UNIOFFA JOURNAL OF INTERDISCIPLINARY STUDIES (U-OJIS)
A Multidisciplinary Journal of Science, Management and Social Sciences
Website: www.unioffajis.com
ISSN:3141-5571
E-ISSN:3141-558X

Machine Learning-Based Runoff Prediction for Flood Forecasting: Integrating Antecedent Rainfall Dynamics in a Data-Scarce Catchment

MUHAMMED USMAN

Department of Civil and Environmental Engineering

Kwara State University

Muhammed.u@unilorin.edu.ng

Professor Abdulganiy Adeniyi Adeogun

Department of Civil and Environmental Engineering

Kwara State University

Adeniyi.A.G@kwasu.edu.ng

Volume 1:2

ABSTRACT

Accurate flood prediction remains a critical challenge in hydrological modelling due to the nonlinear and dynamic interactions between climatic and catchment variables as Nigeria faces annual occurrence of devastating flood. This study presents a machine learning-based framework for runoff prediction using satellite derived and locally improved hydro-meteorological data from Ilorin, Nigeria, spanning the period 1990–2025 to enhance early warning signals and preparedness for inhabitants of Ilorin metropolis. The model incorporated important meteorological variables such as temperature, dew point , relative humidity, rainfall, and an engineered antecedent precipitation index (API_alpha) derived using an exponentially weighted moving average approach. Three models—Linear Regression, Random Forest, and Extreme Gradient Boosting (XGBoost)—were developed and evaluated using standard performance metrics, including Root Mean Squared Error (RMSE) and coefficient of determination (R^2). Results indicate that Linear Regression performs poorly ($R^2 = 0.026$), highlighting its limitations in capturing nonlinear hydrological processes Random Forest achieved near-perfect performance ($R^2 = 1.000$), this result suggests potential overfitting and limited generalizability. In contrast, the XGBoost model demonstrated robust and reliable performance ($R^2 = 0.854$, RMSE = 0.000576), effectively capturing nonlinear relationships between predictors and runoff. The best performing model was used in projecting future run-offs for Ilorin using incorporating ARIMA (Autoregressive Integrated Moving Average) for rainfall and relative humidity projection. The findings emphasize the critical role of antecedent rainfall conditions in flood prediction and demonstrate the superiority of boosting-based models for hydrological applications. This study contributes to the growing body of knowledge on data-driven flood prediction and provides a scalable framework for improving early warning systems in data-scarce regions.

Keywords: Flood Prediction; Machine Learning; Runoff Modelling; Antecedent Precipitation Index; Hydrological Modelling; Climate Variability; Data-Driven Modelling;

INTRODUCTION

Flooding has always been among the most damaging natural hazards that have occurred around the globe; they are associated with so many negative consequences for people, physical built environments, and ecological systems. The Intergovernmental Panel on Climate Change reports that floods account for a large percentage of total losses associated with global disasters, and the frequency and severity of these events have been increasing as a result of climate change and extreme precipitation events (IPCC, 2022). Recent hydrologic studies have also concluded

ISSN:3141-5571
E-ISSN:3141-558X

that the intensification of rainfall patterns due to changing climate and variability are both fundamental causes of the increasing flood risk globally, especially in at-risk regions (Dong et al., 2025).

Flooding processes are very complicated and non-linear in nature, being affected by varying climatic inputs, characteristics of the relevant catchment area, and human activities. This complexity has been noted in the hydrology literature; therefore, one of the main areas of hydrological research has been focused on developing accurate models for the relationship of rainfall to runoff (Wang et al., 2023). While traditional flood risk management strategies, including structural flood mitigation (dams, levees) and non-structural flood mitigation (land-use planning, early warning systems) have been somewhat effective, recent studies demonstrate that these types of strategies have limitations when considering the uncertainty in the climate and the growing pressures associated with urban expansion.

Rapid urbanization, lack of drainage systems and an increasing population make Nigeria one of the most at risk countries for floods. The city of Ilorin has been affected by floods multiple times as a result of the Asa River Basin. The impacts of these floods on the community include economic instability and infrastructure damage. Urban expansion has resulted in more surfaces that water cannot absorb, which leads to an increase in the volume of water that flows over the ground and causes greater risk of flooding, as noted in current literature on urban hydrology (Miller et al., 2014; Beck et al., 2019; Zhang et al., 2023).

Most traditional flood prediction methods are based on using physical models to predict flooding using real-world hydrology and hydraulic models. These models provide good information about the processes that contribute to the flood but often suffer from being very data intensive, high levels of uncertainty in the parameters used and require a large amount of computation to develop the models. Recent studies have shown that it is difficult to develop a physically based flood prediction method for use in ungauged basins or those with little available data because of the challenges of calibrating and estimating parameters (Dong, et al., 2024). Similarly, physically based models require more processing power and computing resources than most areas can realistically obtain for operational use.

Hydrology is becoming increasingly reliant upon the use of remote sensing and climate reanalysis datasets due, in part, to data limitations. Increasingly utilized products, such as Climate Hazard Group Infrared Precipitation with Station data (CHIRPS) and The Fifth generation ECWMF atmospheric reanalysis of global climate (ERA5), provide long-term, high-resolution climate data, making them suitable to support flood modeling in many areas lacking ground-based observational data. There is a growing body of research that has shown that use of predicted models, together with reanalysis datasets, improves forecasting performance, especially when combined with machine learning approaches (Nearing et al., 2024).

Recently, machine learning (ML) methodologies have proven to be effective as a flood prediction option since they are able to model complex non-linear relations without utilizing explicit physical assumptions. Long Short-Term Memory (LSTM) networks have demonstrated their ability to capture temporal dependencies in hydro-logically-based time series data. For instance, LSTM models have been found to simulate stream-flow dynamics and conduct flood frequency analyses (Martel et al., 2025). Furthermore, hybrid modeling approaches that integrate principles of physical hydro-logical modeling frameworks with those of ML models (such as the Ia-LSTM framework) have significantly improved runoff prediction accuracy and reliability (Wang et al., 2023).

Emerging advanced machine learning is the emergence of hybrid deep learning architectures (e.g., CNN–LSTM), as well as probabilistic generative models that can improve extreme flood event prediction and uncertainty quantification. For example, diffusion runoff models outperform traditional deep learning methods for predicting extreme flows and enhancing early warning capabilities (Ou et al. 2024). There are, however, some limitations to purely data-driven models inherent in their limited interpretability and in accurately predicting extreme peak flows; thus, motivating the further development of physics-informed machine learning frameworks (Feng et al., 2022).

Despite machine learning having shown significant promise for predicting floods, the application of these models in Sub-Saharan Africa is extremely limited, especially as it relates to integrating long-term climate data, as well as understanding how

ISSN:3141-5571
E-ISSN:3141-558X

to translate predictive models into operational tools. Most studies have focused on model development with little or no attempt to apply those models to real-time decision support systems. There is, therefore, an urgent need for an integrated framework for collecting data, developing predictive models and integrating them into usable platforms for managing flood risk.

In light of this context, the purpose of this study is to create a data-based methodology to predict floods for the Asa River Basin located in Ilorin. For this task, long-term climatological data were incorporated into the framework with modern machine learning methods to provide predictions of floods caused by runoff experienced in the region. The climatological data used are derived from the 30-year time period of CHIRPS and ERA5 datasets; thus, analyses were conducted on the climatic variability during this period to develop and validate machine-learning methods to predict runoff quantities as well as create predictions of future runoff during climatological variability. As a departure from traditional methods, the developed machine-learning method is successfully integrated as a web-based application to enable real-time flood prediction and is consequently expected to increase its usefulness for flood hazard issues and decision-making processes associated with them.

MATERIALS AND METHODS

Research Design

The approach taken in this research involves using a quantitative, data-driven methodology for predicting flooding by combining hydro-climatic datasets along with machine learning methods to develop flood predictive models. The overall workflow includes data ingestion, data pre-processing, feature engineering, model building, model validation, forecasting, and the deployment of trained models. End-to-end predictive modeling frameworks are now being recommended for use within the field of hydrology to provide greater assurance of model robustness, reproducibility, and operational applicability (Kratzert et al. 2019, Nearing et al. 2021).

Description of Study Area

The region of focus for this study is Ilorin, located in Kwara State, Nigeria. Ilorin is situated in an area that falls within the larger Asa River Basin. The climate of the basin can be classified as tropical with two distinct seasons, wet (rainy) and dry (non-rainy). The bulk of hydrology within this basin is associated with the occurrence of runoff as a result of rainfall. However, rapid growth of urban centers in the basin has resulted in an increase in impervious surface area, which has led to an increase in runoff and flooding.

Google Earth Engine was used to gather hydroclimatic data (Gorelick et al., 2017). The geographic bounding box for the study area is defined by geographic coordinates that encompass both Ilorin and the entire Asa River catchment (Latitude: ~8.3°–8.6°N; Longitude: ~4.4°–4.7°E).

Data Requirements, Preprocessing and Exploratory Analysis

There are two main datasets that were utilized in this study:

(i) The CHIRPS V2.0 (Funk et al., 2015) dataset is a quasi-global precipitation dataset with a spatial resolution of approximately 0.05° and daily temporal resolution. The dataset consists of both satellite observations combined with ground-based (station) precipitation observations.

(ii) The ERA5 Reanalysis dataset from the European Centre for Medium-Range Weather Forecasts is a dataset that provides hourly atmospheric variables at ~0.25° (Hersbach et al., 2020). Daily variables for the years between 1990 and 2025 (30 years' worth of data) were extracted for rainfall, temperature (minimum and maximum), relative humidity, and dew point temperature. Runoff was derived from ERA5-Land variables.

The combination of these two datasets is known to provide the greatest accuracy in predicting hydrology due to complementary characteristics of CHIRPS and ERA5 which provide complementary strengths for estimating precipitation and atmospheric dynamics (Beck et al., 2019; Tarek et al., 2020).

To understand the data distribution, descriptive statistics were generated (mean, standard deviation, etc.). The distribution of the data and pathways were identified through using descriptive statistical analysis and creating a visual representation of the data with the aid of Matplotlib (Hamed, 2008). Time series trend analysis is an important part of hydrography science. Through the use of trend

ISSN:3141-5571
E-ISSN:3141-558X

analysis, long-term variation in the flow of water can be revealed (Hamed, 2008).

Feature Engineering and Selection

The use of seven-day rolling accumulation and Exponentially Weighted Moving Average (EWMA) of rainfall is used to determine antecedent moisture conditions in order to approximate the wetness conditions in each catchment area. The methodology used is consistent with the Antecedent Precipitation Index (API) widely used in hydrology (Brocca et al., 2017). The EWMA formulation is given as:

$$EWMA(t) = \alpha x(t) + (1-\alpha)EWMA(t-1) \quad (1)$$

where $x(t)$ is rainfall at time t, and α is the smoothing factor. In this study, $\alpha = 0.9$ was selected to emphasize recent rainfall events, consistent with studies highlighting the dominance of short-term rainfall memory in runoff generation.

Initially, Pearson correlation analysis was used to assess feature relevance. Since the nature of hydrological systems is nonlinear, however, all variables were retained for model development, allowing the machine learning algorithms to capture these complex interactions (Shen, 2018). This would help eliminate bias associated with feature selection by only using correlations for selection purposes.

Feature Selection

Best features for model training were selected using correlation analysis. The features were correlated with run-off (target variable) to know which is more correlated so as to minimize training time and maximize training output. Features with positive correlation with more than 0.1 correlation value were selected and others with negative correlation were discarded

Model Development

The models incorporated three types of machine learning algorithms from the scikit-learn library (Pedregosa, 2011); the first being linear regression which serves as a base model against which other models will be assessed; and random forest and XGBoost are both considered stronger than linear regression when defining a relationship between two variables. The equation defining linear regression is as follows:

$$y = XW + b \quad (2)$$

where X represents input features, W the coefficient vector, and b the intercept.

The random forest algorithm reduces the risk of overfitting by bootstrapping and using random feature selection (Breiman, 2001).

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \quad (3)$$

XGBoost is a gradient boosting algorithm that iteratively improves prediction by minimizing a loss function using gradient descent (Chen & Guestrin, 2016). It is widely recognized for its superior performance in hydrological prediction tasks (Zhang et al., 2023).

Hyperparameter Tuning and Validation Strategy

To optimize model performance, hyperparameter searches were performed using a combination of grid search and time-series cross-validation. The parameters used for tuning included:

i. Number of trees (n_estimators) and maximum depth (max_depth) for random forest;
ii. Learning rate (η), number of estimators, and maximum depth for XGBoost.

As the data has a temporal dimension, a time-series cross-validation approach was employed to maintain the chronological order of the data and to avoid any potential data leakage (Hyndman & Athanasopoulos, 2021), splitting from 1990 to 2018 for model development and 2018 to 2025 for model deployment.

Model Evaluation Metrics

The model's performance was evaluated through several different metrics in order to confirm its accuracy: Mean Square Error (MSE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Nash-Sutcliffe Efficiency (NSE)

Among those metrics listed above, the NSE metric has been established as a well-recognized and reliable hydrologic modeling performance metric because it

ISSN:3141-5571
E-ISSN:3141-558X

measures how well predicted values correspond to observed values of data (Moriasi et al., 2007).

Predictions for Future Runoff

Future runoff was simulated using a data-driven predictive model incorporating ARIMA. ARIMA was used to predict features like rainfall and dewpoints , while the best performing model from this research was used to project 10 years run-off for the period between 2025 and 2034.

To determine the stationarity of the of the research data , an Augmented Dickey-Fuller (ADF) test was performed. The hypothesis for non-stationarily was assumed to be true given a p-value ≤ 0.05 (Dickey & Fuller, 1979).

The parameters (p, d, q) were chosen for the ARIMA model based on their autocorrelation and partial autocorrelation characteristics. Diagnostic tests were administered on the ARIMA model to verify the independence of the residuals and that the fit to the data was adequate.

Flood Threshold Determination

Historically, statistical thresholds have been used to establish the conditions necessary for flooding. Historically, flood conditions have been defined as rain flows equaling the 75th percentile of previous annual runoff by establishing stations. For many years this has been the common acceptable method for identifying large or extreme rain flows in the absence of flow-based flood data (Blöschl et al., 2020).

RESULTS AND DISCUSSION

Data Characteristics and Exploratory Analysis

Table 1a shows that the dataset includes daily data on temperature (3 PM), dew point temperature (3 PM), relative humidity (3 PM), rainfall, and runoff for a total of 30 years (1990 – 2025). An initial review of the data indicated an incomplete data set with missing values in rainfall and runoff variables. Median substitution, which is a robust technique used widely in hydrological datasets due to its ability to reduce the effect of extreme values and preserve the distributional characteristics of the data, was used to impute missing data (Hyndman & Athanasopoulos, 2021).

Following data preprocessing, the dataset consists of 12,895 total observations for each variable. In addition to this, an engineered variable named API_alpha (calculated using an exponentially weighted moving average (EWMA)) has been created to capture the antecedent rainfall condition (Table 1b). This is consistent with the hydrological literature, where antecedent precipitation indices are commonly used to approximate soil moisture conditions that control runoff generation (Brocca et al., 2017).

Table 1a: Cross-Section of hydro-climatic data

Date	Temperature_3PM_C	DewPoint_3PM_C	RH_3PM_%	Rainfall()	Run-off(mm^3/s)
01-01-90	34.88387451	7.955545044	19.14415312	0	8.05E-07
02-01-90	34.78325806	5.942544556	16.76477789	0	8.34E-07
03-01-90	33.94011841	12.95063171	28.16045722	0	8.34E-07
04-01-90	34.12493286	14.57262573	30.97219468	0	8.05E-07
05-01-90	34.25771484	14.92926941	31.45989288	0	1.04E-06
06-01-90	32.96228943	17.420755	39.64533355	0	1.61E-06
07-01-90	34.97815857	15.62693176	31.61397412	0	9.69E-07
08-01-90	35.23067627	12.26159058	25.05706116	0	9.24E-07
09-01-90	35.37642822	8.328607178	19.10936433	0	7.90E-07
10-01-90	35.24399719	9.756677246	21.1983409	0	8.05E-07
11-01-90	34.74530945	13.94031677	28.72330813	0	7.75E-07
12-01-90	33.8933197	14.77562866	31.78798055	0	9.83E-07
13-01-90	34.42229614	13.84327087	29.05857624	0	9.24E-07
14-01-90	35.62006531	10.09606323	21.24154854	0	8.49E-07

ISSN:3141-5571
E-ISSN:3141-558X

Table 2: Hydro climatic data after adding API-alpha

Date	RH_3PM_%	Run-off (mm^3/s)	Month	API_alpha(mm^3/s
2018-03-31	26.52	1.3411e-07	March	1.3455
2018-04-01	11.33	1.6391e-07	April	0.3802
2018-04-02	9.87	1.1921e-07	April	0.2131
2018-04-03	15.08	2.0862e-07	April	0.7021
2018-04-04	26.68	3.4273e-07	April	1.3398
2018-04-05	43.05	6.4075e-07	April	2.7190
2018-04-06	57.00	3.3379e-06	April	3.5980
2018-04-07	67.32	5.2825e-05	April	5.6262
2018-04-08	48.83	2.1607e-06	April	3.6829

According to the descriptive statistics in Table 2, there is considerable variation in both rainfall (the mean is 3.24 mm with a standard deviation of 5.97 mm) and runoff (mean = 0.000118 m^3/s and the same standard deviation). The fact that both of these variables have a greater level of variability than the mean provides evidence of a great deal of fluctuation from one time period to another and the presence of significant, extreme events. The variability seen in both rainfall and runoff is characteristic of tropical rainfall regimes and has been documented extensively in other flood-prone areas (Blöschl et al., 2020).

On the other hand, temperature showed a much lower degree of variation (mean = 32.24°C, standard deviation = 3.84°C) indicating it has a much more stable climate than both rainfall and runoff. This finding is consistent with results from other studies of tropical climates, which indicate temperature varies less than precipitation does as it is the principal variable that causes hydrologic extremes.

Table 2. Summary of Descriptive Statistics

Statistic	Date (Timestamp)	Temperature_3PM (°C)	DewPoint_3PM (°C)	RH_3PM (%)	Rainfall (mm)	Run-off (mm^3/s)
Count	N/A	12,895	12,895	12,895	12,895	12,895
Mean	2007-08-27 (10:29)	32.243	16.645	45.464	3.238	0.000118
Std Dev	N/A	3.843	6.584	22.347	5.965	0.000118
Min	1990-01-01 (0:00)	21.234	-9.365	4.246	0	0
25% (Q1)	1998-10-29 (12:00)	29.145	12.772	24.607	0	0
50% (Median)	2007-08-27 (0:00)	32.679	19.838	46.545	0	0.000002
75% (Q3)	2016-06-23 (12:00)	35.395	21.588	63.542	4.73	0.000017
Max	2025-12-04 (0:00)	40.495	25.511	98.582	62.897	0.069022

A seasonal analysis shows March has the hottest average monthly temperature at ~34.93°C. Conversely, the average monthly temperature in August is the coldest at ~29.29°C. The seasonal temperature variation in Ilorin is characteristic of the region's climate; peak temperatures occur during the dry season, while the increase in cloud cover and

ISSN:3141-5571
E-ISSN:3141-558X

precipitation during the wet season causes lower average temperatures than during the dry season. This same seasonal temperature variation pattern has been documented in studies of climatic variability in West Africa, with the changes in precipitation impacting temperature variability.

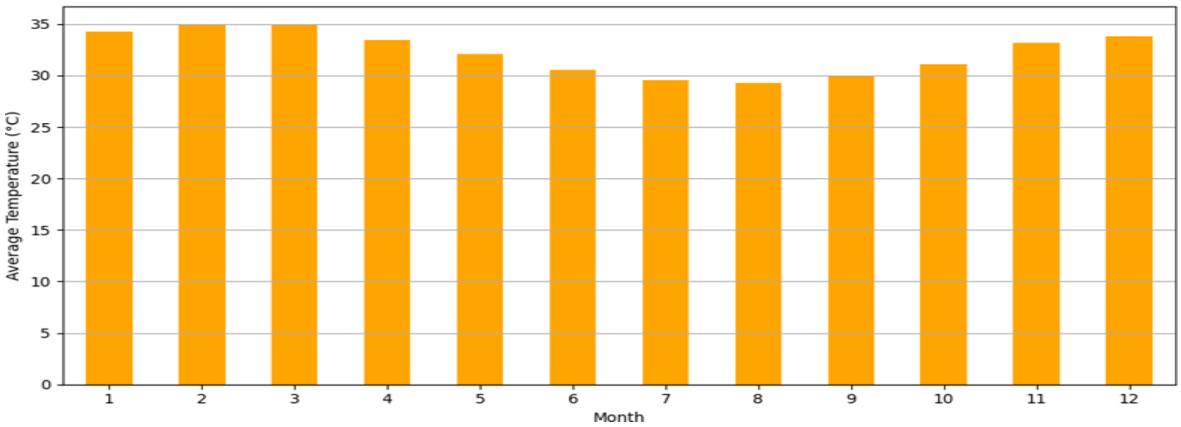


Figure 1: Temperature Distribution in Ilorin

Figure 2 illustrates monthly patterns of precipitation and runoff, showing maximum monthly precipitation and maximum monthly runoff occur from July to September, with the greatest values in September. This indicates a strong relationship between precipitation and runoff, which is consistent with hydrological theory.

The time period between peak precipitation and flood events (which is often into October) indicates that cumulative rainfall and wetness of the catchment are the critical factors influencing flood creation rather than individual events of precipitation. This correlates with other recent research highlighting antecedent moisture conditions' importance in generating floods (Beck et al., 2019).

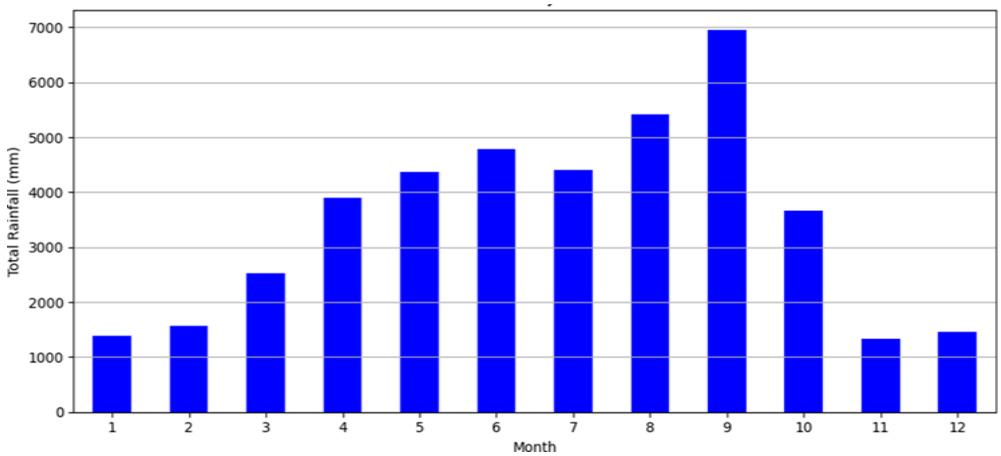


Figure 2: Monthly Average Rainfall Values

Long-term trend analysis has shown an incremental increase in total annual precipitation over the past two decades (Figure 3); with the exception of 2012 (year with highest total) and 2023 (highest total). The highest annual precipitation total occurred in 2023 at ~1531mm; approximately 401mm more than the expected average of ~1130mm.

ISSN:3141-5571
E-ISSN:3141-558X

There has been an evident increase over time in the intensity and variability of rainfall events; which is consistent with other global climate assessment reports that have been published by the IPCC regarding climate change and precipitation extremes being associated with climate change scenarios.

Additional analysis (Figure 4) shows that high-rainfall years (2012 and 2023) coincide with recorded flooding incidents in Ilorin, demonstrating the relationship between the amount of rain and the occurrence of fires caused by flooding; however, flooding is not determined by rainfall alone because of the effect of pre-existing conditions such as land use and characteristics of watersheds, etc.

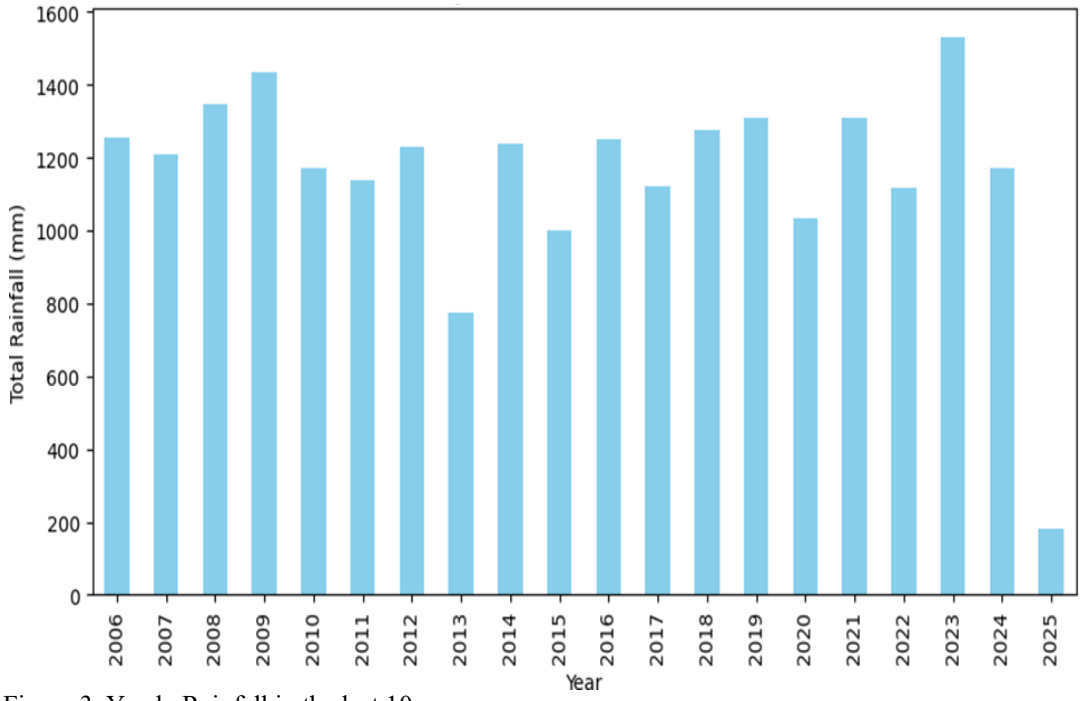


Figure 3: Yearly Rainfall in the last 10 years

Feature Engineering and Hydrological Interpretation

By using EWMA to construct the engineered feature of API_alpha from rainfall, the effect of antecedent rainfall was captured by placing more value on the most recent events of precipitation than on earlier precipitation. Thus, API_alpha enhances our ability to represent soil moisture's memory, which significantly contributes to runoff occurrence.

Additionally, API_alpha's upward trend prior to high runoff events indicates that total precipitation has a large effect in generating floods in the study area. This supports recent hydrological studies which have shown that antecedent wetness conditions frequently have more impact on runoff than instantaneously measured rainfall (Kratzert et al., 2019).

By including this feature in the modelling framework, the study provides greater physical relevance to the machine-learning models, and improves their ability to approximate hydrological reality as compared to purely data-driven models.

ISSN:3141-5571
E-ISSN:3141-558X

Correlation Analysis and Feature Relevance

From the correlation matrix (Figure 5), there is a small positive linear relationship between rainfall and runoff (r = 0.12) and also, a little bit of evidence for a small positive linear relationship between relative humidity and runoff (r = 0.17). Although the correlations are low relative to other studies, and likely due to the nonlinear relationship of hydrological systems, it is expected that these correlations would be lower than other rainfall-related variables, as runoff is very dependent on complex processes such as infiltration, evapotranspiration, soil saturation, subsurface flow, etc., which cannot all be represented through a linear relationship. Recent studies in weather-related machine learning have indicated that if the correlation between two variables is low, that does not mean that they are unrelated, but that nonlinear models can uncover patterns in those variables (Shen (2018). When looking at the relationship between temperature and runoff, there was a negative correlation indicating that on days with high temperatures, there is more evaporation and, thus, less surface water available. On the other hand, the relative humidity had a positive relationship with runoff indicating that on days with higher levels of humidity in the atmosphere, you are also more likely to have rainfall and eventually runoff.

Overall, the correlation analysis shows that the use of linear models in hydrological modeling has limitations and further supports the need for machine learning methods to provide a better understanding of the nonlinear relationships inherent in hydrology.

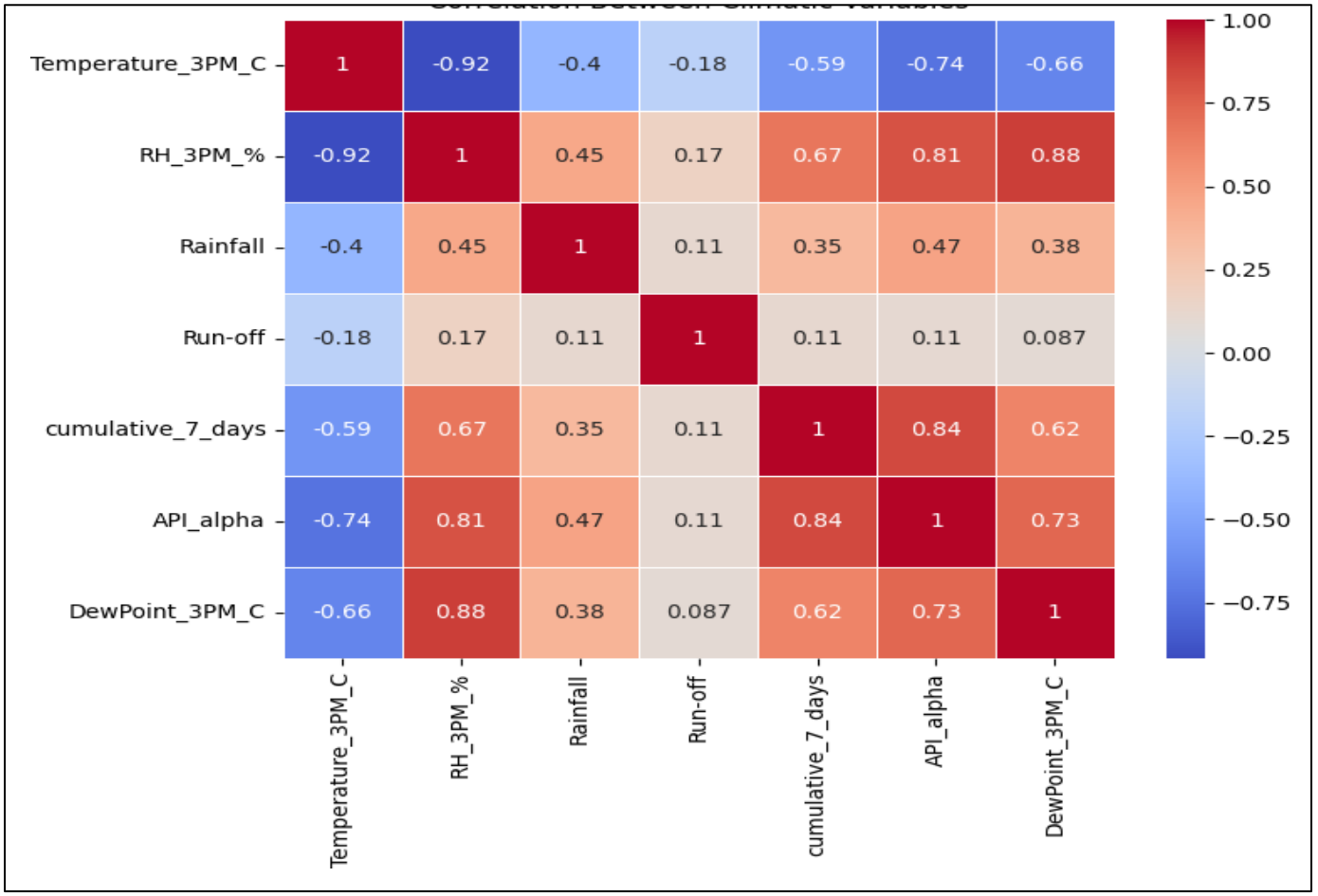


Figure 4: Correlation of Feature

ISSN:3141-5571
E-ISSN:3141-558X

Model Performance Evaluation and Comparative Analysis

The predictive performance of the three machine learning models was quantitatively evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2).

Linear Regression Performance

The first model was implemented using Linear Regression. It used the Mean Square Error (MSE) and the Square Root of the Mean Square Error (RMSE) to calculate the model's performance. After the initial round of training, the MSE was calculated as 0.000002 and the RMSE was calculated as 0.001489. MSE is mathematically defined as the average squared differences between predicted value, or runoff, and the actual rainfall runoff value. Based on the MSE value, it can be concluded that the predicted runoff values were numerically very close to the actual runoff values. With the mean runoff value calculated as 0.000118, nearly 99.865% of the error can be attributed to variability in prediction, indicating several thousand of of ground water runoff predicted will not occur.
Since the values were so highly skewed, the Coefficient of Determination (R2) was also calculated. The calculated R2 value of 0.026223 confirms that only approximately 2.6% of the variance in runoff can be explained by the Linear Regression model, indicating that the model has little explanatory power and may not be defined enough to describe the relationships between variables. A scatter plot in Figure 6 compares the predicted runoff values versus actual values; this plot shows how close or far apart the predicted and actual values are.

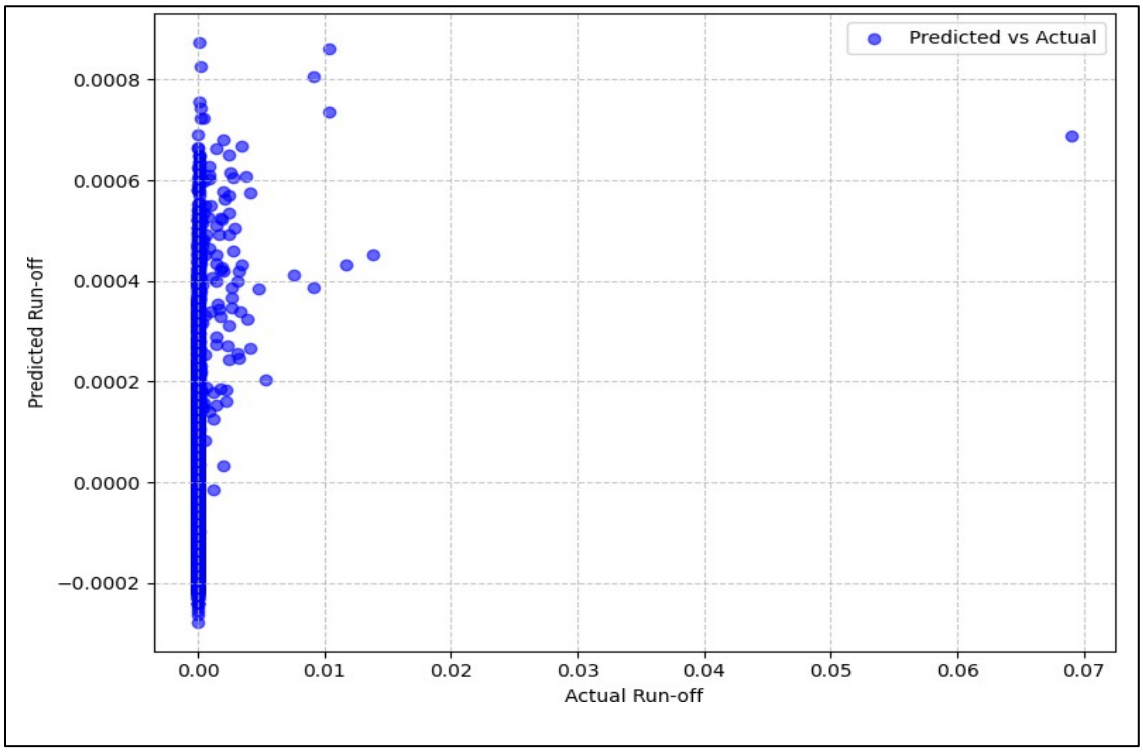


Figure 5: Scatter Plot for Actual and Predicted Run-Offs

This confirmation illustrates the inability of linear methods to accurately describe hydrological processes, which are nonlinear as well as affected by complicated interactions (e.g., soil saturation, infiltration, delayed runoff response). In addition, numerous hydrologists have documented the limitations of linear models when it comes to predicting floods (Shen, 2018).

ISSN:3141-5571
E-ISSN:3141-558X

Random Forest Model Performance

The linear regression model produced an extremely low R^2 value; therefore, a Decision Tree Regressor was utilized for subsequent training rounds. Once again, multiple rounds of training/tuning were performed and the performance metrics (MSE, RMSE, and R^2) from the test data (and corresponding predicted values) were computed. The MSE and RMSE of the Decision Tree Model were both 0.00 and the R^2 of the Decision Tree Model was 1.00. Although this would seem to indicate perfect model performance, these results are also indicative of overfitting, which occurs when a model "remembers" the training data, but does not generalize to new/unseen data. The ability of complex models, such as Decision Trees, to overfit demonstrates the importance of using appropriate regularization/validation techniques to avoid overfitting, particularly when creating highly flexible models (e.g., Random Forests) (Breiman, 2001; Kratzert et al., 2019).

Despite the limitations associated with the overfitting demonstrated in the model outputs, the ability of ensemble, tree-based methods to account for more complex nonlinear relationships found within hydrologic systems, when appropriately configured, is exemplified by the strong performance shown in Figure 7.

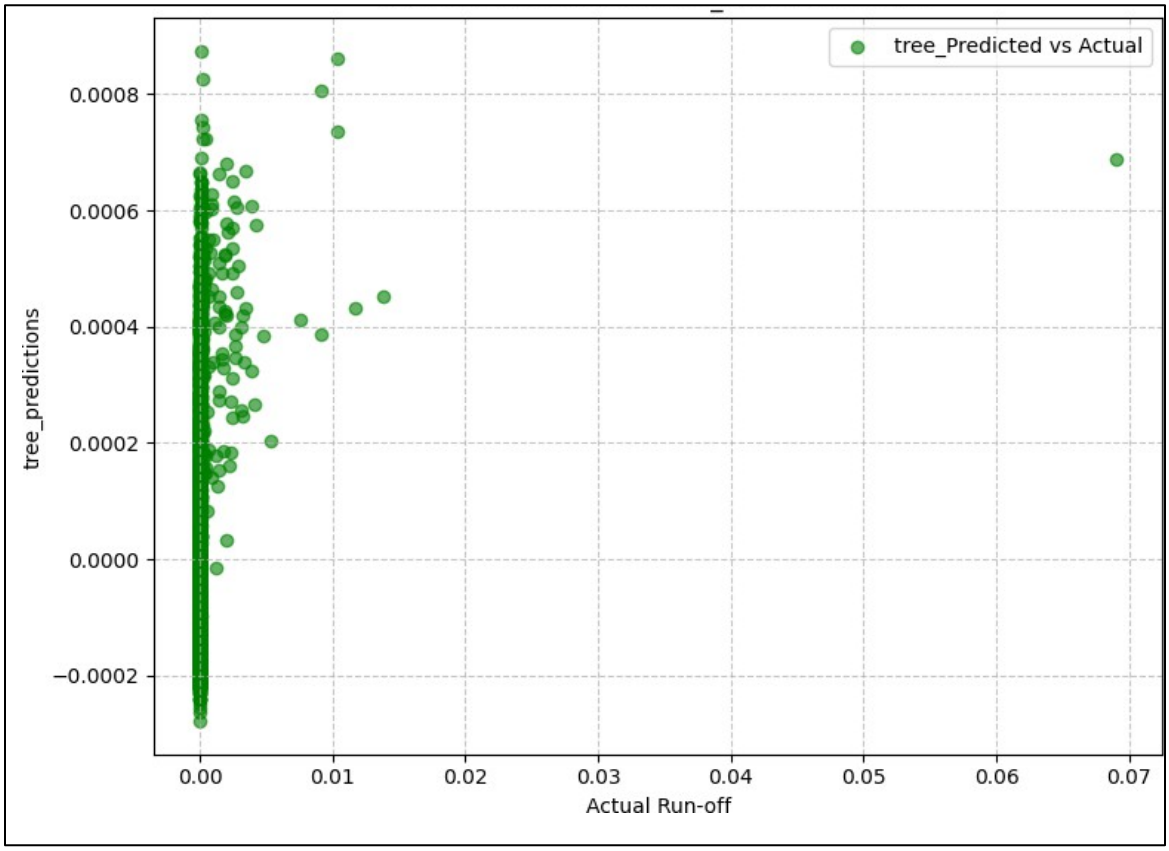


Figure 6: Scatter Plot for Actual and Tree_Regressor_Predicted Run-Offs

ISSN:3141-5571
E-ISSN:3141-558X

XGBoost Model Performance

The performance of the XGBoost model was found to be very good as evidenced by an R^2 of 0.854 and an RMSE of 0.000576. This indicates a strong level of agreement between the predicted and observed runoff values.

On the other hand, XGBoost has a much more balanced model performance compared to Random Forest through the combination of higher accuracy and better generalization ability. In addition, it captures non-linear relationships and interdependencies of input variables well. The strongest relationships between the inputs versus the predicted outputs can be seen with regards to the antecedent precipitation index (API_alpha) and the relative humidity as seen in the comparison of actual runoff to the new predicted runoff values in Figure 8. The improved performance of the XGBoost model is a result of the mechanism behind gradient boosting whereby reducing the error of the model through repeated iterations allows for the better capturing of the data's underlying structure, improving model reliability as well. Findings from other studies have also established, through use of various metrics of performance, that XGBoost is one of the most effective machine learning algorithms for use in hydrological modeling and flood prediction (Chen & Guestrin, 2016; Zhang et al., 2023).

However, as with any machine learning algorithm, the XGBoost model does experience some limitations in regards to accurately predicting extreme runoff values. This is likely due to the lack of data points with extreme runoff values in the dataset provided.

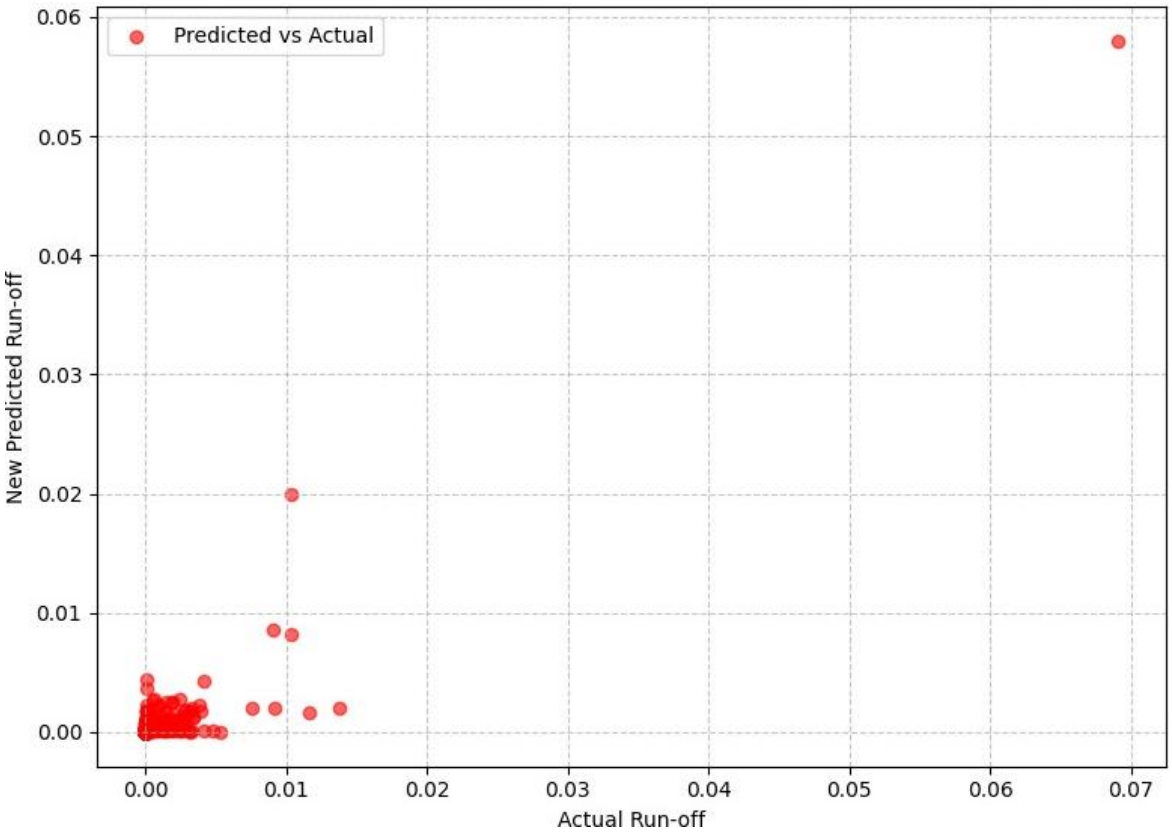


Figure 7: Actual vs New Predicted Run-off

ISSN:3141-5571
E-ISSN:3141-558X

Hydrological Implications of Model Results

One of the main findings of this study was that antecedent rainfall conditions (API_alpha) will significantly impact our ability to predict runoff. The strong ability of the nonlinear models (especially XGBoost) to develop flood-producing models based on cumulative amount of rainfall lead us to believe that flood-producing models are more related to cumulative rainfall than to instantaneous amounts of rainfall. This finding is also consistent with our understanding of hydrology, specifically regarding the role that soil moisture memory and catchment saturation play in determining the way runoff occurs (Beck et al., 2019).

Finally, these results demonstrate the importance of utilizing physically meaningful attributes as inputs into a machine learning framework to improve both predictive accuracy and the ability to interpret these results.

Flood forecasting with XGboost and ARIMA

The forecast result for run-off and rainfall suggested that year 2034 will witness highest runoff of 93 m^3/s with 1299 mm total for rainfall, followed by year 2032 with 81.33 m^3/s of rainfall and 1116.2 mm of rainfall as shown in Figure 9

Out of the years in consideration, year 2030 and 2032 appeared to have the lowest figures for both run off and rainfall.

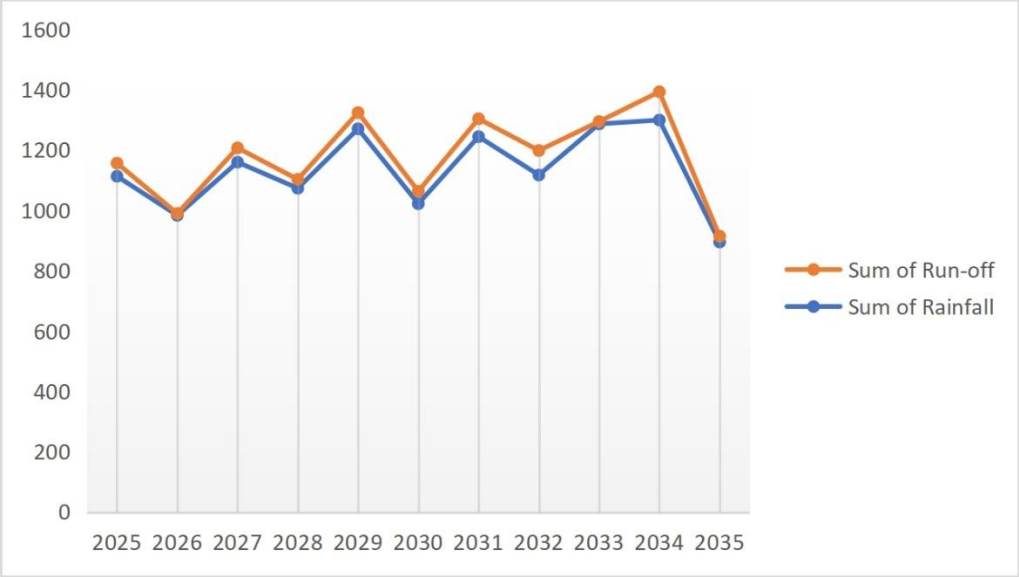


Figure 9 Sum of forecasted Rainfall and Run-off (2025-2035)

Monthly analysis of the forecast rainfall and run-off revealed that the months of September and October are the most likely months to expect flood in the next 10 years with 2114.6 mm of rain-fall amount for October.

It also revealed that the months of September and October are the most likely months to expect flood in the next 10 years with 2114.6 mm of rain-fall amount and 221.2 m^3/s of run-off as well as 2022.45 mm of rain fall and 196.44 m^3/s of run-off for September and October respectively.

ISSN:3141-5571
E-ISSN:3141-558X

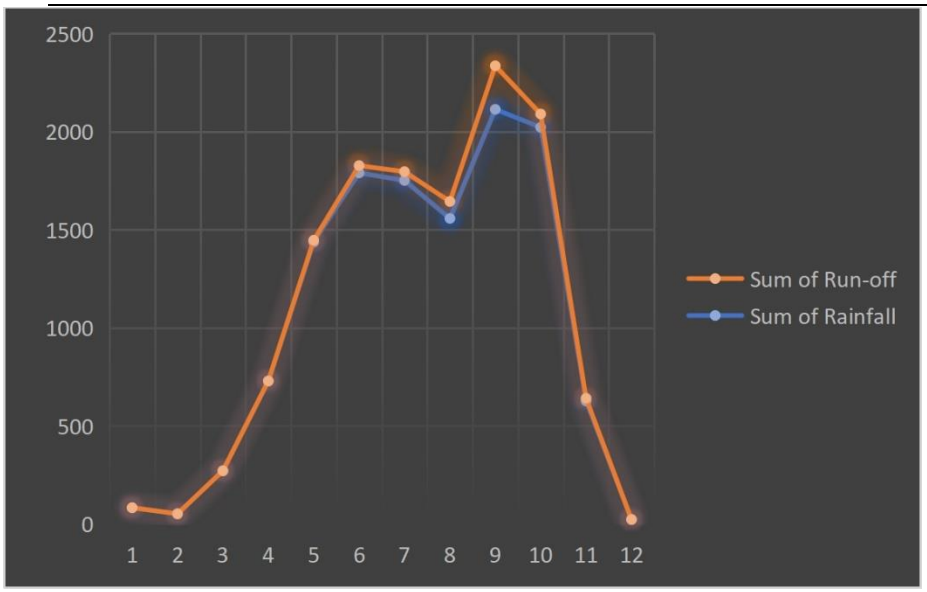


Figure 10 Sum of monthly forecasted Rainfall and Run-off (2025-2034)

CONCLUSION

This study highlights the growing potential of machine learning in transforming flood prediction from purely data-driven approximations to physically informed, scalable, and operational tools. The integration of such approaches into flood risk management systems could significantly improve preparedness and resilience, particularly in rapidly urbanizing and climate-vulnerable regions.

From an applied perspective, the developed framework provides a practical basis for real-time flood prediction and early warning systems in data-scarce environments. Its reliance on widely available reanalysis and satellite-derived datasets enhances its transferability to other regions facing similar hydrological and infrastructural challenges. However, the study also reveals critical limitations, particularly the reduced model sensitivity to extreme runoff events and the constrained feature space, which may limit predictive robustness under highly variable climatic conditions.

Future research should focus on integrating additional hydro-environmental variables such as land use dynamics, soil characteristics, topography, and drainage network properties to improve model comprehensiveness. Furthermore, the application of advanced deep learning architecture such as Long Short-Term Memory (LSTM) networks and hybrid physics-informed machine learning models—offers promising pathways for capturing temporal dependencies and improving extreme event prediction. There is also a need for rigorous cross-validation strategies and uncertainty quantification frameworks to enhance model reliability and support decision-making processes.

REFERENCES

Adelekan, I. O. (2016). Flood risk management in the coastal city of Lagos, Nigeria. *Journal of Flood Risk Management, 9*(3), 255–264. https://doi.org/10.1111/jfr3.12153

Ali, G., & Roy, A. G. (2010). Shopping for hydrologically representative connectivity metrics in a humid temperate forested catchment. *Water Resources Research, 46*(12), W12544. https://doi.org/10.1029/2010WR009442

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & de Roo, A. (2019). MSWEP V2 global 3-hourly precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society, 100*(3), 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1

ISSN:3141-5571
E-ISSN:3141-558X

Blöschl, G., Hall, J., Viglione, A., et al. (2020). Changing climate both increases and decreases European river floods. *Nature, 573*(7772), 108–111. https://doi.org/10.1038/s41586-019-1495-6

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brocca, L., Crow, W. T., Ciabatta, L., et al. (2017). A review of the applications of the antecedent precipitation index for soil moisture estimation. *Journal of Hydrology, 548*, 658–676. https://doi.org/10.1016/j.jhydrol.2017.03.042

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431. https://doi.org/10.1080/01621459.1979.10482531

Feng, D., Fang, K., & Shen, C. (2022). Enhancing streamflow prediction with physics-informed machine learning. *Water Resources Research, 58*(5), e2021WR030915. https://doi.org/10.1029/2021WR030915

Funk, C., Peterson, P., Landsfeld, M., et al. (2015). The climate hazards infrared precipitation with stations (CHIRPS). *Scientific Data, 2*, 150066. https://doi.org/10.1038/sdata.2015.66

Gorelick, N., Hancher, M., Dixon, M., et al. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Hamed, K. H. (2008). Trend detection in hydrologic data: The Mann–Kendall trend test under the scaling hypothesis. *Journal of Hydrology, 349*(3–4), 350–363. https://doi.org/10.1016/j.jhydrol.2007.11.009

Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Kratzert, F., Klotz, D., Herrnegger, M., et al. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research, 55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51–56).

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., et al. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE, 50*(3), 885–900. https://doi.org/10.13031/2013.23153

Nearing, G. S., Kratzert, F., Sampson, A. K., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research, 57*(3), e2020WR028091. https://doi.org/10.1029/2020WR028091

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research, 54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

ISSN:3141-5571
E-ISSN:3141-558X

Tarek, M., Brissette, F. P., & Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling. *Hydrology and Earth System Sciences, 24*, 2527–2544. https://doi.org/10.5194/hess-24-2527-2020

Zhang, J., Li, X., & Wang, Y. (2023). Machine learning approaches for flood prediction: A comparative study of ensemble methods. *Journal of Hydrology, 620*, 129379. https://doi.org/10.1016/j.jhydrol.2023.129379

Miller, J. D., Kim, H., Kjeldsen, T. R., Packman, J., Grebby, S., & Dearden, R. (2014). Assessing the impact of urbanization on storm runoff in a peri-urban catchment using historical change in impervious cover. *Journal of Hydrology, 515*, 59–70. https://doi.org/10.1016/j.jhydrol.2014.04.011