University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Department of Computer Science, University of Ibadan, Ibadan, Nigeria

Volume 14 No. 1, June, 2025

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/

University of Ibadan
Journal of Science and Logics in ICT Research

A Machine Learning Approach to Flood Prediction

[1] D. Otoosakyi., [2*] I. Adinya., and [3] E. S. Taiwo.

1,2Department of Mathematics, University of Ibadan – Nigeria
3Department of Mathematics, University of Ibadan – Nigeria
Faculty of Business and Economics, The University of Winnipeg, Winnipeg, MB R3B 2E9
**Corresponding author's email address: iniadinya@gmail.com*

Abstract

Climate change, driven by both natural processes and human activities, has significantly disrupted living conditions across many countries. Among its most devastating effects is flooding, which impacts millions of people globally. Predicting the timing and severity of future floods remains a major challenge. This study adopts a data-driven methodology, employing machine learning techniques to forecast both the location and magnitude of floods based on historical flood data from Africa. We also investigate the most appropriate probability distribution models for recorded precipitation levels. Our findings indicate that, although Africa is a geographically distinct region that has received limited attention in the literature, its rainfall patterns can be effectively modeled using well-established probability distributions. Additionally, we identify the weeks with the highest and lowest rainfall as significant risk factors among various predictors of flooding. Our analysis further demonstrates that the accuracy of flood predictions is highly dependent on the choice of machine learning algorithm; with the optimal model, we achieve a prediction accuracy of approximately 85% for flood occurrence in targeted areas. These findings suggest that while certain flood predictors in Africa align with those commonly observed in other regions, region-specific factors must still be considered

Keywords: *disaster risk, flood, machine learning, prediction, rainfall analysis*

1. Introduction

The World Health Organization (WHO) identifies floods as the most recurring natural disaster (WHO, 2021). A flood occurs when water inundates land that is typically dry, and its consequences are often catastrophic—causing loss of life, property damage, and environmental degradation (Mind'je et al., 2019). Globally, floods are the most common severe weather event, affecting over 250 million people each year and costing billions of dollars in damages (Matias, 2018). The increasing frequency and impact of flood events pose a growing threat to sustainable development (Nwigwe & Emberga, 2014).

For instance, the 2022 floods in Nigeria had more severe impacts than those in 2012, affecting 2.8 million people and increasing fatalities by approximately 65% (IFRC n.d.; UNDP, 2023). Similarly, the floods induced by Cyclone Idai in 2015 and again in 2019 devastated Malawi (News, 2015). These events illustrate the increasing threat of floods, exacerbated by climate change and human development.

Floods can be categorized into flash floods, river floods, and coastal floods. Each type has distinct causes, including heavy rainfall, snowmelt, and storm surges. Flash floods, the most dangerous, combine high velocity with destructive force. Their likelihood is increased by both natural and human-induced factors, such as topography, impermeable surfaces, inadequate drainage, and deforestation (ResearchClue, 2020).

Between 1998 and 2017, floods accounted for approximately 43% of major global disasters (UNDRR, 2018). Countries with significant populations exposed to flood risks span continents, with notable examples in Asia, Africa, and South America (Luo et al., 2015). In Africa, especially Nigeria, flooding has emerged as a recurrent crisis that impairs

D. Otoosakyi, I. Adinya. and E. S. Taiwo (2025). A Machine Learning Approach to Flood Prediction. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 14 No. 1, pp. 44 - 57

social structures, economic activities, and food security.

Flood damage is not limited to property and infrastructure; it often involves massive human displacement, fatalities, and long-term economic loss. In 2022 alone, floods in Pakistan affected 33 million people and caused $30 billion in damages (Margesson & Kronstadt, 2022). Bangladesh, South Africa, Nigeria, and the Democratic Republic of Congo also experienced significant losses (Abdullah et al., 2022; Jewkes et al., 2023; Oguntola, 2022; UNOCHA, 2022).

Despite technological advancements in weather forecasting and disaster preparedness, floods remain unpredictable due to their stochastic nature. Prediction failures are often linked to poor information dissemination, inadequate infrastructure, and flawed decision-making processes. As such, there is a need for more accurate and localized predictive models.

While existing studies provide robust flood prediction models (e.g., Mosavi et al., 2018), there is a gap in research focusing on Africa. This study aims to fill that gap by applying machine learning algorithms to historical data to predict flood-prone areas and their likely severity. We demonstrate that rainfall patterns in Africa can be modeled using probability distributions such as Normal, Log-normal, Gamma, Gumbel, and Weibull. Our results show that machine learning models can predict floods with high accuracy, offering valuable insights into regional mitigation strategies.

1.1 Flooding Events Around the World

Flooding affects countries across all continents, leading to severe humanitarian and economic crises. According to Luo *et al.* (2015), fifteen countries account for 80% of the global population affected by river floods annually. These include countries in Asia, Africa, and South America (see Table 1). In Bangladesh, for example, one-third of the country is submerged during monsoon flooding (Bhuiyan & Al Baky, 2014; Coca, 2020). India's vulnerability to floods is attributed to monsoon rains, tropical storms, and siltation of rivers, with floods impacting nearly 84% of its GDP annually (Rehman *et al.*, 2019).

In Africa, Egypt and Nigeria suffer from flash and fluvial floods respectively. In Egypt, urban growth and mismanagement have exacerbated flood risks (Saber et al., 2020; Arnous et al., 2022). In Nigeria, structural inadequacies, such as poor drainage and planning laws, contribute significantly to recurring floods (Echendu, 2020).

These patterns reveal that while global flood risk factors may be shared, geographic and socio-economic differences require region-specific predictive strategies. This study focuses on Africa, particularly rainfall-induced fluvial floods, and investigates machine learning methods for prediction and mitigation.

The remainder of this paper is structured as follows: Section 2 reviews related literature and flood events; Section 3 presents methodology; Section 4 details the results and discussion; Section 5 concludes the paper.

2. Related Works

2.1 Flood Prediction Using Machine Learning

Flood prediction research increasingly employs machine learning (ML) algorithms that detect patterns in historical data. Commonly used algorithms include Decision Trees (DT), Random Forests (RF), Linear Regression (LinReg), Logistic Regression (LR), XGBoost, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Mind'je *et al.* (2019) used logistic regression with ten predictors (e.g., elevation, slope, NDVI, rainfall) on historical flood data in Rwanda, achieving a 79.8% prediction accuracy. Talukdar and Pal (2020) applied Markov Chain Cellular Automata and ANN to forecast floodplain transformation in the India-Bangladesh region, achieving AUC values of 84.4% and 86.8%.

Talukdar *et al.* (2020) developed an ensemble model using REPtree, RF, M5P, and Random Tree. The M5P model, combined with bagging, showed superior performance with sensitivity of 86.25%, specificity of 88.75%, and AUC of 0.97. Rahman *et al.* (2019) assessed flood susceptibility in Bangladesh using a combination of ANN, logistic regression, frequency ratio, and AHP. Logistic

regression achieved the highest accuracy (86%).Cui and Cui (2020) employed linear regression to model spring floods in Canada, identifying four significant predictors and achieving an R^2 of 63%.

Machine learning model selection depends heavily on data availability, study scope, and prediction objectives (Obarein & Amanambu, 2019). This study adopts an exploratory approach, evaluating multiple ML models and statistical distributions to optimize flood prediction based on African rainfall data.

Table 1 Annual Expected Population Affected by River Floods

Countries	Population (in millions)	Continent
India	4.84	Asia
Bangladesh	3.48	Asia
China	3.28	Asia
Vietnam	0.93	Asia
Pakistan	0.71	Asia
Indonesia	0.64	Asia
Egypt	0.46	Africa
Myanmar	0.39	Asia
Afghanistan	0.33	Asia
Nigeria	0.29	Africa
Brazil	0.27	South America
Thailand	0.25	Asia
Congo D.R.	0.24	Africa
Iraq	0.19	Asia
Cambodia	0.19	Asia

Source: World Resources Institute [WRI], 2015

3. Methodology

In this section, we analyze the precipitation data used for this study and examine the pattern of the probability distributions with histogram plots and kernel density estimation. Then, we present the most suitable probability distribution for precipitation data.

Our dataset comes from real-world secondary data of historical floods obtained from the repository of Zindi—the first data science competition platform in Africa. The dataset consists of major flooding that hit Southern Malawi with cyclone Idai in 2015 and 2019. The location map is partitioned into approximately 1km2 rectangles, assigned with a target value, a fraction (percentage) of the rectangle flooded in 2015. The data is well-labeled, specifying whether a flood happened in an area and the proportion of the surface area flooded. Thus, the dataset consists of 16466 rows (entries). In this paper, we only consider the 2015 flood extent data. We train our machine learning models with 80% of the dataset and use the remaining 20% to measure the accuracy of the models. The selected features (variables) of interest include the following:

a) Elevation – the mean elevation over the rectangle, based on the NASA Shuttle Radar Topography Mission (SRTM) Digital Elevation 30m dataset in Google Earth Engine.
b) Dominant Land Cover Type – the surface cover on the ground, such as water, vegetation, bare soil, and urban infrastructure.
c) Weekly precipitation – historical rainfall data for each rectangle for 18 weeks beginning two months before the flooding, based on the Tropical Rainfall Measuring Mission (TRMM) dataset in Google Earth Engine.
d) Coordinates – the location's longitude (Y) and latitude (X), representing a rectangle 0.01 degrees on each side, centered on that X − Y location.
e) Target – the proportion of the flooded rectangle, with a value between 0 and 1.

3.1 Analyses of Precipitation Data

A major contributing factor to floods is rainfall (precipitation). Most existing flood prediction models focus on severe rainfall and hurricanes. However, some studies (Cui and Cui 2020) demonstrate that the amount of snow on the ground (snow melt) can also be a significant predictor. Rainfall data used in developing flood prediction models are typically obtained from weather stations or derived from remote sensing datasets. Some researchers argue that gridded and modeled rainfall data may fail to accurately capture climate variability, leading to uncertainties in flood susceptibility modeling when compared to station-based data (Obarein and Amanambu, 2019; Mind'je et al., 2019). Conversely, other studies suggest that remotely sensed rainfall data provide a reliable source, effectively capturing the seasonal patterns of precipitation (Dunning et al., 2016). Nonetheless, the continued use of

precipitation gauges as direct measurement instruments remains strongly recommended.

Distribution Pattern of the Precipitation Data

In some machine learning models, continuous probability distributions are frequently employed, particularly in the distribution of numerical input and output variables and in the spread of model errors.

Normal, Log-normal, Gamma, Gumbel, and Weibull are commonly used probability distributions in rainfall analysis. The Normal distribution is widely used because of its association with the central limit theorem and capacity to describe many natural occurrences. It is well-known that the Gamma distribution is a member of the two-parameter family of continuous probability distributions. The typical "exponential" and "chi-squared" distributions are special cases of the Gamma.

Table 2 presents the summary statistics of precipitations for each week. The table shows that the maximum average rainfall was experienced in Week 9, with an average precipitation of 58.86mm. Following this is a sharp decline in Week 10, with an average rainfall value of 1.25mm, which indicates that the amount of rainfall will decline in the following weeks before the flood (see Figure 1). But this is not so in the following weeks (Week 11 and Week 12). The last week (Week 17) before the flooding event saw the least average precipitation for the entire period in the dataset, with an average precipitation of 0.33mm.

Although the summary statistics reveal potential fluctuations in precipitations, examining the probability distribution of rainfalls for each week in all locations is necessary. Such analyses will enable us to investigate possible changes in the distributional pattern of precipitations over the weeks before a flood event.

Figure 2 presents the histogram plots for precipitation patterns for some selected weeks (e.g., weeks 6, 10, 11, and 14). The plots suggest that weekly precipitations exhibit different probability distributions, most positively skewed. For instance, weeks 10 and 11 seem to have similar distributional trends; they are highly positively skewed.

It is usually challenging to identify the exact probability distribution of a random variable using a histogram plot. Therefore, we estimate the probability distributions using a Kernel Density Estimator (KDE) function.

3.1.1 Kernel Density Estimation (KDE)

KDE, a non-parametric technique, estimates the random variable's probability density function. It utilizes a kernel function, $\mathcal{K}$, to estimate an unknown probability density function. Unlike a histogram that counts the number of data points in random locations, KDE is the sum of a kernel function on each data point.

Table 2 Summary statistics of precipitation data

Precipitation Week	Min. Precip.	Max. Precip.	Mean (μ)	Std. Dev. (σ)
Week 1	0.00	19.35	1.61	4.23
Week 2	0.00	41.02	2.50	8.63
Week 3	0.00	22.02	1.16	4.40
Week 4	1.41	18.87	8.27	4.27
Week 5	3.58	23.04	8.89	3.76
Week 6	1.25	21.75	9.57	4.52
Week 7	7.46	62.43	22.92	13.69
Week 8	15.65	51.20	28.11	7.79
Week 9	30.45	105.28	58.86	16.81
Week 10	0.00	11.10	1.25	1.97
Week 11	14.97	53.01	34.65	7.46
Week 12	13.26	44.34	28.32	8.05
Week 13	0.46	28.56	12.49	7.06

Week 14	0.28	15.72	3.80	2.67
Week 15	6.73	36.97	17.07	6.07
Week 16	3.28	25.71	9.11	4.57
Week 17	0.00	4.95	0.33	1.01

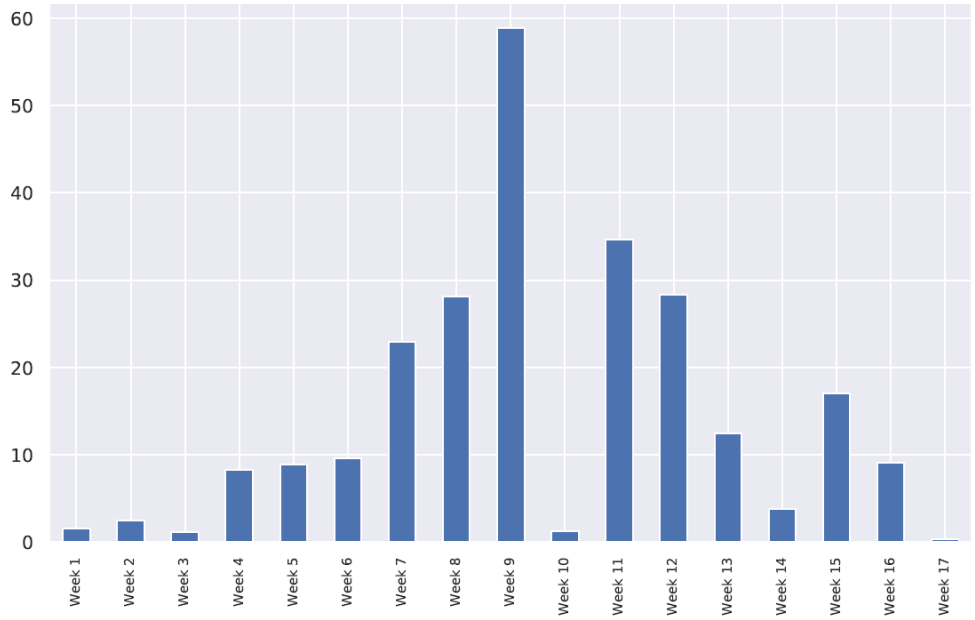


Figure 1 Bar plot of Average Precipitation for each week

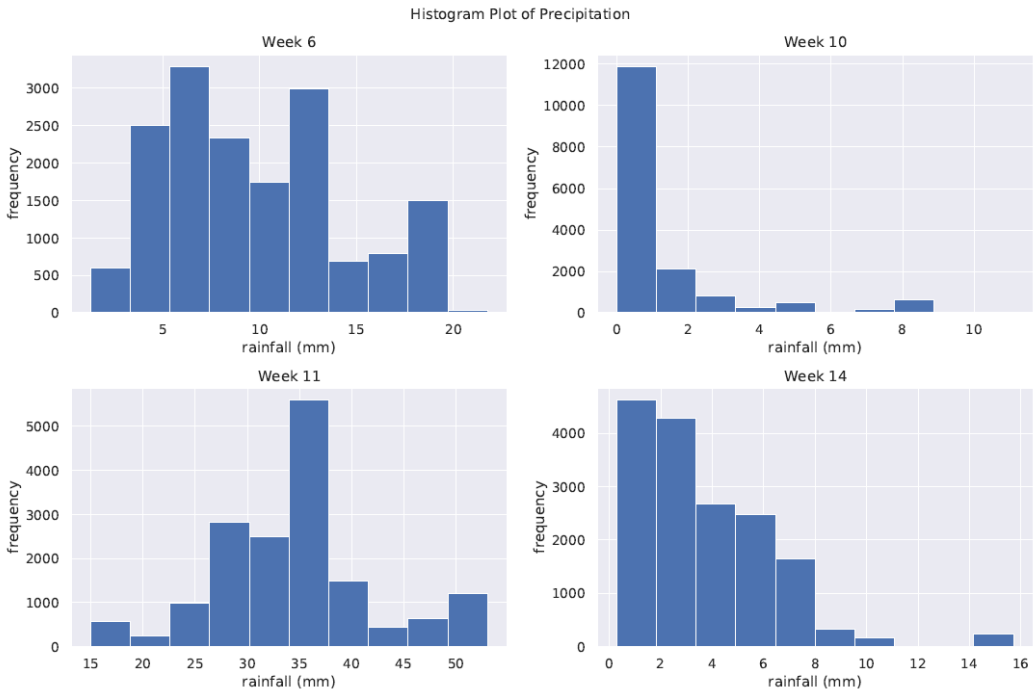


Figure 2 Histogram plot of precipitations

It is a basic data smoothing problem in which population inferences are drawn from a small sample of data.

Consider independent and identically distributed (i.i.d.) samples, $(x_1, x_2, \ldots, x_n)$, sampled from some univariate distribution with an unknown density f at any given point x. We would like to estimate the shape of this function f. Its kernel density estimator is:

$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x - x_i}{h}\right)$$

where $\mathcal{K}$ is the kernel—a non-negative function—and $h > 0$ is a smoothing parameter called the bandwidth. A kernel with subscript h is called the scaled kernel and defined as:

$$\mathcal{K}_h(x) = \frac{1}{h}\mathcal{K}\left(\frac{x}{h}\right)$$

The following are typical properties of the kernel function (Seabold and Perktold 2010):

1. It is symmetrical $\mathcal{K}(x) = \mathcal{K}(-x)$.
2. It can be normalized such that $\int_{-\infty}^{\infty} \mathcal{K}(x)dx = 1$.
3. It is monotonically decreasing, such that $\mathcal{K}'(x) < 0$ when $x > 0$.
4. The expected value equals zero (i.e., $\mathrm{E}[\mathcal{K}] = 0$).
5. Machine learning applications can make use of the Kernel Density Estimation approach. Because parameters in the estimation function define the kernel's scope, a neural network can begin to train itself to correct its estimations and generate more accurate results. The bandwidth and amplitude estimations are continuously updated while the estimation process repeats itself, increasing the accuracy of the calculated probability density curve.

KDE was used to estimate each of the weekly precipitation data in the dataset and the type of probability distribution function for each variable. The density plots in (Figure 3) show the precipitation data estimation for selected weeks (as in Figure 2). The precipitation data shows that all the variables exhibit a non-normal distribution. The result correlates with the outcome of the histogram plots (see Figure 2).

3.2 Fitting Probability Distributions on Precipitation Data

Comparing the histogram of the data with a probability distribution function (pdf) of a known distribution is a fundamental approach typically used to determine the underlying distribution that could have created a data set (e.g., normal distribution). The distribution's parameters, however, are unknown, and there are many different distributions. As a result, an automatic method of fitting many distributions to the data would be beneficial, which is what is implemented here.

We analyze the precipitation data to identify the best-fit probability distribution for each study period. The goal is to find a distribution that suits the data well. The distribution that gives a close fit is supposed to lead to good predictions. The best-fit probability distribution was determined using the least square method and based on the minimum deviation between actual and estimated values.

Finding Best-Fit PDF

The best-fit probability distribution for the precipitation data was obtained using a package in Python. A class in the Fitter package identifies the distribution from which a data sample is created. It employs 80 SciPy distributions and allows the plotting of the results to see which distribution is the most likely and which parameters are the best. The best-fit probability distribution is that with the minimum deviation between actual and estimated values.

Table 3 contains the result of the fitted probability distribution functions for each weekly precipitation. Each table includes the top 5 probability distributions out of the 80 fitted distributions. The best-fit probability distribution (in the first row of each table) is selected based on the least sum of square errors (SSE) as the default metric. The precipitation data for the period originates from one of Lomax, Wald, Double Gamma (dgamma), Skew Normal (skewnorm), Exponential (expon), Cauchy, Laplace, Half Cauchy, Semi Circular, Anglit, and Half Normal (halfnorm) distributions. The precipitation data of some weeks share the same distribution, such as Weeks 4,5 and 11 (Double Gamma), Weeks 7, 16, and 17 (Exponential), Weeks 1 and 2 (Lomax), and Weeks 8 and 15 (Cauchy). In addition, most tables have a particular case of either Gamma or Weibull

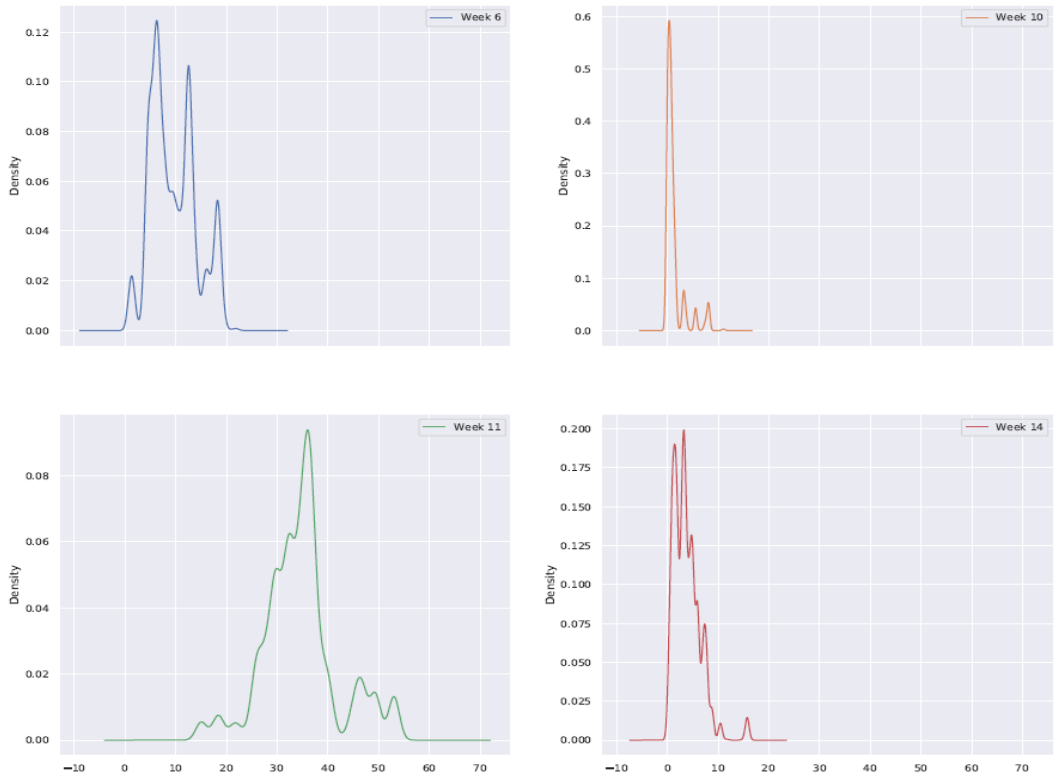


Figure 3 Kernel Density Estimation of the precipitation data

distributions, which are probability distributions commonly used to model rainfall data. Each table presents measures used in selecting the best-fit probability distributions. The Akaike Information Criterion (AIC) evaluates how well each probability density function fits the precipitation data. Based on the maximum likelihood estimate, AIC calculates the relative information value of the model. The AIC is computed using the following:

$$AIC = 2k - 2\ln(\hat{L}),$$

where k is the number of parameters in the density function and L^ is the likelihood estimate.

AIC scores with fewer parameters are better than those with more parameters. For any two models explaining the same amount of variation, the preferred model (i.e., better-fit model) usually has fewer parameters. The Bayesian Information Criterion (BIC) is similar to the AIC in selecting best-fit models. The BIC is obtained by evaluating

$$BIC = k\ln(n) - 2\ln(\hat{L}),$$

where $\hat{L}$ is the value that maximizes the likelihood function of the model, n is the number of data points in the sample size, and k is the number of parameters estimated by the model. A lower BIC value is preferred.

The Kullback-Leibler Divergence (KL_{div}) score, often known as the KL divergence score, measures how much one probability distribution differs from another.

$$\boldsymbol{KL_{div}(p,q)} = \begin{cases} \boldsymbol{p\log(p/q) - p + q} & \boldsymbol{p > 0, q < 0} \\ \boldsymbol{q} & \boldsymbol{p = 0, q \geq 0} \\ \infty & \boldsymbol{Otherwise} \end{cases}$$

Thus, for distributions P and Q of a continuous random variable, (KL_{div}) is defined by the integral

$$\boldsymbol{KL_{div}(p,q) = D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx}$$

The Probability Density Function plot of each precipitation week data in Table 3 is shown in Figure 4. The top legend in each plot indicates the best-fit probability density function.

Table 3 Fitted probability distributions table for precipitation data

PDF	SSE	AIC	BIC	KL div
skewnorm	0.088494	238.810197	−199767.227754	∞
moyal	0.088790	237.147174	−199721.885994	∞
kstwobign	0.089376	235.383197	−199613.665203	∞
gumbel r	0.090281	237.586676	−199447.771977	∞
genlogistic	0.090978	240.133587	−199311.337092	∞

(a) Week 6

PDF	SSE	AIC	BIC	KL div
skewnorm	0.088494	238.810197	−199767.227754	∞
moyal	0.088790	237.147174	−199721.885994	∞
kstwobign	0.089376	235.383197	−199613.665203	∞
gumbel r	0.090281	237.586676	−199447.771977	∞
genlogistic	0.090978	240.133587	−199311.337092	∞

(b) Week 10

PDF	SSE	AIC	BIC	KL div
dgamma	0.021283	295.603862	−223232.035078	∞
dweibull	0.021351	295.363435	−223178.913023	∞
gennorm	0.021503	295.445463	−223062.405693	∞
laplace	0.021566	293.934913	−223023.649583	∞
hypsecant	0.021576	293.733847	−223015.965548	∞

(a) Week 11

PDF	SSE	AIC	BIC	KL div
halfnorm	0.109094	265.923054	−196331.079234	∞
moyal	0.116474	263.189679	−195253.162883	∞
gumbel r	0.117977	274.493330	−195042.100244	∞
rayleigh	0.120038	295.590630	−194756.832298	∞
maxwell	0.124257	302.296821	−194188.095566	∞

(a) Week 14

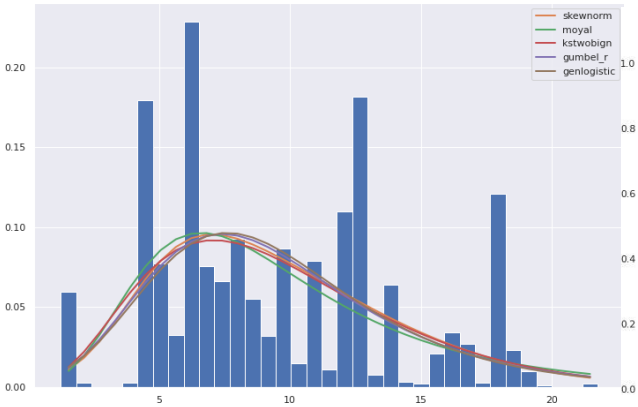
4. Results and Discussion

We partition the dataset into 80:20 train-test split. The precipitation features were standardized using the z−score standardization method after splitting the datasets into train and test sets. We determine the mean (μ) and standard deviation (σ) of the train data, then obtain a standard score for a sample x using

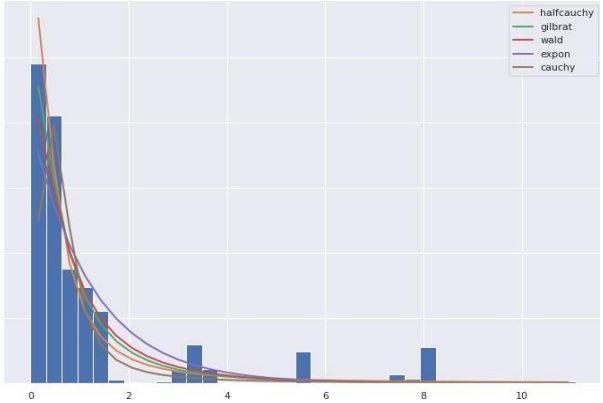
$$z = \frac{x - \mu}{\sigma}$$

We feed the preprocessed data into six machine-learning algorithms to train models suitable for flood prediction. Three of these algorithms are variants of the boosting algorithm: CatBoost, Extreme Gradient Boosting Regressor (XGBoost), and Light Gradient Boosting Regressor (LGB). The other three neural network schemes are Multilayer Perceptron, Support Vector Regressor, and Random Forest.

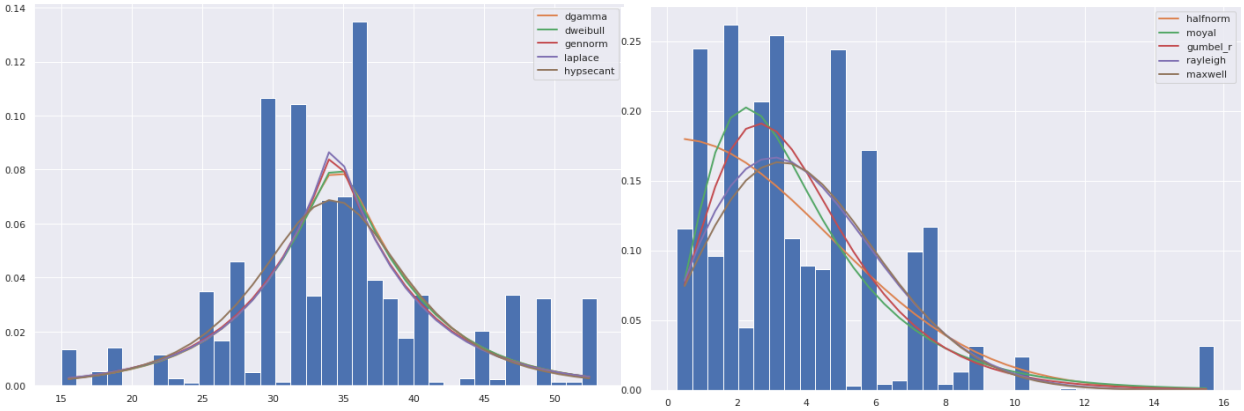
We use the Python package running on a Google Colaboratory cloud notebook to implement the algorithms on the data. We consider the default parameters of each algorithm when training the models. After learning, the trained models were evaluated to see how well the algorithms could learn from the data.



(a) Week 6



(b) Week 10



(c) Week 11 (d) Week 14

Figure 4 Plot of the best-fitted probability density function

4.1. Model Evaluation

Since the target variable of the dataset is continuous, the performance metrics such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Co-efficient of Determination (R^2) were applied to evaluate how well the proposed models predict floods. Table 4 shows the performance of each model based on the selected metrics.

From Table 4, the CatBoost model performs best with the lowest MSE and highest R^2 value, followed by Random Forest. The SVR model has the worst performance. The implication of the R^2 value from the CatBoost model is that the percentage of the flood (target variable) accounted for by the predictors (precipitation data, elevation, and coordinates) is 56.5%. In comparison, other factors may explain the remaining 43.5%; these include temperature, wetlands, and human activities. The RMSE value implies that the prediction that a particular location will be flooded may seem 15.1% off the target. These interesting results reveal the risk factors for floods in Africa and display a low probability of off-target prediction when an appropriate machine-learning algorithm is deployed. However, an improvement procedure is recommended to further reduce the off-target prediction probability.

4.2. Prediction Plots

The prediction plot lets us visualize and compare the actual and predicted values to see how well the models perform. The prediction plot for each of the models is shown in Figure 5. The blue color lines indicate the actual values of the target (y), while the orange color lines ($\hat{y}$) indicate the predicted values. From this plot, the SVR (Figure 5b) did not learn well and thus did not capture the data, thereby underfitting the data. The Catboost model performs best (Figure 5f), capturing the data, although some noise is noticeable

4.3 Feature Importance Plot

In this section, we present the ranking of the importance of feature variables. Based on their relevance, we use the feature importance technique to get a score of the predictors used in training the model. The ranking of the predictors based on the higher score from our best model (CatBoost) is obtained using a wrapper class contained in the algorithm. A higher score implies that the specific feature will significantly impact the model used in predicting the flood extent.

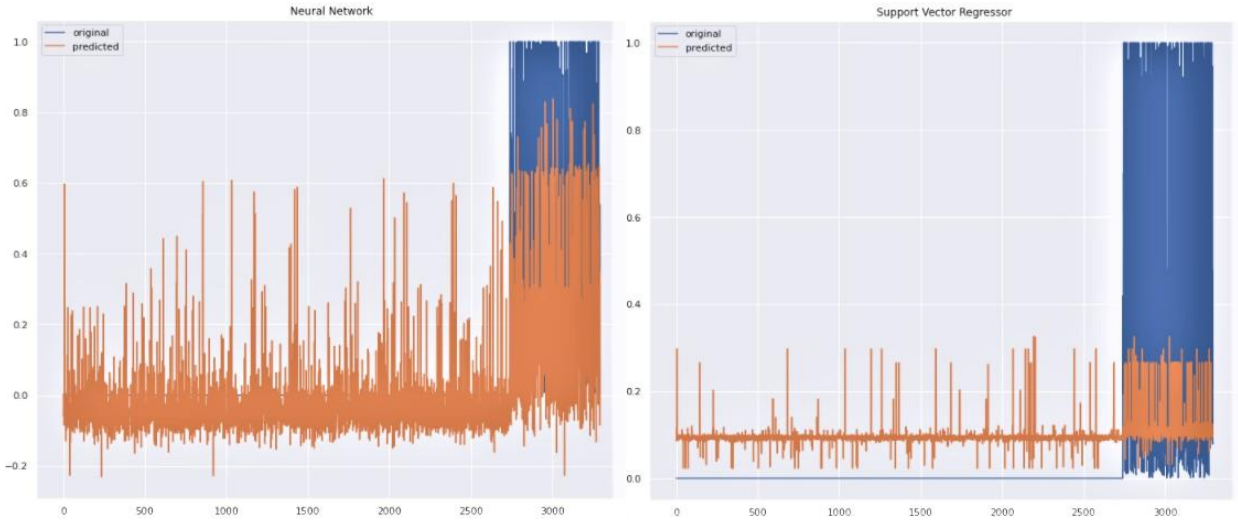
Table 4 Model performance metrics table

Model	**MAE**	**MSE**	**RMSE**	$\boldsymbol{R}^2$
SVR	0.140	0.048	0.220	0.075
XGBoost	0.073	0.027	0.166	0.472
Catboost	0.060	**0.022**	**0.151**	**0.565**
LGBoost	0.061	0.026	0.163	0.494
RandomForest	**0.058**	0.025	0.160	0.51
MultiLayerPerceptron	0.094	0.032	0.179	0.38

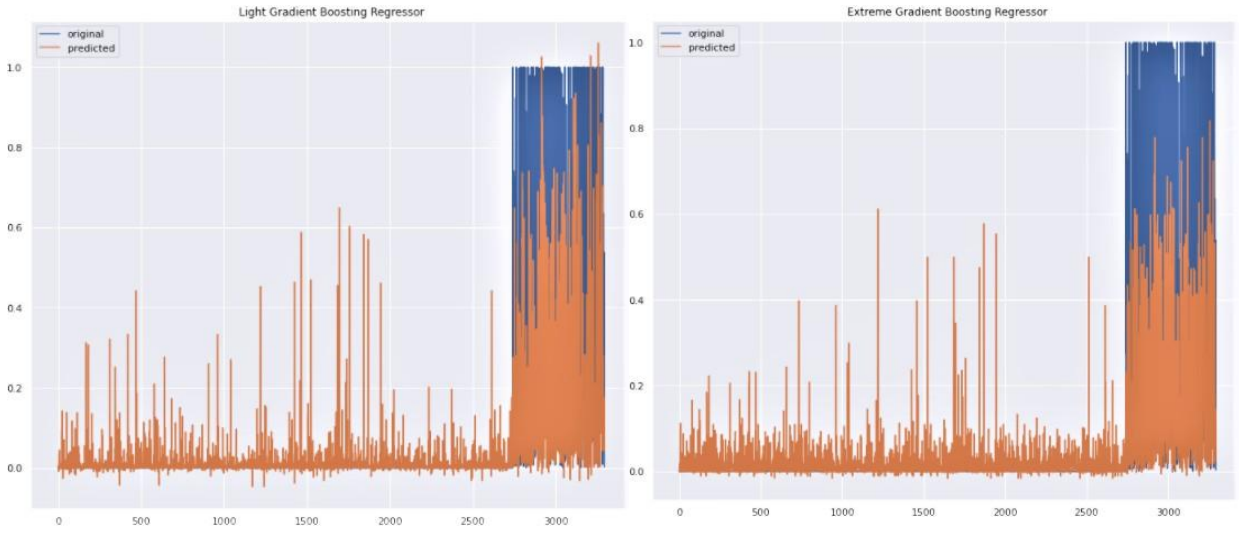
Figure 6 is a visual representation of the critical features of the CatBoost model. It shows the vital contributing predictors in the model in descending order. The most influencing predictor in the model is "elevation," with a score greater than 50. Other important predictors are the location (X and Y coordinates), Week 17 (the week with the lowest average rainfall and the preceding week to the flooding period), Week 9 (the week with the highest average rainfall) precipitation data, and Land Cover Type. Other predictors have a very low score, with Week 10 precipitation data being the least, indicating that the contributions of these low-rank features (predictors) are not influencing the model significantly. Hence, dropping these variables may not necessarily affect the model's performance in predicting flood extent in the targeted location. It may increase the model's efficiency by reducing the prediction error rate (bias).

5. Conclusion

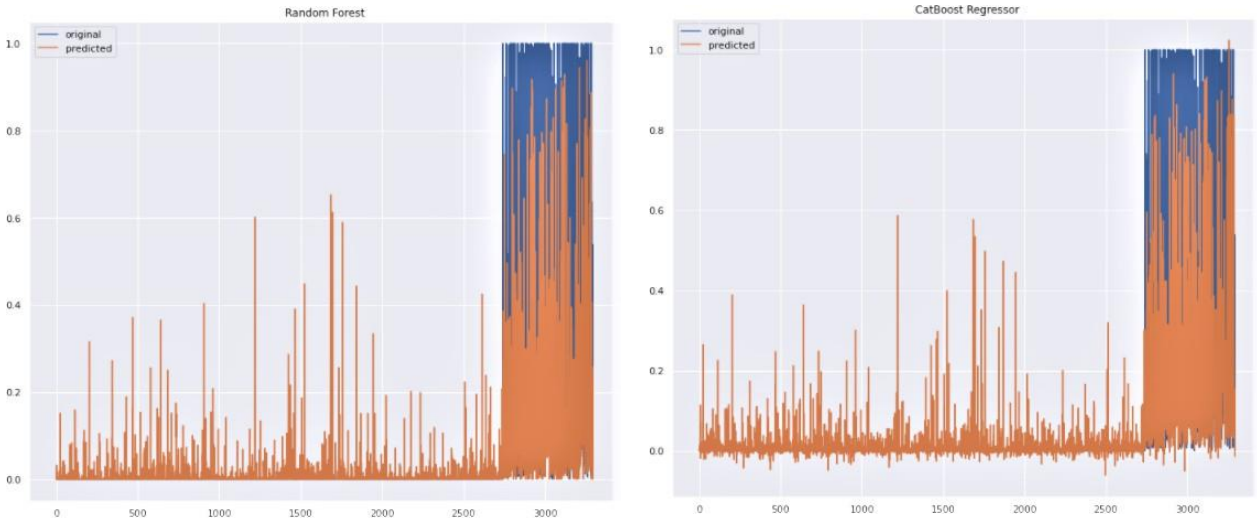
In this paper, we consider mitigating flood occurrences by building predictive machine learning models that learn from historical flood patterns to make predictions of the location and extent of floods. We examine the probability distribution pattern of rainfall with the histogram and kernel density estimation plots and find that most precipitation features have different probability distributions, except for a few that share the same distribution pattern. We train the data using six machine-learning algorithms and evaluate model performances based on four metrics. We find that the Catboost performs best among the selected algorithms; adding topographical information could further improve prediction performance. Our findings reveal that 56.5% of the variability observed in the target variable (proportion of the areas flooded) is explained by the predictors (i.e., elevation, coordinates, precipitation features, and land cover types) in the regression model. Still, predictions of floods at specific locations may be about 15% off the target. We determine that the most critical risk factors are the area's height (elevation) and geographic location (coordinates). Additionally, the week preceding floods with the most and least rainfall is crucial. Therefore, decision-makers should reevaluate their prediction models and update the level of alertness as appropriate during the identified time frame.



(a) Multilayer Perceptron (b) SVR



(c) LGB Regressor (d) XGBoost Regressor



(e) Random Forest (f) Catboost Regressor

Figure 5 Prediction plots comparing actual values against predicted values

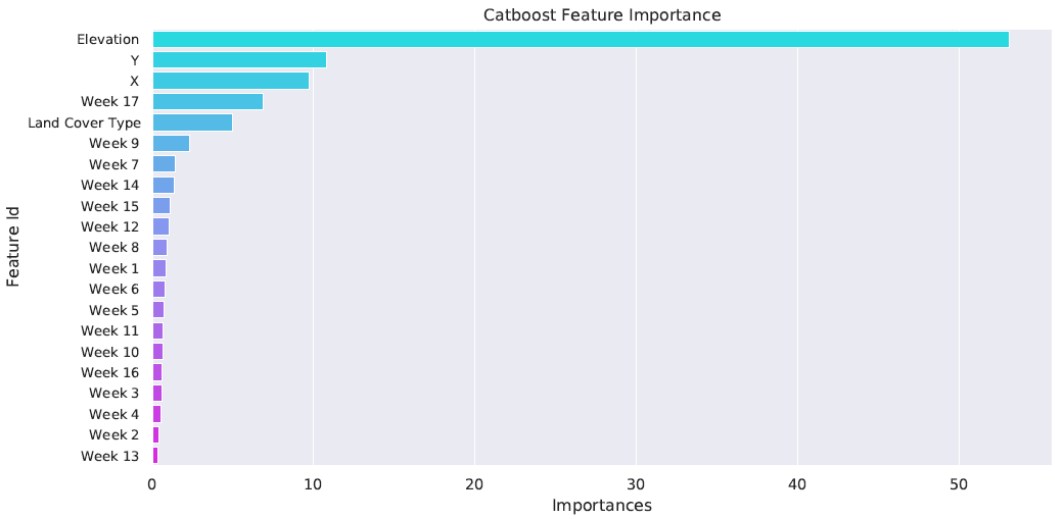


Figure 6 Rank of the important features

This study demonstrates that data-driven techniques for flood prediction in Africa are promising, with sufficient datasets and relevant features. Our results are helpful when determining some cost-effective and optimal evacuation decision policies proposed in the literature (see Taiwo *et al.* 2019) to mitigate the massive loss of lives and properties during floods.

Further improvement can be made to the model for better prediction performance by including other relevant features such as climatic data (e.g., temperature) and topographical data (e.g., distance to wetlands). In addition, predictors with low relative importance (score) may be dropped; this often influences better prediction of machine learning models. While this study focuses on the case of Africa, our approach and insights provide leverage for decision-making at the global level platforms. However, the uniqueness of the case study location may limit the potential for the generalization of our results.

Limitations: Because of the limited access to data, we could not extend our exploration of flooding events on the African continent beyond the case of Malawi. This is a limitation of our study.

References

[1] Abdullah A, Ahmed N, Mahboob A (2022) Community-led housing recovery needs assessment: North and north-eastern flood 2022: Bangladesh. UNDP (December 2022): Accessed 25 December 2022 from https://www.undp.org/bangladesh/publications/community-led-housing-recovery-needs-assessment-north-and-north-eastern-flood-2022-bangladesh.

[2] Arnous MO, El-Rayes AE, El-Nady H, Helmy AM (2022) Flash flooding hazard assessment, modeling, and management in the coastal zone of ras ghareb city, gulf of suez, Egypt. *Journal of Coastal Conservation* 26(6):77.

[3] Bhuiyan SR, Al Baky A (2014) Digital elevation based flood hazard and vulnerability study at various return periods in sirajganj sadar upazila, bangladesh. *International journal of disaster risk reduction* 10:48–58.

[4] Coca N (2020) Flooded asia: Climate change hits region the hardest. nikkei asia. https://asia.nikkei.com/Spotlight/AsiaInsight/Flooded-Asia-Climate-change-hits-region-the-hardest.

[5] Cui C, Cui L (2020) An innovative flood prediction system using improved machine learning approach. *The Canadian Science Fair Journal 2(2)*.

[6] Dunning CM, Black EC, Allan RP (2016) The onset and cessation of seasonal rainfall over africa. *Journal of Geophysical Research: Atmospheres* 121(19):11–405.

[7] Echendu AJ (2020) The impact of flooding on nigeria's sustainable development goals (sdgs). *Ecosystem Health and Sustainability* 6(1):1791735.

[8] IFRC (n.d) Nigeria: Floods. https://www.ifrc.org/emergency/nigeria-floods.

[9] India Today (2015) India is the most flood-prone country in the world. India Today Web Desk https://www.indiatoday.in/education-today/gk-current-affairs/story/india-is-the-most-flood-prone-country-in-the-world-276553-2015-12-10.

[10] Jewkes R, Gibbs A, Mkhwanazi S, Zembe A, Khoza Z, Mnandi N, Washington L, Khaula S, Gigaba S, N¨othling J, et al. (2023) Impact of south africa's april 2022 floods on women and men's lives and gender relations in low-income communities: A qualitative study. *SSM-Mental Health* 4:100255.

[11] Kinghorn J (2017) 3 factors that make flooding in south america worse. AIR https://www.airworldwide.com/blog/posts/2017/4/3-factors-that-make-flooding-insouth-america-worse/.

[12] Levenson M (2021) Severe flooding in guyana prompts extensive relief effort. New York Times. https://www.nytimes.com/2021/06/03/us/guyana-flooding-relief.html.

[13] Luo T, Maddocks A, Iceland C, Ward P, Winsemius H (2015) World's 15 countries with the most people exposed to river floods. World Resources Institute https://www.wri.org/insights/ worlds-15-countries-most-people-exposed-river-floods.

[14] Margesson R, Kronstadt A (2022) Pakistan's 2022 floods and implications for u.s. interests. Congressional Research Service (October 2022): Accessed 25 December 2022 from https://crsreports.congress. gov/product/pdf/IF/IF12211.

[15] Martini A (2020) Weatherwatch: floods across south america after heavy rain. The Guardian. https://www.theguardian.com/news/2020/feb/26/weatherwatch-floods-across-south-america-after-heavy-rain.

[16] Matias Y (2018) Keeping people safe with ai-enabled flood forecasting. Google The Keyword (blog).

[17] Mind'je R, Li L, Amanambu AC, Nahayo L, Nsengiyumva JB, Gasirabo A, Mindje M (2019) Flood susceptibility modeling and hazard perception in Rwanda. *International journal of disaster risk reduction* 38:101211.

[18] Mosavi A, Ozturk P, Chau Kw (2018) Flood prediction using machine learning models: Literature review. Water 10(11):1536

[19] News F (2015) 2015 Floods leave Malawi facing worst food crisis in 10 years. https://floodlist. com/africa/floods-malawi-facing-worst-food-crisis-10-years.
[20] Nwigwe C, Emberga T (2014) An assessment of causes and effects of flood in nigeria. *Standard Scientific Research and Essays* 2(7):307–315.
[21] Obarein OA, Amanambu AC (2019) Rainfall timing: variation, characteristics, coherence, and interrelation- ships in nigeria. *Theoretical and Applied Climatology* 137(3):2607–2621.
[22] Oguntola T (2022) 2022 flood: 603 dead, 1.3m displaced across nigeria – federal govt. Leadership Newspaper (October 2022): Accessed 25 December 2022 from https://leadership.ng/ 2022-flood-603-dead-1-3m-displaced-across-nigeria-federal-govt/.
[23] Rahman M, Ningsheng C, Islam MM, Dewan A, Iqbal J, Washakh RMA, Shufeng T (2019) Flood susceptibility assessment in bangladesh using machine learning and multi-criteria decision analysis. *Earth Systems and Environment* 3(3):585–601.
[24] Rehman S, Sahana M, Hong H, Sajjad H, Ahmed BB (2019) A systematic review on approaches and methods used for flood vulnerability assessment: framework for future research. *Natural Hazards* 96(2):975–998.
[25] ResearchClue (2020) Flooding in nigeria causes, effects and solution, available at https://nairaproject. com/projects/4120.html.
[26] Ritorto D (2013) South american floods: Dozens dead in brazil as mexico also hit. British Broadcasting Corporation (BBC) https://www.bbc.com/news/av/world-latin-america-25514396.
[27] Roy DC, Blaschke T (2015) Spatial vulnerability assessment of floods in the coastal regions of bangladesh. Geomatics, *Natural Hazards and Risk* 6(1):21–44.
[28] Saber M, Abdrabo KI, Habiba OM, Kantosh SA, Sumi T (2020) Impacts of triple factors on flash flood vulnerability in egypt: urban growth, extreme climate, and mismanagement. *Geosciences* 10(1):24.
[29] Seabold S, Perktold J (2010) Kernel density estimator (kde) – statsmodels: Econometric and statistical modeling with Python. 9th Python in Science Conference.
[30] Sowmya K, John C, Shrivasthava N (2015) Urban flood vulnerability zoning of cochin city, southwest coast of india, using remote sensing and gis. *Natural Hazards* 75(2):1271–1286.
[31] Sudha Rani N, Satyanarayana A, Bhaskaran PK (2015) Coastal vulnerability assessment studies over india: a review. *Natural Hazards* 77(1):405–428.
[32] Taiwo E, Adinya I, Edeki S (2019) Optimal evacuation decision policies for benue flood disaster in nigeria. J*ournal of Physics: Conference Series*, volume 1299, 012137 (IOP Publishing).
[33] Talukdar S, Ghose B, Salam R, Mahato S, Pham QB, Linh NTT, Costache R, Avand M, et al. (2020) Flood susceptibility modeling in teesta river basin, bangladesh using novel ensembles of bagging algorithms. Stochastic Environmental *Research and Risk Assessment* 34(12):2277–2300.
[34] Talukdar S, Pal S (2020) Modeling flood plain wetland transformation in consequences of flow alteration in punarbhaba river in india and bangladesh. *Journal of Cleaner Production* 261:120767.
[35] UNDP (2023) Nigeria Flood Impact, Recovery and Mitigation Assessment Report2022-2023. https://www.undp.org/nigeria/publications/nigeria-flood-impact-recovery-and-mitigation-assessment-report-2022-2023.
[36] UNDRR (2018) Disaster data and statistics. PreventionWeb: Accessed 13 April 2021 from https://www. flickr.com/photos/isdr/44062150100/in/photostream/.
[37] UNOCHA (2022) Democratic Republic of the Congo - Flash Update #3: Floods caused by heavy rains in Kinshasa, 31 December 2022. https://www.unocha.org/publications/report/democratic-republic- congo/democratic-republic-congo-flash-update-3-floods-caused-heavy-rains-kinshasa-31-december- 2022.
[38] WHO (2021) World health organisation: Floods. Accessed 28 July 2022: https://www.who.int/health-topics/floods#tab=tab_1

Online Appendices

I. EC.1. SOME DEFINITIONS OF TERMS

The statistical metrics are defined as follows:

- *Mean Square Error (MSE)*

The mean squared error function calculates the expected value of the squared (quadratic) error or loss, a risk indicator. It is defined as:

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}$$

Lower MSE indicates a better fit.

- *Root Mean Square Error (RMSE)*

This is the square root of the mean square error, and it is defined as:

$$\boldsymbol{RMSE} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\frac{(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}}{\boldsymbol{n}}}$$

- *Mean Absolute Error (MAE)*

This function calculates mean absolute error, a risk metric representing the expected magnitude of an absolute error loss. It is defined as:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}|\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}|$$

- *Coefficient of Determination (R^2)*

It represents the fraction of the variation of the outcome variable ($\boldsymbol{y}$) explained by the model's independent variables. The proportion of explained variance indicates the model goodness of fit and, thus, a measure of how well unseen samples are likely to be predicted by the model.

$$\boldsymbol{R^2(y, \hat{y})} = \mathbf{1} - \frac{\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}}{\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{\bar{y}})^{\mathbf{2}}}$$

Where $\boldsymbol{\bar{y}} = \sum_{i=1}^{n}\boldsymbol{y_i}$, $\boldsymbol{n}$ is the number of data points, $\boldsymbol{y}$ is the actual value, and $\boldsymbol{\hat{y}}$ is the predicted value of the $\boldsymbol{i}$-th data point.