EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2026, 27(2), 198–215
https://doi.org/10.12912/27197050/217188
ISSN 2719–7050, License CC-BY 4.0

Received: 2025.01.07
Accepted: 2025.01.28
Published: 2026.02.07

Integration of Sentinel-1 synthetic aperture radar and random forest algorithm for high-precision flood hazard modeling in a data-scarce tropical watershed

Marwiji Muhammad Yusuf Fadhel[1*], Soma Andang Suryana[2], Arif Samsu[3], Rahmat Syaeful[4]

[1] Regional Planning and Development Program, Graduate School, Hasanuddin University, Makassar, South Sulawesi 90245, Indonesia
[2] Department of Forestry, Faculty of Forestry, Hasanuddin University, Makassar, South Sulawesi 90245, Indonesia
[3] Department of Geophysics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, South Sulawesi 90245, Indonesia
[4] Department of Remote Sensing and Geographic Information Systems, Faculty of Vocational Studies, Hasanuddin University, Makassar, South Sulawesi 90245, Indonesia
* Corresponding author's e-mail: myusuffadel@gmail.com

ABSTRACT

Flood hazard management in small tropical river basins faces significant challenges owing to rapid hydrological responses and a lack of hydrometric instrumentation. This study aimed to bridge the technological gap in environmental monitoring by developing a high-precision flood hazard model capable of operating in data-scarce regions where traditional hydrodynamic models fail because of insufficient parameterization. The methodology integrates Sentinel-1 SAR imagery (2019–2025) and the random forest (RF) machine learning algorithm within the Python platform. This study reconstructed historical flood dynamics and predicted spatial hazard zones using ten environmental parameters. The results demonstrated robust model performance with a validation accuracy of 94.45%, an area under curve (AUC) of 0.98, and a sensitivity of 96.80%, significantly outperforming conventional statistical methods, which typically achieve lower accuracy in flashy watersheds. Spatially, the model identified 1,080.24 hectares (11.59% of the total area) as Very High hazard zones concentrated in the downstream alluvial plains. Furthermore, explainable AI (SHAP) analysis revealed that vegetation density (NDVI) and topography are the primary physical determinants of inundation, surpassing the influence of local rainfall variability. These findings provide a scientifically validated framework for precise hazard zoning, confirming that machine learning integration can effectively substitute dense ground gauge networks to develop resilient environmental protection strategies.

Keywords: machine learning, random forest, explainable artificial intelligence, environmental engineering, Sentinel-1 synthetic aperture radar, flood hazard modeling, tropical watershed.

INTRODUCTION

Global climate change has become a major catalyst accelerating the hydrological cycle and triggering an increase in the frequency and intensity of hydrometeorological disasters worldwide. Global warming has altered rainfall patterns, making them more extreme and difficult to predict (Totz et al., 2017). In tropical archipelagic regions such as Indonesia, this threat is compounded by the complex interaction between steep topography and regional climate variability. The increasing recurrence of catastrophic floods in these regions indicates a critical failure of the existing environmental protection infrastructure, suggesting that current mitigation strategies, which are often reactive and reliant on static historical data, are insufficient. This necessitates an urgent paradigm

shift in environmental engineering towards precise, technology-driven spatial modeling that can account for dynamic climatic shifts.

The Takkalasi watershed in South Sulawesi province exemplifies this eco-hydrological vulnerability. As a small-scale catchment (<100 km²), it exhibits a characteristic "flashy" hydrological response, where the combination of steep upstream terrain and a short flow path results in the rapid conversion of rainfall into surface runoff (Ibarreche et al., 2020). Consequently, the watershed frequently experiences hydraulic failure, where peak discharges exceed the channel capacity within hours of a storm event. This rapid onset leaves a dangerously narrow window for warning and evacuation, exposing downstream agricultural lands and settlements to severe environmental degradation and economic losses. The recurrence of these events highlights the urgent need for sustainable environmental monitoring systems that can operate in real time.

Despite the urgency, flood mitigation planning in the Takkalasi watershed faces fundamental methodological limitations. Previous studies and engineering practices in this region have largely relied on conventional hydrologic-statistical methods, such as the rational method or synthetic unit hydrograph (Brunner et al., 2018). While effective for point-based estimation in hydraulic structure design, these deterministic approaches exhibit significant weaknesses for modern disaster management: (1) they are inadequate for mapping the spatial distribution of inundation required for zoning; (2) they depend heavily on rainfall and stream gauge data, which are often sparse or damaged in developing countries (data-scarce regions); and (3) they assume "stationarity," a concept no longer valid in the climate crisis era (Khan et al., 2011).

The absence of accurate risk models owing to data scarcity hinders local governments from formulating adaptive policies. To address this technological gap, computer modeling and IT applications based on machine learning (ML) have emerged as promising paradigms in environmental engineering. Unlike rigid physical models that require extensive parameterization (e.g., river geometry for HEC-RAS), ML algorithms such as random forest (RF) are specifically selected in this study because of their ensemble nature, which effectively handles high-dimensional, non-linear environmental data without succumbing to overfitting, a common pitfall in single-decision tree models (Youssef et al., 2022). Crucially, this approach offers the flexibility to integrate Big Data from remote sensing. Sentinel-1 synthetic aperture radar (SAR) imagery, for instance, provides a decisive advantage by detecting flood inundation through cloud cover, which is a persistent limitation of optical imagery in the tropics, thereby generating objective flood inventory data without reliance on subjective manual reporting.

Moreover, the use of machine learning in environmental engineering goes beyond achieving predictive accuracy; it requires interpretability to inform physical actions (Giudici et al., 2023). The 'black-box' nature of sophisticated algorithms often impedes their use in policymaking. To address this issue, this study utilized explainable artificial intelligence (XAI) with SHapley Additive exPlanations (SHAP) to break down the decision-making process of the Random Forest model. This method enables a detailed assessment of how specific physical factors, such as vegetation density compared to rainfall, affect hydraulic failure. This diagnostic ability is crucial for developing precise bioengineering solutions, shifting from broad hazard zoning to focused landscape restoration.

However, a review of the global literature reveals significant research gaps. Most ML application studies for floods have focused on large-scale watersheds or subtropical regions with abundant data availability (Wang et al., 2024). There is a paucity of research examining the reliability of these advanced methods for small-scale tropical watersheds (<100 km²), which are characterized by rapid hydrological responses but suffer from severe data scarcity. Consequently, these vulnerable areas often bear the brunt of flash floods. This study aims to fill this critical knowledge gap by establishing a novel high-precision flood vulnerability mapping framework specifically tailored for data-scarce small-scale tropical watersheds. The central hypothesis of this study is that by integrating high-resolution SAR imagery with an optimized random forest algorithm, it is possible to achieve a prediction accuracy exceeding 90%, which significantly surpasses traditional models, solely using open-source remote sensing data. Specifically, this study seeks to demonstrate that (1) the proposed ML-based model can accurately identify micro-scale hazard zones that are invisible to coarse regional models, and (2) in small "flashy" catchments, physical land characteristics (such as vegetation density) exert a more dominant control over flood vulnerability than local

rainfall variability, a finding that would fundamentally shift engineering mitigation strategies from structural to ecological approaches.

MATERIALS AND METHODS

Study area

This study focused on the Takkalasi watershed, a critical ecological unit situated in Barru regency, South Sulawesi, Indonesia, geographically centered at 119° 41' 50.082" E. Spanning an area of approximately 9.317 ha, the watershed operates under a tropical monsoon climate regime, exhibiting a pronounced seasonality with a high-intensity wet season typically occurring between November and April. Geomorphologically, the area is characterized by a dramatic topographic gradient (Figure 1), transitioning rapidly from rugged, mountainous upstream regions, which serve as the primary source areas, to the low-lying downstream alluvial plains, where the Takkalasi River eventually discharges into the Makassar Strait. This steep morphological configuration results in a short concentration time, generating a high potential energy for surface runoff during storm events. Furthermore, the land cover gradient, moving from upstream vegetation to downstream intensive agricultural zones and settlements, creates a complex interaction between natural flow paths and anthropogenic activities. When coupled with the intense precipitation characteristic of the monsoon, these hydromorphological features render the downstream communities particularly susceptible to flash floods and riverine inundation, underscoring the necessity for urgent and precise environmental protection measures.

Data acquisition and framework

This study was structured using a comprehensive and systematic methodological workflow designed to ensure reproducibility and robustness in environmental modeling. The framework (Figure 2) operates on a modular "Input-Process-Output" architecture, primarily executed within the google earth engine (GEE) cloud computing environment to efficiently handle large-scale geospatial datasets. The workflow is divided into four main sequential stages: (1) multi-source data acquisition and advanced image pre-processing; (2) construction of physical predictor variables and derivation of a historical flood inventory; (3) development and training of the spatial model using the Random Forest algorithm; and (4) rigorous model performance evaluation and quantitative

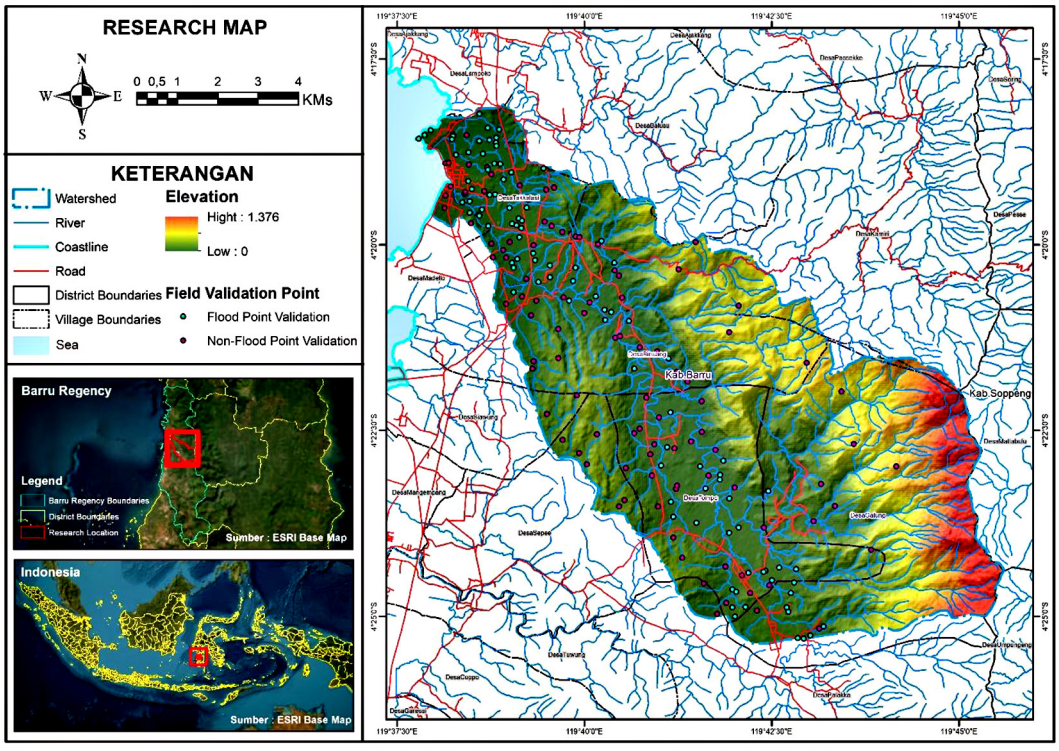

Figure 1. Study area map of the Takkalasi watershed in Barru regency, South Sulawesi

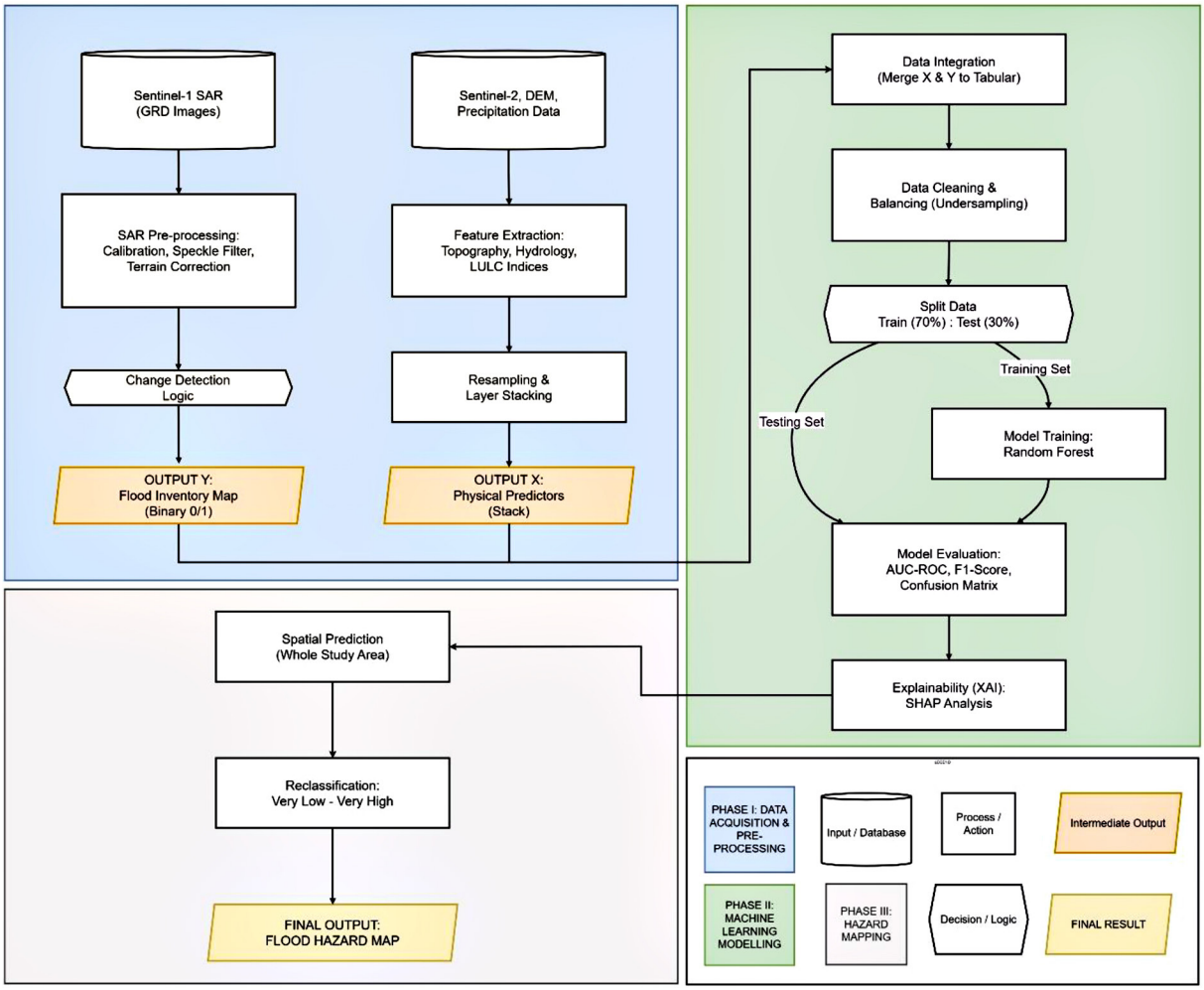

Figure 2. Research methodology flow – integration of remote sensing data and machine learning for flood hazard mapping

analysis of flood determinants using explainable AI. Flood inventory mapping was conducted in GEE using Sentinel-1 SAR imagery and adaptive Change Detection, while the random forest machine learning model was implemented and processed in Python using the Scikit-learn library to ensure flexibility, reproducibility, and integration with advanced statistical validation libraries.

A comprehensive array of validated global datasets was incorporated to underpin this data-driven framework. The target variable (Y), indicative of historical flood occurrences, was derived from high-resolution (10-meter) Sentinel-1 SAR imagery spanning the period 2019–2025, employing an adaptive Change Detection method. The predictor variables (X) were organized into four principal environmental clusters to encapsulate the multifaceted nature of flood risk: Topographic data – high-precision variables, including elevation, slope, topographic wetness index (TWI), and height above nearest drainage (HAND), were derived from the 8-meter resolution National Digital Elevation Model (DEMNAS). Land physical characteristics: land cover and vegetation density (NDVI) were extracted from 10-meter resolution Sentinel-2 MSI optical imagery. Climatic Variables – daily rainfall data were obtained from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) dataset, which provides a spatial resolution of 0.05° (~5.5 km). Anthropogenic and soil factors – distance to rivers and road networks were derived from the 1:25,000 scale Indonesian Topographic Map (RBI), and soil properties were obtained from a 1:250,000 scale RePPProT map. To ensure spatial consistency and compatibility for the machine learning analysis, all raster datasets were resampled and aligned to a uniform 10-meter pixel resolution using the nearest-neighbor method. This rigorous data standardization process ensured that the

model input represented a coherent spatial matrix, facilitating precise pixel-based classification.

Flood inventory using SAR imagery

Precise flood susceptibility mapping (FSM) relies heavily on the quality of the training data. To overcome the limitations of optical imagery in cloud-prone tropical regions, this study utilized the GEE platform to process Sentinel-1 ground range detected (GRD) imagery. The inventory was constructed by analyzing seven significant flood events that occurred between 2019 and 2025 (Table 1). The image processing workflow included: (1) thermal noise removal and radiometric calibration; (2) speckle filtering using the Refined Lee filter to reduce noise while preserving water body edges; and (3) terrain correction to rectify geometric distortions. Flood detection was performed using a Change Detection approach by comparing the "Target Period" (flood event) and "Baseline Period" (dry conditions). A random forest classifier trained with samples from the JRC Global Surface Water dataset was applied to distinguish flood inundation from permanent water bodies. The final dataset consisted of 20,000 stratified sample points (10.000 flood and 10.000 non-flood) split into 70% for training and 30% for testing.

Flood conditioning factors

Flooding occurs as a result of complex interactions between landforms, water flow, surface characteristics, weather, and anthropogenic activities (Lyu et al., 2018). To comprehensively model flood risk in the Takkalasi watershed, this study selected ten variables representing topographic, hydrological, and ecological engineering factors processed using GEE at a 10-meter resolution. Topographic factors include elevation (Figure 3a), which indicates vulnerability to water accumulation in lowland areas owing to gravity, and slope (Figure 3b), which controls the speed of water flow, where flat terrain slows runoff and triggers ponding (Hou and Gao, 2019). The analysis also included the topographic wetness index (Figure 3c) to identify soil saturation zones, as well as the height above nearest drainage (Figure 3d), which measures the vertical proximity of areas to drainage channels. From a hydrological perspective, the distance to river (Figure 3e) was used as the main indicator of fluvial overflow risk. Physical land characteristics are represented by the normalized difference vegetation index (Figure 3f), indicating the surface roughness of vegetation; land cover (Figure 3g), distinguishing between permeable and impermeable areas; and soil type (Figure 3h), which determines the natural infiltration capacity (Dahigamuwa et al., 2016).

In addition to the physical factors of the land, meteorological and anthropogenic elements are also considered triggers and differentiators of risk. Climatic factors are represented by the average rainfall (Figure 3i) from the CHIRPS dataset, which highlights the importance of precipitation variability as a primary input in the hydrological system (Cerón et al., 2020). Finally, the influence of human activity was examined using the variable distance to road (Figure 3j). Road infrastructure is often correlated with built-up areas that have low infiltration, and the structure of road embankments can alter natural flow patterns, acting as barriers that exacerbate local floods (Douglas et al., 2008). All of these spatial variables were normalized and integrated into the machine learning model to precisely detect flood vulnerability patterns.

Random forest hazard modeling

Spatial prediction was performed using the RF algorithm, an ensemble learning method that is robust against overfitting in high-dimensional

Table 1. Sentinel-1 data acquisition period for the flood inventory

Year	Target period (flood events)	Baseline period (dry conditions)
2019	20 January–31 January	01 August–31 August
2020	18 December–28 December	20 November–30 November
2021	01 December–20 December	01 August–25 August
2022	01 December–30 December	01 August–30 August
2023	01 January–28 February	01 August–30 August
2024	01 January–30 January	01 August–30 August
2025	01 November–25 December	01 July–31 July

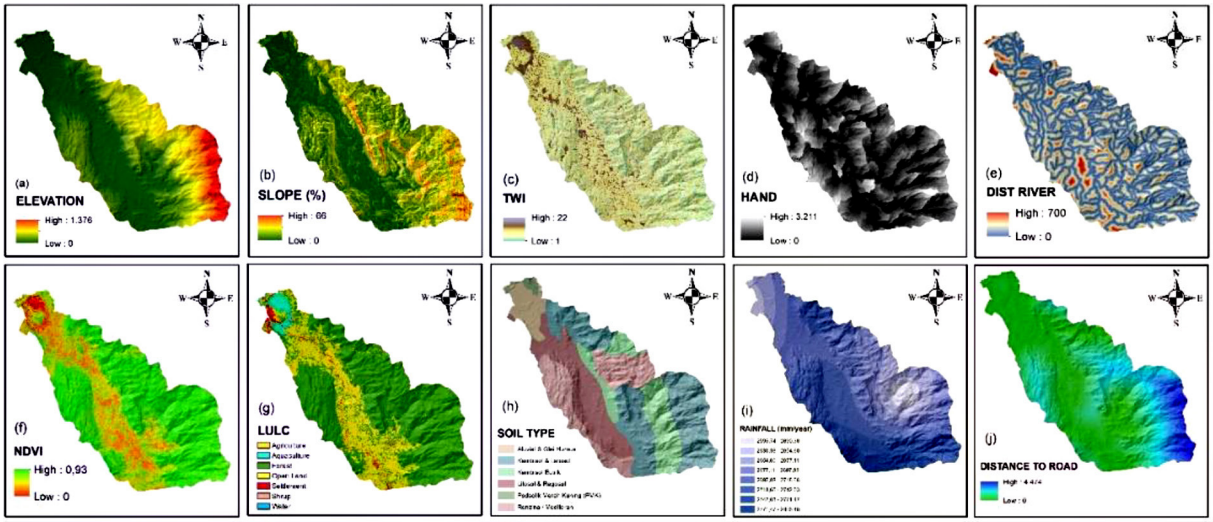

Figure 3. Spatial distribution of the ten flood conditioning factors used in the random forest model: (a) elevation, (b) slope, (c) topographic wetness index (TWI), (d) height above nearest drainage (HAND), (e) distance to river, (f) normalized difference vegetation index (NDVI), (g) soil type, (h) average rainfall, and (i) distance to road

data (Wu et al., 2024). Unlike standard implementations, this study explicitly optimized the algorithm for spatial hazard modeling by implementing a balanced sampling strategy using stratified sampling with a 1:1 ratio between flood and non-flood pixels in the training dataset. The model configuration included fine-tuning hyperparameters: the number of trees (*n_estimators*) was set to 500 to ensure stability, and the maximum tree depth (*max_depth*) was limited to 25 to prevent overfitting while capturing complex, nonlinear interactions. The Gini Impurity criterion was used for node splitting. The prediction function for a new class (x') is formulated as Equation 1:

$$y = \text{mode}\{T_b(x')\}_{b=1}^{B} \quad (1)$$

where: a $T_b(x')$ is the class output of the b-th tree, and (B) is the total number of trees.

The configuration and training of the model in the Python script (Scikit-learn) are carried out by implementing a balanced sampling strategy using stratified sampling with a 1:1 ratio to address data imbalance between flood and non-flood pixels. Meanwhile, the optimized hyperparameters set the number of trees (n_estimators) to 500, with the tree depth (max_depth) limited to 25 to prevent overfitting, and the Gini Impurity criterion was used for splitting. To enhance the transparency and credibility of the methodology, the actual computational implementation is shown in Figure 4. This figure displays the Python code snippet utilized for configuring the random forest classifier and SHAP analysis, confirming the application of specific hyperparameters (class_weight='balanced', n_estimators=500) rather than default settings.

Model evaluation and validation

Model validation was conducted using 70% of the training set and 30% of the test set. The evaluation employed A confusion matrix was used to compute the performance metrics.

Confusion matrix

The performance of the classification model was fundamentally assessed using a confusion matrix, which compared the predicted classes with the actual ground truth (Pomme et al., 2022). It consists of four key components: true positive (TP), which represents flood pixels correctly identified by the model; true negative (TN), which denotes non-flood pixels correctly classified as safe; false positive (FP), which indicates non-flood pixels incorrectly predicted as flood (Type I error); and false negative (FN), which refers to actual flood pixels that the model failed to detect (Type II error).

Accuracy

The percentage of accurate predictions across all regions serves as an indicator of the model's efficacy in distinguishing areas of high and low

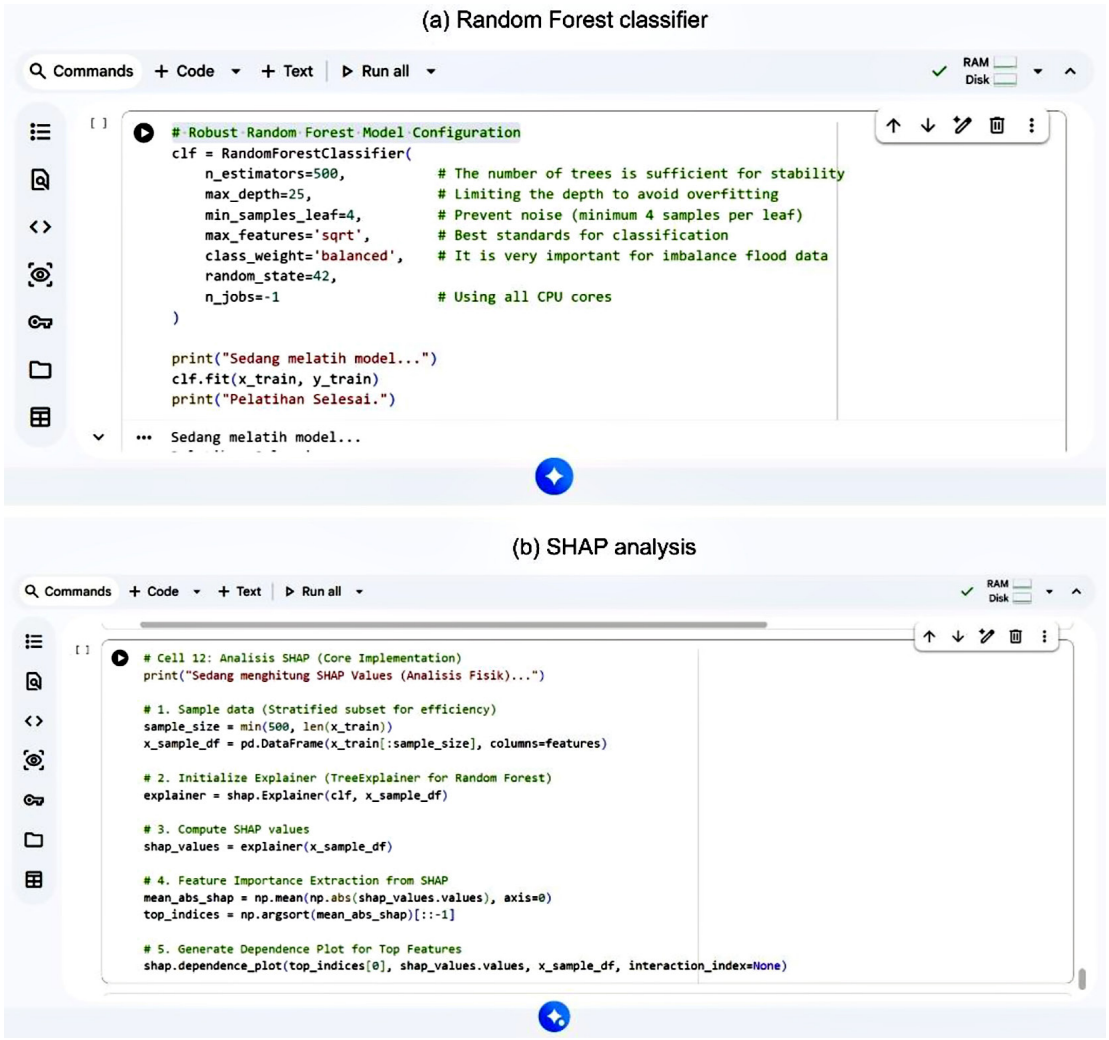

Figure 4. Python implementation screenshot: (a) configuration of the optimized random forest model and (b) SHAP analysis

vulnerability (Williams et al., 2024), as defined in Equation 2:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Precision

The accuracy of positive predictions is crucial for minimizing false alarms when forecasting floods, as defined in Equation 3:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall

Recall is defined as the ratio of correctly identified flood-prone areas to the total number of actual flood-prone areas, serving as an indicator of the effectiveness of a method in identifying all genuine flood-prone areas. It is computed by dividing the number of flood-prone areas accurately identified by the method by the total number of actual flood-prone areas (Taherizadeh et al., 2023), as shown in Equation 4:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1-Score

The F1-score assesses a model's balance by incorporating both precision and recall, thereby facilitating the accurate identification of flood-prone areas by reducing false positives and negatives (Tsumita et al., 2025), as expressed in Equation 5:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

AUC-ROC

The receiver operating characteristic (ROC) technique was employed to evaluate the model's performance by calculating the area under the curve (AUC). The ROC technique is among the most prevalent methods for assessing the accuracy of predictive models (Centor and Schwartz, 1985). The ROC curve compares the true positive rate (sensitivity) with the false-positive rate (1-specificity) at different thresholds. This provides a complete picture of the model accuracy, as shown in Equations 6 and 7:

$$\text{FPR} = 1 - \text{specificity} = \frac{FP}{FP + TN} \quad (6)$$

$$\text{ROC} - \text{AUC} = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR}) \quad (7)$$

Ground check sampling

To physically validate the map results in the field, the minimum sample size (n) was determined using the Cochran Formula to ensure statistical representativeness at a 90% confidence level (Zhao et al., 2014). is defined as in Equation 8:

$$n = \frac{Z^2 \cdot p \cdot q}{e^2} \quad (8)$$

Flood predictors significance

Variable importance

The significance of the flood predictors was evaluated by examining the feature importance characteristics of the Random Forest model. This analysis employed the *feature_importances_* attribute available in the Scikit-learn Python library, which offers insights into the relative contribution of each predictor variable to the predictive performance of the models (Ghanim et al., 2023). This methodology is essential in flood studies because it allows researchers to identify the most influential factors contributing to flood susceptibility and quantify their impact on flooding.

SHapley additive exPlanations (SHAP)

SHAP, developed by Lundberg and Lee, identifies the most significant input features and their interactions. In contrast to earlier methodologies, SHAP uniquely guarantees 'local accuracy,' thereby addressing issues inherent in other feature importance measures (Wang et al., 2024). It provides a more nuanced understanding than broader perspectives. This study used the Shapley value to assess feature significance in flood risk and ranked features by their contributions. Each point on the graph represents a sample's Shapley value (ϕ_i) with red points indicating higher values and blue points denoting lower values. A broader distribution of points indicates a significant influence on flood risk. The features are arranged vertically by importance, with the most critical features at the top. The average Shapley value indicates the impact of each feature on flood risk. The bar graph shows the feature importance, with taller bars indicating substantial positive effects and shorter bars indicating a lesser impact. The Shapley value is computed as defined in Equation 9.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \quad (9)$$

where: ϕ_i represents the Shapley value for feature i, indicating its contribution to the model's output. F is the complete set of input features, while S denotes a subset of features excluding i.

The term $f(S)$ is the prediction of the model with the feature subset S, and $f(S \cup \{i\})$ is the prediction when feature i is included. The weighting factor $\frac{|S|!\,(|F| - |S| - 1)!}{|F|!}$ accounts for all possible permutations of feature introduction, ensuring a fair allocation of the prediction value among the constituent features.

RESULTS

Historical flood event inventory

The extraction results from Sentinel-1 SAR imagery over a seven-year period (2019–2025) using the Adaptive Time-Window Change Detection method successfully mapped the spatial dynamics of significant flood events in the Takkalasi watershed. Data analysis (Figure 5, Table 2) showed that fluctuations in the extent of inundation were strongly correlated with regional climate anomalies. The peak of extreme flood events was recorded in 2022, with an inundation area of 539.38 ha, contributing 21.52% to the total historical flood footprint. This massive event coincided with a moderate La Niña phenomenon that

triggered an annual rainfall increase of 3.546 mm. Conversely, the lowest inundation extent was recorded in 2023 at 64.45 ha (2.57%), reflecting dry conditions due to the El Niño phase. These findings confirm that radar-based methods can capture the sensitivity of watershed hydrological responses to climate variability.

Physical predictor variables construction

The training dataset was formed by integrating the historical flood event map/Y variable (Figure 6a) with the stack of predictor variables X (Figure 6b), resulting in 20,000 stratified sample points (50% flood, 50% safe) from a total population of 931,793 pixels. The Balanced Stratified Random Sampling strategy was used to address class imbalance, ensuring that the Machine Learning model could recognize both flood and safe area patterns evenly. Each sample point was extracted from locations that were consistently flooded or not throughout 2019–2025 and assigned values from the 10 main physical variables.

The distribution analysis shown in Figures 7a–7j provides critical insights into the flood mechanisms specific to small tropical watersheds (<100 km²). Unlike larger basins, where regional hydrology dominates, flood events in the Takkalasi watershed are strictly confined by micro-topographical and land cover constraints. Most flood points were concentrated at low elevations (<25 masl, Figure 7a) and flat slopes (<8%, Figure 7b); however, they heavily overlapped with sparse vegetation areas (NDVI 0.1–0.3, Figure 7f). This distinct pattern confirms the "flashy" nature of the watershed, where the absence of upstream vegetative buffers leads to a rapid downstream accumulation. The consistency between the distribution of the training samples and the total population strengthened the validity of the proposed Random Forest model, ensuring that the algorithm learned from physically meaningful patterns rather than statistical noise.

Determination of flood causing factors using random forest results

Variable importance

The analysis of variable importance provides quantitative insights into the physical parameters that are the main drivers of disasters. Based on the computational results presented in Table 3, a clear hierarchy of influence among the ten predictor variables has emerged. The NDVI (Vegetation Index) ranked highest, with a Mean Importance score of 0.3902, followed by slope (0.1988), HAND (0.1264), and elevation (0.1174). These four variables form the "Primary Determinant Cluster", contributing more than 83% of the model's predictive power. The dominance of NDVI underscores that flooding in the Takkalasi watershed is ecologically controlled by vegetation cover, which determines

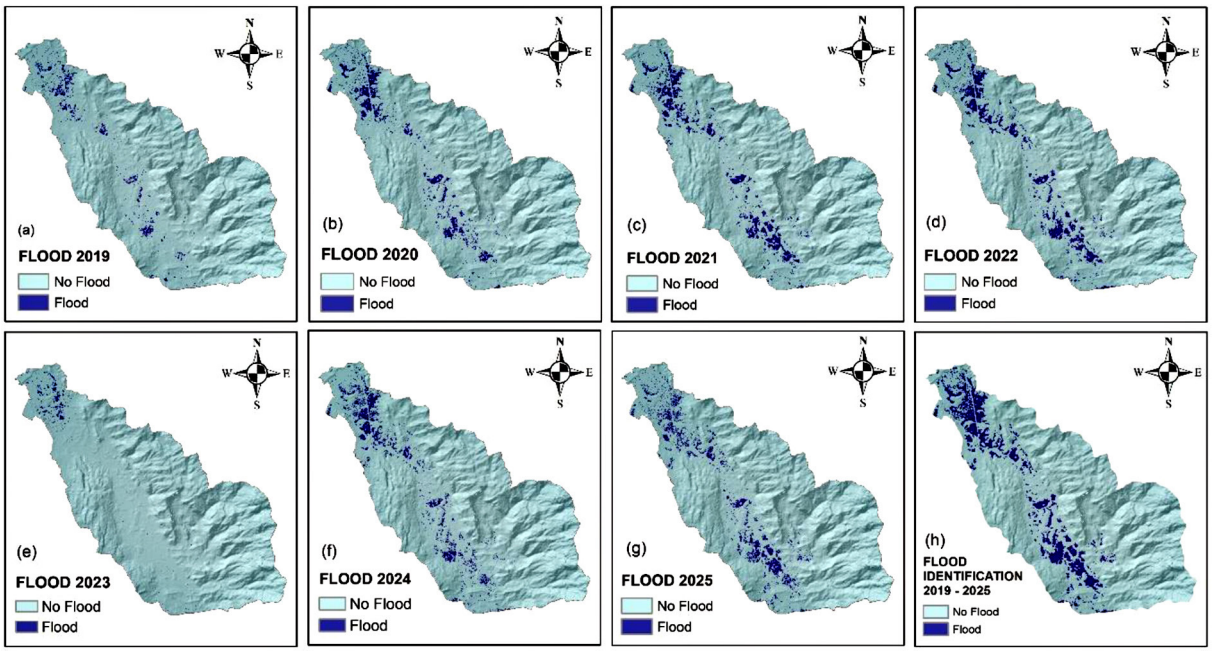

Figure 5. Maps of annual flood distribution (a-g) and composite inventory results (h) for the period 2019–2025

Table 2. Relationship between flood inundation areas and average annual rainfall (2019–2025)

Year	Flood area (ha)	Average rainfall (mm/year)
2019	175.22	1869.59
2020	385.10	2814.35
2021	483.11	3134.54
2022	539.38	3546.24
2023	64.45	1779.64
2024	412.45	2758.43
2025	447.03	2814.89

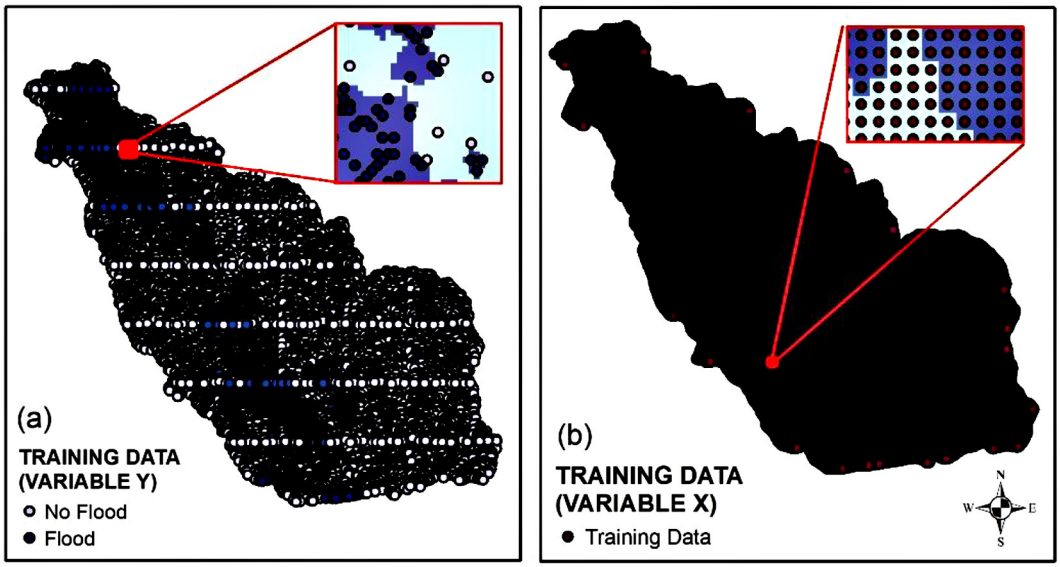

Figure 6. Spatial distribution of training data: (a) distribution of binary sample points (flood/safe); (b) extraction of multiparameter values (variable X) at sample points

infiltration capacity. In contrast, rainfall ranked lowest (0.0142), indicating that in small watersheds with homogeneous meteorological conditions, flood locations are more determined by the land's inability to absorb water (static factors) than by the variability of rainfall itself (Zhou et al., 2021).

SHAP dependence analysis

The analysis of the SHAP dependence plot (Figure 8) elucidates the nonlinear physical mechanisms and identifies critical tipping points in flood control. The NDVI dependence graph indicates a significant increase in extreme risk within the range of 0.1–0.3, corresponding to open land or sparse vegetation, which is associated with the highest positive SHAP value. Conversely, the risk markedly decreased to negative at high vegetation densities (>0.7), corroborating the role of vegetation in mitigating surface runoff. A comparable pattern was observed in the slope gradient, where the risk was concentrated at slopes of less than 8% (flat terrain), which were identified as points of flow stagnation. Furthermore, the anthropogenic variable, distance to road, demonstrated an increase in risk within a radius of 100–500 m, suggesting an impoundment effect due to road infrastructure disrupting natural drainage patterns. In contrast, the dependency plot for rainfall showed very small SHAP value fluctuations around zero, confirming its limited role as a spatial differentiator at the local level.

Spatial flood hazard analysis

The concluding phase of hazard modeling entails transforming flood probability data (Figure 9) into practical zoning maps. Utilizing the probability figures produced by the random forest algorithm for each 10-meter pixel, reclassification was performed using the Natural Breaks (Jenks)

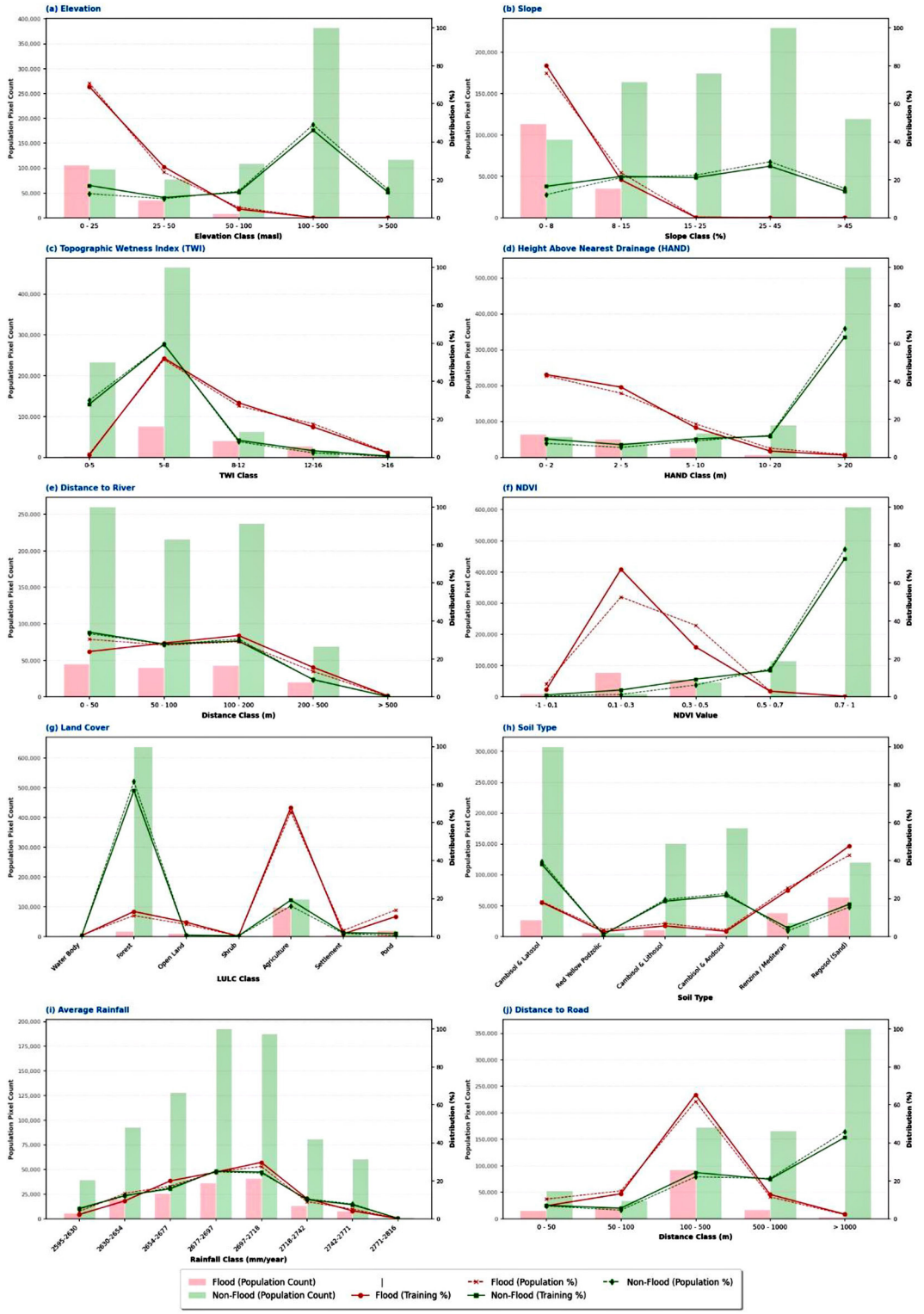

Figure 7. Comparison of data distribution between training sample and total population: (a) elevation, (b) slope, (c) topographic wetness index (TWI), (d) height above nearest drainage (HAND), (e) distance to river, (f) normalized difference vegetation index (NDVI), (g) land cover, (h) soil type, (i) average rainfall, (j) distance to road

method. This process organizes the data into five standard interval categories, illustrating the risk gradient from "very low" to "very high." Figure 8b shows the resulting flood vulnerability map.

Spatial statistical analysis (Table 4) showed unique risk distribution characteristics in the Takkalasi watershed. Most of the area (79.79% or 7.434 ha) is classified as a safe zone (very low),

Table 3. Ranking of physical predictor variables based on random forest importance and SHAP values

Rank	Physical variable	Mean importance (RF)	Standard deviation (RF)	Mean SHAP value
1	NDVI (Vegetation)	0.3902	0.2886	0.2109
2	Slope	0.1988	0.2459	0.1088
3	HAND	0.1264	0.2006	0.0578
4	Elevation	0.1174	0.1817	0.0650
5	Land cover (LULC)	0.0551	0.1250	0.0295
6	Distance to road	0.0408	0.0522	0.0202
7	TWI	0.0207	0.0245	0.0144
8	Soil type	0.0191	0.0573	0.0108
9	Distance to river	0.0171	0.0054	0.0100
10	Rainfall (RF)	0.0142	0.0066	0.0089

generally located in the upstream region with steep topography that functions as a runoff zone. However, zones with very high vulnerability constitute a significant proportion, reaching 11.59% (1,080.24 ha). This extreme polarization between the safe and high-hazard zones is a typical indicator of a watershed with a rapid hydrological response, where water from the upstream area is directly accumulated in the downstream flat area without a wide transitional zone.

Model performance evaluation

Ground truth

Field validation was conducted at 188 sample locations using Cochran's formula with 90% confidence to confirm the accuracy of the flood inventory map from Sentinel-1 imagery. This process ensures that the training data accurately reflect real-world conditions. According to the confusion matrix from the field validation (Figure 10), this radar-based inventory method achieved an overall accuracy of 91.5%. Of the 89 points identified as flood areas, 81 were verified as actual flood sites (true positive) through water marks and resident interviews. Only eight points were detection errors (false positives). Analysis showed that these false positives occurred mainly in irrigated rice fields during inundation, exhibiting radar reflections similar to river overflow floods (Phan et al., 2019). This low error rate (<10%) confirms that the Sentinel-1 inventory data are valid for training the random forest model.

Random forest model evaluation

Following physical validation, the random forest model underwent a statistical assessment using a 70:30 data split on a separate test dataset to evaluate its ability to generalize. As explicitly demonstrated in the receiver operating characteristic (ROC) analysis (Figure 11), the model achieved an exceptional area under curve (AUC) score of 0.9849. This statistical metric confirms the model's superior capability to distinguish between flood and non-flood classes with high precision, far surpassing the standard benchmarks. The predictive performance was further substantiated by an overall accuracy of 94.45% and a precision score of 92.45%, indicating a very low rate of false-positive alarms.

Crucially, the model achieved a recall value of 96.80%. In the context of disaster management, this metric is of paramount importance, as it signifies that the model successfully detected nearly 97% of all actual flood events, minimizing the risk of "missed targets" (false negatives) that could have catastrophic consequences for community safety. The high recall of the Random Forest model demonstrates its robustness in capturing spatial patterns and environmental signals that distinguish flood-prone areas from safe zones. This was reinforced by the AUC-ROC score of 0.9849 (Figure 12), which indicates excellent separability between the flood and non-flood classes. Such a high AUC value confirms the model's accuracy across probability thresholds, making it adaptable to various scenarios. The combination of high accuracy, precision, recall, and near-perfect AUC establishes the Random Forest model as a powerful tool for early warning systems. This provides decision-makers with confidence that the model can be deployed in flood management strategies, supporting proactive mitigation, resource allocation, and community preparedness in the Takkalasi watershed and beyond (Taherizadeh et al., 2023).

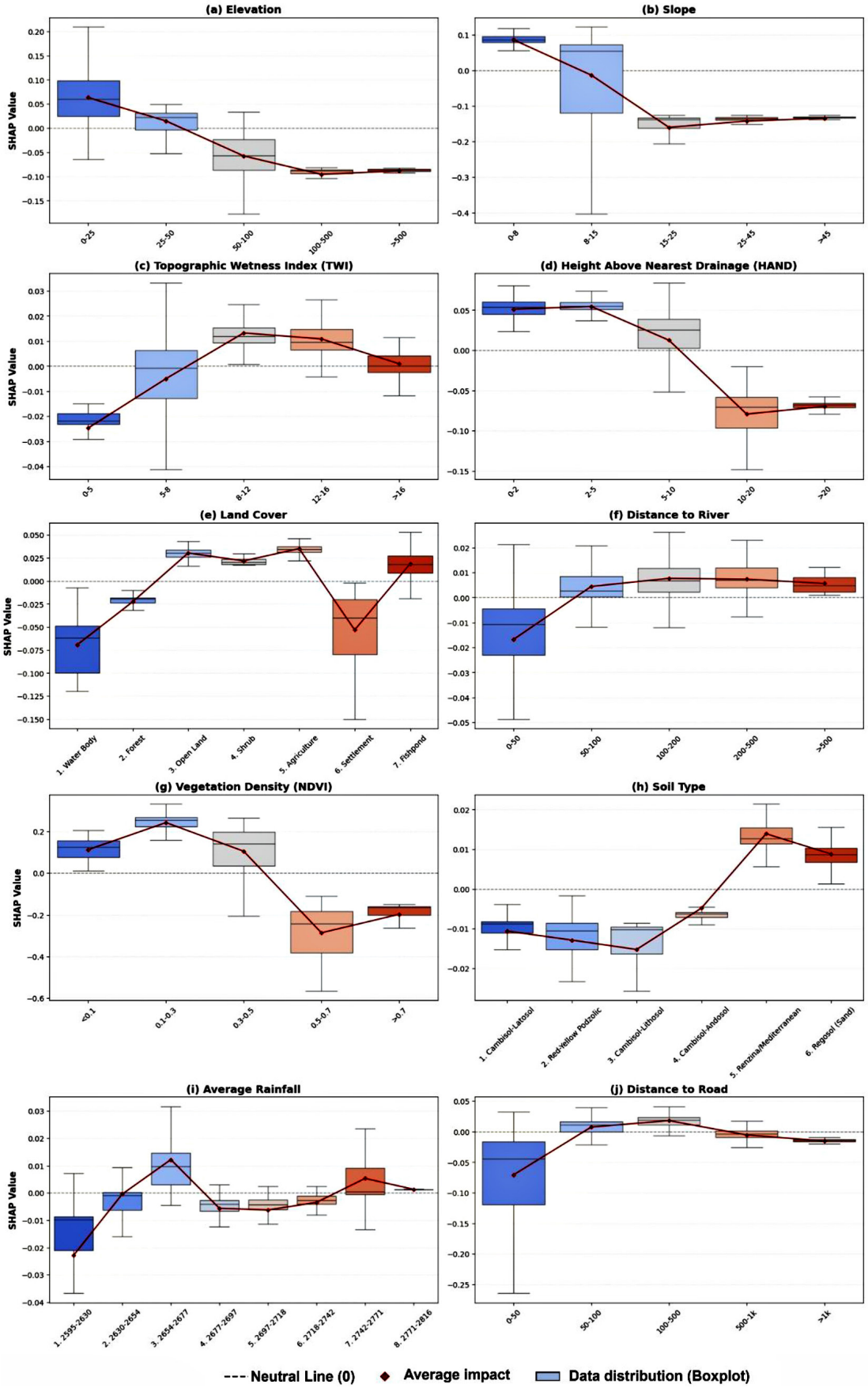

Figure 8. SHAP dependence analysis plots of 10 variables related to flooding: (a) elevation, (b) slope, (c) TWI, (d) HAND, (e) land cover, (f) distance from river, (g) NDVI, (h) soil type, (i) average rainfall, (j) distance from road

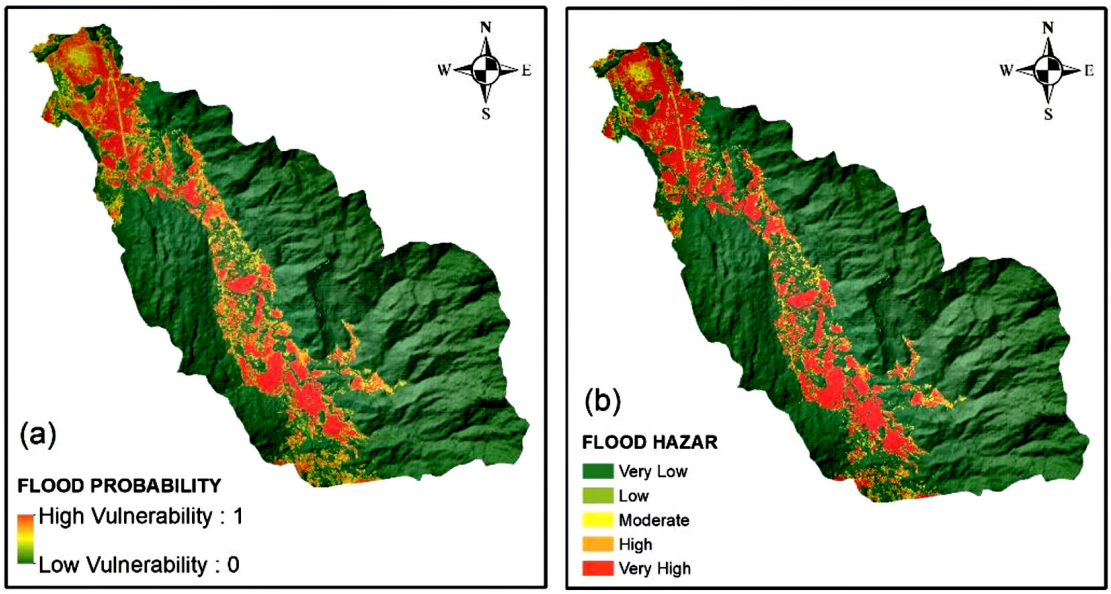

Figure 9. Machine learning-based flood hazard mapping results: (a) spatial probability distribution, (b) flood vulnerability zoning

Table 4. Distribution of area and percentage of flood hazard levels

Hazard level	Area (Ha)	Percentage (%)	Probability classification
Very low	7,434.33	79.79%	Probability < 0.20
Low	279.12	3.00%	0.20 ≤ Probability < 0.40
Moderate	237.47	2.55%	0.40 ≤ Probability < 0.60
High	286.41	3.07%	0.60 ≤ Probability < 0.80
Very high	1,080.24	11.59%	Probability ≥ 0.80
Total	9,317.57	100%	

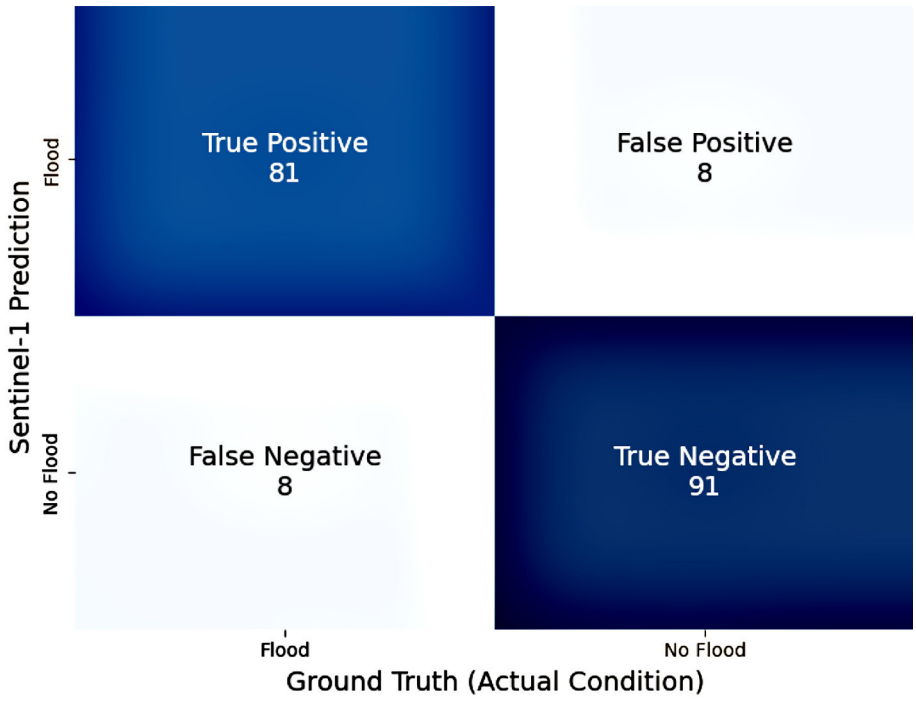

Figure 10. Confusion matrix of field validation results for flood inventory

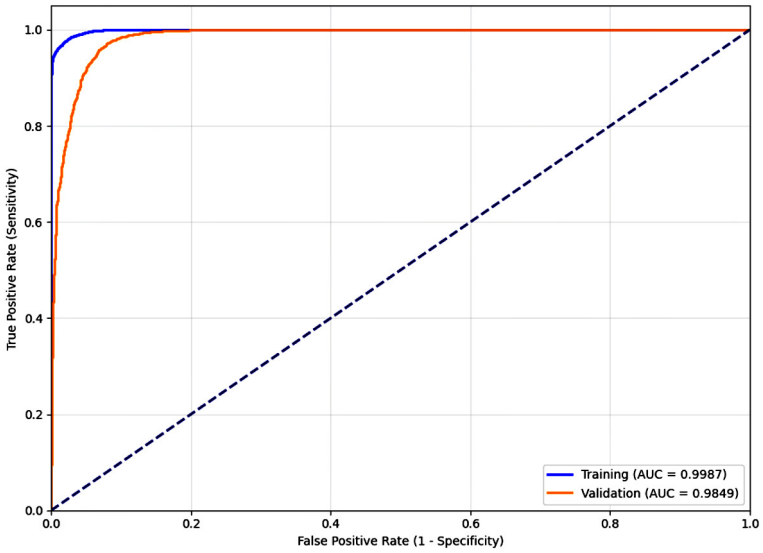

Figure 11. Comparison curve of receiver operating characteristic (ROC) for training and validation

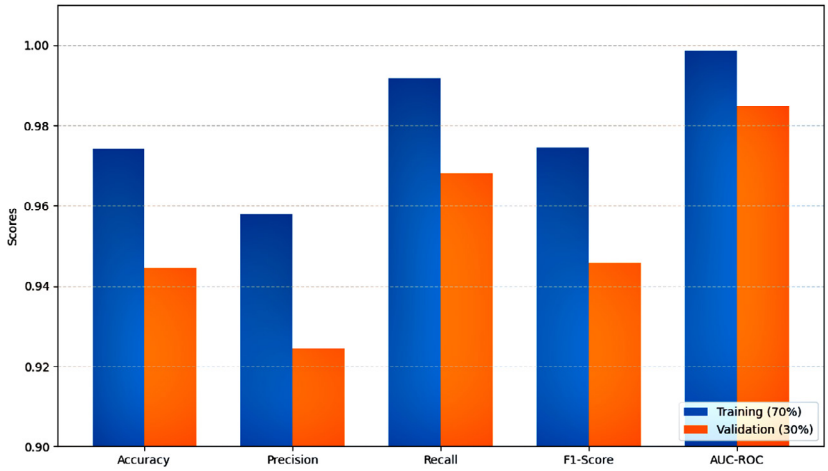

Figure 12. Comparison of the model training and validation performance

DISCUSSION

Reliability of radar-based environmental monitoring in data-scarce regions

This study demonstrates that a data-driven approach using the random forest algorithm integrated with multi-source remote sensing data can provide precise solutions for flood hazard mapping in regions with very limited hydrological data (data-scarce regions). The model validation accuracy, which reached 94.45% with an AUC value of 0.98, proves that this model is highly reliable and robust (Abdelkader and Csámer, 2025). Its primary advantage lies in its high generalization capability and superior recall value (96.80%). In the context of disaster risk management, this high sensitivity is crucial to ensure that virtually no flood-prone areas are missed during detection, thereby minimizing the risk of a false sense of security for the community (Ahmad and Afzal, 2022).

Compared with traditional hydrological methods, which are often constrained by the scarcity of rain and stream gauge stations in small-scale watersheds, this approach offers a more adaptable environmental monitoring system. The utilization of flood inventory data derived from Sentinel-1 SAR imagery has proven effective as a surrogate for discharge data and is capable of objectively documenting inundation extent without human reporting bias (Misra et al., 2025). Furthermore, the resulting 10-meter spatial resolution facilitates the identification of micro-hotspots at the granularity of residential blocks or rice field plots, offering a significant improvement over typical regional risk maps.

Ecological engineering implications: Vegetation as primary defense

Explainable AI (SHAP) analysis revealed critical hydrological insights into the flood dynamics of the Takkalasi watershed. The identification of NDVI (0.39) as the primary determinant highlights that the watershed's main defense against flooding is its land retention capacity, which is ecologically controlled by the vegetation density (Durigon et al., 2014). The SHAP dependence graph indicates a critical "tipping point" at NDVI values between 0.1 and 0.3; below this range, the soil's ability to intercept and absorb water diminishes significantly, leading to a rapid transformation of rainfall into surface runoff (Sharma et al., 2021). This confirms the characteristic of a "flashy" response watershed, where upstream land cover degradation directly impacts the downstream peak discharge.

An intriguing finding was the relatively low ranking of rainfall's influence in the spatial model compared to static physical parameters. This observation warrants a nuanced interpretation: although rainfall is undeniably the meteorological trigger, it functions primarily as a temporal catalyst rather than a spatial discriminator within the context of a small catchment. In watersheds smaller than 100 km², the rainfall distribution during storm events tends to be spatially homogeneous. Therefore, while precipitation determines when the system is stressed, the physical land characteristics, specifically the interplay between vegetation density (NDVI), topographic gradients, and soil saturation potential, strictly dictate the specific locations where hydraulic failure and inundation occur (Ganjirad and Delavar, 2023). Consequently, flood mitigation strategies in the Takkalasi watershed should not rely solely on downstream civil engineering solutions (e.g., levees). Instead, priority must be given to Ecological Engineering approaches, specifically bioengineering and the rehabilitation of critical upstream areas, to restore the natural retention capacity of the watershed.

Technogenic factors and engineering recommendations

The flood hazard map delineates a pronounced dichotomy in risk levels between relatively secure upstream regions and significantly vulnerable downstream areas. The identification of Tompo, Takkalasi, and Binuang Villages as areas of extreme risk (very high) offers a robust scientific foundation for local governments to amend their Regional Spatial Plans. These regions, constituting 11.59% of the total watershed, are recommended to be designated as Strict Protection Zones or wetland agricultural areas (rice fields), with stringent restrictions on new residential development (Reis et al., 2017).

Additionally, the significant influence of the "distance to road" variable highlights a technogenic factor in flood risk assessment. Roads constructed across natural contours appear to function as artificial embankments that exacerbate local flooding. Therefore, urgent engineering interventions are recommended, including (1) the technical evaluation of existing road infrastructure, (2) the improvement of cross-drainage systems (culverts), and (3) the implementation of permeable road technologies to mitigate the damming effect and restore natural flow paths.

CONCLUSIONS

This study successfully achieved its primary objective: the development of a high-precision, automated flood hazard modeling framework that effectively overcomes the data scarcity limitations inherent in small tropical watersheds. By integrating Sentinel-1 SAR imagery with an optimized Random Forest algorithm, this study delivered a novel scientific contribution: establishing that machine learning models can substitute for physical hydrodynamic models with a validation accuracy of 94.45% and a recall of 96.80%, a level of precision previously unattainable with conventional methods in this region. This study fills a significant gap in hydrological science by quantifying the dominant role of land retention capacity (NDVI) over local rainfall variability in driving flood risks within "flashy" catchments (<100 km²). The identification of a critical vegetation threshold (NDVI 0.1–0.3) provides a new empirically derived parameter for ecological engineering interventions. This opens promising prospects for the widespread adoption of nature-based solutions in disaster mitigation, shifting the focus from reactive infrastructure to proactive landscape restoration. The established framework offers a replicable and cost-efficient protocol for environmental protection practitioners worldwide, particularly in developing nations facing similar climatic and data challenges.

Acknowledgments

The authors express their sincere gratitude to the European Space Agency (ESA) for supplying the Sentinel-1 SAR data and to the USGS and NASA for the Sentinel-2 and CHIRPS datasets used in this study. We further extend our appreciation to the Indonesian Geospatial Information Agency (BIG) for making the high-resolution DEMNAS data publicly accessible and to the developers of the Google Earth Engine platform for enabling cloud-based processing of these multitemporal datasets. The authors also acknowledge the anonymous reviewers for their constructive feedback, which significantly enhanced the quality of this study.

REFERENCES

1. Abdelkader, M. M., Csámer, Á. (2025). Comparative assessment of machine learning models for landslide susceptibility mapping: a focus on validation and accuracy. *Natural Hazards*, *121*(9), 10299–10321. https://doi.org/10.1007/s11069-025-07197-0
2. Ahmad, D., Afzal, M. (2022). Flood risk public perception in flash flood-prone areas of Punjab, Pakistan. *Environmental Science and Pollution Research International*, *29*(35), 53691–53703. https://doi.org/10.1007/s11356-022-19646-5
3. Brunner, M. I., Sikorska, A. E., Favre, A.-C., Furrer, R., Viviroli, D., Seibert, J. (2018). Synthetic design hydrographs for ungauged catchments: a comparison of regionalization methods. *Stochastic Environmental Research and Risk Assessment*, *32*(7), 1993–2023. https://doi.org/10.1007/s00477-018-1523-3
4. Centor, R. M., Schwartz, J. S. (1985). An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. *Medical Decision Making*, *5*(2), 149–156. https://doi.org/10.1177/0272989x8500500204
5. Cerón, W. L., Molina-Carpio, J., Kayano, M. T., Canchala, T., Andreoli, R. V., Ayes Rivera, I. (2020). A principal component analysis approach to assess CHIRPS precipitation dataset for the study of climate variability of the La Plata Basin, Southern South America. *Natural Hazards*, *103*(1), 767–783. https://doi.org/10.1007/s11069-020-04011-x
6. Dahigamuwa, T., Gunaratne, M., Yu, Q. (2016). Feasibility study of land cover classification based on normalized difference vegetation index for landslide risk assessment. *Geosciences*, *6*(4), 45. https://doi.org/10.3390/geosciences6040045
7. Douglas, I., Alam, K., Mcdonnell, Y., Mclean, L., Maghenda, M., Campbell, J. (2008). Unjust waters: climate change, flooding and the urban poor in Africa. *Environment and Urbanization*, *20*(1), 187–205. https://doi.org/10.1177/0956247808089156
8. Durigon, V. L., Carvalho, D. F., Antunes, M. A. H., Oliveira, P. T. S., Fernandes, M. M. (2014). NDVI time series for monitoring RUSLE cover management factor in a tropical watershed. *International Journal of Remote Sensing*, *35*(2), 441–453. https://doi.org/10.1080/01431161.2013.871081
9. Ganjirad, M., Delavar, M. R. (2023). Flood risk mapping using random forest and support vector machine. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *X-4/W1-2022*, 201–208. https://doi.org/10.5194/isprs-annals-x-4-w1-2022-201-2023
10. Ghanim, A. A. J., Shaf, A., Ali, T., Zafar, M., Al-Areeq, A. M., Alyami, S. H., Irfan, M., Rahman, S. (2023). An improved flood susceptibility assessment in Jeddah, Saudi Arabia, using advanced machine learning techniques. *Water*, *15*(14), 2511. https://doi.org/10.3390/w15142511
11. Giudici, P., Gramegna, A., Raffinetti, E. (2023). Machine learning classification model comparison. *Socio-Economic Planning Sciences*, *87*, 101560. https://doi.org/10.1016/j.seps.2023.101560
12. Hou, W., Gao, J. (2019). Simulating runoff generation and its spatial correlation with environmental factors in Sancha River Basin: The southern source of the Wujiang River. *Journal of Geographical Sciences*, *29*(3), 432–448. https://doi.org/10.1007/s11442-019-1608-z
13. Ibarreche, J., Aquino, R., Edwards, R. M., Rangel, V., Pérez, I., Martínez, M., Castellanos, E., Álvarez, E., Jimenez, S., Rentería, R., Edwards, A., Álvarez, O. (2020). Flash flood early warning system in Colima, Mexico. *Sensors*, *20*(18), 5231. https://doi.org/10.3390/s20185231
14. Khan, S. I., Policelli, F., Adler, R. F., Habib, S., Yilmaz, K. K., Irwin, D., Wang, J., Hong, Y., Brakenridge, G. R., Gourley, J. J. (2011). Satellite remote sensing and hydrologic modeling for flood inundation mapping in Lake Victoria Basin: Implications for hydrologic prediction in Ungauged Basins. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(1), 85–95. https://doi.org/10.1109/tgrs.2010.2057513
15. Lyu, H.-M., Cheng, W.-C., Arulrajah, A., Xu, Y.-S. (2018). Flooding hazards across southern china and prospective sustainability measures. *Sustainability*, *10*(5), 1682. https://doi.org/10.3390/su10051682
16. Martinis, S., Ćwik, K., Plank, S. (2018). The use of Sentinel-1 time-series data to improve flood monitoring in arid areas. *Remote Sensing*, *10*(4), 583. https://doi.org/10.3390/rs10040583
17. Misra, A., White, K., Nsutezo, S. F., Straka, W., Lavista, J. (2025). Mapping global floods with 10

years of satellite radar data. *Nature Communications*, *16*(1). https://doi.org/10.1038/s41467-025-60973-1

18. Phan, A., N Ha, D., D Man, C., T N Nguyen, T., Q Bui, H., T Nguyen, T. (2019). Rapid assessment of flood inundation and damaged rice area in Red River Delta from Sentinel 1A imagery. *Remote Sensing*, *11*(17), 2034. https://doi.org/10.3390/rs11172034
19. Pomme, L.-E., Bourqui, R., Giot, R., Auber, D. (2022). *Relative Confusion Matrix: Efficient Comparison of Decision Models*. 98–103. https://doi.org/10.1109/iv56949.2022.00025
20. Reis, V., Hermoso, V., Hamilton, S. K., Ward, D., Fluet-Chouinard, E., Lehner, B., Linke, S. (2017). A global assessment of inland wetland conservation status. *BioScience*, *67*(6), 523–533. https://doi.org/10.1093/biosci/bix045
21. Sharma, M., Bangotra, P., Gautam, A. S., Gautam, S. (2021). Sensitivity of normalized difference vegetation index (NDVI) to land surface temperature, soil moisture and precipitation over district Gautam Buddh Nagar, UP, India. *Stochastic Environmental Research and Risk Assessment*, *36*(6), 1779–1789. https://doi.org/10.1007/s00477-021-02066-1
22. Taherizadeh, M., Niknam, A., Nguyen-Huy, T., Mezősi, G., Sarli, R. (2023). Flash flood-risk areas zoning using integration of decision-making trial and evaluation laboratory, GIS-based analytic network process and satellite-derived information. *Natural Hazards*, *118*(3), 2309–2335. https://doi.org/10.1007/s11069-023-06089-5
23. Totz, S., Pfeiffer, K., Coumou, D., Tziperman, E., Cohen, J. (2017). Winter precipitation forecast in the European and mediterranean regions using cluster analysis. *Geophysical Research Letters*, *44*(24). https://doi.org/10.1002/2017gl075674
24. Tsumita, N., Piyapong, S., Kaewkluengklom, R., Jaensirisak, S., Fukuda, A. (2025). Flood susceptibility mapping of urban flood risk: comparing autoencoder multilayer perceptron and logistic regression models in Ubon Ratchathani, Thailand. *Natural Hazards*, *121*(15), 17833–17867. https://doi.org/10.1007/s11069-025-07494-8
25. Wang, H., Liang, Q., Hancock, J. T., Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00905-w
26. Wang, W., Lu, L., Xu, M., Liu, Y., Zhao, Q., Shao, G., Sang, G. (2024). Prediction of flash flood peak discharge in hilly areas with ungauged basins based on machine learning. *Hydrology Research*, *55*(8), 801–814. https://doi.org/10.2166/nh.2024.004
27. Williams, S. A., Megdal, S. B., Zuniga-Teran, A. A., Quanrud, D. M., Christopherson, G. (2024). Mapping groundwater vulnerability in arid regions: A comparative risk assessment using modified DRASTIC models, land use, and climate change factors. *Land*, *14*(1), 58. https://doi.org/10.3390/land14010058
28. Wu, Y., Zhang, Z., Qi, X., Hu, W., Si, S. (2024). Prediction of flood sensitivity based on logistic regression, extreme gradient boosting, and random forest modeling methods. *Water Science and Technology: A Journal of the International Association on Water Pollution Research*, *89*(10), 2605–2624. https://doi.org/10.2166/wst.2024.146
29. Youssef, A. M., Pourghasemi, H. R., El-Haddad, B. A. (2022). Advanced machine learning algorithms for flood susceptibility modeling - performance comparison: Red Sea, Egypt. *Environmental Science and Pollution Research*, *29*(44), 66768–66792. https://doi.org/10.1007/s11356-022-20213-1
30. Zhao, Y., Gong, P., Yu, L., Hu, L., Li, X., Li, C., Zhang, H., Zheng, Y., Wang, J., Zhao, Y., Cheng, Q., Liu, C., Liu, S., Wang, X. (2014). Towards a common validation sample set for global land-cover mapping. *International Journal of Remote Sensing*, *35*(13), 4795–4814. https://doi.org/10.1080/01431161.2014.930202
31. Zhou, Z., Smith, J. A., Baeck, M. L., Wright, D. B., Smith, B. K., Liu, S. (2021). The impact of the spatiotemporal structure of rainfall on flood frequency over a small urban watershed: an approach coupling stochastic storm transposition and hydrologic modeling. *Hydrology and Earth System Sciences*, *25*(9), 4701–4717. https://doi.org/10.5194/hess-25-4701-2021