. Okkan, "Assessing the effects of climate change on monthly precipitation: proposing of a downscaling strategy through a case study in Turkey," *KSCE J. Civ. Eng.*, vol. 19, no. 4, pp. 1150–1156, 2015.
[18] H. B. Galadima, Y. A. Geidam, B. U. Shamaki, H. I. Abdulrahman, B. Ibrahim, and I. A. Gulani, "Screening of Antimicrobial Residue in Commercial Eggs in Maiduguri Metropolis, Borno State," *Annu. Res. Rev. Biol.*, vol. 25, no. 1, pp. 1–6, 2018.
[19] Z. Ibrahim, P. Tulay, and J. Abdullahi, "Multi-region machine learning-based novel ensemble approaches for predicting COVID-19 pandemic in Africa," *Environ. Sci. Pollut. Res.*, vol. 30, no. 2, pp. 3621–3643, 2023.
[20] U. U. Aliyu *et al.*, "Optimizing biomedical waste generation modeling using quantum machine learning and economic development indicators," *Biomass Bioenergy*, vol. 204, p. 108312, 2026.
[21] U. U. Aliyu, A. M. Jibrin, A. S. Baba, I. A. Aminu, S. Chaki, and R. Kumar, "Advanced Data Driven Prediction of BOD in the Ganga River Using Multivariate Regression and Nonlinear Bilayered Neural Network Ensembles," *Techno-Comput. J.*, vol. 1, no. 3, pp. 1–15, 2025.