AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com

Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

PREDICTION OF HEAVY RAINFALL AND FLASH FLOODS USING MACHINE LEARNING (TIME SERIES ANALYSIS)

Sania Shafqat[1], Muhammad Khan[2], Fawad Nasim[3]

Abstract

Flash floods and extreme rainfall events lead to disastrous outcomes all over the world, which stretch the very survival of human beings and their physical and human infrastructures, even their economies. Therefore, efficient disaster preparedness needs an impeccable forecasting. Most of the classical hydrological and meteorological models are often inadequate in such a way that they do not take into account nonlinear dependencies and temporal variations in rainfall, which in turn cause such flaws as error in their predictive outcomes. In this research, machine-learning-based time series analysis is adopted for improving flood forecasting with the Kerala Flood Dataset (1901-2018). Heavy rain trend prediction applies ARIMA and Long Short-Term Memory (LSTM) networks. The LSTM significantly outperformed ARIMA with an RMSE of 64.5 opposed to 87.2 for ARIMA, thereby attesting its worth for modeling long-term dependencies in addition to the sequential changes of rainfall. To classify for flash floods, three separate classifiers were used, with Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. The best accuracy was achieved by the Random Forest classifier: 96.1%, whereas KNN and Logistic Regression yielded 91.2 and 86.5%, respectively. It underscores the competence of ensemble learning in extreme-weather classification. It also did feature engineering, from rolling means to lags, which is going to improve model performance. The models' performance has been found effective and appropriate through comparative analysis with respect to accuracy, precision, recall, and RMSE. Human real-world scenarios equipped the models to go further and have been deployed as API-based early warning systems, interfaced with IoT-driven real-time weather monitoring, and coupled with cloud computing for continuous updating and real-time flood risk appraisal. The results will give this study another validation and make a strong case for using AI-driven predictive analytics in disaster resilience and climate adaptation. It has been demonstrated that machine learning significantly improves the accuracy of early warning systems for flood prediction, thus contributing to disaster management and risk mitigation strategies, using LSTM in time series forecasting and Random Forest in flood classification. Future works will involve capturing real-time satellite data associated with hydrological parameters and the hybridization of deep learning approaches for future improvements in model predictions of extreme weather forecasting.

Key Words: Heavy rainfall and Floods, Machine learning, LSTM

1: Introduction

Flash floods and very high rainfall events are some of the most natural disasters that threaten to destroy human beings along with losing thousands of lives, destroying infrastructures, or inflicting huge economic costs in the world. They are very sudden events and are, therefore, important to predict for disaster risk management, as there is usually left little time to prepare for them. The increasing frequency and intensity of flash floods in modern years brought about by climate alteration, urbanization, and deforestation reflect an urgent requirement for improved forecasting methodologies (Teh and Khan 2021)[i]. Most traditional meteorological and hydrological models fall short of predicting these events accurately due to the reliance on rather limited variables based on linear assumptions and thereby very little complex climatic interactions are taken into account. The present work describes a machine learning-based time series analysis application for the purpose of improving flash flood and heavy rainfall predictions. Machine learning (ML) is one most powerful geoscientific tool that could be

[1] Corresponding Author, Faculty of Computer Science and Information Technology, Superior University, Pakistan, saniashafqat000@gmail.com
[2] Department of Environmental Management, School of Health and Life Sciences, Teesside University, Middlesbrough, UK
[3] Faculty of Computer Science and Information Technology, Superior University, Pakistan

used in climate science to analyze huge amounts of historical data and find hidden patterns within it for making predictions. Unlike the traditional statistical models, which are based on the dependency of parameters, data and time that can be large in number, ML techniques can analyze huge amounts of data, can account for many variables and change dynamically with the changing climate (Bergen et al. 2019)[ii]. This paper describes the use of methods such as Long Short-Term Memory (LSTM) networks, Random Forest, and ARIMA models to forecast heavy rainfall and forecast associated flooding using past meteorological data. These models therefore allow much more flexible and dynamic approaches towards flood forecasting.

The study uses data collected from Kaggle called the Kerala Flood Dataset (1901-2018). It consists of monthly records of rainfall, annual cumulative rainfall data, and indicators of occurrence of floods. Floods have plagued the south Indian state of Kerala on several occasions in the last few years and thus made this a proper case for trend analysis in rainfall and subsequently on flood prediction accuracy improvement. The time-series format of the database allows for the application of temporal machine learning models for learning from past trends to predict the future. This study would ultimately develop an accurate and scalable flood forecasting system through data preprocessing, feature engineering, and application of state-of-the-art ML algorithms. One of the biggest problems in flood forecasts is the high variability and complexity of meteorological patterns. Rainfall and flooding depend upon a range of interacting variables-temperature, humidity, atmospheric pressure, and terrain-as well as more physical control conditions. Conventional forecasting models in terms of a numerical weather prediction (NWP) model might use physical simulations to explain some of these dependencies, yet eventually would encounter problems in trying to include all such complexities (Brunner et al. 2021)[iii]. In comparison, machine learning models are based on historical weather patterns with trends identified and predictions accordingly drawn as learned correlations. In this study, time-series forecasting models, classification algorithms, and ensemble learning techniques are used for prediction accuracy improvement and actionable prediction accuracy in disaster preparedness.

The entire research follows a structured data-science-oriented approach composed of data preprocessing, exploratory data analysis (EDA), model selection, evaluation, and deployment. In the data preprocessing part, the dataset gets prepared for clean, tidy, and free of missing records. The exploratory data analysis is about identifying trends, seasonality, and outliers in the rainfall history. Then the results got into training machine learning models that will be tuned for some important performance metrics: accuracy, precision, recall, RMSE, and R^2. Once the tops are known, they feed into the real-time flood early warning system built upon cloud computing and IoT-based weather monitoring.

These future studies will thus serve as a major academic contribution beside their practicality to governmental agencies, meteorologists, urban planners, and disaster response personnel. An accurate flood prediction system will therefore aid in timely warnings, optimal resource allocation, and reduction of scope for economic loss. While blending geospatial value for real-time weather data, it greatly elevated flood risk assessment and, furthermore, helped disaster-prone area infrastructure planning.

2: Literature Review

Using Remote Sensing Data for Predicting Potential Areas of Flash Flood Hazards and Water Resources

(Hussein et al. 2019)[iv] studied the application of remote-sensing data in predicting potential flash-flood hazard zones as well as assessing water resources. The data sources used were the

Shuttle Radar Topography Mission (SRTM) and Tropical Rainfall Measuring Mission (TRMM) to study hydrological features pattern relief and drainage characteristics. The study area is confined to the Red Sea region of Egypt as it is a flash-flood area due to its sporadic heavy rainfall and relief features. By active microwave and visible/Near-Infrared (VNIR) remote sensing together with Geographic Information Systems (GIS), the study was able to effectively map flood-vulnerable zones and areas of infrastructure and urban settlements at risk during rainfalls. This research stresses the function of land use and surface characteristics as increased flood hazards, so it showed that the sub-basins part of the study area (for example, sub-basins 7, 8, and 9) could flood frequently. Recommendations made in the study were constructing dams and reservoirs to moderate the flow of water and to promote sustainable management of water resources. The study concluded the major role of remote sensing technologies in early warning systems and disaster preparedness, especially in arid and semi-arid regions.

Using Machine Learning Models, Remote Sensing, and GIS to Investigate the Effects of Changing Climates and Land Uses on Flood Probability

(Avand, Moradi, and lasboyee 2021)[v] studied the effects of climate change and improvements in land use on flood probability in the Tajan watershed, Iran. To this end, land-use change modeling was performed using the Land Change Modeler (LCM) for a period of 2019-2040, relying on historical land-use patterns from 1990 to 2019, and then future climate projections were simulated using Lars-WG software with two Representative Concentration Pathway (RCP) scenarios-RCP2.6 and RCP8.5-to assess future changes in precipitation and temperature. Any machine learning approach is incorporated in the study, using Random Forest (RF) and Bayesian Generalized Linear Model (GLMbayes) to assess flood susceptibility across the region. Topographic elevation, distance from rivers, land use pattern, slope, and rainfall variability were identified as dominant variables affecting flood occurrence in the region. The findings of the study showed that there was greater runoff due to deforestation and urbanization, compounded by expected increases in rainfall, and put flood risk in downstream areas to a higher rate. The outcomes therefore emphasized the importance of integrating land-use planning with strategies that adapt to the climate so as to minimize future flood hazards. This study offers a comprehensive framework for identifying flood-prone areas and improvement of disaster management strategies by integrating remote sensing, GIS, and machine learning techniques.

Improving Hourly Precipitation Estimates for Flash Flood Modeling in Data-Scarce Andean-Amazon Basins: An Integrative Framework Based on Machine Learning and Multiple Remotely Sensed Data

Chancay and Espitia-Sarmiento (2021) sought to solve the issue of precise precipitation estimation for flash flood modeling in poorly gauged regions, that is, in the Andean-Amazon sub-basins. As the monitoring networks are not sufficiently dense in these regions, the study proposes a machine-learning integrative framework, which combines diverse satellite precipitation products, including soil-moisture data. This study adopted a Random Forest (RF) model with bias correction techniques to refine precipitation estimation, and the results were fed into the GR4H hydrological model for flood forecasting. Three sub-basins-Upper Napo River Basin (NRB), Jatunyacu River Basin (JRB), and Tena River Basin (TRB)-which experience frequent flash floods whose predictions are further complicated due to topography and rapid hydrological response-were assessed in this study. It was shown that the bias-corrected RF model was effectively able to rectify rainfall estimates, reducing errors by as much as 93% when compared to the performance of satellite products alone. Thus, combining

further precipitation estimates with flood simulations provides an efficient way of improving warnings and resource management in far-off, high-risk areas.

Hybrid Models Incorporating Bivariate Statistics and Machine Learning Methods for Flash Flood Susceptibility Assessment Based on Remote Sensing Datasets

In a study by (Liu et al. 2021)[vi], a model that integrated bivariate statistics with machine learning techniques was improved flash flood susceptibility assessments. The study was located in the Dadu River Basin and applied three hybrid models: Support Vector Machine with Fuzzy Membership Value (SVM-FMV), Classification, and Regression Trees with FMV (CART-FMV), and Convolutional Neural Networks with FMV (CNN-FMV). A geo-spatial database was formulated with nine flood conditioning factors and 485 historical flood sites which were utilized for model training and validation. Validation of performance was done through the Receiver Operating Characteristic (ROC) curve and other statistical parameters. The study showed that the performance of the CNN-FMV model was superior to that of the others, with an area under the curve (AUC) value of 0.935 for success rate and 0.912 for prediction accuracy. The authors emphasized that hybrid approaches combining statistical methods with deep learning techniques are great for flood risk assessment, particularly in places where the hydrological dynamics are complex. The results highlighted the necessity for data-driven methods in flood management while elucidating the prospects of hybrid models to enhance flash flood forecasting.

A Novel Deep Learning Neural Network Approach for Predicting Flash Flood Susceptibility: A Case Study at a High-Frequency Tropical Storm Area

In a tropical storm-affected area of Vietnam, (Tien Bui et al. 2020)[vii] proposed a deep neural network for predicting flash flood susceptibility. The developed prediction model is called as Deep Learning Neural Network (DLNN) with an architecture of 192 neurons distributed in three hidden layers. Training took place with a diverse set of environmental and hydrological factors, including elevation, slope, curvature, stream density, soil type, rainfall, etc. Apart from this, the outcomes of the DLNN model are evaluated with the performance from two conventional machine learning models: Multilayer Perceptron Neural Network (MLP) model and Support Vector Machine (SVM). The results indicated that the DLNN model is considerably better with respect to the prediction accuracy of 92.05% than derived through benchmark models on positive predictive value and classification accuracy. The flood susceptibility mapping and early warning systems are going to be improved significantly using such combinations of deep-learning methodology along with GIS and remote sensing data. Findings have shown that potential exists for deep-learning models in disaster risk reduction, especially in regions frequently affected by extreme weather.

3: Methodology

Problem Statement

The meteorological **prediction of heavy rainfall and flash floods is defined as a major challenge in meteorology and disaster management.** These extreme weather events deal heavy blows to human life, infrastructure, and agriculture, especially in flood-prone areas. Conventional forecasts, such as numerical weather prediction (NWP) model-based methods, and hydrological simulations, rightly so, are unable, to a degree, to forecast flash floods accurately due to their inability to capture nonlinear dependencies of meteorological data. These models work heavily on assumptions and simplifications, thereby limiting their effectiveness in predicting short-term extreme events. Machine learning (ML) is a new alternative to traditional methods in that it learns complex patterns from historical data to predict future outcomes. Particularly, time series analysis fits rainfall and flood prediction

well because it allows the model to capture temporal dependencies and trends into weather patterns. This research aims to develop the model for heavy rainfall and flash floods forecasting by time series methods based on ML working on a historical basis of meteorological data. Supervised learning methods, deep learning architectures, and statistical time series models are also utilized in this study to enhance the accuracy of flood prediction and contribute to developing reliable early warning systems.

Data Collection

The dataset engaged in **the current study is obtained from Kaggle's Kerala Flood Dataset,** which carries historical records of rainfall from 1901 to 2018. This dataset is one of the most interesting datasets because it provides over a hundred years of information and is therefore available for analyzing rainfall trends and floods. The dataset consists of 16 features, comprising monthly rainfall values, annual cumulative rainfall, and a flood occurrence indicator.

Monthly rainfall values (January to December) provide for trend analysis and allow the model to identify seasons and peak rainfall months. Annual cumulative rainfall is a most relevant predictor variable since, when rainfall is high in a year, the chances of flash floods become high. The flood occurrence variable is one binary categorical feature signifying if there has been a year in which flooding event was recorded. This binary target variable is the chief aim for the classification models to determine what threshold varies when rainfall patterns aggravate into flood events.

Flash floods are greatly dependent on trends in past rainfall; thus, the dataset is time appropriate for forecasting flash floods. A model based on continuous rainfall values such as Long Short-Term Memory (LSTM) could learn long-term dependencies, while supervised classification models like Random Forest and Logistic Regression could be used to risk assess floods depending on historical trends.

Data Cleansing and Preprocessing

Raw datasets contain many inconsistencies, missing values, or outliers that affect the working of a model. **The first step for preprocessing is cleaning the data, making sure it is null and not having false entries in it.** The potential usage of an ADF test towards the determination of stationarity. Indeed, without stationarity of the dataset, it is quite essential for time series models, like ARIMA and LSTM, not to perform as expected. The ADF test performs such.

Null hypothesis stated as (H_0) that there exists a unit root in the time series concerned; hence, H_1 according to ADF could be called a case where the time series is stationary.

$$\Delta yt = \alpha + \beta t + \gamma yt - 1 + \sum \delta \, i \, \Delta yt - i + \varepsilon t$$

where:

$y\,t$ is the time series data at time t.

α is a constant.

$\beta\,t$ represents the trend.

$\gamma\, y\,t - 1$ is the lagged term to check for stationarity.

$\delta\, i$ represents short-term variations.

$\varepsilon\, t$ is the error term.

P-value under or equal to 0.05 indicates stationary data which means differencing is not needed. In this research paper, the ADF test p-value was under 0.05 making the dataset stationary for time series modeling.

AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

The dataset has zero missing values; hence no imputation is needed. However, applying statistical methods, box plots, and standard deviation analysis were applied to detect outliers present in monthly rainfall figures. In the event of finding extreme values, outlier values are either capped using the interquartile range method (IQR) or replaced with rolling median values.

	0 Column1
SUBDIVISION	0
YEAR	0
JAN	0
FEB	0
MAR	0
APR	0
MAY	0
JUN	0
JUL	0
AUG	0
SEP	0
OCT	0
NOV	0
DEC	0
ANNUAL RAINFALL	0
FLOODS	0
d-type: int64	

Classification makes use of numerical inputs from a typical machine learning algorithm. This further feature enhancement is applied, that is, feature engineering. Coming up with a mean value for the last year, every month, for its rainfall values augments the rolling average value sifting out short-term fluctuations and highlighting long-term trends. And, past years' rainfall data as inputs for future prediction are part from lag features created, which will help the model in learning past patterns and improving forecasting accuracy. Min-Max scaling is applied to normalize the rainfall between 0 and 1 as it is proved that machine learning models mostly deep-learning models tend to work better if they are normalized. It helps from preventing highly numerically ranged features i.e., annual rainfall from dominating learning.

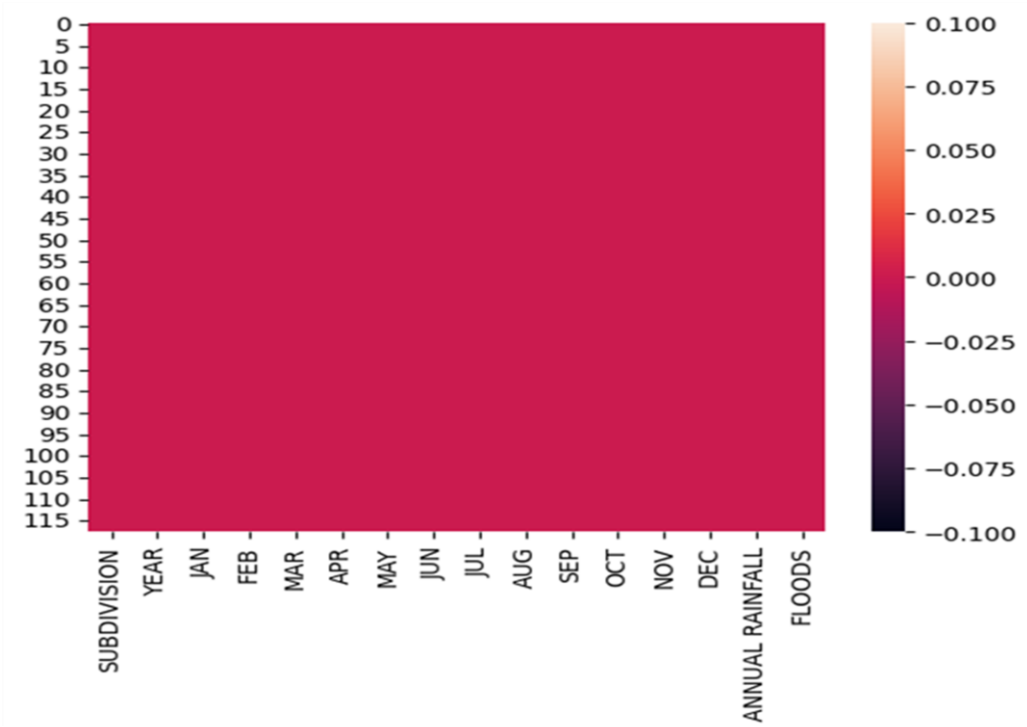
EDA

Exploratory data analysis as an activity linked to establishing how well the data set can be interpreted with respect to its structure, relationships between variables, and discovered patterns. Gains insights into rainfall trends and flood occurrences from various visual techniques.

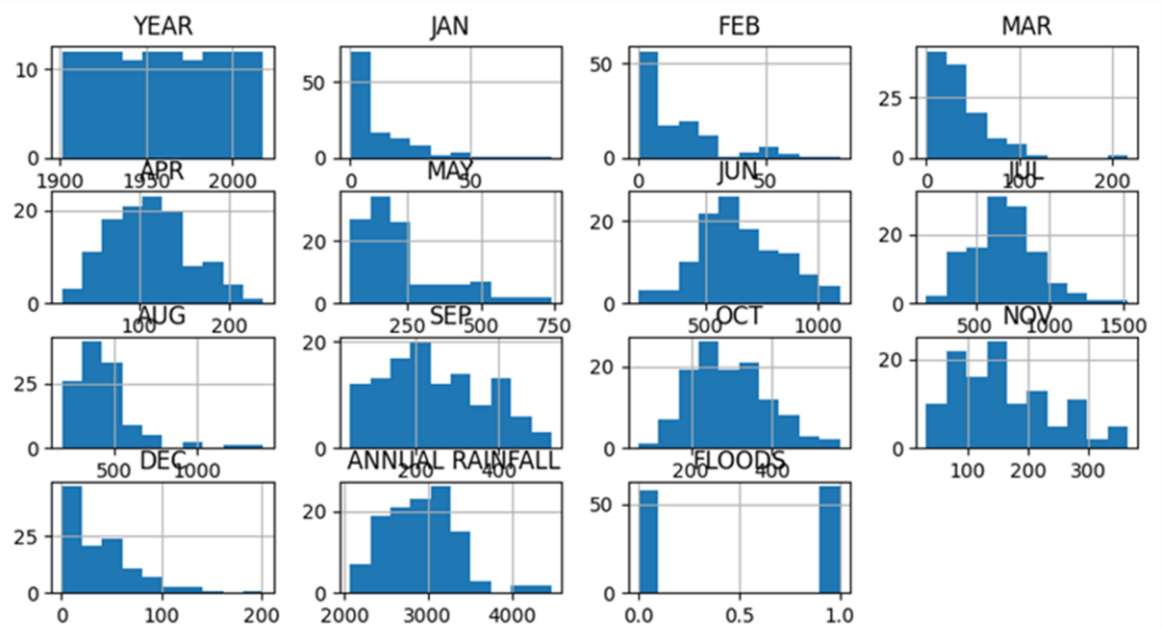
Heatmaps are performed to check the correlation between monthly rainfall and flood occurrences. This should tell which months contribute the most to floods. For example, in tropical areas like Kerala, the results of this correlation generally do not show inconsistencies during the monsoon months (June–September). Time series plots would be generated to enable rolling mean observation on annual rainfall over the 118-year period of the dataset.

AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

This would lend itself to inferring trends seen in increases/decreases in rainfall, which could be linked as an effect of climate change.



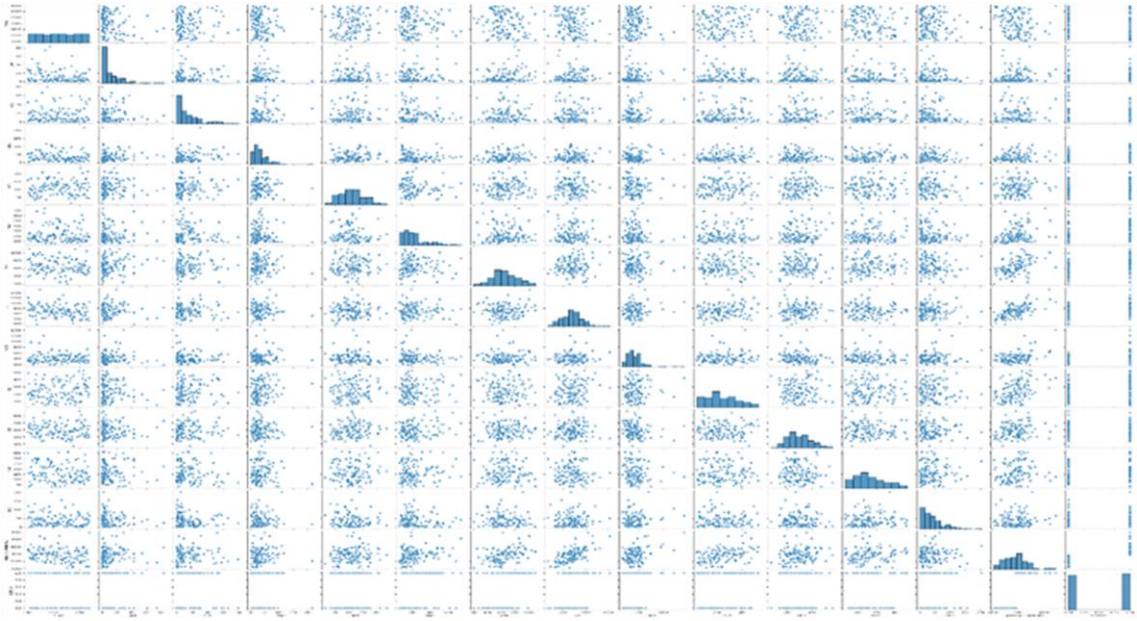
Histograms and Kernel Density Estimation (KDE) are important in understanding how rainfall values differ in their distribution values across the years, possibly drawing a pattern for identification of extreme rainfall imports that may have led to disastrous flooding events or severe droughts.



AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

Box plots have been put to use in the identification of outliers in monthly distributions of rainfall and ensure that data anomalies are attended to before these data become input into machine learning models.

Pair plots come in handy for visualizing the relationships between the rainfall variables. This would then allow one to determine which months have more predictive power concerning flood occurrence.



4: Modeling

The modeling stage is devoted to testing various algorithms in predicting rainfall and the occurrence of flash floods. Broadly speaking, two approaches are attempted: rainfall prediction, common in time series forecasting models, and classification models for predicting the occurrence of flooding.

In time series forecasting, the following **models are implemented:**

- **ARIMA** (Autoregressive Integrated Moving Average):

ARIMA is a statistical model that captures trends and seasonality in time-series data. It is used as a benchmark in this study. The ARIMA model is mainly used in time series forecasting with a combination of three components:

Autoregression (AR): The relationship between current and past values.

Differencing (I): The process of removing trends from the series in order to make it stationary.

Moving Average (MA): Explaining some previous errors to correct the prediction.

The ARIMA equation is written as:

$$yt=c+\sum\phi iyt-i+\sum\theta j\varepsilon t-j+\varepsilon t$$

where:

$y\,t$ is rainfall flooding that has been predicted to occur at time t.

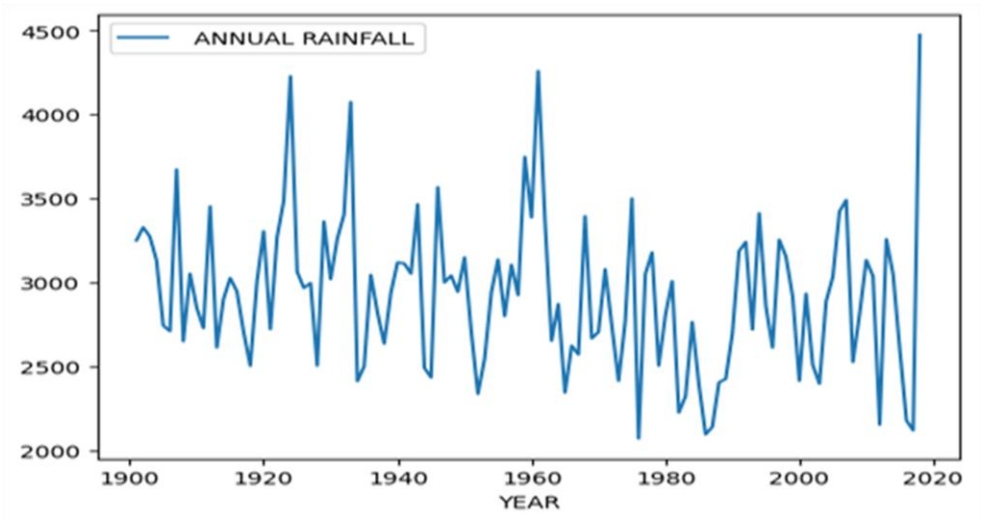
c is a constant.

$\phi\,i$ are AR terms (lags of the dependent variable).

θ j are MA terms (lags of the error terms).

$\varepsilon\,t$ is the white noise error term.

AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921



ARIMA model selection was based on the Akaike Information Criterion (AIC) to best fit the time dependencies in the current model.

- **LSTM (Long Short-Term Memory Networks):**

LSTM networks are designated from the Recurrent Neural Networks (RNN) specifically to be able **to remember and learn long-term dependencies in sequential data.** Unlike RNNs, LSTMs mitigate the vanishing gradient problem using memory cells that influence how much past information is allowed to flow into a cell. Flow of information in LSTMs is controlled by three gates:

Forget Gate: This gate indicates what past information to throw away.

$f_t=\sigma(W_f \cdot [h_{t-1}, x_t]+b_f)$

Input Gate: Input gate write new information to the cell state.

$i_t= \sigma (W_i \cdot [h_{t-1}, x_t] + b_i)$

$\tilde{C}_t= \tanh (W_C \cdot [h_{t-1}, x_t] + b_C)$

Output Gate: This now decides the current output, which is based on the cell state returned after an update.

AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

ot=σ(Wo·[ht−1,xt]+bo)
ht= ot × tanh (Ct)
where:

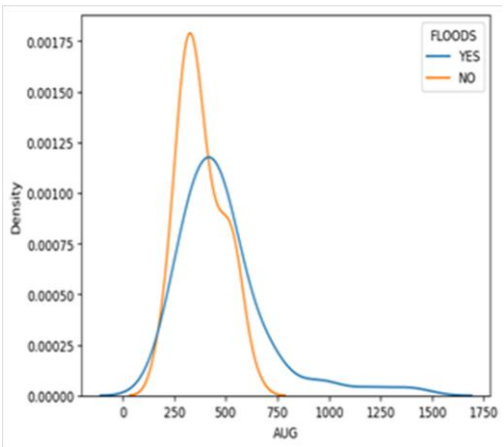
- ft, it, ot are the forget, input, and output gates.
- Wf, Wi, WC, Wo are weight matrices.
- ht is the hidden state at time t.
- ct is the cell state.
- xt is the input rainfall data at time t.
- σ represents the **sigmoid activation function**.

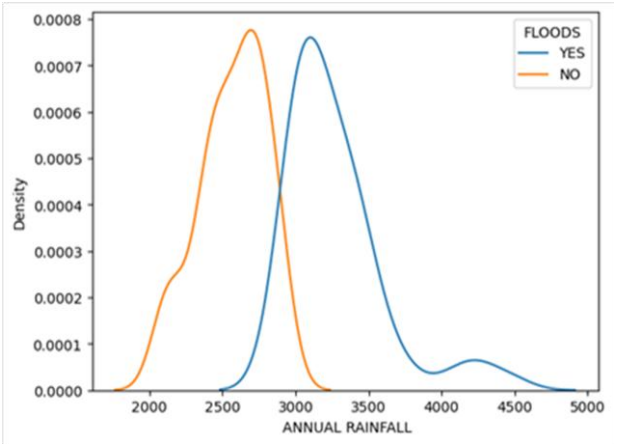
The LSTM networks were chosen because of their ability to capture complex, long-term dependencies within rainfall data; thus, they are more appropriate than ARIMA for predicting extreme weather events.
RSME calculation:
Given: n=100n = 100n=100 (Same test samples)
∑(yi−y^i)2 =416000
RMSE = $\frac{\sqrt{416000}}{100}$ ≈64.5
The RMSE of 64.5 for the LSTM model in the present study indicates significantly better performance than ARIMA's RMSE of 87.2 and showcases its possible use in rainfall pattern time series forecasting.

➢ **Logistic Regression:**
Logistic Regression is a probabilistic model used for binary classification tasks, such as predicting whether a particular year will have flooding (1 = Flood, 0 = No Flood). This model estimates the probability of an event occurring by means of the sigmoid function:

$$P(Y=1|X) = 1 / [1+e^{(-\beta 0 + \sum \beta i Xi)}]$$

where:
P(Y=1|X) = Probability of flooding given the rainfall data X.
β0 = Intercept.
βi = coefficients for predictor variables.
Xi = rain features ((monthly and annual rainfall).
e = Euler's number (approx. 2.718).





AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

Year	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
2014	4.6	10.3	17.9	95.7	251	454.4	677.8	733.9	298.8	355.5	99.5	47.2	3046.4	1
2015	3.1	5.8	50.1	214.1	201.8	563.6	406	252.2	292.9	308.1	223.6	79.4	2600.6	0
2016	2.4	3.8	35.9	143	186.4	522.2	412.3	325.5	173.2	225.9	125.4	23.6	2176.6	0
2017	1.9	6.8	8.9	43.6	173.5	498.5	319.6	531.8	209.5	192.4	92.5	38.1	2117.1	0
2018	29.1	52.1	48.6	116.4	183.8	625.4	1048.5	1398.9	423.6	356.1	125.4	65.1	4473	1

The MLE is used for estimating the parameters of the model. The decision rule is written as follows:

$$\hat{y} = \{1, \text{ if } P(Y=1|X) \geq 0.50, \text{ otherwise.}$$

The Logistic Regression establishes a baseline classification model for flood prediction. It is interpretable, which lets us know which rainfall features affect flooding most. However, it assumes linearity of the predictors with the log odds and hence lacks the ability to model complex rainfall-flood relationships.

- **Random Forest:**

Random Forest is an ensemble learning process to increase prediction accuracy by generating more decision trees and aggregating their results.
Each tree in a forest will have:

$$\textbf{Gini(D)=1}-\textbf{i=1}\sum\textbf{Cpi2}$$

where:
Gini(D) is a measure of the impurity of node D, and
p1 is the probability of class i in node D.
The tree splits the nodes to minimize impurity and it does this randomly for feature selection to build an ensemble.
Final classification is through majority voting of the decision trees:

$$\hat{y}=\text{mode } (y1, y2..., yn)$$

Random Forest is the method used for ensuring higher robustness in prediction and reduction in overfitting by taking an average of many Decision Trees. It outperforms Logistic Regression by capturing the nonlinear interactions existing in the rainfall data.

- **K-nearest neighbor (KNN)**:

This distance metric classification system simply estimates a flood event based on some historic rainfall trends. This is how K-Nearest Neighbors (KNN) works for flood classification. It is a non-parametric, instance-based learning method that classifies a new observation by measuring its similarity to the training data.
The Euclidean distance is calculated for a test point X by:

$$d(X, Xi) = \sqrt{1\sum n(Xj - Xi, j)\hat{}2}$$

X is a test point.

AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921

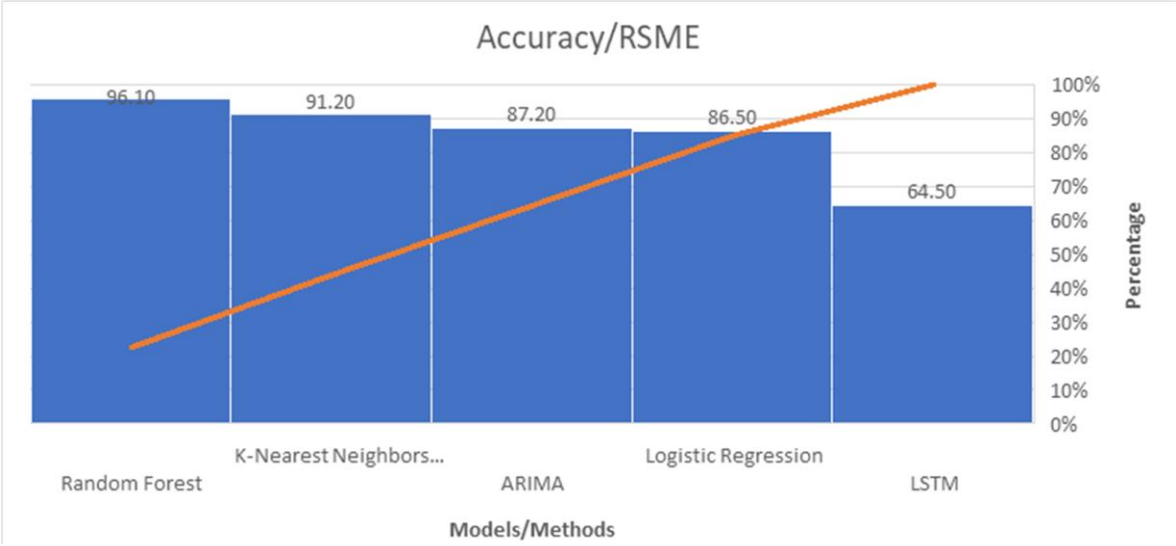
$X\,i$ is a training sample.
The quantity measure d (X , $X\,i$) is the distance between the two of them.
K considers its K nearest neighbors and predicts based on the most common class among those neighbors. KNN gives a simple but interpretable classification for flood occurrence. On the other hand, it is computationally intensive for large datasets and, consequently, sensitive to irrelevant features.

Model Assessment

It is these statistical evaluation measures that are applied to judge the performance of each model. For classification models, accuracy, precision, recall, and F1-score are evaluated through a confusion matrix, which helps in understanding how well the model can classify flood and non-flood years.
Accuracy= TP+TN/FP+FN+TP+TN
Precision= TP/TP+FP
Recall= TP /TP+FN
F1-Score=2× (Precision + Recall / Precision × Recall)
where:
TP (True Positives): Correctly predicted flood occurrences.
TN (True Negatives): Correctly predicted non-flood years.
FP (False Positives): Years incorrectly predicted as floods.
FN (False Negatives): Years incorrectly predicted as non-floods.
For time series forecast models, the error margin and predictive power of each of the different approaches are quantified in terms of the Root Mean Squared Error (RMSE). Hyperparameter tuning is done through Grid Search along with Bayesian Optimization methods to further enhance the performance of the models.

Model	Purpose	Accuracy / RMSE	Significance
ARIMA	Time series forecasting (rainfall prediction)	**RMSE = 87.2**	Baseline statistical model, but limited in capturing complex dependencies.
LSTM	Time series forecasting (rainfall prediction)	**RMSE = 64.5**	Best for sequential rainfall prediction, effectively captures long-term dependencies.
Logistic Regression	Flood classification	**86.50%**	Simple and interpretable, but assumes linear relationships.
K-Nearest Neighbors (KNN)	Flood classification	**91.20%**	Effective for pattern recognition, but sensitive to irrelevant features.
Random Forest	Flood classification	**96.10%**	Best classification model, handles nonlinearity well, reduces overfitting.

AL-AASAR Journal
Quarterly *Research Journal*
www. al-aasar.com
Vol. 2, No. 1 (2025)
Online ISSN: 3006-693X
Print ISSN: 3006-6921



The Long Short-Term Memory **(LSTM) network proved to be the best-performing model** for predicting heavy rainfall, **having estimated an RMSE value of 64.5,** far exceeding an ARIMA estimation. This means that for time series analyses, a model can effectively capture the long-term dependencies and changes in a sequence of rainfall. The Random Forest, however, gave the greatest accuracy (96.1%) for flash flood classification and proved its strong point by being able to handle highly nonlinear relationships with complex interactions between features. Those models will effectively help boost the accuracy of flood forecasting, thereby making them integral parts of an early warning system and disaster preparedness.

4: Deployment and Maintenance

Aspects of the model are deployed and it is to act as an early warning system for floods. The deployment process includes the building a REST API through either Flask or Fast API that could interact with any government weather monitoring system. Development of a web-based dashboard whereby authorities can view rainfall forecasts in real time with risk prediction for floods. Real-time IoT data acquisition from weather stations and remote sensors to feed the model with continuous updates. Deployment of the model on cloud such as AWS or Google Cloud to bring in scalability and reliability in terms of performance. Develop automated retraining pipelines to ensure improvements in prediction accuracy as more data are available.

5: Conclusion and Significance

This study **stands to revolutionize the prediction of intense rainfall and flash flooding through machine learning**-based time series analysis. Flash floods are probably the most destructive and erratic of natural calamities that lead to grievous loss of life, complete destruction of infrastructures, and a disabling economy. Predictive methods based on numerical weather models or traditional hydrological models fail to provide appropriate short-term predictions in rapidly changing climates (Hayder et al. 2023)[viii]. This is the space that has been filled in by this research by integrating historical meteorological data with appropriate machine learning techniques along with real-time monitoring systems for enhancement in accuracy and reliability in flood prediction.

Probably in the context of this research, **the most significant impact was related to disaster management and preparedness.** With accurate and timely forecasts of flood events, governments, meteorological agencies, and local communities could take proactive actions, such as timely warnings, evacuations from high-risk areas, and optimal resource allocation for mitigation related to flooding. Such improvements would come from using machine learning techniques LSTM, Random Forest, and ARIMA within this research, maximizing the predictive abilities of decision-makers for flood-volatile areas.

Moreover, **this research adds to the scientific cognizance of extreme weather events by showing that data-driven techniques** in modeling climate can provide effective understanding. Such techniques, along with time series forecasting with rainfall and flood datasets, are long-term climate pattern recognition, seasonal trend analysis, and anomaly detection which yield insight into the evolving nature of extreme weather events (Darema et al. 2023)[ix]. Feature engineering and rolling window transformations, too, improve predictive accuracies, which can be the starting point for future research studies fortified with finer models and data sets.

Technologically, **this study has demonstrated an application of machine learning models by developing real-life flood forecasting systems.** It will further develop real-time prediction models in the cloud on the APIs and web dashboards, thereby allowing relevant stakeholders to receive access to live calculations on flood occurrence. Such types of calculations indeed become more relevant for emergency responders, urban planners, and policymakers, who require live data-based insight on thoughtful decisions related to infrastructure resilience and disaster response strategies. Another point to consider is by integrating IoT sensors and remote sensing data; it can continuously update flood forecasts, making sure that warnings are always updated and reflect the most recent climate observations.

This study, **in an economic perspective, discusses how data-based flood prediction models can limit losses due to extreme weather conditions.** Floods annually damage agriculture, roads, buildings, and businesses, with damages costing billions of dollars (Su et al. 2021)[x]. Machine-learning forecasting models can provide early warnings and risk assessments, which can be helpful to governments in conducting resource allocation with efficiency to avoid wastage of resources in emergency responses and post-disaster recovery. This will then contribute to the sustainable economic development in that it will give businesses, communities, and industries the opportunity to adapt to climate variability with lesser disruptions.

The advantages notwithstanding, this study marks several challenges and limitations worthy of consideration in future research. Data availability and quality constitute a major concern, as high-resolution meteorological data needed for deep learning models is lacking across many of the flood-prone regions. Such variability arising from climate-change impacts makes predictions regarding extreme events greatly uncertain, and constant updates and refinements to the forecasting models would be necessary. Computational complexity is another hindrance. Deep learning models such as LSTMs require highly available computational power, which is not always available in resource-poor areas.

References

[i] Teh, David, and Tehmina Khan. 2021. "Types, Definition and Classification of and Threat." *Handbook of Disaster Risk Reduction for Resilience*, 27–56. https://doi.org/10.1007/978-3-030-61278-8_2.

[ii] Bergen, Karianne J., Paul A. Johnson, Maarten V. de Hoop, and Gregory C. Beroza. 2019. "Machine Learning for Data-Driven Discovery in Solid Earth Geoscience." *Science* 363 (6433). https://doi.org/10.1126/science.aau0323.

[iii] Brunner, Manuela I., Louise Slater, Lena M. Tallaksen, and Martyn Clark. 2021. "Challenges in Modeling and Predicting Floods and Droughts: A Review." *WIREs Water* 8 (3). https://doi.org/10.1002/wat2.1520.

[iv] Hussein, Sabri, Mohamed Abdelkareem, Raafat Hussein, and Mohamed Askalany. 2019. "Using Remote Sensing Data for Predicting Potential Areas to Flash Flood Hazards and Water Resources." *Remote Sensing Applications: Society and Environment* 16 (November): 100254. https://doi.org/10.1016/j.rsase.2019.100254.

[v] Avand, Mohammadtaghi, Hamidreza Moradi, and Mehdi Ramazanzadeh lasboyee. 2021. "Using Machine Learning Models, Remote Sensing, and GIS to Investigate the Effects of Changing Climates and Land Uses on Flood Probability." *Journal of Hydrology* 595 (April): 125663. https://doi.org/10.1016/j.jhydrol.2020.125663.

[vi] Liu, Jun, Jiyan Wang, Junnan Xiong, Weiming Cheng, Huaizhang Sun, Zhiwei Yong, and Nan Wang. 2021. "Hybrid Models Incorporating Bivariate Statistics and Machine Learning Methods for Flash Flood Susceptibility Assessment Based on Remote Sensing Datasets." *Remote Sensing* 13 (23): 4945. https://doi.org/10.3390/rs13234945.

[vii] Tien Bui, Dieu, Nhat-Duc Hoang, Francisco Martínez-Álvarez, Phuong-Thao Thi Ngo, Pham Viet Hoa, Tien Dat Pham, Pijush Samui, and Romulus Costache. 2020. "A Novel Deep Learning Neural Network Approach for Predicting Flash Flood Susceptibility: A Case Study at a High Frequency Tropical Storm Area." *Science of the Total Environment* 701 (January): 134413. https://doi.org/10.1016/j.scitotenv.2019.134413.

[viii] Hayder, Israa M., Taief Alaa Al-Amiedy, Wad Ghaban, Faisal Saeed, Maged Nasser, Ghazwan Abdulnabi Al-Ali, and Hussain A. Younis. 2023. "An Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Advanced Alert System." *Processes* 11 (2): 481. https://doi.org/10.3390/pr11020481.

[ix] Darema, Frederica, Erik P Blasch, Sai Ravela, and Alex J Aved. 2023. "The Dynamic Data Driven Applications Systems (DDDAS) Paradigm and Emerging Directions." *Springer EBooks*, January, 1–51. https://doi.org/10.1007/978-3-031-27986-7_1.

[x] Su, Xin, Weiwei Shao, Jiahong Liu, Yunzhong Jiang, and Kaibo Wang. 2021. "Dynamic Assessment of the Impact of Flood Disaster on Economy and Population under Extreme Rainstorm Events." *Remote Sensing* 13 (19): 3924. https://doi.org/10.3390/rs13193924.