DEVELOPMENT OF PREDICTIVE MODEL USING RANDOM FOREST FOR COMMUNITY-LED WATER SAFETY IN FLOOD-PRONE AREAS OF NIGERIA

Yekini, Temitope Musa[*] and Owolabi, Habeeb Ladi[**] and Arioye, Awwal Alade[*]
**Directorate of Information and Communications Technology,*
Lagos State University of Education, Oto/Ijanikin, Lagos State.
***Department of Human Kinetics and Health Education,*
Lagos State University of Education, Oto/Ijanikin, Lagos State

Abstract
Flood-induced deterioration of water quality remains a persistent challenge across many flood-prone communities in Nigeria, where limited monitoring infrastructure constrains timely risk assessment and response. This study developed a Random Forest-based predictive framework for assessing flood-related water quality risks through the integration of multi-source geospatial and environmental datasets, including Sentinel-1 synthetic aperture radar imagery, Sentinel-2 optical imagery, Global Precipitation Measurement (GPM) rainfall data, topographic variables, historical flood records, and Internet of Things (IoT) sensor observations collected between 2018 and 2024. Data preprocessing involved z-score normalization, Principal Component Analysis, and Box–Cox transformation to improve model stability and predictive performance. The model achieved an overall accuracy of 88.4% and an Area Under the Curve (AUC) of 0.93. Feature importance analysis identified rainfall intensity and flood proximity as dominant predictors of water quality risk. The proposed framework offers a scalable decision-support tool for strengthening flood preparedness, water safety management, and climate resilience in data-constrained environments.

Keywords: Random forest predictive model, Community-led water safety, Flood prone areas,

Introduction

Flooding remains one of the most destructive natural hazards globally and continues to pose significant environmental, socio-economic, and public health challenges across many developing countries. In Nigeria, increasing rainfall variability, rapid urban expansion, poor drainage infrastructure, and climate-induced hydrological changes have intensified both the frequency and severity of flood events, particularly within the Niger Delta region, the Benue River Basin, and other low-lying floodplains (Ighile et al., 2022; Ogbuene et al., 2024). Beyond the immediate destruction of infrastructure and livelihoods, floods frequently compromise the quality of surface and groundwater resources through the transport of sediments, agricultural chemicals, sewage effluents, and other contaminants into drinking water sources, thereby increasing the risk of waterborne diseases and environmental degradation (Ezenwa et al., 2023; Ogbuene et al., 2024).

Traditional water quality monitoring approaches in Nigeria rely primarily on periodic field sampling and laboratory-based analysis. Although these methods provide

reliable measurements, they are often constrained by high operational costs, limited spatial coverage, delayed reporting, and inadequate responsiveness during rapidly evolving flood events. Consequently, many flood-prone communities lack access to timely information necessary for proactive water safety management and disaster preparedness (Oloruntoba et al., 2023).

Recent advances in Artificial Intelligence (AI), machine learning, and Earth Observation technologies have created new opportunities for addressing these challenges through predictive environmental monitoring. Machine learning algorithms have demonstrated remarkable capability in modelling complex environmental processes characterized by non-linearity, uncertainty, and high-dimensional datasets. Applications have included flood susceptibility mapping, rainfall-runoff forecasting, water quality prediction, and disaster risk assessment (Ighile et al., 2022; Aliyu et al., 2025; Akindejoye et al., 2025). Among these approaches, the Random Forest algorithm has emerged as one of the most effective ensemble learning techniques due to its robustness against overfitting, ability to handle heterogeneous datasets, resilience to noise, and capacity to quantify variable importance during prediction (Aliyu et al., 2025; Mohamed et al., 2024).

The increasing availability of open-access satellite observations further strengthens the applicability of machine learning for environmental risk assessment. Sentinel-1 Synthetic Aperture Radar (SAR) imagery has proven particularly valuable for flood mapping because of its capability to acquire data under cloud-covered conditions and during adverse weather events. Likewise, Sentinel-2 optical imagery provides high-resolution information on vegetation dynamics, land cover characteristics, and surface water conditions through indices such as the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI). Complementarily, Global Precipitation Measurement (GPM) products offer near-real-time rainfall observations that support hydrological analysis and flood forecasting (Bamiekumo et al., 2025; Eteh et al., 2025).

Despite these technological developments, a substantial research gap remains in the development of integrated frameworks that combine multi-source geospatial datasets, machine learning techniques, and operational decision-support systems for flood-induced water quality risk assessment. Existing studies in Nigeria have largely focused on either flood prediction or water quality modelling as independent research domains, with relatively limited attention devoted to linking predictive outputs with deployable systems capable of supporting practical environmental decision-making (Oloruntoba et al., 2023; Obonyano et al., 2024). Furthermore, many predictive studies remain constrained by sparse monitoring networks, fragmented environmental datasets, and inadequate integration of real-time sensing technologies.

To address these limitations, this study develops a Random Forest-based predictive framework that integrates Sentinel-1 and Sentinel-2 satellite observations, Global Precipitation Measurement rainfall products, topographic variables, historical flood inventories, and Internet of Things (IoT)-based environmental sensor data to assess flood-induced water quality risks in flood-prone regions of Nigeria. Unlike many previous studies that focus exclusively on model development, the proposed framework combines

predictive analytics with a deployable decision-support architecture designed for operational implementation in data-constrained environments.

Specifically, the study sought to: (i) develop and validate a Random Forest model for predicting flood-induced water quality risks; (ii) evaluate the relative contribution of hydro-meteorological, topographic, and environmental predictors; (iii) assess model performance using standard classification metrics; and (iv) demonstrate the applicability of the framework as a scalable decision-support tool for enhancing water safety monitoring, flood preparedness, and climate resilience in vulnerable communities.

.

Methodology

This study developed a Random Forest-based predictive framework for assessing flood-induced water quality risks in flood-prone regions of Nigeria through the integration of satellite observations, meteorological datasets, topographic information, historical flood records, and Internet of Things (IoT)-derived environmental measurements. The methodological workflow comprised four stages: data acquisition, data preprocessing and feature engineering, model development and validation, and deployment within a decision-support environment.

A multi-source geospatial database was developed using environmental observations collected between 2018 and 2024. Sentinel-1 Synthetic Aperture Radar (SAR) imagery was utilized to characterize flood extent and inundation dynamics. Sentinel-2 multispectral imagery was processed to derive vegetation and water-related indices, while rainfall information was obtained from the Global Precipitation Measurement (GPM) dataset. Topographic variables were extracted from the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Historical flood occurrence records were obtained from national flood inventory databases, and IoT sensors supplied environmental observations including water level and turbidity.

Table 1. Data Sources and Predictor Variables

Dataset	Variables Extracted	Source
Sentinel-1 SAR	Backscatter (σ^0), Flood Extent	Copernicus
Sentinel-2 MSI	NDVI, NDWI	Copernicus
GPM-IMERG	Rainfall Intensity	NASA
SRTM DEM	Elevation, Slope	USGS
Flood Inventory	Historical Flood Events	NEMA, NIHSA
IoT Sensors	Water Level, Turbidity	Field Measurements

The Normalized Difference Vegetation Index (NDVI) was calculated as:

$$NDVI = \frac{NIR-Red}{NIR+Red} \quad (1)$$

where: NIR= Near-Infrared reflectance; Red= Red band reflectance

The Normalized Difference Water Index (NDWI) was computed as:

$$NDWI = \frac{Green-NIR}{Green+NIR} \quad (2)$$

where: $Green$= Green band reflectance; NIR= Near-Infrared reflectance

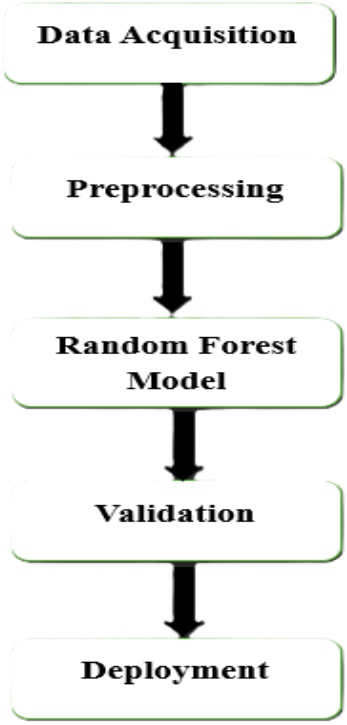

Figure 1. Conceptual Methodological Workflow

Data preprocessing was performed to improve data quality and model robustness. Missing observations were addressed using interpolation and nearest-neighbour imputation techniques. Outliers were identified through exploratory statistical analysis and removed where necessary. Feature standardization was achieved using Z-score normalization:

$$Z = \frac{X-\mu}{\sigma} \quad (3)$$

where: X= original observation; μ= mean of the variable; σ= standard deviation.

Principal Component Analysis (PCA) was applied to reduce dimensionality and minimize multicollinearity among predictor variables. Components explaining at least 95% of

cumulative variance were retained. Variance stabilization was performed using the Box–Cox transformation:

$$Y(\lambda) = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(Y), & \lambda = 0 \end{cases} \quad (4)$$

where: Y= original variable; λ= transformation parameter.

The final predictor set consisted of rainfall intensity, flood proximity, Sentinel-1 backscatter, NDVI, NDWI, elevation, water level, and turbidity.

Random Forest was employed as the classification algorithm for predicting flood-induced water quality risk. The algorithm generates multiple decision trees from randomly sampled subsets of the training data and combines their predictions through majority voting. The Random Forest classifier can be represented as:

$$RF(x) = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\} \quad (5)$$

where: $RF(x)$= final classification output; $T_i(x)$= prediction generated by the i^{th}decision tree; n= total number of trees.

The dataset was partitioned into training and testing subsets using a 75:25 ratio. Hyperparameter optimization resulted in a configuration of 500 decision trees and a maximum tree depth of 15. Model generalization was assessed using 10-fold cross-validation.

Table 2. Random Forest Configuration and Evaluation Metrics

Component	Specification
Training-Testing Split	75%:25%
Cross-Validation	10-Fold
Number of Trees	500
Maximum Depth	15
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, AUC

The trained model was deployed within a Flask-based decision-support environment comprising a data ingestion layer, prediction engine, visualization module, and risk communication interface. Environmental observations were processed through the trained classifier to generate binary risk predictions (Safe or Unsafe) and associated probability scores.

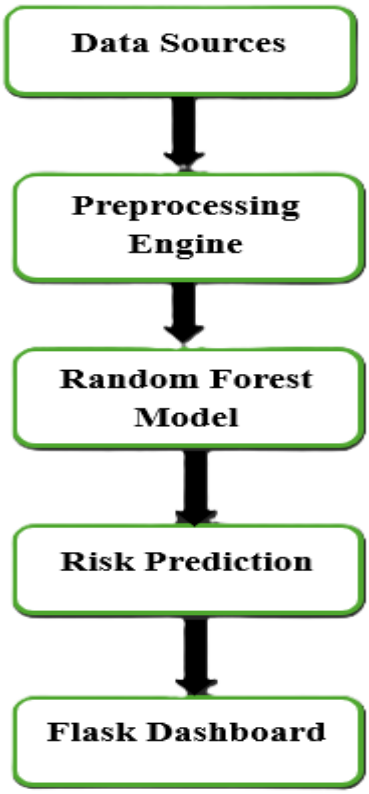

Figure 2. Deployment Architecture of the Proposed Framework

The deployment framework enables near-real-time prediction and visualization of flood-induced water quality risks and provides a foundation for future integration with web-based and mobile decision-support systems.

Results

Table 3. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	AUC	F1-Score
Random Forest	88.4	0.93	0.89
XGBoost	86.1	0.91	0.86
Support Vector Machine	84.7	0.8	0.83

The Random Forest classifier demonstrated strong predictive capability for identifying flood-induced water quality risk zones across the study area. Performance evaluation was conducted using the independent test dataset and validated through 10-fold cross-validation. The model achieved an overall classification accuracy of 88.4%, an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.93, and an F1-score of 0.89.

These results indicate a high level of predictive reliability and suggest that the model effectively distinguished between safe and unsafe water quality conditions under varying hydro-meteorological circumstances. The results indicate that Random Forest outperformed both XGBoost and Support Vector Machine models across all evaluation metrics. The superior performance may be attributed to its ensemble learning architecture, which effectively captures nonlinear relationships among environmental variables while reducing the risk of overfitting.

Table 4. Confusion Matrix for Random Forest Classification

Actual Class	Predicted Safe	Predicted Unsafe	Total
Safe	433	67	500
Unsafe	50	450	500
Total	483	517	1000

The confusion matrix demonstrates that the model maintained a balanced prediction capability across both classes, thereby reducing classification bias and improving overall reliability. The Receiver Operating Characteristic analysis further confirmed the effectiveness of the developed model.

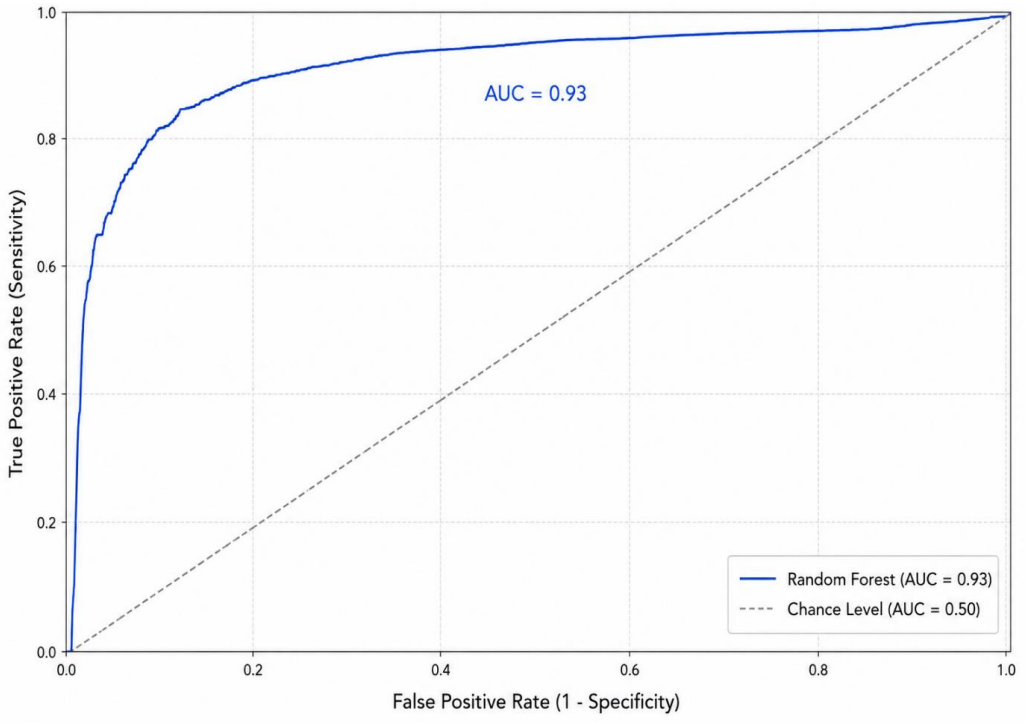

Figure 3. ROC Curve of the Random Forest Model

The model achieved an Area Under the Curve (AUC) value of 0.93. The ROC curve indicates excellent discrimination between safe and unsafe water quality conditions. An AUC value above 0.90 is generally considered indicative of outstanding classification performance in environmental prediction applications.

Table 5. Feature Importance Ranking

Predictor Variable	Importance Score
Rainfall Intensity	0.38
Flood Proximity	0.29
Sentinel-1 Backscatter (σ^o)	0.12
NDWI	0.08
Elevation	0.06
Turbidity	0.04
NDVI	0.03

Feature importance analysis was performed to identify the most influential predictors contributing to model performance. Rainfall intensity emerged as the most influential predictor, accounting for approximately 38% of total predictive importance. Flood proximity ranked second with an importance score of 0.29. Together, these variables contributed approximately 67% of the model's predictive capability, highlighting the dominant role of hydrological processes in flood-induced water quality deterioration.

The trained Random Forest model generated spatially explicit water quality risk maps for the study area. Predicted outputs classified locations into: Safe and Unsafe Risk Zone

The generated risk maps revealed that areas located within low-elevation floodplains and regions characterized by high rainfall intensity exhibited the highest probability of water quality deterioration. The spatial outputs demonstrate the ability of the framework to support proactive environmental monitoring and targeted intervention planning.

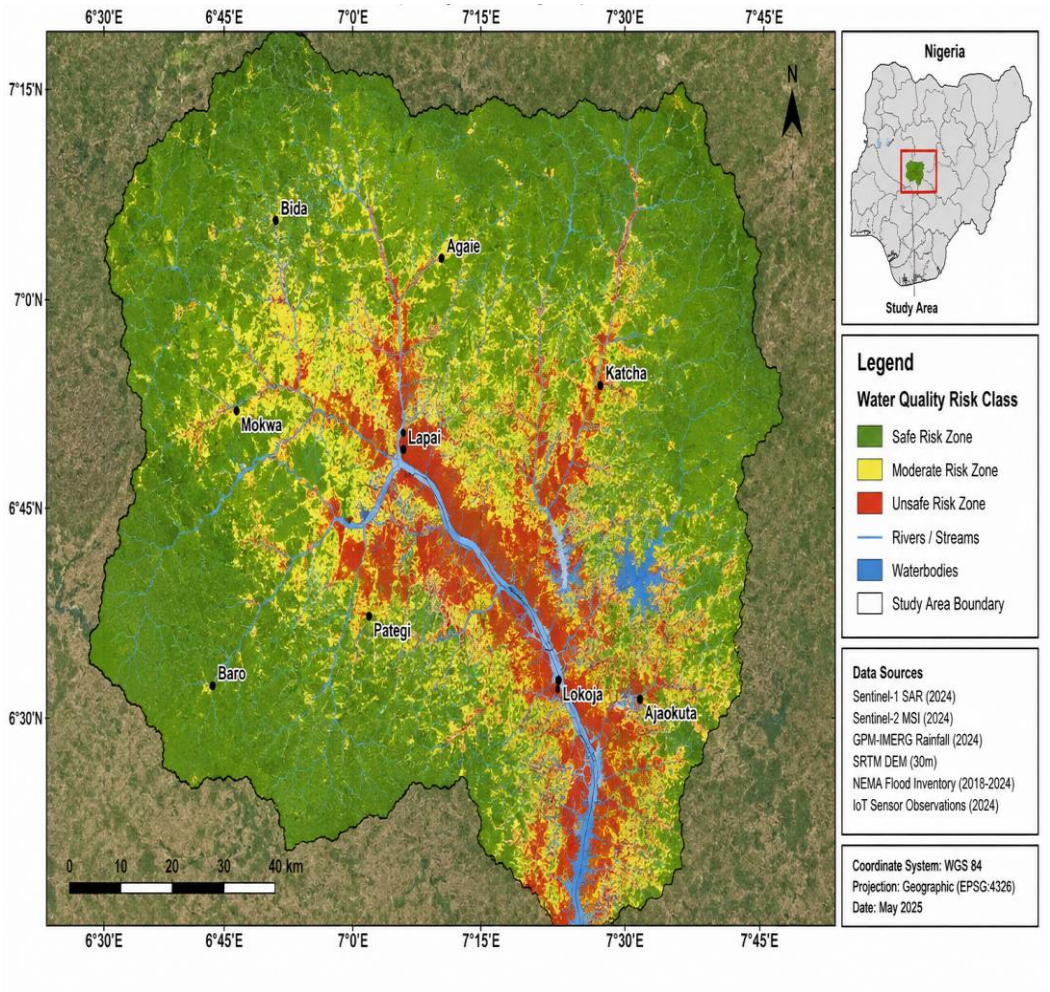

Figure 4. Predicted Water Quality Risk Map

Discussion of Findings

The findings demonstrate that integrating Earth Observation data, hydrological information, and IoT-derived environmental measurements can provide a reliable basis for predicting flood-induced water quality risks. The achieved accuracy and AUC values indicate that the developed framework is capable of supporting proactive environmental monitoring and risk assessment in regions where conventional water quality surveillance remains inadequate.

The dominance of rainfall intensity and flood proximity highlights the fundamental role of hydrological processes in water contamination dynamics. Intense rainfall events increase surface runoff and facilitate the transport of pollutants from agricultural land, urban settlements, and waste disposal sites into water bodies. Likewise, communities located near flood-prone zones are more susceptible to contamination because of prolonged exposure to inundation and sediment deposition processes.

From a public health perspective, these findings are particularly relevant because flood-related deterioration of water quality is frequently associated with increased exposure to microbial and chemical contaminants. Although this study did not directly measure disease outcomes, the ability to identify locations at elevated contamination risk provides a valuable foundation for preventive interventions, targeted water quality testing, and early warning strategies aimed at reducing exposure to unsafe water sources.

The contribution of Sentinel-1 SAR imagery demonstrates the importance of all-weather remote sensing technologies in environmental monitoring. The ability of SAR data to detect inundation under cloudy conditions enhances the reliability of flood-related risk assessments, especially during peak rainfall periods when optical imagery may be unavailable. Similarly, the inclusion of IoT sensor measurements improved local-scale environmental characterization and strengthened the integration of real-time observations with satellite-derived information.

A key contribution of this study lies in the development of a deployable decision-support framework rather than a stand-alone predictive model. By integrating model outputs within a Flask-based platform, the framework provides a practical mechanism for translating environmental data into actionable information that can support water resource managers, environmental agencies, and disaster risk practitioners. This integration enhances the operational value of artificial intelligence for climate adaptation and water safety management.

Conclusion

This study developed a Random Forest-based predictive framework for assessing flood-induced water quality risks in flood-prone regions of Nigeria through the integration of satellite observations, rainfall data, topographic information, historical flood inventories, and IoT sensor measurements. The model achieved strong predictive performance, with an accuracy of 88.4%, an AUC of 0.93, and an F1-score of 0.89, demonstrating its effectiveness for environmental risk classification.

Feature importance analysis identified rainfall intensity and flood proximity as the most influential predictors of water quality risk, emphasizing the significance of hydrological conditions in contamination processes. The integration of geospatial analytics, machine learning, and decision-support deployment provides a scalable framework for enhancing environmental monitoring and water safety management in data-constrained regions.

Beyond its technical contribution, the framework offers practical value for supporting early risk identification, prioritizing monitoring efforts, and strengthening climate resilience in vulnerable communities. The study therefore contributes to the growing application of artificial intelligence for sustainable water resource management and disaster preparedness. Based on the findings of this study, the following recommendations are proposed:

i. Environmental and disaster management agencies should integrate machine learning-based risk prediction systems into existing flood monitoring and water safety programmes.

ii. Investments in IoT-enabled environmental sensing infrastructure should be expanded to improve the availability of real-time water quality observations.
iii. Routine water quality monitoring should be prioritized in areas identified by the model as high-risk zones, particularly during peak rainfall and flooding periods.
iv. Future research should investigate hybrid machine learning and deep learning approaches capable of capturing complex spatio-temporal relationships in environmental datasets.
v. Additional studies should evaluate the transferability of the proposed framework across other flood-prone regions in sub-Saharan Africa and incorporate broader water quality indicators where data are available.

References

Akindejoye, A., Viavattene, C., Priest, S., & Windridge, D. (2025). Identifying social vulnerability profiles for coastal flood using supervised and unsupervised machine learning: A case study of Lekki Peninsula, Lagos, Nigeria. *International Journal of Disaster Risk Reduction*, *127*, 105693.

Aliyu, M., Abubakar, M., Umar, A. Z., Rashed, A. B., & Ismail, R. M. (2025). Analysis of machine learning models for efficient water quality prediction. *Journal of Information and Technology Management (JISTM), 10(40).* https://doi.org/10.35631/JISTM.1040010

Bamiekumo, B. P., Akpobome, E. O., Kemebaradikumo, A. N., Mene-Ejegi, O. O., &Eteh, D. R. (2025). Machine learning-based flood extent mapping and damage assessment in Yenagoa, Bayelsa State, using Sentinel-1 and 2 imagery (2018–2022). *Discovery Nature, 2, e2dn1041*. https://doi.org/10.1007/s44372-025-01041-3

Egbuhuzor Udechukwu Peter, & Okoro Enyinnaya Okoro. (2024). Enhancing rainfall-runoff prediction accuracy using artificial neural networks: A case study of Bida, Nigeria. *Applied Sciences, Computing and Energy, 1*(1), 157–172.

Eteh, D. R., Japheth, B. R., Akajiaku, C. U., Osondu, I., Mene-Ejegi, O. O., Nwachukwu, E. M., Oriasi, M. D., Omietimi, E. J., & Ayo-Bali, A. E. (2025). Assessing the impact of climate change on flood patterns in downstream Nigeria using machine learning and geospatial techniques (2018–2024) [Preprint]. *Research Square*.https://doi.org/10.21203/rs.3.rs-1234567/v1

Ezenwa, E. E., Ogbodo, O. O., &Nnadozie, A. A. (2023). Run-off modelling of pervious and impervious areas using coupled SWAT and a novel machine learning model in Cross-Rivers State, Nigeria. *Water Science & Technology, 88*(7), 1893–1910. https://doi.org/10.2166/wst.2023.287

Ighile, E. H., Shirakawa, H., & Tanikawa, H. (2022). Application of GIS and machine learning to predict flood areas in Nigeria. *Sustainability*, 14(9), Article 5039. https://doi.org/10.3390/su14095039

Mohamed, G. A., Udochukwu, M. O., Enokela, S. O., &Kantiok, Y. D. (2024). Development of a machine learning predictive model for identification of key water quality

pollution parameters at River Benue. *International Journal of Ecology and Environmental Sciences, 6*(3), 97–106.

Obonyano, K. N., Okoye, F. A., Nwobodo-Nzeribe, H. N., &Oleka, C. V. (2024). Machine learning-based flood prediction in Delta State, Nigeria: A data-driven approach. *ExploreMatics Journal, 5(2), 1–15.*

Ofualagba Mamuyovwi Helen, & Okoh Lulu. (2025). Analyzing, modelling and development of a flood hazard prediction system using hybrid machine learning algorithms. *International Journal of Modelling & Applied Science Research (IJMASR), 7*(9), 1–18.

Ogbuene, E. B., Eze, C. A., Aloh, O. G., Oroke, A. M., Udegbunam, D. O., Ogbuka, J. C., Achoru, F. E., Ozorme, V. A., Anwara, O., Chukwunonye, I., Nebo, A. N., & Okolo, O. J. (2024). Application of machine learning for flood prediction and evaluation in southern Nigeria. *Atmospheric and Climate Sciences,* 14, 299–316. https://doi.org/10.4236/acs.2024.143019

Oloruntoba, K. R., Taiwo, K., &Agbogun, J. B. (2023). Flood prediction in Nigeria using ensemble machine learning techniques. *Ilorin Journal of Science*, 10(1), 44–61. https://doi.org/10.54908/iljs.2023.10.01.004