International Journal of Computational Science and Engineering Research
ISSN: 3107 -8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 2, Issue. 2 (April – June), 2025 , Pages: 54 - 59
Received: 28 January 2025 ; *Accepted:* 29 April 2025 ; *Published: 30* June 2025
https://doi.org/10.63328/IJCSER-V2RI2P11
Original Paper : Applied Research – Compare and Contrast Research Paper

Flood and Landslide Prediction Using AI and Ensemble Machine Learning Models

P GangadharaReddy[1], J Pradeep [2] , P Naga Subba Rayudu [3] , T Ramashri [4]

[1-4] Department of ECE, Aditya College of Engineering, Madanapalle, Andhra Pradesh, India.

** Corresponding Author : P Gangadhar Reddy ; gangadharreddy.p@gmail.com*

Abstract: Floods and Landslides are among the most devastating natural disasters, causing significant loss of life, damage to infrastructure, and disruption of livelihoods. With climate change, rapid urbanization, and environmental degradation, the frequency and intensity of these disasters have increased globally. Effective prediction and early warning systems are critical in mitigating their impacts and improving disaster preparedness. This work proposes a Machine Learning- based approach for flood and landslide prediction by analyzing environmental factors such as monsoon intensity, deforestation levels, urbanization, topographical changes, and climatic variations. The dataset utilized for this work includes multiple influencing parameters. Various Machine Learning models, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, Ridge Regression, XGBoost, and Gradient Boosting, are employed to predict flood and landslide probabilities. The predictive Modeling process begins with feature selection, where significant environmental variables contributing to floods and landslides are identified. Data preprocessing techniques such as Normalization and Standardization are applied to improve model efficiency. Several performance metrics, including Accuracy, Precision, Recall, F1-score, and RMSE, are used to assess model effectiveness. Results from the study indicate that Logistic Regression performs best in classifying flood-prone areas, achieving a good accuracy. Similarly, Ridge Regression and Gradient Boosting models are effective in estimating the severity of landslides.

Keywords: SVM), K- Nearest Neighbors (KNN), Naïve Bayes, Ridge Regression, XGBoost, and Gradient Boosting.

1. Introduction And Background

Floods and Landslides are among the most frequent and devastating natural disasters affecting millions of people worldwide. These catastrophic events cause significant loss of life, damage to infrastructure, destruction of agricultural land, and economic disruption. The unpredictable nature of rainfall patterns can give rise to extreme weather events, such as prolonged droughts or deviating floods, which can have far-reaching consequences for ecosystem, agriculture, and human population [1]. According to the National Centers for Environmental Information, the projected global average precipitation for 2021 stands at 2.66 millimeters per day, slightly below the 40-year climatological mean of 2.69millimeters per day [2].

Floods occur when an overflow of water submerges land that is usually dry. They can be caused by heavy rainfall, storm surges, dam failures, rapid snowmelt, or poor drainage systems. Techniques such as Regression Analysis, Artificial Neural Networks have proven to be effective in climate prediction [3].Flash floods are particularly dangerous due to their sudden onset and high velocity. Landslides involve the movement of rock, soil, and debris down a slope due to gravity. These can be triggered by intense rainfall, earthquakes, volcanic activity, or human activities such as deforestation and improper land use.

The traditional approach to Flood and Landslide prediction relies on physical and statistical models that use historical data, meteorological observations, and hydrological analysis. Machine learning algorithms have been extensively studied for their effectiveness in rainfall prediction [4]. However, these models often fail to capture the complexity of environmental interactions, leading to inaccurate predictions and delayed responses. The advent of Artificial Intelligence (AI) and Machine Learning (ML) has introduced new possibilities for improving disaster forecasting by leveraging large datasets, pattern recognition, and adaptive learning capabilities.

Machine learning-based prediction models analyze various environmental factors, including rainfall patterns, soil moisture levels, deforestation rates, land topography, and climate variations. These models simulate the complex interactions of atmospheric, oceanic, and

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 2, Issue 2 , 2025, https://doi.org/10.63328/IJCSER-V2RI2P11

environmental factors to predict changes in the global climate system in response to initial atmospheric conditions [5].By learning from past data, these models can predict the likelihood of floods and landslides more accurately than traditional methods. This approach not only improves early warning systems but also enhances decision-making for disaster management and resource allocation. Early warnings allow authorities to implement evacuation plans, reinforce infrastructure, and mobilize emergency response teams.

This work is driven by the need to enhance disaster preparedness, minimize economic losses, and safeguard vulnerable communities through data-driven. The predictive framework employs both classification and regression techniques to determine flood probability and estimate the expected number of landslides. This classification is achieved through multiple machine learning models, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes. The system evaluates these models based on their accuracy, recall, precision, and F1-score to select the best-performing classifier for flood prediction. The Regression is achieved through multiple ML models including Linear regression , Ridge Regression, LASSO Regression, XG Boost. The system Evaluates based on Evaluation metrics such as MSE, MAE,RMSE,R2.

2. Literature Survey

Z. He et.al. conducted a comprehensive study on rain forecasting using an active learning system. They utilized a large dataset from Kaggle, consisting of historical weather data, and employed entropy sampling as the query approach [6]. This study focuses on assessing the efficiency of the machine learning algorithms—Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM), KNN, GNB for flood susceptibility.

A. G. Nolan et.al. discussed a decision-tree model to predict weather outcomes, specifically focusing on forecasting rainy days [7]. By analyzing these factors, the study aims to predict areas at higher risk of landslides.

M. S. Balamurugan et.al. conducted a comparative study to evaluate the performance of machine learning techniques in rainfall prediction compared to statistical methods [8]. An approach involves developing a cutting-edge computer vision model capable of real-time detection of landslides in social media image streams.

O.Ejike et.al. explored the application of logistic regression modeling to predict rainfall uses a combination of traditional machine learning methods (SVM, LR, RF) with CNN (Convolutional Neural Network) [9].

A. D. Kumarasiri et.al. developed neural network models for rainfall prediction at different time scales [10]. This process begins with the creation of a large dataset of landslide images.

N. J. Ria et.al. conducted a rain prediction study using machine learning models based on a dataset, gathered information about things like the slope of the land, how close it is to streams and roads, and other factors that might affect landslides [11].

S. Neelankandan et.al.proposed a CA-SVM-based prediction model for rainfall forecasting using real-time data sets [12]. Final landslide susceptibility maps (LSMs) are generated using these hybrid models, alongside maps produced using individual ML methods for comparison.

S. Sankaranarayanan utilized machine learning methods, including Artificial Neural Networks (ANN) with a single hidden layer, to predict floods based on various factors such as precipitation, temperature, water velocity, water level, and humidity [13].

3. ML Methods Used In Flood and Landslide Prediction

Flood Prediction Techniques

Traditional Hydrological Models: Hydrological models have been used extensively for flood prediction. These models rely on physical principles to simulate the movement, distribution, and quality of water.

Common hydrological models include: HEC-RAS (Hydrologic Engineering Center's River Analysis System),SWAT (Soil and Water Assessment Tool).

Although these models offer accurate predictions, they require extensive calibration and high-quality input data, making them resource-intensive.

Machine learning approaches have revolutionized flood prediction by enabling automated learning from historical data. Some widely used ML models include: Logistic Regression (LR),Support Vector Machines (SVM),Random Forest (RF),K-Nearest Neighbours (KNN). These models have demonstrated improved efficiency in predicting flood probability, especially when combined with feature selection and optimization techniques. Deep learning models have recently been applied for flood prediction due to their ability to capture complex patterns in large datasets. Some notable models include: Convolutional Neural Networks (CNNs),Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM),Transformer-Based Models.

Landslide Prediction Techniques

Geospatial and Geological Models*:* Landslide susceptibility mapping relies on geospatial data and geological surveys to determine high-risk areas. These models consider: Slope and Elevation Data: Landslides are more frequent in steep

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 2, Issue 2 , 2025, https://doi.org/10.63328/IJCSER-V2RI2P11

regions, Soil Composition: Clayey soils tend to be more prone to landslides, Rainfall and Seismic Activity: High precipitation and seismic movements can trigger landslides. Various geospatial analysis tools, such as GIS and remote sensing, have been used to map and predict landslides.

Machine Learning-Based Models: Similar to flood prediction, ML models have been widely applied to landslide prediction, with some commonly used algorithms including: Decision Trees and Random Forest, Gradient Boosting Methods (XG Boost, Light GBM),Artificial Neural Networks (ANNs).

Hybrid Models: To improve prediction accuracy, researchers have explored hybrid models that combine multiple techniques. For instance:

Integrating Hydrological and ML Models: Combining traditional hydrological models with ML algorithms improves overall flood and landslide prediction accuracy.

Hybrid Deep Learning Models: Using CNNs,these models have shown promising results in capturing both static geological features and dynamic patterns.

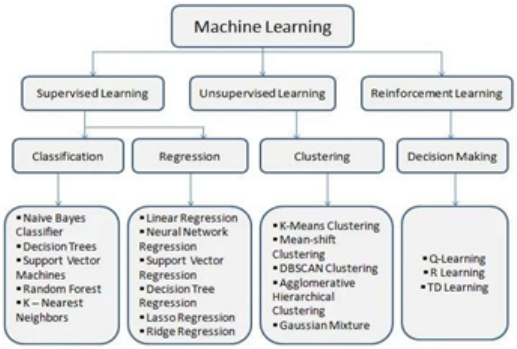

Figure. 1 Machine Learning

In this work we use Python as a software tool.

Python: It is an object-oriented, high-level programming language . Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages.

Figure. 2 Python

4. Proposed Methodology

The proposed system aims to develop an efficient and robust machine learning-based predictive model for flood and landslide detection by analyzing multiple environmental and human-influenced factors. The system is designed to process large volumes of structured data, consisting of parameters such as monsoon intensity, topography drainage, river management, deforestation levels, urbanization rates, climate change indicators, siltation levels, agricultural practices, encroachments, drainage systems, coastal vulnerability, watersheds, deteriorating infrastructure, population density, wetland loss, inadequate planning, and political factors. These features collectively determine the probability of flood occurrence and landslide susceptibility in a given region.

The predictive framework employs both classification and regression techniques to determine flood probability and estimate the expected number of landslides. The classification is achieved through multiple machine learning models, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and Naïve Bayes. The system evaluates these models based on their accuracy, recall, precision, and F1-score to select the best-performing classifier for flood prediction. The classification results are further validated using confusion matrices and ROC-AUC analysis to understand the trade-offs between true positive and false positive rates..

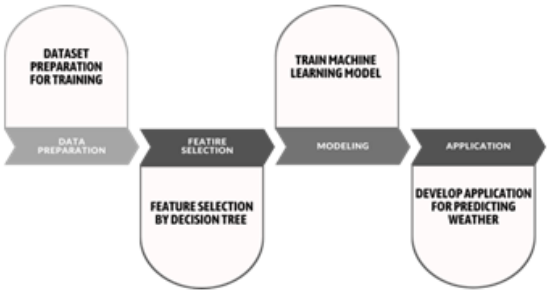

Figure. 3 Overview of proposed model

For regression-based landslide prediction, the proposed system integrates data-driven insights with real-time applications, allowing urban planners, and disaster management agencies to make informed decisions. By employing effective feature engineering techniques, the researchers aimed to identify the most informative variables that would contribute to the construction of highly accurate prediction models. The ability to predict floods and landslides enables early warning mechanisms, resource allocation strategies, and preventive measures to mitigate disaster impacts.

The system can be extended to incorporate real-time sensor data, satellite imagery, and Geographic Information System (GIS) data for enhanced prediction accuracy. Future enhancements may also include deep learning-based models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture complex spatiotemporal patterns in environmental changes. Overall, the proposed system offers a comprehensive, data-driven approach to disaster prediction, combining statistical modeling with machine learning to enhance resilience against natural calamities.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 2, Issue 2 , 2025, https://doi.org/10.63328/IJCSER-V2RI2P11

5. Implementation

The implementation phase begins with pre-processing the data to make it suitable for modeling, followed by feature engineering to highlight key patterns. Once the data is prepared, model development and training take center stage. Fine-tuning hyper parameters and evaluating the model's performance ensures that the solution meets the set objectives. Lastly, deploying the trained model and integrating it into the target application completes the process. Implementing flood prediction using machine learning (ML) involves using historical data, weather patterns, and other factors to develop predictive models, allowing for early warnings and minimizing the impact of floods. The model can also assist urban planners in designing infrastructure resilient to environmental hazards. Furthermore the researches can extends this study by deep learning and IOT based real time techniques.

The following machine learning algorithms are implemented using python.

Classification Models: There are five types of classification models are there. They are

Logistic Regression:Logistic Regression is a supervised learning classification algorithm used to predict the probability of a target variable.

Decision Tree: Decision tree is a hierarchical tree structure starts with one main question at the top called a node which further branches out into different possible outcomes.

Random Forest : The Random forest is used for classification, regression, and other tasks using decision trees.

Support Vector Machine (SVM):SVM is a powerful classifier that finds the optimal decision boundary to distinguish between classes.

Naive Bayes Classifier : The Naive Bayes Classifier is a simple probabilistic classifier that can predict at a faster speed than other classification algorithms.

Regression Models

Linear Regression : The main goal of the linear regression model is to find the best-fitting straight line (often called a regression line) through a set of data points.

Ridge Regression: Ridge regression—also known as L2 regularization,is one of several types of regularization for Linear regression models.

XG Boost:XG Boost, or extreme Gradient Boosting, is a XG Boost algorithm in machine learning algorithm under ensemble learning.

Gradient Boosting: Gradient Boosting is ensemble learning method used for classification and regression tasks.

6. Results

Classification Model

The flood prediction was formulated as a binary classification problem, where six different classification algorithms were employed: Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Gaussian Naive Bayes (GNB). Among these models, Logistic Regression achieved the highest accuracy of 92.61%, followed by SVM with 91.95%. In contrast, the Decision Tree and Random Forest classifiers performed relatively poorly, with accuracy values around 70.6% and 70.2%, respectively. This suggests that simpler models, such as Logistic Regression and SVM, were better suited for this specific dataset compared to tree-based models.

Table. 1 Results For Classification Model

	ACCURACY	PRECISION	RECALL	FL- SCORE
LR	0.9261	0.9162	0.9154	0.9158
SVM	0.9195	0.9052	0.9106	0.9079
DT	0.7061	0.6495	0.6704	0.6597
RF	0.7020	0.6438	0.6656	0.6545
KNN	0.8280	0.7398	0.8487	0.7905
GNB	0.9103	0.8618	0.9286	0.8940

Observations

Logistic Regression (Best Performance)

Achieved highest accuracy (92.6%), meaning it made the most correct predictions.

Balanced precision (91.6%) and recall (91.5%), ensuring minimal false positives and false negatives.

The best model for flood classification.

SVM (Slightly Lower than Logistic Regression)

Achieved 91.9% accuracy but had lower precision (90.5%). Slightly lower recall than logistic regression.

Decision Tree and Random Forest (Poor Performance)

Low accuracy (~70%), meaning they failed to classify floods correctly. Over fitting might have led to poor generalization.

KNN (Good Recall, Low Precision)

Achieved high recall (84.9%), meaning it detected most flood-prone areas. Precision was low (73.9%), indicating it

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 2, Issue 2 , 2025, https://doi.org/10.63328/IJCSER-V2RI2P11

misclassified some areas as flood-prone when they were not.

Gaussian Naïve Bayes (Balanced Performance)

Performed well with 91% accuracy and 92.9% recall. More sensitive to class imbalance.

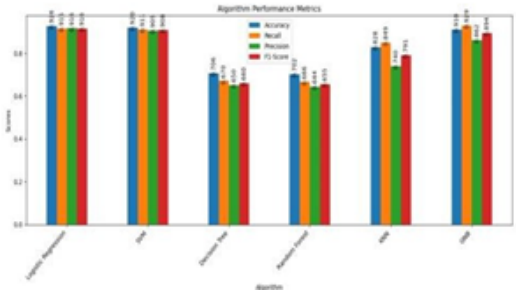

Figure. 4 Classification model performance

From the above figure, observe that Logistic Regression achieved the highest accuracy.

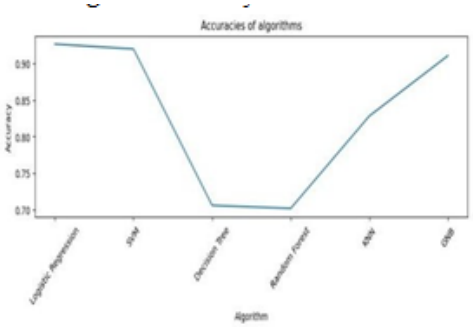

Figure. 5 Accuracy for classification model

Regression model:

The landslide prediction was modelled as a regression problem, where six regression techniques were tested. Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, XG Boost Regression, and Gradient Boosting Regression. Among these, Linear Regression performed the best, achieving an R2 score of 0.086, indicating that it could explain only 8.6% of the variance in landslide occurrences. Ridge Regression achieved an identical score, while Lasso Regression and Elastic Net Regression had slightly lower R2 values.

Table. 2 Results for Regression model

	MAE	MSE	RMSE	R2	AVERAGE
LR	1.0767	4.4664	2.1132	0.0860	2.0969
RR	1.7067	4.4669	2.1133	0.8060	2.0971
Lasso	1.7627	4.7629	2.1824	0.0254	2.1833
XG	1.7538	4.7486	2.1791	0.0283	2.1775
GB	1.7363	4.5927	2.1430	0.0602	2.1330

Observations

Linear & Ridge Regression (Similar Performance)

Both had an R2R^2R2 of 0.086, indicating low explanatory power. Performed similarly with minimal over fitting.

LASSO & Elastic Net (Poor Performance)

R2R^2R2 close to 0, meaning they struggled to capture variance. Over-penalized certain coefficients.

XG Boost & Gradient Boosting (Best Performers)

Gradient Boosting performed best (R2=0.060R^2 = 0.060R2=0.060). Had the lowest MSE(4.59) meaning lower prediction.

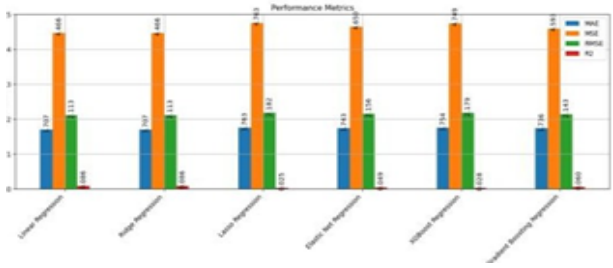

Figure. 6 Regression Model Performance

Among ensemble-based regression models, XGBoost Regression and Gradient Boosting Regression performed better than the simple regression techniques, yet their R2 scores remained low, at 0.028 and 0.060, respectively. The high Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values across all models indicate that predicting the exact number of landslides remains a challenging task, possibly due to the presence of complex and non-linear dependencies in the dataset that these models failed to capture.

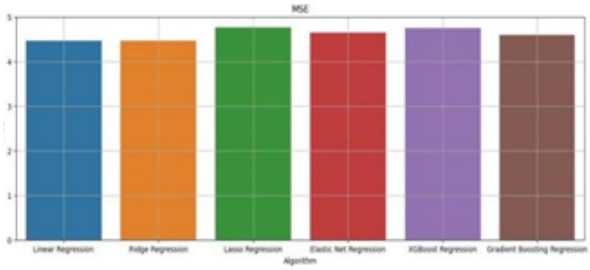

Figure. 7 MSME

From the above figure we observe that LASSO Regression achieved highest MSE.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 2, Issue 2 , 2025, https://doi.org/10.63328/IJCSER-V2RI2P11

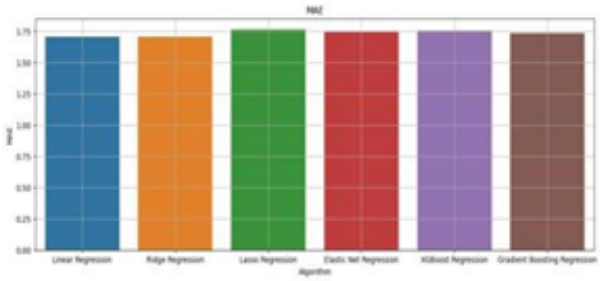

Figure. 8 MAE

From the above figure we observe that LASSO Regression and XG Boost similarly achieved highest MAE.

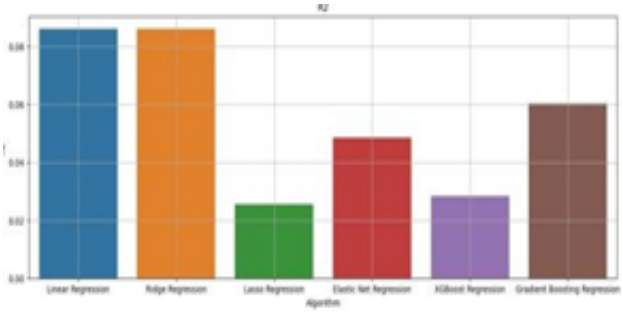

Figure. 9 R2

From the above figure we observe that Linear Regression and Ridge Regression has same R2 values.

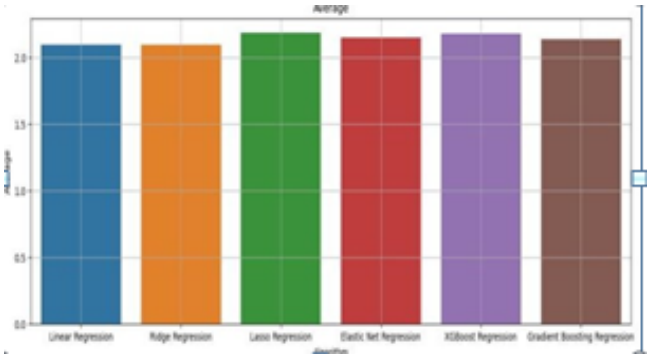

Figure. 10 Average

From the above figure we observe that based upon the average the Linear Regression achieved Lowest Error.

7. Conclusion

The Flood and Landslide prediction system presented in this work demonstrates a comprehensive and data-driven approach to tackling two of the most devastating natural disasters faced globally. By leveraging the power of machine learning, the system effectively identifies the likelihood of floods and estimates the number of landslides that may occur based on a wide range of environmental, geographical, and anthropogenic factors. These factors include critical parameters such as monsoon intensity, topography, river management efficiency, deforestation, urbanization levels, climate change indicators, quality of dams, siltation, and numerous others, all of which contribute significantly to disaster susceptibility in vulnerable regions. Logistic Regression emerged as the most effective in terms of accuracy, precision, recall, and F1-score, suggesting its robustness and reliability in binary classification tasks involving disaster risk. Gradient Boosting Regression and Linear Regression showed the most promising results, offering a low error margin and a good balance between complexity and accuracy.

The future scope for flood and landslide prediction using AI and ensemble machine learning models is highly promising, driven by advancements in data availability, computational power, and algorithmic development. Ensemble learning models, which combine the strengths of multiple algorithms, offer improved accuracy and robustness in predicting complex natural disasters like floods and landslides.

References

[1] P. K. Srivastava, A. Mehta, M. Gupta, S. K. Singh,and T. Islam, ‘‘Assessing impact of climate change on Mundra mangrove forest ecosystem, Gulf of Kutch, Western Coast of India: A synergistic evaluation using remote sensing,’’ *Theor. Appl. Climatol.*, vol. 120, nos. nos. 3–4, pp. 685–700, May 2015.

[2] *Annual 2021 Global Climate Report National Centers for Environmental Information (NCEI)*, Nat. Centers Environ. Inf., USA, 2022.

[3] R. Janarthanan, R. Balamurali, A. Annapoorani, and V. Vimala, ‘‘Prediction of rainfall using fuzzy logic,’’ *Mater. Today, Proc.*, vol. 37, pp. 959–963, Jan. 2021.

[4] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, ‘‘An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives,’’ *Exp. Syst. Appl.*, vol. 85, pp. 169–181, Nov. 2017.

[5] S.-C. Yang, Z.-M. Huang, C.-Y. Huang, C.-C. Tsai, and T.-K. Yeh, ‘‘A case study on the impact of ensemble data assimilation with GNSS-zenith total delay and radar data on heavy rainfall prediction,’’ *Monthly Weather Rev.*,vol. 148, no. 3, pp. 1075–1098, Mar.2020.