

ORIGINAL ARTICLE OPEN ACCESS

A Machine Learning Framework for Flood Risk Assessment With Imbalanced Data

Alexia Tsouni¹ | Constantinos F. Panagiotou²  | Charalampos Konstantinou^{3,4} | Vasiliki Pagana¹ | Diofantos Hadjimitsis^{2,5} | Nikos Mamassis⁶ | Demetris Koutsoyiannis⁶ | Charalampos Kontoes¹

¹Operational Unit “BEYOND Center of Earth Observation Research and Satellite Remote Sensing”, Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing (IAASARS), National Observatory of Athens (NOA), Athens, Greece | ²Department of Resilient Society, ERATOSTHENES Centre of Excellence (ECOE), Limassol, Cyprus | ³Department of Civil and Environmental Engineering, University of Cyprus (UCY), Nicosia, Cyprus | ⁴Faculty of Engineering, American University of Beirut – Mediterraneo (AUBM), Paphos, Cyprus | ⁵Department of Civil Engineering and Geomatics, Cyprus University of Technology, Limassol, Cyprus | ⁶Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens (NTUA), Athens, Greece

Correspondence: Constantinos F. Panagiotou (constantinos.panagiotou@eratosthenes.org.cy) | Charalampos Konstantinou (konstantinou.charalampos@ucy.ac.cy)

Received: 28 November 2025 | **Revised:** 28 May 2026 | **Accepted:** 10 September 2026

Keywords: citizens' emergency calls | flood risk assessment | high dimensional class-imbalanced data | machine learning | Mandra River basin | sustainable development goals

ABSTRACT

This study presents a machine-learning framework for identifying areas of elevated flood risk using imbalanced, high-dimensional geospatial datasets. Using the Mandra River Basin in Greece as a case study, an extreme gradient boosting (XGBoost) classifier was trained on 11 flood-related features including hydrological, meteorological, soil and topographic data, whereas 454 geolocated citizens' emergency calls were used as target information. Class imbalance was addressed through a weighted-only strategy to preserve the integrity of the original dataset, while adjusting class contributions during the training process. Bayesian optimization and stratified 10-fold cross-validation were combined to tune and evaluate the model performance. Threshold-sweep analysis enables the identification of multiple operational settings based on precision–recall trade-offs and cost-sensitive criteria. The results revealed physically interpretable feature importance patterns, with terrain elevation (TE) (45%) and accumulated rainfall (25%) emerging as the dominant predictors according to feature importance analysis, followed by curve number (16%) and the Manning coefficient (9.2%). The proposed framework provides a reliable, data-driven tool for flood risk assessment, and supports a flexible, two-tier decision strategy that differentiates between emergency escalation and broader situational awareness.

1 | Introduction

Flooding is a physical phenomenon that can cause grave disasters, including human and animal fatalities, as well as significant damage to human health, the environment, cultural heritage, property, and infrastructure. According to the European Commission (2025), floods are the most common and costliest disasters in Europe and are responsible for several

devastating effects, such as endangering lives, leading to heavy economic losses, releasing pollutants, destroying wetland areas, and reducing biodiversity.

In the Centre for Research on the Epidemiology of Disasters (Centre for Research on the Epidemiology of Disasters 2025), 142 floods are recorded worldwide, killing 5883 people and affecting 48.8 million people, resulting in economic losses of 32.8

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Journal of Flood Risk Management* published by Chartered Institution of Water and Environmental Management and John Wiley & Sons Ltd.

billion USD. In Europe, 24 floods were recorded that directly led to 345 human deaths, while severely affecting the lives of half a million people, resulting in economic losses of 19.6 billion USD. Therefore, in Europe, compared with the entire world, 17% of the flood events that occurred caused 6% of the deaths, 1% of the affected people, and 60% of the economic losses. These findings reveal that, in Europe, compared with the global total, floods have a much weaker effect on human activity on the one hand but a much greater economic impact on the other hand. This, in turn, means that European urban environments are more vulnerable to floods, possibly due to the historical evolution of urban sprawl in flood-prone areas; as described by Sargentis et al. (2025), who assessed urbanization trends, cities have expanded into floodplains, increasing the frequency and severity of floods, and concluded that urban expansion may be a more significant contributor to flood risk than natural flood events. This is verified by Wang et al. (2023), who examined 910 severe flood events that occurred from 2000 to 2018 and reported that human distancing from floods greatly mitigated flood fatalities and displacements.

The severe impact of floods on society is due mainly to the strong variability of rainfall and streamflow, known as Long-Memory processes or Long-Range Dependent processes (Beran et al. 2013), and to the water cycle in general (Dimitriadis et al. 2021; Koutsoyiannis 2020). Beran et al. (2013) discussed the mathematical and probabilistic foundations and statistical methods, emphasizing their practical motivation and mathematical justification, and provided proofs of the main theorems and data examples, including rainfall and streamflow. Moreover, Koutsoyiannis (2020) processed information from gridded ground observations, satellite data and reanalyses, and concluded that the established climatological hypotheses according to which the hydrological cycle should intensify because of global warming are not confirmed; instead, monotonic trends, fluctuations from intensification to deintensification, and vice versa, are observed. Furthermore, Dimitriadis et al. (2021) conducted a global-scale investigation of stochastic similarities in the marginal distribution and dependence structure of key hydrological-cycle processes using an extended collection of several billions of data values from hundreds of thousands ground stations worldwide and confirmed the universal stochastic view of the hydrological cycle, with marginal structures and dependence structures exhibiting roughness (fractal) behavior at small scales, long-term persistence at large scales, and a transient behavior at intermediate scales.

The EU has adopted the Floods Directive (EU Directive 2007), which requires all EU countries to create and regularly update flood hazard maps and flood risk maps. Flood hazard maps should cover geographical areas that could be flooded, while flood risk maps should show the potential adverse consequences associated with these flood scenarios. On the basis of these maps, flood risk management plans should be drafted. Given that flood risk management is an integral part of integrated river basin management, the Floods Directive is closely coordinated with the Water Framework Directive (EU Directive 2000) in terms of flood risk management plans, river basin management plans, and public participation procedures. At a global scale, the UN has adopted the Sendai Framework for Disaster Risk Reduction 2015–2030 (United Nations 2015a), which provides Member

States with concrete actions to protect development gains from the risk of disaster. It outlines seven clear targets and four priorities for action to prevent new and reduce existing disaster risk: (i) understanding disaster risk; (ii) strengthening disaster risk governance to manage disaster risk; (iii) investing in disaster reduction for resilience; and (iv) enhancing disaster preparedness for effective response and to “Build Back Better” in recovery, rehabilitation and reconstruction. Moreover, flood risk management and reduction are relevant to the Sustainable Development Goals (SDGs) adopted by the UN in the 2030 Agenda for Sustainable Development (United Nations 2015b), most directly to SDG 11 (Sustainable Cities and Communities), SDG 13 (Climate Action), and SDG 6 (Clean Water and Sanitation). However, the effective adaptation of the EU and UN frameworks is challenging because of the high complexity of this physical phenomenon, which is driven by the combined interactions of several physical factors, such as topography, hydrogeology, meteorology, and nonphysical factors, such as land cover change and human interventions, among others. An additional source of complexity arises from the strong variability and persistence exhibited by rainfall, streamflow, and the hydrological cycle in general, commonly referred to as long-term persistence or the Hurst phenomenon (Beran et al. 2013; Dimitriadis et al. 2021; Hurst 1951; Koutsoyiannis 2020), which further amplifies the unpredictability of flood-generating processes across multiple spatial and temporal scales. Furthermore, these factors are characterized by strong spatiotemporal heterogeneities that occur at a wide range of scales. Therefore, it is crucial to understand how these factors interact and quantify their impact on flood dynamics prior to formulating effective policies to mitigate flood risk in urban, semiurban, and rural areas, thereby necessitating the development of advanced techniques for sustainable management and long-term emergency planning.

A plethora of modeling tools and data sources such as physically-based models (Nkeki et al. 2022; Pakati et al. 2025; Tsouni et al. 2025), multicriteria decision analysis (Feloni et al. 2020; Panagiotou 2025; Panagiotou, Feloni, et al. 2025), and machine learning models (MLMs) (Abedi et al. 2022; Chen et al. 2021; Tsangaratos et al. 2023; Zhuang et al. 2026) are used to simulate the spatiotemporal evolution of flood events and quantify their connections with flood drivers and mechanisms. MLMs are being increasingly used for conducting nonsubjective flood risk assessments and developing effective mitigation and adaptation strategies because of their notable ability to capture complex associations among driving mechanisms and effectively handle large high-dimensional datasets. Nevertheless, flood events are rare compared with nonflood events, resulting in highly imbalanced datasets that can significantly degrade the performance of machine learning algorithms (Zhuo and Ge 2021). Therefore, different solutions have been proposed to tackle imbalanced classification issues that frequently occur in field applications, which can be grouped into two main categories: algorithm-level and data-level. The former focuses on adjusting the learning process through class weights or cost-sensitive loss functions (Khan et al. 2018), whereas the latter category aimed at balancing the sizes of the different classes via either oversampling or under-sampling (Abdi and Hashemi 2015). However, very few studies have reported the use of such strategies to address the inherent uncertainty of datasets in flood-related applications, such as streamflow observed data (Pizarro et al. 2022; Wu et al. 2020),

flood risk mapping in urban areas (Wang et al. 2024), early detection of distillation column flooding (Adebayo et al. 2025), and flood damage (Liu et al. 2024). More specifically, Pizarro et al. (2022) reported that in the case of streamflows, there is large variability and intermittency across scales, from fine-scale records (on the order of minutes) to multicentennial paleoclimatic reconstructions (on the order of hundreds of years), and identified a common stochastic structure for the streamflow process, based on the Pareto-Burr-Feller marginal distribution and a generalized Hurst-Kolmogorov (HK) dependence structure. Wu et al. (2020) reported that historical flood data, which are essentially imbalanced, cause machine learning classifiers to be more biased toward patterns with majority samples, resulting in poor classification of patterns with minority samples.

Nishat et al. (2025) proposed a robust machine learning approach for assessing the severity of flood predictions in the Brahmaputra Basin (Bangladesh). Several shallow MLMs were trained and compared on a highly imbalanced dataset that contained 4554 samples and 13 distinct features. To enhance the reliability of the model outputs, the authors applied the Synthetic Minority Oversampling Technique (SMOTE) in different formats of the dataset. Their results achieved more accurate predictions for both multiclass and binary formats when the SMOTE method was applied, revealing the potential of using the resulting models in real-world applications. Wang et al. (2024) focused on urban flooding and its impact on urban management in Zhengzhou city (North China Plain). The authors proposed a novel hybrid resampling method for handling imbalanced datasets by (i) applying the Borderline-SMOTE algorithm, (ii) generating synthetic datasets by undersampling strategies, (iii) feeding the resulting datasets into a random forest classifier, and (iv) training the classification models with a genetic algorithm. Their results achieved excellent agreement with respect to flood risk levels for several classification performance metrics (Accuracy, Recall, G-mean and F1-score). The need for resampling techniques such as SMOTE and Borderline-SMOTE arises from the inherently imbalanced nature of flood-related datasets, in which high-severity or extreme-flow events are heavily underrepresented relative to normal or low-flow conditions. Standard ML classifiers trained directly on such data tend to be biased toward the majority class, and exhibit high overall accuracy but poor recall of the very events that matter most for risk assessment, which justifies the use of oversampling, undersampling, or hybrid resampling schemes to restore class balance and improve generalization to minority classes.

Motivated by previous works, this study aims to develop a machine learning approach for predicting flood risk using the calls to the Fire Service as outputs. Particularly, an extreme gradient boosting algorithm was developed based on hydrometeorological, topographical, and soil information. An extensive record of geolocated citizens' emergency calls was used to train and validate the models. The output is a probability map that aims to inform experts on inspection schemes under two scenarios: (i) routine inspections that focus on areas showing consistently high flood susceptibility, and (ii) event-driven inspections that help direct field teams during severe rainfall. This provides a more structured, data-driven basis for guiding inspection priorities than what is typically available, which serves as the main novelty of this study.

2 | Description of the Study Area and Collected Data

The Mandra River Basin, which is located in the Region of Attica, Greece (38.08°N, 23.46°E), covers an area of approximately 56.8 km². Its hydrographic network contains both natural and artificial parts, including the streams of Agia Ekaterini and Soures that cross the city of Mandra (Figure 1). Historical data have reported the occurrence of extreme precipitation events in this area, which is classified as a flood-prone area in the national implementation of the EU Floods Directive. Particularly, the most severe flash flood event occurred on November 15th 2017, resulting in 24 fatalities and substantial material damage. The rainfall duration was approximately 6 h and led to more than 200 mm of accumulated rainfall, whereas the mean annual rainfall in this area was 373 mm (National Observatory of Athens 2017).

The FloodHub team of the Operational Unit BEYOND/IAASARS/NOA was immediately activated to analyse the flood event. Following a multiparameter analysis, the areas affected by the flood were accurately derived using, on the one hand, detailed flood extent mapping by processing high-resolution satellite imagery and conducting photo interpretation. On the other hand, detailed hydraulic simulations were performed by integrating data collected during the field visits, as well as additional published data. Two-dimensional hydraulic modeling of the maximum flood extent was performed via open-source HEC-RAS software, which was proven to provide accurate predictions even in high-density urban areas (Tsouni et al. 2023).

A series of flood protection technical works were constructed after this event, including a partial diversion of the Agia Ekaterini stream to the Soures stream, which was located upstream of Mandra city. The HEC-RAS 2-D hydraulic model was accordingly adapted to incorporate all the new technical works, and the spatially distributed rainfall (rain-on-grid) method was used to include the cumulative rainfall in the updated hydraulic models. Finally, hydraulic simulations were conducted to map the flood depth under different flood scenarios (50, 100, and 1000-year return periods), assuming wet antecedent soil moisture conditions (CN III) and a 12-h rainfall duration (Tsouni et al. 2025).

Table 1 lists the characteristics of the datasets collected from multiple sources. Particularly, 11 flood factors were selected on the basis of their relevance and impact on the flood phenomenon. Owing to the presence of different spatial resolutions among these factors, the collected information was resampled to a common spatial resolution (25 m) prior to performing the statistical analysis.

The above mentioned flood factors were selected because each reflects an important distinct physical parameter that significantly influences the generation of surface runoff and the likelihood of flooding. Moreover, these factors are commonly used in the literature for assessing flood susceptibility. The TE was estimated via a digital elevation model (DEM), which was updated so that the constructed flood mitigation technical works and the data collected during the experts' field visits

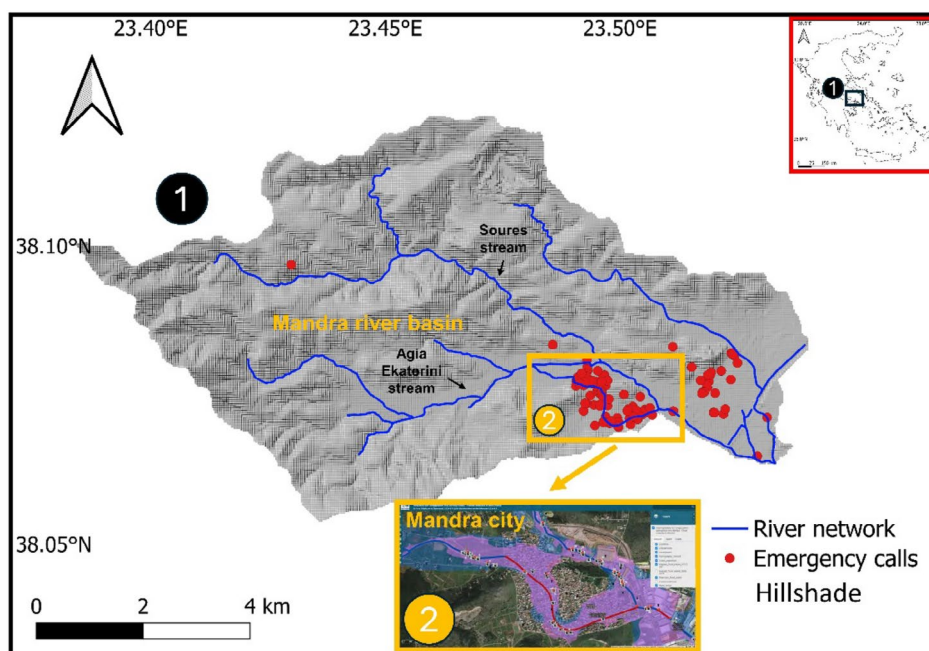


FIGURE 1 | (1) Geographical location of the Mandra River Basin, including the terrain elevation (hillshade background), river network (blue lines) and citizens' emergency calls (red filled circles). (2) Satellite view of Mandra city with mapping of the maximum flash flood extent on November 15th, 2017, which is based on high-resolution Sentinel-2 satellite data (magenta color) and the HEC-RAS hydraulic simulation (blue color).

are integrated to better represent reality and provide more reliable results. Four DEM-derived parameters, particularly the terrain aspect (TA), pan curvature (PC), flow direction (FD), and flow accumulation (FA), were selected because of their strong influence on determining water flow paths, flow intensity, and potential flood accumulation areas. The topographic wetness index (TWI) is another important factor because it combines the upslope drainage area and the steepness of the terrain slope to identify locations where water is expected to accumulate because of gravitational forces. Furthermore, the Euclidean distance from the drainage network (DS) was selected to represent the proximity of a location to the nearest stream, given that regions close to a river have greater flood susceptibility. Drainage density (DR) also contributes to flooding since the presence of a dense network of streams and channels can quickly accumulate runoff water during heavy rainfall. Areas with high DR tend to drain water rapidly, hence they are more susceptible to flash floods. Finally, the accumulated rainfall (RF) provides the main physical trigger for flooding, whereas the Curve Number (CN) and Manning's roughness coefficient (MC) represent how soil and land cover conditions affect infiltration and the resistance of the surface to flow, respectively. Together, these factors give machine-learning models all the essential information they need to identify areas that are more prone to flooding. The geographical distribution of the selected flood factors is shown in the "Supporting Information" file (Figure S1).

With respect to the target variable, information regarding citizens' emergency calls to the Hellenic Fire Service was utilized as a strong indicator of high-risk flood areas. This was possible given that the Fire Service registers every emergency phone call they receive for assistance, including rescue and water pumping operations during flood events. More specifically, the registry includes the date and time of the phone calls

as well as the postal addresses; thus, timing and geospatial information were available to identify the flood-affected people and properties. The collected dataset comprises 454 emergency calls for flood events that occurred in the Mandra River Basin between 2006 and 2020.

3 | Methodological Framework

This study focused on developing a machine learning approach suitable for highly imbalanced datasets for the identification of high flood risk areas. An extreme gradient boosting (XGBoost) model was trained directly on the imbalanced dataset, with Bayesian inference used to optimize its hyperparameters. Model performance was evaluated via a stratified 5-fold outer cross-validation procedure. Feature-importance analysis was then performed on the tuned model to quantify the influence of each explanatory variable on the predictions, improving the understanding of how flood-related factors shape the model output. Finally, the validity of the model outcomes was assessed against existing flood risk maps. Following preliminary testing at 10, 15, and 25 m spatial resolutions, the 25 m grid was selected as an optimal compromise between computational efficiency and predictive performance, while maintaining comparable model accuracy. As a result, data are collected from 90,858 spatial locations which are used to train and test the machine learning models.

3.1 | Multicollinearity Detection for Constructing the Feature Dataset

In classification or regression tasks, the explanatory variables should be as uncorrelated as possible to avoid inflated standard errors of the model coefficients, high statistical uncertainty, and even numerical instabilities (Chan et al. 2022). This situation is

TABLE 1 | Details of the collected datasets. The Coordinate Reference System (CRS) for all parameters is the GGRS87 (EPSG:2100).

Flood factors	Data type	Data description	Data sources
TE	Raster	Terrain elevation set equal to the updated DEM, following modifications based on the constructed flood mitigation technical works and the data collected during the experts' field visits (2 m resolution)	1. National Cadastre of Greece 2. Approved final technical studies of the flood mitigation works as constructed 3. In situ data collection from experts' field visits
TA	Raster	Terrain aspect derived from the updated DEM (2 m resolution)	GIS processing of the updated DEM
PC	Raster	Plan curvature derived from the updated DEM (2 m resolution)	GIS processing of the updated DEM
FD	Raster	Flow direction derived from the updated DEM (2 m resolution)	GIS processing of the updated DEM
FA	Raster	Flow accumulation derived from Flow Direction (2 m resolution)	GIS processing of the updated DEM
TWI	Raster	Topographic wetness index derived from Terrain slope and Flow accumulation (2 m resolution)	GIS processing of the updated DEM
DS	Raster	Distance from drainage network (Euclidean distance) (2 m resolution)	GIS processing of the hydrographic network (which is derived from the updated DEM)
DR	Raster	Drainage density derived from the hydrographic network (2 m resolution)	GIS processing of the hydrographic network (which is derived from the updated DEM)
RF	Raster	Accumulated rainfall (200 m spatial resolution) of the extreme flash flood event on 15/11/2017	XPOL meteorological radar of Institute for Environmental Research and Sustainable Development (IERSD), National Observatory of Athens (NOA)
CN	Vector	Runoff Curve Number Type II (average antecedent soil moisture conditions) 2023	Ministry of Environment 2023
MC	Vector	Manning Roughness Coefficient (n) derived from the Copernicus Urban Atlas (2018)	Copernicus Urban Atlas (2018)

Abbreviations: CN, curve number; DR, drainage density; DS, distance from the drainage network; FA, flow accumulation; FD, flow direction; MC, Manning coefficient; PC, plan curvature; RF, rainfall; TA, terrain aspect; TE, Terrain elevation; TWI, topographic wetness index.

known as the “multicollinearity problem”, which can hinder the ability of a model to assess the individual impacts on the dependent variable that is being estimated. In this study, correlation and multicollinearity analyses were conducted to identify potential multicollinearity issues, enabling the construction of the feature dataset that was used to construct the machine learning models. This is accomplished by calculating the Pearson (linear) and Sperman (ranking) coefficients for all pairs of flood indicators to quantify their pairwise correlations, whereas the variance inflation factor (VIF) and tolerance index (TOL) were used to identify potential multicollinearity effects.

$$VIF_j = \frac{1}{(1-R_j)^2}, \quad TOL_j = \frac{1}{VIF_j} = (1-R_j)^2, \quad j = 1, \dots, n_f, \quad (1)$$

where n_f is the number of flood factors and R_j is the regression coefficient of determination of the j -th factor with respect to the remaining variables. As a thumb of role, VIF values larger than 5 and TOL values less than 0.2 indicate the presence of strong multicollinearity (Tsangaratos et al. 2023).

In addition to the construction of the feature dataset, the binary target variable (*Result*) was defined by assigning a value of 1 to pixels containing at least one geolocated emergency call during the November 2017 flood event, and 0 otherwise, while multiple calls within the same pixel were treated as a single occurrence; nevertheless, these labels should be viewed as indirect indicators of flood impacts, as they may also reflect differences in residents' reporting behavior and may overlook affected areas where no calls were made.

3.2 | Model Development

3.2.1 | Extreme Gradient Boosting (XGBoost) Algorithm

The classification task was addressed using Extreme Gradient Boosting (XGBoost) (Chen and He 2018), a gradient-boosted decision tree method that belongs to the family of random forests and is widely used for water resources or fluid flow in porous media applications (Konstantinou and Papanastasiou 2025;

Konstantinou and Stoianov 2020; Konstantinou and Wang 2024; Konstantinou et al. 2024; Panagiotou et al. 2024). XGBoost is used to construct an ensemble of decision trees in an additive manner. At each boosting iteration, a new tree is built to approximate the negative gradient of the loss function with respect to the model's predictions in the current iteration step. This iterative procedure allows the algorithm to reduce errors made by previous trees, thereby improving accuracy in a progressive manner (Chen and Guestrin 2016).

The objective function of XGBoost combines a differentiable loss function with a regularization term. In this study, logistic loss is employed to model the probability of class membership since this function is often used in binary classification problems. The regularization term penalizes tree complexity by discouraging overly deep or unbalanced trees, thus enhancing generalization and reducing the risk of overfitting. Moreover, XGBoost includes advanced engineering optimizations such as parallel tree construction, the handling of missing data, and the use of approximate algorithms for split finding (e.g., histogram-based methods). These features make XGBoost highly efficient and well suited for imbalanced datasets.

Given a training set of N samples with feature vectors x_i and binary labels $y_i \in \{0,1\}$, the XGBoost prediction after T boosting rounds is calculated by the following equation:

$$\hat{y}_i = \sum_{t=1}^T f_t(x_i), \quad f_t \in F, \quad (2)$$

where F is the functional space of the regression trees. The regularized objective minimized at iteration t is as follows:

$$L(t) = \sum_{i=1}^N \ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3)$$

where $\ell(y, \hat{p}) = -[y \log \sigma(\hat{p}) + (1-y) \log(1 - \sigma(\hat{p}))]$ denotes the logistic loss, $\sigma(\cdot)$ denotes the sigmoid function, and $\Omega(f)$ denotes the regularization term, defined as follows:

$$\Omega(f) = \gamma T_{\text{leaf}} + \frac{1}{2} \lambda \sum_{j=1}^{T_{\text{leaf}}} w_j^2, \quad (4)$$

where T_{leaf} is the number of leaves in tree f , w_j is the weight associated with leaf j , γ denotes the complexity penalty per leaf, and λ denotes the L_2 regularization coefficient applied to the leaf weights (Ridge regularization). The histogram-based split-finding implementation was used throughout, with binary:logistic as the objective function and AUPRC as the evaluation metric.

3.2.2 | Class Imbalance

The dataset analyzed in this study was characterized by extreme class imbalance, where the positive class (Calls) accounted for only a very small proportion of the total observations. In highly imbalanced settings, conventional models tend to achieve deceptively high accuracy by predominantly predicting the majority class. However, this results in poor recall for the minority class, which is the target of interest. Therefore, direct training on imbalanced data can lead to models that fail to detect rare yet critical positive events.

To address this challenge, a single class-imbalance handling strategy was used in this study: “weighted-only” XGBoost training, where the original data are kept intact and imbalance is handled through per-class weights (Velarde et al. 2024). This approach allows the data distribution to remain unchanged while still penalising misclassifications of the minority class, ensuring that the model properly accounts for rare flood risk instances without introducing synthetic samples.

The weighted-only strategy operates entirely at the algorithmic level. Instead of altering the data distribution through resampling, the imbalance is compensated by modifying the relative importance of each class in the training process. This is accomplished by assigning a larger weight to minority-class samples. The positive-class weight is defined as the ratio of majority to minority instances in the training partition. Formally, if N_{neg} and N_{pos} represent the numbers of negative and positive samples, respectively, then the weight applied to each positive instance was computed as follows:

$$w_{\text{pos}} = \frac{N_{\text{neg}}}{\max(1, N_{\text{pos}})} \times \text{WEIGHT_MULT}. \quad (5)$$

Here, WEIGHT_MULT is a scalar multiplier (set to 1.0 in this study) that can be adjusted if more or less emphasis on the minority class is desired. Negative samples were uniformly assigned a weight of one. Consequently, misclassifications of positive samples were penalized approximately $N_{\text{neg}}/N_{\text{pos}}$ times more often than misclassifications of negative samples were. This ensures that the optimization procedure treats positive instances as equally important despite their scarcity.

The computed weights were directly in XGBoost, which internally applied them to the gradient and Hessian calculations at each boosting iteration. Therefore, the training process is guided to focus more strongly on learning patterns associated with the minority class without introducing artificial samples or discarding existing data.

Although the weighted-only configuration was retained as the recommended option, its selection was validated by implementing 11 alternative imbalance-handling strategies, each independently tuned using the same Bayesian-optimisation technique and evaluated on the same spatial test split (see Section 2 in the Supporting Information). The selection process was based on the following criteria: (i) preservation of the integrity of the original dataset, thereby avoiding potential biases introduced by oversampling or undersampling; (ii) computational efficiency, since no additional preprocessing steps are needed; (iii) integration with the XGBoost optimization routine, providing a stable and interpretable approach to handling imbalance; and (iv) performance on the spatial test set. The weighted-only configuration was found to be competitive with and at most operating points superior to the remaining resampling-based alternatives.

3.2.3 | Data Partitioning and Spatial-Leakage Control

Since the underlying data are inherently spatial, observations correspond to raster pixels indexed by planar coordinates (X, Y),

and a random rowwise train/test partition would place pixels from the same spatial neighborhood on both sides of the split, leading to inflated estimates of out-of-sample performance due to spatial autocorrelation. To mitigate this risk while preserving target-class balance, a block-based partitioning scheme was applied prior to any model fitting.

The study area was divided into a regular rectangular grid, where the numbers of cells in the X- and Y-directions, n_x and n_y , respectively, were chosen such that $n_x \times n_y \approx N_{\text{blocks,target}}$ (set to 50 in this study) and $n_x/n_y \approx \Delta x/\Delta y$, where Δx and Δy denote the spatial extent of the dataset along each axis. Each observation was then assigned to the grid cell containing its (X, Y) location. The non-empty cells define a set of spatial blocks $B = \{B_1, \dots, B^M\}$.

The spatial blocks were subsequently partitioned into training and test groups, with a target test fraction of approximately 0.30. The assignment was performed through a randomized search over candidate test-block sets, minimizing a composite objective that combines three normalized deficit terms: (i) the deviation between actual and target row counts, (ii) the deviation between actual and target positive-class counts, and (iii) the deviation between actual and target block counts. Each deficit was normalized by its corresponding total to ensure comparability across the scales. The composite score is defined as follows:

$$S = w_n \cdot \frac{(|N_{\text{train}} - N_{\text{train}}^*| + |N_{\text{test}} - N_{\text{test}}^*|)}{N_{\text{total}}} + w_{\text{pos}} \cdot \frac{(|P_{\text{train}} - P_{\text{train}}^*| + |P_{\text{test}} - P_{\text{test}}^*|)}{P_{\text{total}}} + w_{\text{blocks}} \cdot \frac{(|B_{\text{train}} - B_{\text{train}}^*| + |B_{\text{test}} - B_{\text{test}}^*|)}{B_{\text{total}}} + Q, \quad (6)$$

where N , P , and B denote row, positive-class, and block counts respectively; the starred quantities denote the partition-weighted targets; $w_n = 1$, $w_{\text{pos}} = 4$, and $w_{\text{blocks}} = 1$ assign the greatest priority to positive-class balance; and Q is a penalty term applied when a candidate split violates the minimum-block constraints (at least 40% of blocks assigned to training and at least 20% assigned to testing) or when either partition contains zero positive samples.

A total of 20,000 candidate partitions were generated. Each candidate begins with a random subset of blocks of size, followed by one to ten random add, remove, or swap operations to locally refine the balance. The candidate yielding the lowest composite score was retained as the final spatial split.

This procedure produced a spatial train/test split in which (i) no spatial block contributed to both partitions, (ii) the row, positive-class, and block-count proportions closely matched the target 70/30 ratio, and (iii) the test set consisted of multiple blocks scattered across the study area rather than a single contiguous region. In the final split used in this study, the test partition contained 27,385 rows distributed across 13 spatial blocks, including 35 positive samples (test fraction equal to 30.1%), with positive-class proportions of 0.129% and 0.128% in the training and test partitions, respectively.

3.2.4 | Model Tuning

Hyperparameter optimization was performed via Bayesian optimization. This approach was selected because it efficiently explores the hyperparameter space by constructing a surrogate probabilistic model of the objective function and iteratively updating it with new evaluations (Putatunda and Rama 2018). In this study, the surrogate model was based on Gaussian processes with expected-improvement (EI) acquisition. The optimiser was initialised with 20 Latin-hypercube sample points and then performed 25 sequential acquisition iterations, with up to three independent restarts to recover from rare Gaussian-process fitting failures.

The search space included the following hyperparameters: tree depth, learning rate (η), minimum child weight, subsampling ratio, column sampling ratio, and regularization parameter γ . The bounds used for the optimization are summarized in Table 2.

For each sampled configuration, 10-fold stratified cross-validation was performed on the training set, with folds stratified on the target variable. The performance metric optimized was the area under the precision-recall curve (AUPRC), which is more appropriate for highly imbalanced datasets than AUROC or accuracy. This is because under extreme class imbalance the precision-recall curve is far more sensitive to model improvements than the ROC curve is. AUROC can remain deceptively high even when nearly all the predicted positives are false alarms because the very large negative class dominates the false-positive-rate denominator; AUPRC, by construction, focuses on the rank ordering of the rare positive class. Additionally, AUPRC is threshold-independent and therefore appropriate for hyperparameter selection that should not be tied to a particular operational decision threshold.

An early stopping criterion was also applied: training was terminated if the cross-validated AUPRC did not improve for 100 consecutive boosting rounds, and the model complexity (number of trees) was set to the iteration that achieved the highest AUPRC. The maximum allowed boosting depth in any single CV run was 2000 rounds.

The combination of parameters that maximized AUPRC was retained. The corresponding optimal number of boosting rounds was also recorded. These tuned parameters were then reused consistently across all outer folds to ensure comparability and reproducibility.

TABLE 2 | Bayesian optimization search space.

Hyperparameter	Lower	Upper	Type
Max_depth	3	7	Integer
Eta (η)	0.01	0.20	Continuous
Min_child_weight	1.0	6.0	Continuous
Subsample	0.50	1.00	Continuous
Colsample_bytree	0.50	1.00	Continuous
Gamma (γ)	0.0	5.0	Continuous

To quantify the sensitivity of the final result to the stochastic components of Bayesian optimization, the complete tuning procedure was repeated using five independent random seeds, each generating a new Latin-hypercube initialization and a new fold partition. The variability in the selected hyperparameters and the corresponding cross-validated AUPRC across these repetitions was recorded, providing a measure of the reproducibility of the optimization procedure.

3.2.5 | Model Evaluation and Feature Importance Analysis

The evaluation process involved stratified 10-fold outer cross-validation (Kagalkar et al. 2025; Rana et al. 2024), with folds drawn from the training partition and stratified on the target variable. In each fold, the model was trained on a weighted training subset and evaluated on a validation subset that retained the natural class distribution. The Out-of-fold (OOF) probabilities were collected for the training data, whereas predictions were also generated for the “unseen” test set.

The predicted probabilities were subsequently converted into binary classifications through a threshold-sweep analysis over the unit interval [0,1] with a step size of 0.001. For each threshold, predictions were binarized and a set of diagnostic metrics, particularly precision, recall, and specificity, were computed to evaluate model behavior under different decision rules. Multiple F scores were also used, defined as follows:

$$F_{\beta} = (1 + \beta^2) \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \quad \beta = 0.5, 1, 2, 3 \quad (7)$$

where parameter β controls the relative importance between Precision and Recall. The Balanced accuracy is defined as follows:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}, \quad (8)$$

whereas Youden's J index is defined as follows:

$$J = \text{Sensitivity} + \text{Specificity} - 1. \quad (9)$$

A cost-sensitive metric was also evaluated, defined as follows:

$$\text{Cost} = C_{\text{FN}} \cdot \text{FN} + C_{\text{FP}} \cdot \text{FP}, \quad (10)$$

where false negatives and false positives were penalized at $C_{\text{FN}} = 50$ and $C_{\text{FP}} = 1$, respectively. Following the computation of all metric values across thresholds, nine threshold-selection criteria were applied to capture different operational and risk-management priorities. These included the thresholds that maximized precision, recall, balanced accuracy, the F_1 -score, and the F_2 -score; the threshold that minimized the cost function; the threshold yielding the highest precision while maintaining a recall of at least 0.5; the threshold at which precision and recall were equal; and the highest-recall threshold associated with zero false negatives.

All thresholds were determined exclusively from the out-of-fold predictions generated on the training partition. The test set was held aside and was used solely for the final evaluation

of the selected thresholds. Consequently, the reported test metrics constitute an unbiased estimate of generalization performance.

Confusion matrices were generated for both training and test predictions at each selected threshold. These were stored both in aggregated tabular form and as individual 2×2 matrices to ensure reproducibility. In addition, performance summaries were reported in terms of the area under the receiver operating characteristic curve (AUROC) to measure the effectiveness of the model.

To assess the spatial homogeneity of model performance, two additional diagnostic outputs were produced. First, per-fold OOF AUPRC and AUROC were tabulated across the 10 cross-validation folds, providing a measure of model stability under different training subsets. Second, the test partition was decomposed back into its constituent spatial blocks, and AUPRC and AUROC were computed separately for each test block. The spread of these per-block metrics was used to verify that the headline test performance was not driven by a single easy region of the study area.

To further check the spatial split robustness, the complete training-and-evaluation procedure was repeated for five independent spatial partitions of the blocks, each generated by a different random seed for the randomized block assignment search. For every repetition, the previously tuned hyperparameters were retained, and the model was retrained on the new training partition and evaluated on the new test partition. The distribution of the test AUPRC and AUROC across these five spatial partitions provides a direct estimate of the variability that would result from drawing a different held-out test set.

The feature importance in the weighted-only XGBoost configuration was assessed using both the model's internal importance metrics and the SHapley Additive exPlanations (SHAP) values. XGBoost provides several measures of variable influence, most commonly Gain, Cover, and Frequency. Gain represents the average improvement in the loss function produced by splits involving a given feature and is generally regarded as the most informative internal metric because it reflects how strongly a variable contributes to reducing classification error. Cover measures the proportion of training samples falling into branches where the feature is used for splitting, whereas Frequency counts how often the feature appears across all the trees in the ensemble.

To complement these internal indicators, SHAP values were computed to obtain a unified, game-theoretic explanation of feature contributions (Aldrees et al. 2024; Konstantinou et al. 2025; Shapley 1953a, 1953b). SHAP assigns each prediction an additive contribution value for every feature, indicating how strongly that feature shifts the prediction toward or away from the positive class. Aggregating these values across all samples yields the mean absolute SHAP value, which ranks features according to their overall influence on model predictions, and the mean SHAP value, which indicates whether their influence is generally positive or negative with respect to the target outcome.

To complement the global feature-importance summaries with local interpretability information, three additional

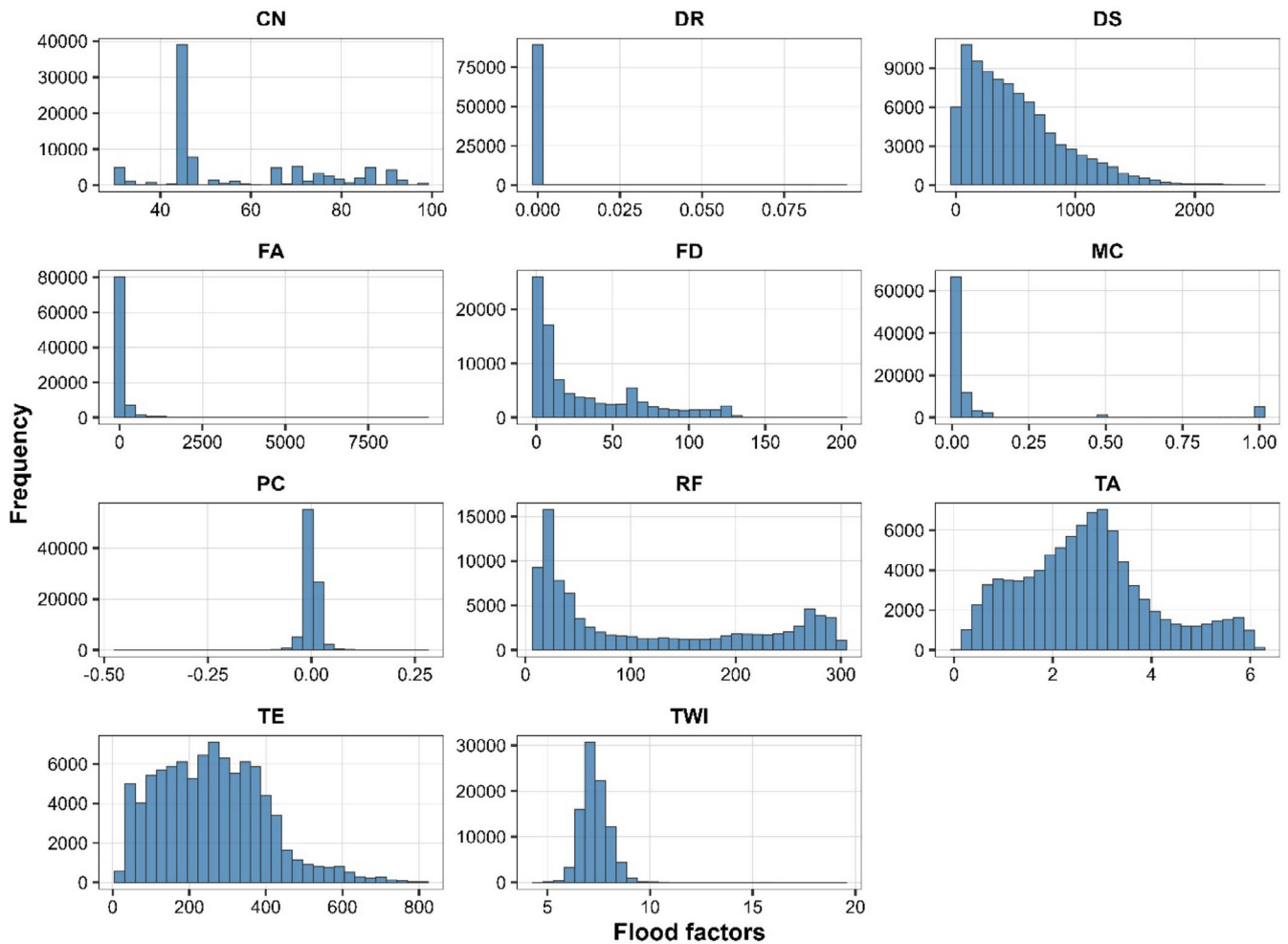


FIGURE 2 | Frequency distributions (histograms) of the selected flood parameters.

SHAP-based outputs were generated. First, SHAP dependence data were recorded for the top-ranked features according to the mean $|SHAP|$, and each observation's feature value was paired with its corresponding SHAP contribution. Second, pairwise SHAP interaction values were computed using the TreeSHAP interaction extension. The resulting mean-absolute-interaction matrix together with the ranked list of the strongest interacting feature pairs was used to identify localized joint dependencies between predictors. Third, the complete SHAP contribution matrix for the test set was retained, together with the most influential drivers of each individual test prediction. The latter enable each test observation to be interpreted in terms of the features that most strongly shifted its predicted probability, thereby providing a row-level audit trail that complements the ranking based on the global gain importance metrics.

4 | Results and Discussion

4.1 | Exploratory Analysis

Eleven physical parameters were used as input variables, namely CN, DR, distance from the DS, FA, FD, Manning coefficient (MC), PC, accumulated rainfall (RF), TA, TE, and TWI. The distributions of these factors are generated by extracting

their values from each location (pixel) of the common numerical grid (see Section 3).

Highly-skewed profiles are observed for all flood factors (Figure 2), especially for DR, FA, MC, and PC, because of the presence of discontinuous distributions and outliers. This observation highlights the difficulty of effectively handling this feature dataset to provide robust model predictions.

Table 3 shows the pairwise correlations among the flood parameters. The upper triangular part of the correlation table displays the Pearson (linear) coefficient, whereas the lower triangular part of the table displays the Spearman (rank) coefficient. The highest correlation is observed for the pair [RF, TE] (0.773 and 0.857 for Pearson and Spearman, respectively), indicating a strong connection between the high-altitude regions and the rainfall events. Moderate (positive) correlations are observed between TE and DS (0.578 and 0.509 for Pearson and Spearman, respectively), whereas moderate (negative) correlations are observed for pairs [TE, CN], [PC, FA], and [MC, TE]. Overall, the majority (80%) of the (absolute) pairwise correlations are less than 0.25. The correlation plot is also shown in Figure SM2.

Moreover, the data in Table S1 reveal that the VIF and TOL values for all the parameters do not exceed the corresponding thresholds. In particular, all the criteria exhibited VIF values below

4.1, remaining well under the commonly adopted thresholds of 5 or 10, whereas all Tolerance Indicator (TOI) values exceeded 0.26, which is far above the 0.1 threshold generally considered to indicate of problematic multicollinearity. Consequently, all the parameters are used to construct the feature dataset which is used in the machine learning analysis.

4.2 | Model Performance and Inference

Figure 3 shows the Receiver Operating Characteristic (ROC) curve for the training and testing dataset obtained using the weighted-only XGBoost model. The ROC curve plots the true positive rate

against the false positive rate for all the thresholds. The curve's trajectory increases well above the diagonal line (which represents random guessing) and approaches the top-left corner, which demonstrates that the model effectively separates positive and negative cases. The high AUROC values (0.979 for training and 0.941 for testing) confirm that weighted-only XGBoost can distinguish flood-related calls from nonflood ones with high confidence, maintaining reasonable class separability across thresholds.

The sensitivity of the Bayesian-optimization outputs to stochastic initialization was assessed by repeating the entire tuning pipeline using five independent random seeds, each generating new Latin-hypercube initialization and a new

TABLE 3 | Pearson (upper triangular part of the table) and Spearman (lower triangular part of the table, green cells) correlation coefficients for all the flood parameters. Bold values indicate the highest correlation for each parameter.

	DR	DS	FA	FD	TE	RF	TWI	TA	PC	CN	MC
DR		−0.143	0.021	−0.013	−0.086	−0.039	−0.039	0.007	−0.11	0.126	−0.01
DS	−0.24		0.003	0.058	0.578	0.25	0.141	0.008	−0.01	−0.35	−0.16
FA	0.022	0.067		0.002	0.034	0.07	−0.008	−0.029	−0.27	0.029	−0.06
FD	0.038	0.001	−0.041		0.113	0.098	−0.022	−0.393	−0.01	−0.145	−0.01
TE	−0.11	0.509	0.138	0		0.773	0.19	−0.052	−0.03	−0.448	−0.39
RF	−0.05	0.264	0.167	0.001	0.857		0.186	−0.11	−0.05	−0.32	−0.26
TWI	−0.05	0.139	−0.119	0.002	0.243	0.252		−0.059	0.071	0.057	0.157
TA	0.016	−0.001	−0.042	−0.241	−0.058	−0.112	−0.101		0.003	−0.012	0.003
PC	−0.10	−0.007	−0.504	−0.014	−0.015	−0.037	0.077	−0.01		−0.006	0.05
CN	0.144	−0.395	−0.042	−0.078	−0.464	−0.309	0.043	0.004	−0.02		0.387
MC	0.01	−0.144	−0.089	−0.029	−0.433	−0.327	0.061	−0.009	0.019	0.167	

Abbreviations: CN, Curve number; DR, Drainage density; DS, Distance from drainage network; FA, Flow accumulation; FD, Flow direction; MC, Manning coefficient; PC, Plan curvature; RF, Accumulated rainfall; TA, Terrain aspect; TE, Terrain elevation; TWI, Topographic wetness index.

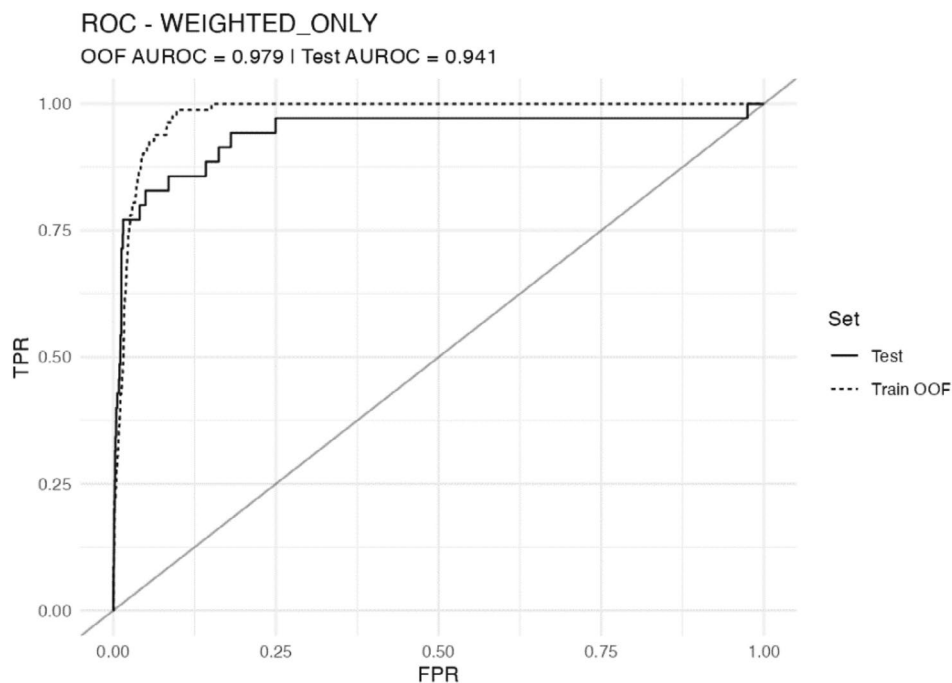


FIGURE 3 | Receiver operating characteristics.

cross-validation fold partition. The cross-validated AUPRC was very stable across runs (mean = 0.964, SD = 0.005, coefficient of variation 0.42%), confirming that the optimiser converges to essentially the same training-time performance regardless of initialization. The selected hyperparameters were also internally consistent: `max_depth` converged to 4 in all five runs; `subsample` (mean 0.83, CV 13.8%) and `colsample_bytree` (mean 0.68, CV 16.4%) were tightly clustered; and the learning rate and gamma showed the largest spread (CV 97% and 84% respectively), but with compensating variation in `best_nrounds`, indicating that BO explores the equivalence class of ($\eta \times nrounds$) configurations that yield the same loss minimum. The test AUROC was likewise stable (mean = 0.927, SD = 0.023, CV 2.5%). The test AUPRC was the only metric with high relative variability (mean = 0.034, SD = 0.023, and CV 67%), a direct and unavoidable consequence of having very few positive pixels in any single test partition: the AUPRC denominator is then dominated by a handful of high-leverage points. The detailed seed-by-seed hyperparameter values are reported in Table SM5.

A second, independent source of stochastic variability arises from the selection of the spatial blocks assigned to the test partition. Repeating the entire training-and-evaluation procedure across five independent spatial partitions (29 training blocks and 13 test blocks in each partition, while preserving 82 training positives and 35 test positives) yielded highly stable test AUROC values (mean = 0.960, SD = 0.011, and CV = 1.2%) and moderately variable test AUPRC values (mean = 0.066, SD = 0.007, and CV = 11.2%). The variability in the test AUPRC was substantially smaller than the across-seed variability observed for a single split, thereby confirming that the reported headline performance is not driven by a single favorable partition of the study area. Per-repeat values are reported in Supplementary Table SM6.

Operating at very high thresholds (e.g., maximum precision, threshold = 0.883) yields an extremely selective model. Here, the classifier almost never issues a false alarm, with only 28 false positives among 27,350 negatives, but it identifies only 4 true positives while 28 actual flood calls are missing. This behavior is akin to a “confirm-only” emergency check, where the system triggers an alert only when the probability of a flood-related call is extremely high.

Such a conservative threshold would be suitable in resource-limited emergency response settings where each alarm triggers a costly or disruptive intervention (e.g., dispatching inspection teams, activating flood barriers, or mobilizing rescue units).

The priority in these cases is reliability over coverage, ensuring that almost every issued alert truly corresponds to an emergency, even if this means that most real flood events go undetected automatically and require manual follow-up.

This mirrors real-world triage logic where emergency control centers might use such a classifier as a verification layer to confirm incidents flagged by other systems (e.g., citizen reports or rainfall sensors), ensuring that false positives do not waste critical resources.

In contrast, operating at low thresholds (e.g., maximum recall or the no-false-negative criterion, threshold = 0.116) produces a model that captures almost all true flood calls (TP = 34 out of 35) but at the cost of a large number of false positives (FP = 8291).

This configuration prioritizes completeness over precision: it ensures that nearly every possible emergency is detected, even if many nonemergency calls are also flagged for review.

In the context of flood risk management, this approach would be valuable for early-warning or mass-screening systems, where missing an event is far more costly than investigating a false alarm. The additional false positives can then be filtered downstream, for example, by human dispatchers or automated verification modules using auxiliary data (e.g., rainfall radar and sensor networks).

This mirrors a “safety-first” strategy, which is consistent with the operational philosophy of emergency preparedness: it is preferable to check and dismiss a nonevent rather than to miss a real flood emergency.

Intermediate thresholds (e.g., `min_cost` at 0.781) provide a practical compromise between the two extremes. For instance, the minimum cost criterion achieves 26 true positives and only 9 false negatives, while limiting the number of false positives to 409. This corresponds to a manageable workload for verification teams and a high probability of capturing real emergencies. From a decision-support standpoint, this threshold could be interpreted as the optimal operational setting where the balance between missed emergencies and unnecessary interventions aligns with the organization's defined cost ratios Table 4.

Within an emergency-management framework, these results delineate two complementary operational roles for the same model. The first corresponds to an emergency-verification mode operating at high thresholds, which minimizes false alarms and is therefore suitable for confirmatory decision layers, especially when human resources or dispatch capacity are constrained. The second corresponds to an early-warning or monitoring mode operating at low thresholds, which maximizes detection coverage and supports the early detection of potential flood events, even if many alerts require postfiltering. The final operational threshold depends on the acceptable trade-off between response cost and the risk of missed-events. In practice, emergency-management agencies could adopt a two-tiered strategy in which the low-threshold configuration flags all potential high-risk calls for preliminary review, thereby providing broad coverage, while the high-threshold configuration triggers automatic alerts or escalations, serving as an emergency-verification layer. A layered approach based on the findings of the XgBoost algorithm is shown in Figure 4. The predicted probabilities were grouped into four classes: very low [< 0.116], low [$0.116, 0.201$], moderate [$0.201, 0.781$], and high [> 0.781]. These thresholds were defined based on the various threshold-selection criteria presented previously. The model appears to capture almost all flood-related calls, while it flags the downstream area in the lower part of the basin near the vicinity of the cells where a call was registered as more vulnerable to flooding. However, two

TABLE 4 | Confusion matrices for weighted XgBoost across various criteria.

Objective	Threshold	TN	FP	FN	TP	Precision	Recall	F1	F2	Balanced accuracy	Cost	
Maximum precision	0.886		27,348	2	35	0	0.000	0.000	—	—	0.500	1752
Maximum recall	0.116		19,059	8291	1	34	0.004	0.971	0.008	0.020	0.834	8341
Maximum balanced accuracy	0.201		22,091	5259	2	33	0.006	0.943	0.012	0.030	0.875	5359
Maximum F1-score	0.883		27,322	28	31	4	0.125	0.114	0.119	0.116	0.557	1578
Maximum F2-score	0.883		27,322	28	31	4	0.125	0.114	0.119	0.116	0.557	1578
Minimum Cost ($C_{\text{FN}} = 50$, $C_{\text{FP}} = 1$)	0.781		26,941	409	9	26	0.060	0.743	0.111	0.226	0.864	859
Precision at recall ≥ 0.5	0.823		27,083	267	19	16	0.057	0.457	0.101	0.189	0.724	1217
Precision = recall	0.888		27,350	0	35	0	—	0.000	—	—	0.500	1750
No false negatives	0.116		19,059	,291	1	34	0.004	0.971	0.008	0.020	0.834	8341

Note: (1) Thresholds are the operating points selected on the training OOF probabilities for each objective and then applied to the spatially independent test set. (2) Cost = $50 \cdot FN + 1 \cdot FP$, with the 50:1 weighting matching the manuscript's cost-sensitive criterion ($C_{FN} = 50$, $C_{FP} = 1$). (3) Bal. Acc. = (Sensitivity + Specificity)/2; F2 weights recall four times more than precision relative to F1. (4) — marks NaN cells, which occur when an objective's optimal threshold produces zero predicted positives on the test set (Precision/F1/F2 then undefined).

important sources should be acknowledged. First, flooding may have occurred in some areas even if no emergency calls were recorded. Second, emergency calls may have been reported in locations where they did not actually occur. This discrepancy could be explained by the fact that in some cases, the recorded call location does not always correspond to the precise location of the flood event, especially when the flooded area is not accessible to the citizens.

The feature importance analysis based on the XGBoost Gain metric (Figure 5a) revealed that TE was by far the most influential predictor, contributing nearly half of the total model gain. These findings confirm that TE is the dominant control on the spatial distribution of flood-related emergency calls, as low-lying areas are naturally more prone to surface-runoff accumulation and inundation. Cumulative rainfall (RF) ranked second, reflecting the critical role of hydrology in terms of the magnitude and spatial distribution of rainfall during extreme flash-flood events. The CN, which represents the soil infiltration capacity and land-cover characteristics, and Manning's roughness coefficient (MC) emerged as the next most influential factors, reflecting the importance of both soil conditions and hydraulic behavior.

To investigate whether the dominant predictors of flood-related emergency calls operate consistently across the basin, the SHAP outputs were examined at three increasing levels of locality: (i) the directional decomposition of global feature effects

(Figure 5b), (ii) the strongest pairwise feature interactions, and (iii) the per-pixel attribution of individual high-risk predictions on the test set.

The mean absolute SHAP values indicated that the terrain elevation (TE, 0.909), accumulated rainfall (RF, 0.702), Curve Number (CN, 0.455), Manning's roughness coefficient (MC, 0.147), and distance from the drainage network (DS, 0.075) were the five most influential predictors. Together, the first three variables account for approximately 89% of the total mean |SHAP| magnitude across the 11 features, confirming the dominant role of terrain, rainfall, and infiltration-related controls already identified by the Gain index in Figure 5a. The Spearman correlation between each feature value and its corresponding SHAP contribution provides a physically interpretable measure of the directionality of the effect. TE was negatively correlated with its SHAP contribution ($\rho = -0.234$), indicating that lower elevations shift predictions toward the flood class, which is consistent with the gravitational accumulation of surface runoff in low-lying parts of the basin. Distance from the DS behaved similarly ($\rho = -0.424$), implying that proximity to drainage channels increases the predicted probability of flooding. The Curve Number, which encodes the infiltration capacity of soil and land cover, showed the strongest positive directional relationship ($\rho = +0.707$), indicating that higher CN values, corresponding to less permeable and more urban surfaces, increase the predicted flood risk. The Topographic Wetness Index (TWI; $\rho = +0.693$) and the flow-direction encoding ($\rho = +0.810$) also

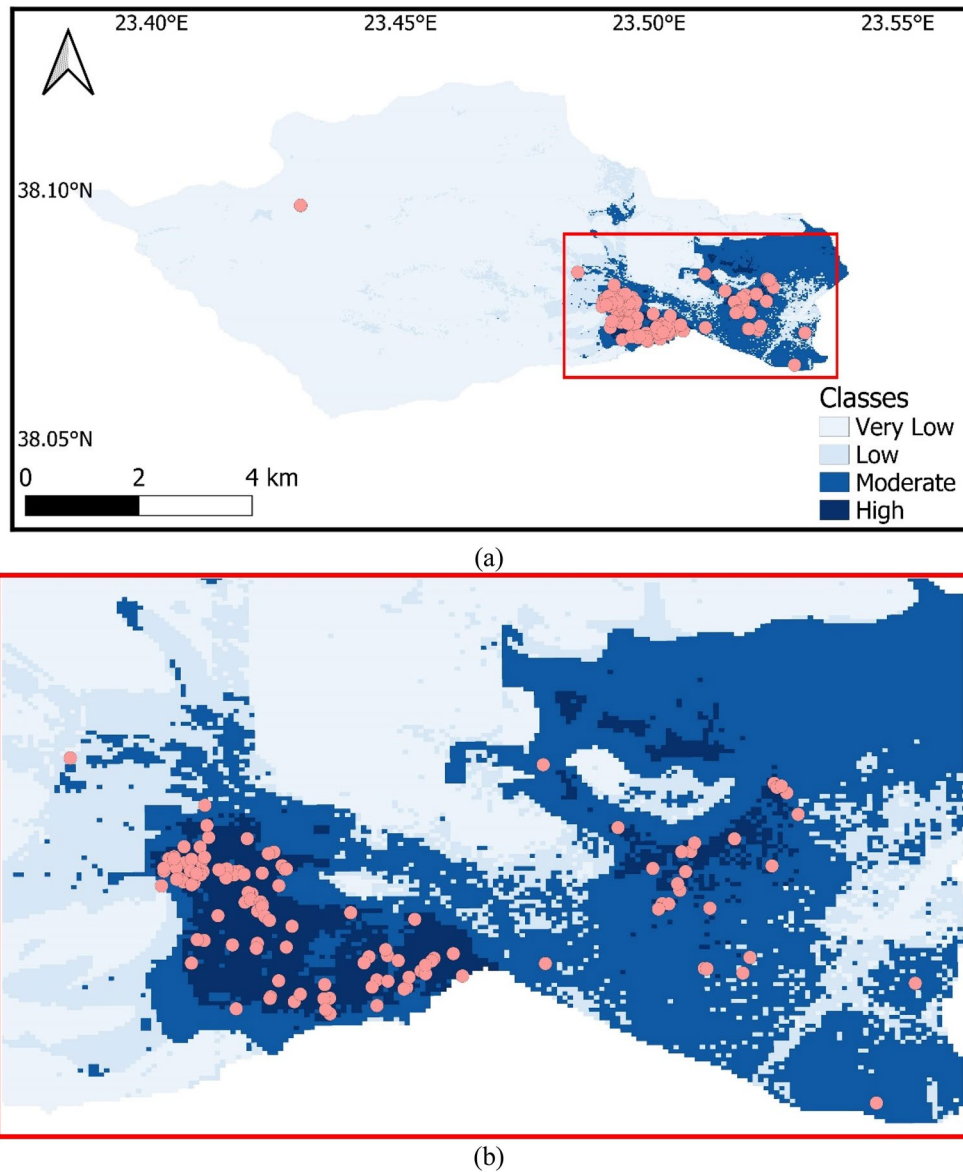


FIGURE 4 | (a) Geographical distribution of the probability of emergency calls occurring based on the predictions of the weighted XgBoost model. (b) Zoomed window covering the region where the emergency calls are present. The predictions are grouped into 4 classes: Very Low [<0.116], Low [$0.116, 0.201$], Moderate [$0.201, 0.781$], and High [>0.781]. Pink solid circles denote the locations where the flood-related emergency calls were recorded.

exhibited monotonic relationships consistent with the expected hydrological behavior. The negative directionality observed for accumulated rainfall ($\rho = -0.161$) initially appears counterintuitive. However, this result is consistent with the strong positive correlation between rainfall and TE in the Mandra basin (Spearman $\rho = +0.857$). After TE captures the dominant topographic signal, the residual rainfall variability effectively behaves as an altitude-related proxy with the opposite sign. The agreement between the SHAP directionalities and the underlying hydrological and hydraulic processes provides an independent post hoc validation of the model's internal logic.

In addition, the identified importance of TE, rainfall, and surface roughness is in line with previous findings by Dimitriadis et al. (2016), who also highlighted topographic gradients, rainfall-driven inflows, and roughness parameters as dominant controls on flood behavior and uncertainty.

The full TreeSHAP interaction tensor was also computed on a representative subsample of training pixels, yielding a matrix of mean absolute pairwise interaction strengths (Table SM8 and Figure SM3). The three strongest interactions in the entire matrix occurred among the three dominant features: TE $\times$ RF (0.200), RF $\times$ CN (0.143), and TE $\times$ CN (0.055). The combined magnitude of these three interaction terms (0.398) is of the same order as the mean absolute SHAP value of CN itself (0.455) and is more than five times greater than the mean |SHAP| of the next-ranked individual feature (Manning's roughness coefficient). These results indicate that the model does not treat the three principal drivers as purely additive. Instead, high CN values occurring in low-elevation and high-rainfall locations are associated with substantially higher predicted risk than would be expected from the linear sum of their individual effects. This localized nonlinear amplification is consistent with the physical dynamics of urban flash flooding in the Mandra basin,

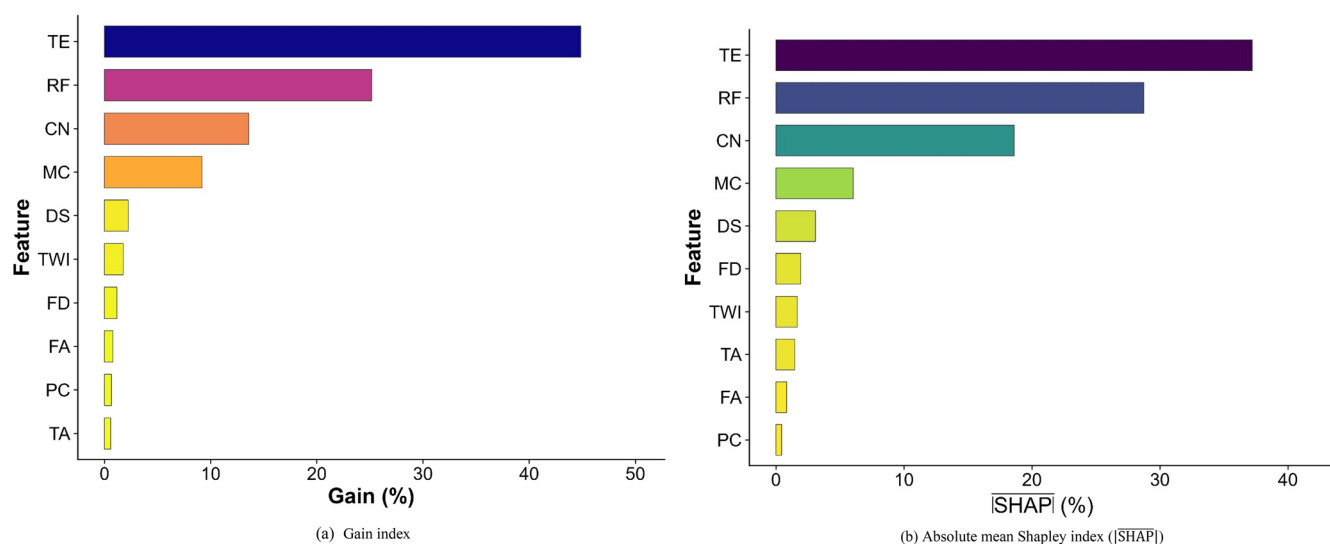


FIGURE 5 | Feature importance analysis using (a) Gain index and (b) the $|\overline{\text{SHAP}}|$ index with the weighted-only XGBoost model.

where impermeable surfaces located near drainage pathways intensify the response of runoff generated by extreme rainfall. The fourth- and fifth-strongest interaction terms ($\text{FD} \times \text{TE}$ and $\text{DS} \times \text{TE}$, both approximately 0.020) were an order of magnitude weaker, confirming that the interaction structure of the model is concentrated around a small number of dominant hydrological controls rather than being diffusely distributed across all the predictors.

Pixel-level SHAP decomposition further renders the model's logic auditable at any specific location of operational interest. Across all 35 positive test pixels, TE, CN, and RF appeared among the three strongest contributing features in 100% of the cases, with Manning's coefficient replacing RF in only a single instance. Among the 30 true-positive predictions (positive pixels assigned a predicted probability ≥ 0.5), TE was the single most influential feature in 22 cases (73%), whereas accumulated rainfall dominated in the remaining 8 cases (27%). This distribution mirrors the global feature-importance ranking and indicates that the model relies on the same physically meaningful mechanisms across the entire population of correctly identified flood pixels, rather than on idiosyncratic feature combinations confined to few locations.

The same decomposition also reveals the localized structure of the false-positive predictions. The 3155 high-probability false positives ($\text{Prob} \geq 0.5$, true negative label) were spatially concentrated within three contiguous blocks of the test partition: B_{09_03} (1245 pixels), B_{10_03} (1511 pixels), and B_{07_02} (421 pixels). These blocks share the same hydrological signature as the true flood-affected blocks do, namely, low elevation, high CN values, and a dense local drainage network, despite the absence of recorded emergency calls. Inspection of the corresponding per-pixel SHAP attributions in these blocks reveals that the model identified these areas using the same physical criteria as those associated with true-positive predictions. Specifically, TE was the dominant driver in 94% of the false-positive cases, with elevation-related SHAP contributions typically ranging between +0.95 and +1.00, values essentially indistinguishable from those observed for true positives. This pattern does not necessarily

indicate model overfitting. Rather, the close agreement between the SHAP-derived attributions of true and false positives suggests that these blocks are genuinely flood-prone according to the same physical criteria characterizing the documented flood-affected areas but did not produce emergency calls during the recorded events. This interpretation is consistent with the earlier discussion regarding the limitations of citizen emergency calls as an indirect and incomplete proxy for actual flood occurrence. The local SHAP attributions therefore not only confirm that the model has internalised physically meaningful flood dynamics, but also help to identify locations where the target variable itself may underrepresent the true spatial extent of flooding.

5 | Conclusion

Floods remain among the most destructive hazards, threaten lives, damage infrastructure, and challenge the resilience of communities worldwide. Therefore, the identification of locations where flood events are most likely to occur and cause significant damage is essential for effective preparedness, emergency response, and planning (including human distancing from floods). In this study, a machine learning framework based on the XGBoost algorithm was developed using 11 spatially distributed physical parameters together with geolocated emergency-call records to assess flood risk in the Mandra River Basin (Greece).

The proposed framework successfully captured relationships consistent with well-established hydrological and hydraulic processes. Terrain elevation, accumulated rainfall, surface roughness, and drainage-related characteristics consistently emerged as the main drivers of flooding within the Mandra River Basin. The highest correlation was observed for the pair Accumulated Rainfall & Terrain Elevation, with correlation values 0.773 and 0.857 for Pearson and Spearman coefficients, respectively, indicating a strong connection between the high-altitude regions and the rainfall events. The feature importance analysis using the Gain index demonstrated that the most influential flood parameters were the TE (44.9%), the Accumulated Rainfall (25.2%), the Curve Number (13.5%) and the Manning's roughness coefficient

(9.2%). When using the Absolute mean Shapley index, the top flood parameters were the TE (37.2%), the Accumulated Rainfall (28.7%), the Curve Number (18.6%) and the Manning Coefficient (6%). Therefore, in total, the most important flood parameters are the TE, the Accumulated Rainfall, the Manning Coefficient and the CN.

By translating these relationships into a continuous flood-probability map, the model offers a practical tool capable of complementing traditional flood dynamics studies and supporting flood risk management under real-world operational constraints. Although the developed models are influenced by site-specific characteristics, their broader applicability could be progressively enhanced through the integration of data from additional case studies, enabling the development of more generalized and transferable models. In this context, transferability could be supported through full retraining on labeled data from the target basin, fine-tuning of the existing model using limited local observations, or preliminary domain-adaptation diagnostics based on covariate-shift analysis between source and target predictor distributions.

The resulting flood risk map can be adapted to different operational requirements. It can guide emergency managers in identifying areas where damage is more likely to occur during extreme flood events, while also assisting planners in prioritizing preventive measures, maintenance activities, and targeted interventions in vulnerable locations. Moreover, the probability structure of the model provides substantial operational flexibility: it can inform both conservative strategies that prioritize minimizing missed events and more selective strategies aimed at reducing unnecessary mobilization of emergency-response resources.

Acknowledgments

This work was supported by the Programming Agreement “Earthquake, fire and flood risk assessment in the Region of Attica” (Part A) between the Prefecture of Attica and the National Observatory of Athens. This research was also partially funded by the “EXCELSIOR”: ERATOSTHENES: Excellence Research Centre for Earth Surveillance and Space-Based Monitoring of the Environment H2020 Widespread Teaming project (<https://www.excelsior2020.eu>, accessed on 13 July 2026). The “EXCELSIOR” project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 857510, from the Government of the Republic of Cyprus through the Directorate General for the European Programmes, Coordination and Development and the Cyprus University of Technology.

Funding

This work was supported by the “EXCELSIOR” project, which has received funding from the European Union's Horizon 2020 Research and Innovation Program (grant agreement no 857510).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available upon request from the corresponding author. The data are not publicly available due to privacy restrictions.

References

- Abdi, L., and S. Hashemi. 2015. “To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling and Boosting Techniques.” *Soft Computing* 19, no. 12: 3369–3385.
- Abedi, R., R. Costache, H. Shafizadeh-Moghadam, and Q. B. Pham. 2022. “Flash-Flood Susceptibility Mapping Based on XGBoost, Random Forest and Boosted Regression Trees.” *Geocarto International* 37, no. 19: 5479–5496.
- Adebayo, O., S. Imtiaz, and S. Ahmed. 2025. “Machine Learning for Early Detection of Distillation Column Flooding.” *Chemical Engineering Science* 312: 121603.
- Aldrees, A., M. Khan, A. T. B. Taha, and M. Ali. 2024. “Evaluation of Water Quality Indexes With Novel Machine Learning and SHapley Additive ExPlanation (SHAP) Approaches.” *Journal of Water Process Engineering* 58: 104789.
- Beran, J., Y. Feng, S. Ghosh, and R. Kulik. 2013. *Long-Memory Processes: Probabilistic Properties and Statistical Methods*. Springer Berlin Heidelberg.
- Centre for Research on the Epidemiology of Disasters (CRED). 2025. “2024 Disasters in Numbers: A Hot and Stormy Year. Centre for Research on the Epidemiology of Disasters (CRED), Brussels, Belgium, 8 pp.”
- Chan, J. Y.-L., S. M. H. Leow, K. T. Bea, et al. 2022. “Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review.” *Mathematics* 10, no. 8: 1283.
- Chen, J., G. Huang, and W. Chen. 2021. “Towards Better Flood Risk Management: Assessing Flood Risk and Investigating the Potential Mechanism Based on Machine Learning Models.” *Journal of Environmental Management* 293: 112810.
- Chen, T., and C. Guestrin. 2016. “XGBoost.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–94. New York, NY, USA: ACM.
- Chen, T., and T. He. 2018. “xgboost: eXtreme Gradient Boosting.” R package.
- Copernicus Urban Atlas. 2018. *Urban Atlas Land Cover/Land Use 2018 (vector), Europe*. European Environment Agency, European Union's Copernicus Land Monitoring Service. <https://doi.org/10.2909/949683b7-5795-4c72-845f-77d049010649>.
- Dimitriadis, P., D. Koutsoyiannis, T. Iliopoulou, and P. Papanicolaou. 2021. “A Global-Scale Investigation of Stochastic Similarities in Marginal Distribution and Dependence Structure of Key Hydrological-Cycle Processes.” *Hydrology* 8, no. 2: 59.
- Dimitriadis, P., A. Tegos, A. Oikonomou, et al. 2016. “Comparative Evaluation of 1D and Quasi-2D Hydraulic Models Based on Benchmark and Real-World Applications for Uncertainty Assessment in Flood Mapping.” *Journal of Hydrology* 534: 478–492.
- EU Directive. 2000. “Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy.” <https://eur-lex.europa.eu>.
- EU Directive. 2007. “Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the Assessment and Management of Flood Risks (Text With EEA Relevance).” <https://eur-lex.europa.eu>.
- European Commission Energy Climate change Environment. 2025. “Floods.” <https://environment.ec.europa.eu>.
- Feloni, E., I. Mousadis, and E. Baltas. 2020. “Flood Vulnerability Assessment Using a GIS-Based Multi-Criteria Approach—The Case of Attica Region.” *Journal of Flood Risk Management* 13, no. S1: e12563. <https://doi.org/10.1111/jfr3.12563>.

- Hurst, H. E. 1951. "Long-Term Storage Capacity of Reservoirs." *Transactions of the American Society of Civil Engineers* 116, no. 1: 770–799.
- Kagalkar, A., S. Dharaskar, N. Chaudhari, V. Vakharia, and R. R. Karri. 2025. "Enhanced Metal Ion Adsorption Using ZnO-MXene Nanocomposites With Machine Learning-Based Performance Prediction." *Scientific Reports* 15, no. 1: 15563.
- Khan, S. H., M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. 2018. "Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data." *IEEE Transactions on Neural Networks and Learning Systems* 29, no. 8: 3573–3587.
- Konstantinou, C., M. Giakoumi, and C. F. Panagiotou. 2025. "A Game-Theoretic Framework for Sustainable Water Allocation in Agriculture: Enhancing Economic and Water Efficiency in Akrotiri District (Cyprus)." *Science of the Total Environment* 966: 178691.
- Konstantinou, C., C. Jara-Arriagada, and I. Stoianov. 2024. "Investigating the Impact of Cumulative Pressure-Induced Stress on Machine Learning Models for Pipe Breaks." *Water Resources Management* 38, no. 2: 603–619.
- Konstantinou, C., and P. Papanastasiou. 2025. "A Data-Driven and Machine Learning-Assisted Interpretation of Hydraulic Fracturing Experiments in Various Formations." *Geomechanics for Energy and the Environment* 43: 100707.
- Konstantinou, C., and I. Stoianov. 2020. "A Comparative Study of Statistical and Machine Learning Methods to Infer Causes of Pipe Breaks in Water Supply Networks." *Urban Water Journal* 17, no. 6: 534–548.
- Konstantinou, C., and Y. Wang. 2024. "Statistical and Machine Learning Analysis for the Application of Microbially Induced Carbonate Precipitation as a Physical Barrier to Control Seawater Intrusion." *Journal of Contaminant Hydrology* 263: 104337.
- Koutsoyiannis, D. 2020. "Revisiting the Global Hydrological Cycle: Is It Intensifying?" *Hydrology and Earth System Sciences* 24, no. 8: 3899–3932.
- Liu, C.-F., L. Huang, K. Yin, S. Brody, and A. Mostafavi. 2024. "FloodDamageCast: Building Flood Damage Nowcasting With Machine-Learning and Data Augmentation." *International Journal of Disaster Risk Reduction* 114: 104971.
- National Observatory of Athens. 2017. "Press Release, 20/11/2017." <https://www.noa.gr>.
- Nishat, F. Z., N. Nahar, F. I. Joti, S. Islam, N. Adri, and M. U. Ahmed. 2025. "Flood Severity Classification in Bangladesh: A Comprehensive Analysis of Historical Weather and Water Level Data Using Machine Learning Approaches." *Natural Hazards* 121, no. 9: 10195–10224.
- Nkeki, F. N., E. I. Bello, and I. G. Agbaje. 2022. "Flood Risk Mapping and Urban Infrastructural Susceptibility Assessment Using a GIS and Analytic Hierarchical Raster Fusion Approach in the Ona River Basin, Nigeria." *International Journal of Disaster Risk Reduction* 77: 103097.
- Pakati, S. S., C. Shoko, and T. Dube. 2025. "Integrated Flood Modelling and Risk Assessment in Urban Areas: A Review on Applications, Strengths, Limitations and Future Research Directions." *Journal of Hydrology: Regional Studies* 61: 102583.
- Panagiotou, C. F. 2025. "Copula-Based Assessment of Flood Susceptibility in the Island of Cyprus via Stochastic Multicriteria Decision Analysis." *Science of the Total Environment* 979: 179469.
- Panagiotou, C. F., E. Feloni, K. Aristidou, and M. Eliades. 2025. "Probabilistic Assessment of Flood Susceptibility via a Coparticipative Multicriteria Decision Analysis." *Environmental Processes* 12, no. 2: 22.
- Panagiotou, C. F., C. Konstantinou, and A. Chekirbane. 2024. "A Generalized Machine Learning Approach for Cost-Effective Monitoring of Irrigation Suitability: A Demonstration Case in El Fahs Aquifer (Tunisia)." *Groundwater for Sustainable Development* 27: 101324.
- Pizarro, A., P. Dimitriadis, T. Iliopoulou, S. Manfreda, and D. Koutsoyiannis. 2022. "Stochastic Analysis of the Marginal and Dependence Structure of Streamflows: From Fine-Scale Records to Multi-Centennial Paleoclimatic Reconstructions." *Hydrology* 9, no. 7: 126.
- Putatunda, S., and K. Rama. 2018. "A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost." *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, pp. 6–10. New York, NY, USA: ACM.
- Rana, M., D. Patel, V. Vakharia, and S. K. Singh. 2024. "UAV-Based DEM Augmentation Using ConSinGAN for Efficient Flood Parameter Prediction With Machine Learning and 1D Hydrodynamic Models." *Physics and Chemistry of the Earth, Parts A/B/C* 135: 103675.
- Sargentis, G.-F., R. Ioannidis, M. Koukia, et al. 2025. "Do Floods Attack Cities or Cities Invade Flood Plains? Case Study: Athens, Greece." In *Proceedings of the International Conferences on Digital Technology Driven Engineering 2024*, 216–224. Springer.
- Shapley, L. S. 1953a. "The Value of an N-Person Game."
- Shapley, L. S. 1953b. "A Value for n-Person Games. Contributions to the Theory of Games." In *Contributions to the Theory of Games II*. Princeton University Press.
- Tsangaratos, P., I. Ilia, A.-A. Chrysafi, I. Matiatos, W. Chen, and H. Hong. 2023. "Applying a 1D Convolutional Neural Network in Flood Susceptibility Assessments—The Case of the Island of Euboea, Greece." *Remote Sensing* 15, no. 14: 3471.
- Tsouni, A., S. Antoniadi, E. Ieronimidi, et al. 2023. "Multiparameter Analysis of the Flood of November 15, 2017 in West Attica Using Satellite Remote Sensing." In *Geoinformatics for Geosciences*, 325–357. Elsevier.
- Tsouni, A., S. Sigourou, V. Pagana, et al. 2025. "Multiparameter Flood Risk Assessment and Management Planning at High Spatial Resolution in the Region of Attica, Greece." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18: 25966–25979.
- United Nations. 2015a. "Sendai Framework for Disaster Risk Reduction 2015–2030." <https://www.undrr.org>.
- United Nations. 2015b. "The 2030 Agenda for Sustainable Development." USA. <https://docs.un.org>.
- Velarde, G., M. Weichert, A. Deshmunkh, et al. 2024. "Tree Boosting Methods for Balanced and Imbalanced Classification and Their Robustness Over Time in Risk Assessment." *Intelligent Systems With Applications* 22: 200354.
- Wang, H., Y. Meng, H. Xu, et al. 2024. "Prediction of Flood Risk Levels of Urban Flooded Points Though Using Machine Learning With Unbalanced Data." *Journal of Hydrology* 630: 130742.
- Wang, N., F. Sun, D. Koutsoyiannis, et al. 2023. "How Can Changes in the Human-Flood Distance Mitigate Flood Fatalities and Displacements?" *Geophysical Research Letters* 50, no. 20: e2023GL105064. <https://doi.org/10.1029/2023GL105064>.
- Wu, Y., Y. Ding, and J. Feng. 2020. "SMOTE-Boost-Based Sparse Bayesian Model for Flood Prediction." *EURASIP Journal on Wireless Communications and Networking* 2020, no. 1: 78.
- Zhuang, Y., T. Gong, J. Fang, et al. 2026. "Integrating Social Media Data and Machine Learning Methods for Flash Flood Susceptibility Mapping in China." *Journal of Hydrology* 664: 134397.
- Zhuo, Y., and Z. Ge. 2021. "Auxiliary Information-Guided Industrial Data Augmentation for Any-Shot Fault Learning and Diagnosis." *IEEE Transactions on Industrial Informatics* 17, no. 11: 7535–7545.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** jfr370257-sup-0001-FigureS1-S3-TableS1-S9.docx. **Figure SM1:** Geographical distribution of the selected flood indicators. **Figure SM2:** Correlation (Pearson) plot of the selected flood indicators. **Figure SM3:** Barplot showing the pairwise SHAP interactions in the weighted-only XGBoost model, computed via the TreeSHAP interaction tensor on a 2000-pixel sub-sample of the training partition. **Table SM1:** Variance Inflation Factor (VIF) and Tolerance Indicator (TOL) for all flood features. VIF values larger than 5 and TOL values less than 0.2 indicate strong multicollinearity. **Table SM2:** Class-imbalance strategies compared in this study. **Table SM3:** The tuned parameters for the main model used in this work. **Table SM4:** The results obtained for the various class imbalance strategies. **Table SM5:** Bayesian-optimisation stability on a fixed spatial split. All five runs share the same training rows ($N_{\text{train}} = 63,519$, 82 positives), the same test rows ($N_{\text{test}} = 27,211$, 35 positives), and the same fold assignments; only the BO initialization seed differs. **Table SM6:** The values of the various models developed for various testing sets. **Table SM7:** Global SHAP-based feature importance for the weighted-only XGBoost model, ranked by mean absolute SHAP value over the training partition. **Table SM8:** Fifteen strongest pairwise SHAP interactions in the weighted-only XGBoost model, computed via the TreeSHAP interaction tensor on a 2000-pixel sub-sample of the training partition. **Table SM9:** Frequency of each feature as the top-ranked (Rank-1) SHAP driver across three classes of test-set predictions: true positives (TP, Prob ≥ 0.5 with True = Call, $N = 30$), false positives (FP, Prob ≥ 0.5 with True = No call, $N = 3155$), and false negatives (FN, Prob < 0.5 with True = Flood Call, $N = 5$). Counts are absolute; percentages are within-class. Terrain elevation dominates as the top-1 driver across all three classes, indicating that the model uses the same physical mechanism to flag both correctly- and incorrectly-predicted positives.