water

MDPI

Article

A Similar-Basin Learning Framework for Flood Susceptibility Prediction Using Historical Flood Inventories

Ufuk Yükseler [1], Ömerul Faruk Dursun [1,*] and Abdolmajid Mohammadian [2,*]

[1] Department of Civil Engineering, Faculty of Engineering, Inonu University, 44280 Malatya, Türkiye; ufukyukseler12@gmail.com
[2] Department of Civil Engineering, Faculty of Engineering, Ottawa University, Ottawa, ON K1N 6N5, Canada
* Correspondence: faruk.dursun@inonu.edu.tr (Ö.F.D.); amohamma@uottawa.ca (A.M.)

Abstract

Flood susceptibility prediction mapping has become an essential component of disaster risk reduction; however, the spatial transferability of machine learning models remains insufficiently explored, particularly under geographically independent validation conditions. This study evaluates the generalization capability of three tree-based machine learning algorithms, namely M5P Regression Tree (M5PRT), Extremely Randomized Trees (Extra Trees), and Gradient Boosting (GB), for flood susceptibility prediction in hydro-climatically similar basins of the Eastern Black Sea Region, Türkiye. A flood inventory derived from historical flood events recorded in Rize Province between 1973 and 2023 was combined with ten flood-conditioning factors representing topographic, hydrological, climatic, geological, and land-cover characteristics. Model performance was first assessed using a Leave-One-Basin-Out (LOBO) validation framework to quantify cross-basin generalization within the training region. Subsequently, an independent external validation was conducted using flood events from Ordu (2018), Giresun (2015), Gümüşhane (2017), and Trabzon (2019), none of which were included during model development. The LOBO analysis revealed consistent predictive performance across most basin scenarios, indicating that the selected predictors effectively captured transferable flood-conditioning relationships. External validation produced Area Under the Curve (AUC) values ranging from 0.77 to 0.89, with the highest predictive performance observed in Trabzon (AUC = 0.89) and Giresun (AUC = 0.86), while comparatively lower performance was obtained in Gümüşhane (AUC = 0.79). The observed performance gradient closely reflected the degree of hydro-climatic similarity between the training and validation basins, highlighting the influence of regional environmental characteristics on model transferability. The findings demonstrate that tree-based machine learning models can maintain robust predictive skill beyond the calibration domain when applied to hydro-climatically comparable basins. By integrating basin-wise cross-validation with geographically independent external testing, the proposed framework provides a rigorous assessment of model generalization and offers a transferable approach for regional-scale flood susceptibility prediction assessment.

Keywords: artificial intelligence; flood hazard mapping; similar-basin approach; machine learning

Check for updates

Academic Editors: Xiekang Wang and Ataur Rahman

Received: 2 July 2026
Revised: 6 October 2026
Accepted: 24 November 2026
Published: 27 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Floods are among the most destructive natural hazards worldwide, causing substantial environmental, social, and economic losses every year. The frequency and severity of flood events have increased considerably over recent decades due to climate change, rapid

urbanization, land-use alterations, and the growing occurrence of extreme precipitation events [1,2]. Consequently, the identification of flood-prone areas has become a fundamental component of disaster risk reduction, sustainable watershed management, and climate adaptation planning. Reliable flood susceptibility prediction assessments not only support mitigation strategies but also contribute to the protection of human life, infrastructure, and natural resources. The increasing exposure of communities to flood hazards has therefore intensified the demand for accurate, transferable, and operational flood prediction frameworks [3,4].

In Türkiye, particularly within the Eastern Black Sea Region, flood disasters frequently occur in steep and geomorphologically complex watersheds characterized by intense precipitation, short hydrological response times, and dense drainage networks. Provinces such as Rize, Trabzon, Giresun, and Artvin have experienced numerous destructive flood events that have resulted in significant loss of life and widespread infrastructure damage. The combination of mountainous terrain, high annual rainfall, narrow valley systems, and rapid runoff generation creates highly favorable conditions for flash flooding. Despite the high frequency of flood disasters, the availability of long-term hydrological observations remains limited in many watersheds, constraining the application of conventional flood assessment approaches and increasing the need for alternative data-driven methodologies.

Traditionally, flood hazard assessments have been conducted using physically based and semi-empirical approaches. Hydraulic models such as HEC-RAS and MIKE are widely employed to simulate flood processes and delineate inundation extents; however, their successful implementation generally requires detailed topographic information, reliable hydrological records, and appropriate boundary conditions [5,6]. Similarly, multi-criteria decision-making (MCDM) approaches have been extensively applied to identify flood-prone areas by integrating multiple flood-conditioning factors through expert-based weighting schemes [7–9]. Although these methods provide valuable insights into flood processes, their applicability is often restricted by data requirements, subjective assumptions, and limited transferability across different environmental settings.

Recent advances in remote sensing, geographic information systems, and artificial intelligence have led to increasing interest in machine learning (ML)-based flood susceptibility prediction assessment. Unlike traditional approaches, ML algorithms can capture complex nonlinear relationships between flood occurrence and environmental variables without requiring explicit physical assumptions. Historical flood inventories have therefore emerged as valuable sources of information for developing predictive flood models, enabling the extraction of flood-conditioning relationships directly from observed events [10,11]. Numerous studies have demonstrated the effectiveness of ML techniques for flood susceptibility prediction mapping. Wang et al. (2015) [12] highlighted the capability of Random Forest models to identify influential flood-conditioning factors, whereas Kundzewicz et al. (2017) [13] emphasized the importance of incorporating climate-related variability into flood prediction frameworks [12,13]. Similarly, Yükseler et al. (2023) [14] reported that tree-based machine learning models outperform conventional statistical approaches in flood susceptibility prediction studies conducted within the Eastern Black Sea Region of Türkiye [14]. Although Yükseler et al. (2023) [14] demonstrated the effectiveness of tree-based machine learning for flood prediction in the Eastern Black Sea Region, the present study extends this approach by explicitly focusing on spatial transferability. Rather than limiting model evaluation to the basin used for model development, the proposed framework combines iterative LOBO validation across historical Rize basins with independent testing in Ordu, Giresun, Gümüşhane, and Trabzon. Thus, the main novelty lies in evaluating whether flood-conditioning relationships learned in Rize can be transferred to hydrologically independent basins and neighboring regions [14]. More recently, ensemble learning

techniques and deep learning architectures have demonstrated substantial improvements in predictive performance across diverse geographical and environmental conditions [15,16].

Despite these methodological advances, an important limitation remains unresolved within the existing literature. Most machine learning-based flood susceptibility prediction studies adopt basin-specific modeling strategies, where model training and validation are performed using data derived from the same watershed. Although such approaches often yield high predictive performance, they provide limited evidence regarding model transferability and generalization capability beyond the calibration basin. Consequently, it remains uncertain whether flood-conditioning relationships learned from historical events in one watershed can be successfully transferred to hydrologically independent basins exhibiting similar environmental characteristics. This limitation is particularly important in ungauged or data-scarce regions, where comprehensive flood inventories are unavailable and transferable modeling frameworks are required.

The Eastern Black Sea Region provides an appropriate setting for addressing this research gap. Rize Province, despite covering a relatively small area of approximately 3920 km^2, has experienced numerous destructive flood events over the last five decades. The availability of historical flood records spanning the period 1973–2023 offers a unique opportunity to develop a comprehensive multi-event flood inventory and investigate the extent to which flood-conditioning relationships can be generalized across independent basins. Furthermore, neighboring provinces located within the same hydro-climatic region provide suitable test environments for evaluating model transferability under geographically independent conditions.

In this context, the present study proposes a similar-basin validation framework for flood susceptibility prediction using historical flood inventories collected from multiple flood events recorded in Rize Province. Three tree-based machine learning algorithms representing different learning paradigms, namely M5P Regression Tree (M5PRT), Extremely Randomized Trees (Extra Trees), and Gradient Boosting (GB), were comparatively evaluated [17,18]. In this study, "learning framework" refers to the organization of model training and cross-basin evaluation using existing tree-based algorithms, rather than to a newly developed learning algorithm. Its contribution is an empirical assessment of predictive performance across three complementary validation settings: basin-wise LOBO validation, an unseen-event holdout, and geographically external testing. Environmental similarity is examined as a descriptive aid for interpreting performance differences; the analysis does not establish the physical mechanisms or causal determinants of model transferability. Unlike conventional basin-specific studies, model robustness was assessed through a hierarchical validation framework. First, a Leave-One-Basin-Out (LOBO) validation strategy was implemented to evaluate cross-basin generalization within the training region. Subsequently, geographically independent flood events from neighboring provinces were employed for external validation. By integrating multi-basin learning, basin-wise validation, and geographically independent testing, this study aims to evaluate the transferability of flood-conditioning relationships across hydro-climatically similar basins and contribute to the development of operational flood susceptibility prediction frameworks for data-scarce environments.

2. Study Area

Rize Province, located in the northeastern region of Türkiye along the Black Sea coast, serves as the study area. With an area of 3920 km^2, the region features diverse landforms, including mountainous terrains, deep valleys, and narrow plains. The Kaçkar Mountains rise to 3927 m, while the lowest elevation is at sea level. Prominent rivers such as Fındıklı, Arılı, Fırtına, Pazarsuyu, Büyükdere, Taşlıdere, and İyidere drain the area. Twelve major

flood events considered in this study occurred in Rize Province between 1973 and 2023, including 11 historical events used for model development and the 2023 event reserved for independent testing (Figure 1).

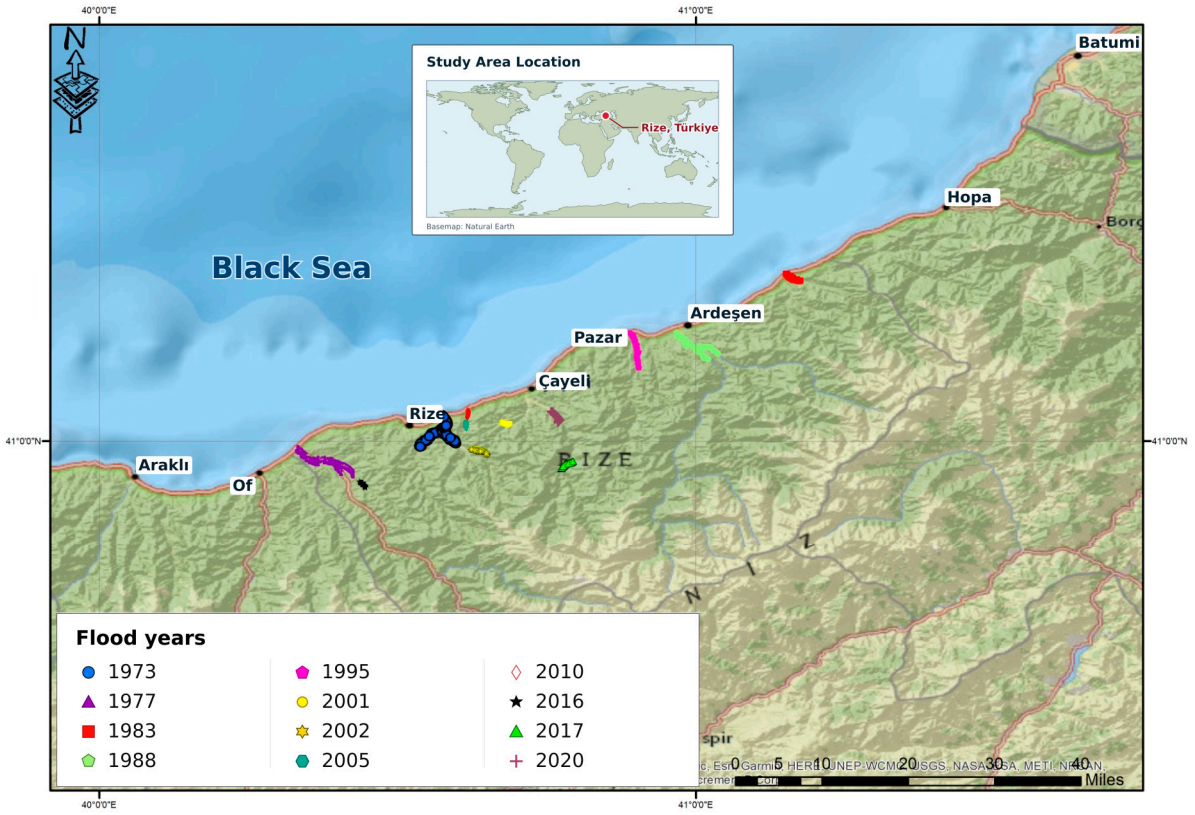


Figure 1. Study Area and Basin-Based Distribution of Historical Flood Events in Rize Province (Türkiye).

Between 1970 and 2023, Many disasters occurred in Rize, and these disasters resulted in many losses of life and property. Factors contributing to the region's flood vulnerability include steep river gradients, high annual rainfall, and low evaporation rates. On average, Rize records an annual temperature of 14 °C, with extremes ranging from −7 °C to 38 °C. Rainfall is unevenly distributed throughout the year, peaking in September, October, and November (Figure 2). Over the past 50 years, an average of 172.6 rainy days per year has been observed, with March being the month with the most rainy days. These climatic and geographic features make Rize particularly prone to flood events.

Rize Province is highly prone to floods due to its steep riverbeds, the highest annual rainfall in Türkiye, and low evaporation rates. Over recent decades, flood disasters in the region have resulted in considerable loss of life and widespread damage to infrastructure and agricultural areas, with a total of approximately 180 fatalities reported.

Between 1973 and the early 2000s, several major flood events affected multiple districts, including Güneysu, Kalkandere, Pazar, Hemşin, Fındıklı, Ardeşen, and Çayeli. These events were generally triggered by intense rainfall leading to the overflow of small and steep streams such as Taşlıdere, İyidere, Fırtına, Hemşin, Bodasarı, and Arılı, causing extensive damage to settlements, cultivated lands, buildings, and transportation structures.

Subsequent flood events, particularly those recorded after 2000, were frequently accompanied by landslides and occurred over short time periods with extremely high rainfall intensities. Notable examples include the floods in 2010, 2016, and 2017, during which rainfall amounts exceeded critical thresholds and resulted in severe structural damage and loss of life. More recently, the flood event in July 2023 was triggered by approximately

250 kg/m^2 of rainfall within 24 h, leading to widespread flooding, landslides, and significant disruptions to transportation and infrastructure. Official assessments indicated that hydraulic inadequacies in existing structures further intensified the impacts of this event (Figure 3).

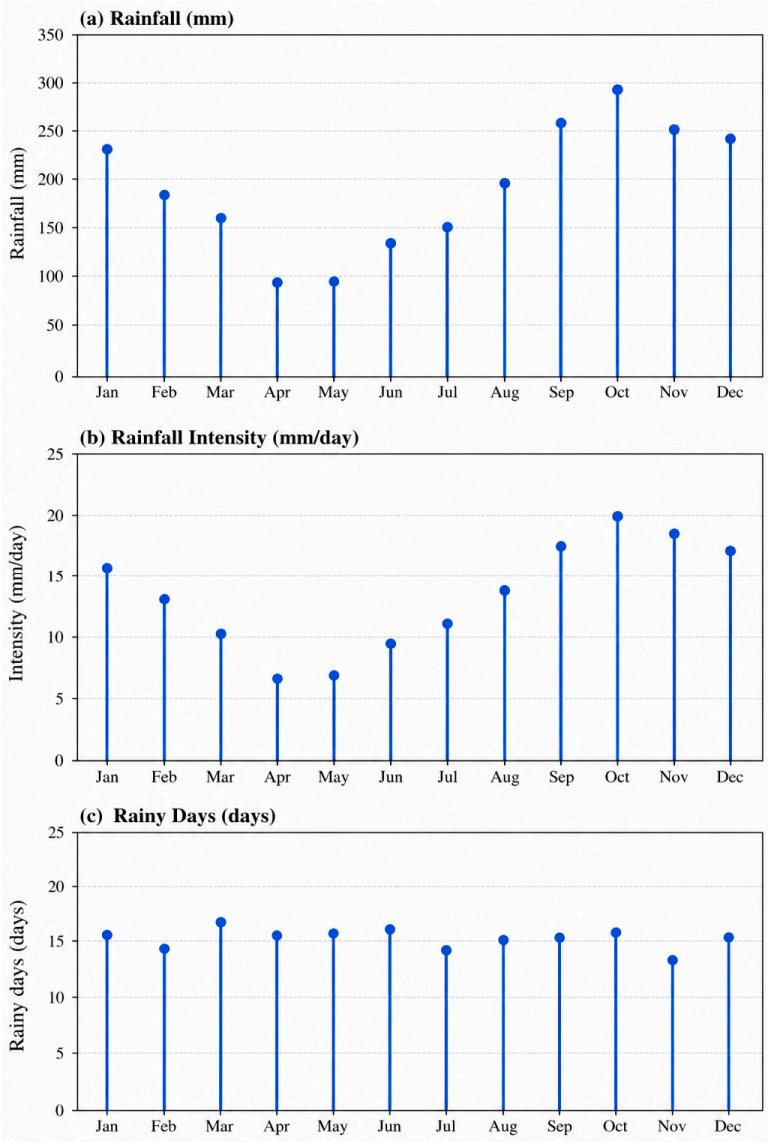


Figure 2. Climatic Characteristics of Rize Province: Rainfall Distribution, Intensity, and Rainy Days (values from 1970 to 2023).

Sediment transport plays a crucial role in exacerbating the impact of flood disasters, significantly increasing their destructive power. In the 2010, 2016, and 2017 flood events, excessive rainfall not only triggered severe flooding but also caused widespread landslides. The destabilization of slopes led to structural collapses, particularly affecting building foundations, resulting in loss of life. During the 2017 disaster, the closure of key highways connecting roads further hampered emergency response efforts, delaying aid and isolating affected communities (Figure 4). Representative examples of flood impacts observed during the 2010 İyidere flood (Figure 4a,b), the 2017 Çayeli flood (Figure 4c,d), and the 2020 Çayeli flood (Figure 4e,f) are presented to illustrate the severity of recent flood disasters in the study area.

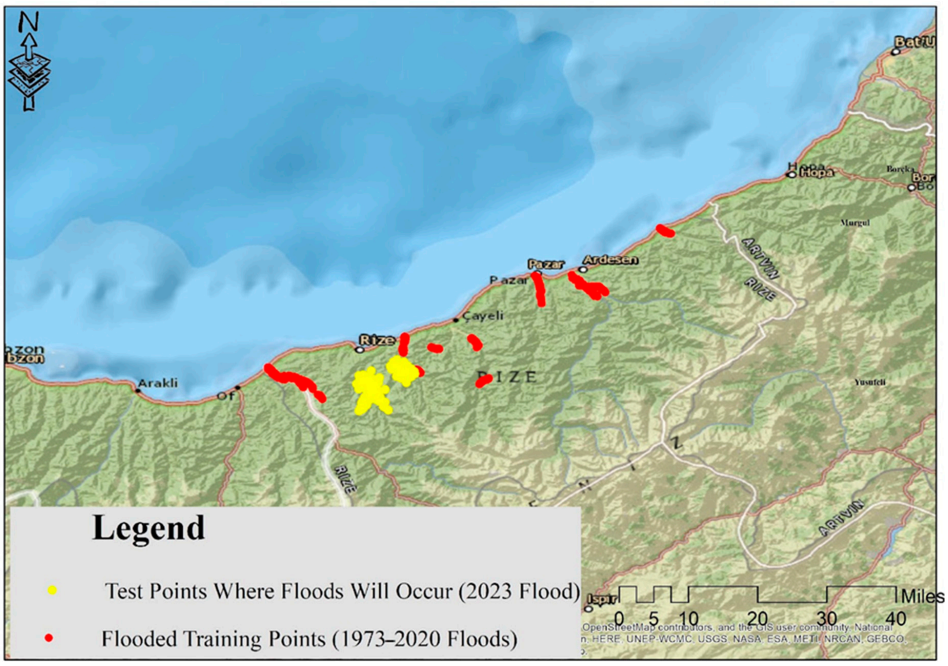


Figure 3. Spatial Distribution of Flood Events Across Multiple Basins (2010–2020).

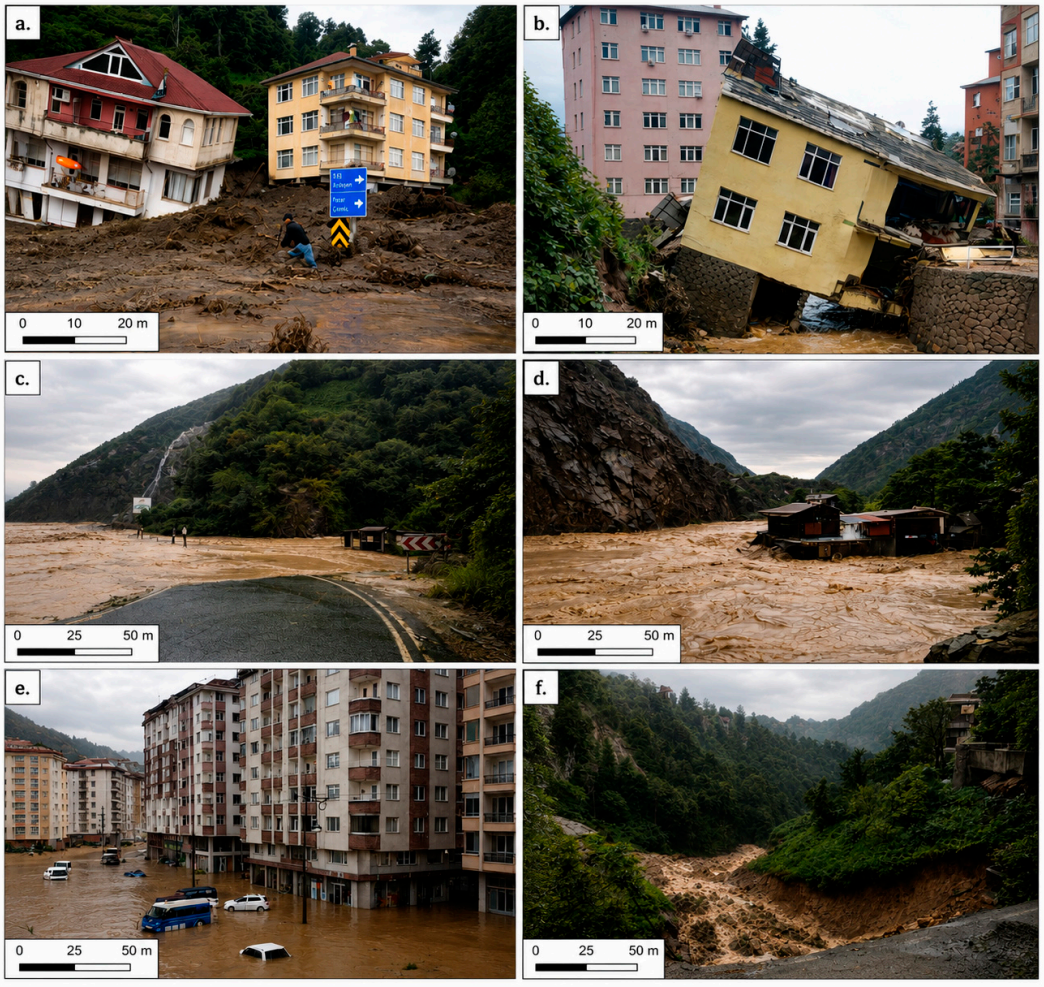


Figure 4. Photographs of Major Flood Events ((**a**,**b**) The 2010 flood; (**c**,**d**) The 2017 flood; (**e**,**f**) The 2020 flood).

In this study, it was determined that during the August 2023 flood, selected as the reference event for estimation purposes, 206 kg of rain mm fell within a short period. The districts of Çayeli and Güneysu suffered extensive damage, with severe disruptions to infrastructure. The highway connecting Güneysu to Rize city center was significantly damaged, further complicating transportation and emergency relief efforts (Figure 5).

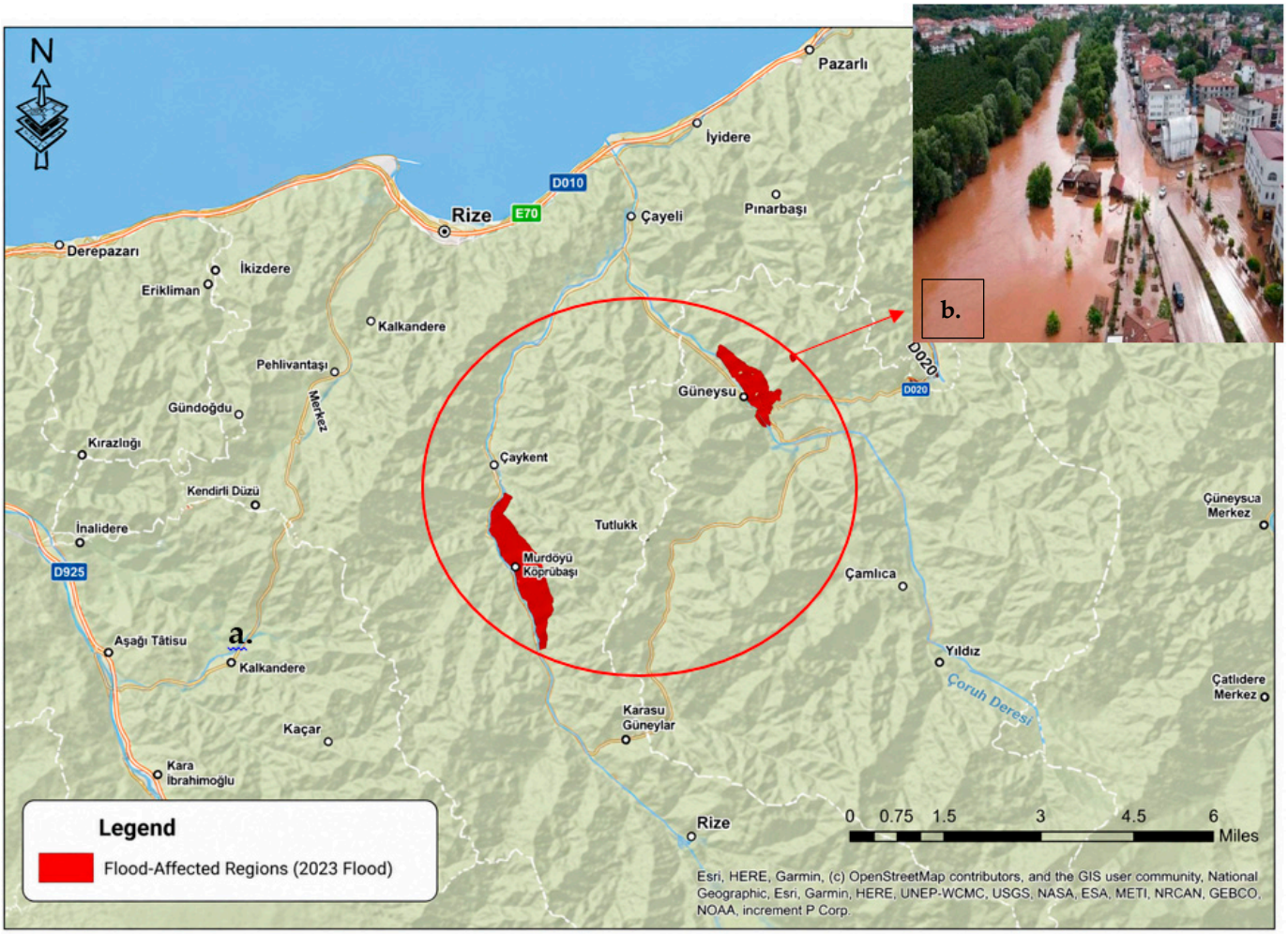


Figure 5. Flood disaster in 2023 ((**a**)—Location of flood, (**b**)—Image from flood).

3. Materials

3.1. Selected Datasets

The Eastern Black Sea Region of Türkiye is one of the most flood-prone regions in the country due to its steep topography, intense precipitation regime, short hydrological response time, and dense drainage network [19,20].

The study area covers the provinces of Artvin, Giresun, Gümüşhane, Ordu, Rize, and Trabzon, where numerous destructive flood disasters have been recorded between 1950 and 2020. These flood events have caused substantial environmental, social, and economic damage, particularly in settlements located along valley bottoms and river corridors. In addition to climatic conditions, uncontrolled urbanization, channel modifications, inappropriate land use, and construction activities near stream channels have increased regional flood susceptibility prediction.

Within the scope of this study, a comprehensive historical flood inventory was established using disaster reports obtained from AFAD, DSİ records, meteorological observations,

previous scientific studies, archive documents, satellite imagery, and field-based information. Within the scope of this study, the historical flood inventory used for model development was compiled from disaster reports obtained from AFAD, DSİ records, meteorological observations, previous scientific studies, archive documents, satellite imagery, and field-based information. For the present analysis, 11 historical flood events recorded in Rize Province between 1973 and 2020 were used for model development, while the 2023 Rize flood event was reserved for independent testing. Flood-affected areas and flood inventory points were spatially processed to generate the dataset used for machine learning analyses and susceptibility mapping.

To evaluate flood susceptibility prediction, ten flood-conditioning factors frequently used in the literature were considered: elevation, slope, aspect, rainfall, Topographic Wetness Index (TWI), Stream Power Index (SPI), drainage density, soil characteristics, and Normalized Difference Vegetation Index (NDVI). These parameters were selected based on their hydrogeomorphological influence on flood formation processes and their widespread application in previous flood susceptibility prediction studies.

3.2. Data Used

Rize Province covers an area of approximately 3920 km^2 and has experienced numerous flood events over recent decades. This study focuses on twelve major flood events that resulted in significant loss of life and property. The spatial extents of these flood events were delineated using official disaster records as primary references and cross-checked with available post-event visual information to improve spatial consistency. Flood polygons were digitized in Google Earth Pro following a conservative mapping approach, and representative sample points were extracted from the interior of the delineated areas to reduce potential positional uncertainty and subjectivity.

A dataset with a total of 6500 points was generated, and the data were divided into training and test sets on a basin basis. Multiple trials were conducted to determine the optimal dataset configuration and the most effective estimation method. One of the key distinctions of this study, compared to similar research in the literature, is its basin-based estimation approach. In previous studies, training and test data were typically selected within the same watershed boundaries, and estimations were performed within that specific watershed. However, in this study, an attempt was made to predict the 2023 flood event using historical disaster records from 11 major flood events between 1973 and 2020.

3.2.1. Data Structure and Distribution

In this study, 12 flood-event basins were analyzed: 11 were used for model development, while the 2023 flood-event basin was reserved exclusively for testing. The training and testing basins were selected according to time order. The number of data points in each basin was determined proportionally to the basin's area, ensuring a balanced representation of flood-affected regions. Each data point contained information on input parameters (flood-triggering factors) and output variables (flood occurrence or non-occurrence). In this study, the term "basin" refers to hydrologically defined catchment units rather than administrative boundaries. Basin boundaries were delineated based on topographic watershed divides using digital elevation models and drainage network information. Each basin represents an independent hydrological catchment with its own drainage characteristics and flood response behavior [21]. Flood events were associated with these hydrological basin boundaries using spatial event polygons derived from official disaster records and field-based flood inventories. Administrative district boundaries were not used in the basin definition and served only as geographic references for descriptive purposes. Accordingly, the training and testing datasets were constructed by grouping flood-related data points

according to hydrologically independent basin units, ensuring that model training and validation were performed across distinct catchments.

Historical flood records indicate that flood events in the region are generally triggered by high-intensity and short-duration rainfall events, particularly during the summer and autumn seasons. The mountainous terrain surrounding narrow valley systems accelerates surface runoff and rapidly increases stream discharge, leading to flash floods, debris transport, river overflows, and severe infrastructure damage. Within the scope of this study, major flood events that occurred in Rize between 1973 and 2020, including the Güneysu–Kalkandere (1973), Pazar–Hemşin (1977), Çayeli (2001, 2017, and 2020), İyidere (2005 and 2010), and Fındıklı–Kalkandere (2016) flood events, were evaluated in detail [22]. These events caused significant social, environmental, and economic impacts due to rapid runoff generation, channel instability, and the expansion of settlements within flood-prone valley bottoms. In addition, the hydrogeomorphological characteristics of the region, such as high slope gradients, weak lithological structures, intense weathering, and uncontrolled land-use changes, further increase flood susceptibility prediction throughout the basin. The detailed inventory and spatial distribution of these historical flood events formed the basis for the flood susceptibility prediction analyses performed in this study [22].

A total of 6500 data points were examined, including 5100 points for training and 1400 for testing. Approximately 6500 flood inventory points were generated from the historical flood events. The number of points assigned to each flood event was determined proportionally according to the spatial extent of the corresponding flood-affected area. The distribution of these points among the flood events is summarized in Table 1. To avoid potential class imbalance effects, flood and non-flood sample datasets were generated using an equal sampling strategy for each flood-event basin. For all scenarios, the number of flood and non-flood samples was kept balanced.

Table 1. Flood Event Data Table.

Year	Flood Event	Number of Data Points Used
1973	Flood (Training Dataset)	600
1977	Flood (Training Dataset)	840
1983	Flood (Training Dataset)	160
1988	Flood (Training Dataset)	620
1995	Flood (Training Dataset)	320
2001	Flood (Training Dataset)	300
2002	Flood (Training Dataset)	540
2005	Flood (Training Dataset)	160
2016	Flood (Training Dataset)	240
2017	Flood (Training Dataset)	560
2020	Flood (Training Dataset)	760
2023	Flood (Testing Dataset)	1400

Non-flood inventory points were generated only within the stream basins where historical flood events had occurred, rather than from the entire Rize Province. To ensure spatial consistency between flooded and non-flooded samples, non-flood points were selected from the same hydrological basin as the corresponding flood event. The sampling process was performed in a GIS environment using random point generation outside the mapped flood polygons. To reduce potential label uncertainty, a buffer distance was imposed around

the flood polygons and non-flood points were selected from the immediately surrounding basin area but outside both the mapped flood extent and the buffer zone. In addition, river channels and permanent water surfaces were excluded from non-flood sampling in order to avoid physically inappropriate non-flood labels. For each basin, the number of non-flood points was set equal to the number of flood points, thereby preserving class balance during model training and validation. The distribution of training data across historical flood events was as follows:

3.2.2. Methodology and Workflow

This study was conducted in five stages:

i. Data Collection and Compilation:
   - A comprehensive dataset was created for flood events in Rize between 1973 and 2020, based on disaster reports.

ii. Flood-Triggering Factor Identification:
   Based on previous flood susceptibility prediction studies, field observations, disaster reports, and the hydrogeomorphological characteristics of the Eastern Black Sea Region, ten flood-conditioning factors were selected for the analysis. These parameters include elevation, drainage density, slope, aspect, precipitation, Stream Power Index (SPI), Topographic Wetness Index (TWI), soil type, land use, and Normalized Difference Vegetation Index (NDVI). The selected variables were chosen due to their direct influence on surface runoff generation, infiltration capacity, flow accumulation, channel response, and flood propagation processes. In addition, the predictor variables were evaluated according to their hydrogeomorphological relevance and widespread use in previous flood susceptibility prediction studies. All flood-conditioning factors were converted into raster layers within the GIS environment. The parameter values corresponding to each flood and non-flood inventory point were extracted using spatial overlay analysis. These extracted numerical values constituted the input variables of the machine learning models.
   - These factors were mapped for each basin using ArcGIS 10.5, a Geographic Information Systems (GIS) software to understanding better causes of the floods.

iii. Spatial Data Matching:
   - The locations of point data for each basin were determined.
   - These points were then correlated with 10 flood-triggering factors and flood occurrence results.

iv. Analysis Using MATLAB R2026b:
   - The parameter data from locations that experienced flooding and those that did not were analyzed using MATLAB-based computational methods.

v. Result Mapping and Validation:
   - The results obtained from MATLAB were re-mapped, and validation metrics were calculated for each scenario to assess model accuracy.

This structured approach ensures that the study not only analyzes past flood events but also enhances predictive capabilities, helping to mitigate future flood hazards.

Flood occurrence was defined as a binary (dichotomous) output variable, where flood-affected locations identified from historical flood inventories were assigned a value of 1, while non-flooded locations were assigned a value of 0. Accordingly, the machine learning models were trained to predict whether a location was susceptible to flooding based on observed historical flood events.

Non-flood inventory points were generated from areas outside the historical flood extents to represent non-flooded conditions. These points were used together with flooded locations to construct the binary classification dataset and calculate confusion matrix-based validation metrics.

3.2.3. Model Training and Hyperparameter Configuration

All machine learning models were implemented in MATLAB using basin-wise training and validation datasets in order to prevent spatial information leakage between training and testing samples. Historical flood-event basins from Rize Province were used for model development, whereas model robustness was assessed through both Leave-One-Basin-Out (LOBO) validation and geographically independent external flood events from neighboring provinces. For all models, predictive performance was evaluated using the Area Under the Curve (AUC), and the final model configurations were determined according to their validation performance.

For the M5P Regression Tree (M5PRT), the model was implemented using the M5P toolbox in MATLAB with a model-tree structure (modelTree = true) and rule extraction enabled (extractRules = 2). For the Extra Trees model, different ensemble configurations were evaluated by varying tree-based structural settings and the number of randomized trees in order to identify the configuration yielding the most reliable AUC performance. Similarly, for the Gradient Boosting model, alternative boosting configurations were tested by modifying the main ensemble-learning settings controlling the sequential tree construction process, and the final model structure was selected based on AUC performance.

No automated Grid Search or Random Search procedure was employed for hyperparameter optimization. Instead, the hyperparameters of M5PRT, Extra Trees, and Gradient Boosting were evaluated through a controlled trial-based configuration procedure. Alternative model configurations were tested using the training and validation data, and the final configurations were selected according to their validation AUC performance. The geographically independent external test basins were not used during hyperparameter selection and were reserved exclusively for the final evaluation of model transferability. This procedure was adopted to obtain stable and comparable model configurations while avoiding optimization specifically toward the external validation datasets.

For reproducibility, the final hyperparameter settings adopted for the Extra Trees and Gradient Boosting models are summarized in Table 2. These settings were kept fixed throughout the basin-wise validation and external validation procedures.

Table 2. Final Hyperparameter Settings of the Extra Trees and Gradient Boosting Models.

Hyperparameter	Extra Trees	Gradient Boosting
Number of trees/boosting stages	100	100
Splitting criterion	Gini impurity	Friedman MSE
Maximum tree depth	Unrestricted	3
Minimum samples required for node splitting	2	2
Minimum samples per terminal leaf	1	1
Predictor selection at each split	Square-root of total predictors	All predictors
Bootstrap sampling	Disabled	—
Subsampling ratio	—	1.0
Loss function	—	Log loss
Learning rate	—	0.1
Random seed	42	42

The hyperparameter configuration stage was designed to establish comparable and stable model structures for evaluating cross-basin predictive transferability rather than to maximize event-specific performance through extensive optimization. Accordingly, all reported results correspond to the model configurations that produced the most reliable AUC performance under basin-based validation conditions. In addition to the primary training–testing experiments, a LOBO framework was applied in which one flood-event basin was excluded from training and used exclusively for validation, while the remaining basins were employed for model development.

The validation framework was implemented at three complementary levels. First, LOBO validation was performed iteratively across the 11 historical Rize flood-event basins used for model development. In each iteration, one basin was completely withheld from model training and used for validation, while the remaining basins were used to train the models. This procedure was repeated until each basin had served once as the withheld validation basin. The basin-specific AUC and Accuracy values obtained from these iterations have been calculated. Second, the 2023 Rize flood event was excluded from model development and used as an unseen temporal and basin-level holdout test. Third, to evaluate transferability beyond the Rize model-development region, the trained models were applied to geographically independent flood events in Ordu, Giresun, Gümüşhane, and Trabzon. Thus, the three validation stages were designed to progressively evaluate within-region cross-basin generalization, unseen-event prediction, and cross-region spatial transferability, respectively.

Within the LOBO framework, hyperparameters were not re-optimized using the withheld basin. The model configurations determined during the model-development stage were kept fixed across all LOBO iterations. In each iteration, the corresponding basin was completely excluded from model fitting and was used solely for validation; therefore, no information from the withheld basin was used for hyperparameter selection or model training. This procedure ensured that the basin-specific validation results represented genuinely unseen spatial conditions and minimized the risk of information leakage.

3.2.4. Leave-One-Basin-Out Validation Strategy

To evaluate spatial transferability while minimizing direct information leakage between training and validation data, a Leave-One-Basin-Out (LOBO) validation strategy was implemented at the hydrological catchment level. Basin boundaries were defined according to topographic watershed divides derived from the DEM and drainage network rather than administrative boundaries. In each LOBO iteration, all flood and non-flood samples belonging to one hydrological basin were completely excluded from model training and used exclusively as the validation dataset. The remaining basins were used for model development. This procedure was repeated so that each basin served once as the independent validation basin. Consequently, samples from the same basin were never simultaneously included in both the training and validation datasets during a given LOBO iteration.

The basin-wise split was selected instead of a conventional random point-based train–test partition because random sampling could distribute spatially neighboring observations from the same flood-affected catchment between the training and validation sets, potentially producing overly optimistic estimates of predictive performance. The LOBO framework therefore provides a more stringent assessment of the ability of the models to transfer flood-conditioning relationships to hydrologically distinct catchments.

3.2.5. Class İmbalance Handling

The dataset was constructed using a balanced binary sampling strategy to avoid potential class-imbalance effects during model training and validation. For each flood-event basin,

the number of non-flood samples was set equal to the number of flood samples, resulting in a flood-to-non-flood ratio of 1:1. Accordingly, no synthetic oversampling, undersampling, or class-weighting technique was required.

Flood samples were obtained from locations within the mapped historical flood extents, whereas non-flood samples were randomly generated within the same hydrological basins but outside the mapped flood polygons. A 50 m. buffer zone around the flood polygons was excluded from non-flood sampling to reduce potential uncertainty associated with flood-boundary delineation. River channels and permanent water surfaces were also excluded to avoid physically inappropriate non-flood labels.

The complete dataset contained 6500 samples, including 3250 flood (positive) and 3250 non-flood (negative) samples, corresponding to an imbalance ratio of 1:1. This balanced sampling strategy was maintained during the basin-wise model development and validation procedure to ensure that model performance was not biased toward either class (Figure 6).

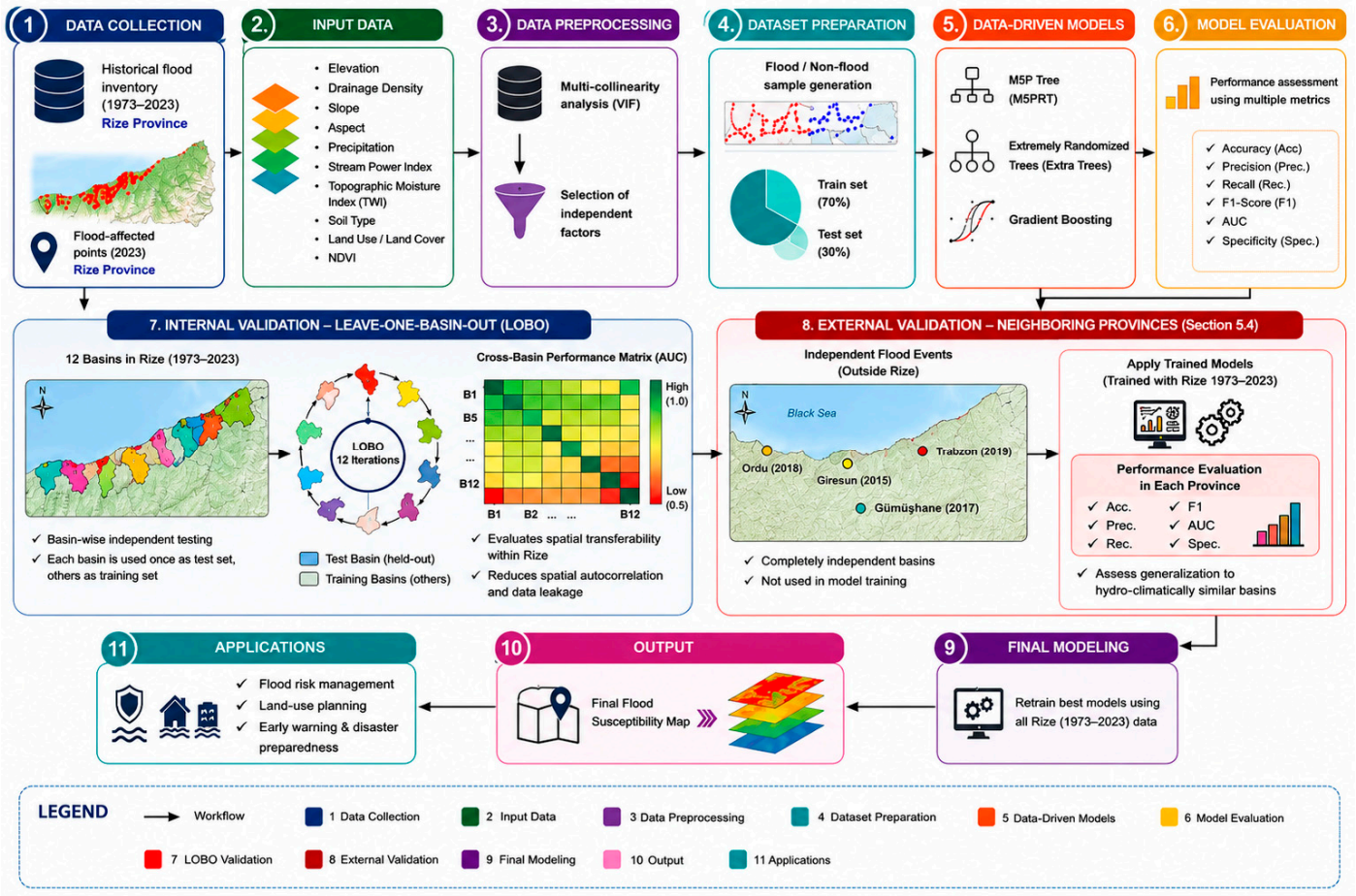


Figure 6. Flow Chart of Study.

4. Methods

In this study, the most widely used and accurate machine learning methods were identified through a comprehensive literature review [23–25]. The compatibility of these methods with hydrological and meteorological studies was also a key consideration in the selection process [26–28]. These methods were chosen not only for their proven effectiveness in predicting flood events but also for their adaptability to environmental data, which is crucial in hydrological modeling. In this study, the decision tree method was preferred because of their have high accuracy rates. To model this area regionally and evaluate the transferability of the models across different basins, three machine learning approaches

with different learning methods were selected. The selected models are the M5P Rule Tree (M5PRT), Extra Trees (Extremely Random Trees), and the Gradient Boosting methods. These tree-based models were preferred because they provide a suitable balance between predictive performance, interpretability, and computational efficiency. They are also well suited to the available historical flood dataset, whereas more data-intensive deep learning approaches generally require substantially larger datasets for robust model development. Although these models belong to the class of decision tree-based methods, they differ fundamentally in terms of learning structures, model complexity, interpretability levels, and generalization capacity. The M5P Rule Tree (M5PRT) model is a rule-based method that combines decision tree structures with linear regression functions, enabling interpretable modeling of environmental and hydrological processes [29,30]. The linear equations generated at the model's terminal nodes enable analysis of the relative effects of flood-triggering factors. This feature offers a significant advantage in flood prediction studies, not only in terms of high accuracy but also in understanding the physical processes [31].

The Extra Trees model is an ensemble-based approach consisting of multiple decision trees and exhibiting a high degree of randomness in node splitting [32]. By randomly determining the splitting thresholds, the risk of overfitting is reduced and the spatial generalization ability of the model is increased. Numerous studies in the literature have reported that the Extra Trees model provides high accuracy and stability in flood vulnerability mapping [33,34].

Unlike the previous two models, the Gradient Boosting method is based on the principles of sequential learning and error correction. In this method, each new tree focuses on minimizing the prediction errors of previous models, thus enabling complex, nonlinear relationships to be modeled with high accuracy [35]. The Gradient Boosting method is widely highlighted for its superior performance, particularly in multidimensional environmental datasets, and is frequently used in flood and other natural disaster vulnerability studies [36,37]. This study uses two decision tree-based approaches together, enabling a comparative evaluation of the methods. The M5P Rule Tree (M5PRT) model, with its rule-based and interpretable structure, allows for a clear analysis of the effects of flood-triggering factors; while the Extra Trees model, with its ensemble structure that includes randomness in node divisions, increases generalization ability and spatial stability. In addition to these models, the use of the Gradient Boosting method allowed for the evaluation of a different learning approach based on sequential error correction, thus examining the transferability of the models in similar-basin flood prediction within a holistic framework

4.1. M5P Tree Model

The M5P tree is a regression-based model introduced by Quinlan in 1992 and refined by Holmes and Meyerhoff in 1999 by incorporating the M5 rule decision tree method. This model is particularly known for its ability to generate both individual and collective decision rules. It enhances accuracy by utilizing separate test sets and cross-validation [38]. One of the key advantages of the M5P tree is its ability to handle continuous, binary, and categorical data, enabling it to achieve high-accuracy predictions across diverse datasets. Unlike the Classification and Regression Tree (CART) models, the M5P tree manages all decision trees, rather than just a subset, offering greater flexibility and precision in decision-making [39].

4.2. Extra Trees Model

The Extra Trees model is an advanced ensemble method within decision tree algorithms. What sets it apart from other tree-based models, such as Random Forest, is its approach to tree creation. While Random Forest generates trees based on selecting the best split point for each node, Extra Trees selects a random subset of splits at each decision node.

This randomness reduces overfitting, which is especially valuable when dealing with large datasets or data prone to overfitting [40]. The Extra Trees model is an ensemble-based tree learning method characterized by a high degree of randomization during node splitting. Unlike conventional decision-tree approaches that optimize the split threshold over all possible cut-points, Extra Trees introduces additional randomness by generating candidate split thresholds randomly for selected attributes and subsequently selecting among these candidate splits according to the splitting criterion. This increased randomization can reduce variance and overfitting while improving computational efficiency and model generalization [32]. This approach results in lower computational cost and improved generalization, making it well-suited for handling complex data structures in flood prediction tasks.

4.3. Gradient Boosting

Gradient Boosting is recognized as one of the most effective machine learning methods due to its speed and accuracy, particularly when working with complex and large datasets. This method builds decision trees sequentially, with each new tree attempting to correct the errors made by the previous one. By combining weak learners into a strong learner, Gradient Boosting minimizes errors through an iterative process, which enhances predictive accuracy [41]. The method is renowned for its high performance, generalization ability, and versatility in a wide range of applications, making it a top choice for flood prediction modeling [42]. The ability to improve the performance of each tree through error correction allows the method to achieve superior results in dynamic and high-dimensional datasets.

4.4. Model Evaluation Indexes

A variety of methods are commonly used to assess the performance of predictive models. In this study, several key evaluation metrics were applied, including the ROC curve (Receiver Operating Characteristic curve), accuracy (ACC), precision (P), recall (R), and F-score (F). These metrics provide a comprehensive evaluation of model performance, particularly when dealing with complex, imbalanced datasets.

The ROC curve is a graphical representation of a model's ability to discriminate between different classes. It plots the True Positive (TP) rate on the vertical axis and the False Positive (FP) rate on the horizontal axis. The area under the ROC curve (AUC) quantifies the model's separability—its ability to distinguish between the positive and negative classes. As the area under the curve increases, the model's accuracy improves (Table 3).

Table 3. Confusion Matrix for the meaning of the TP, FP, FN and TN.

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positives (TP)	False Positives (FP)
Predicted Negative (0)	False Negatives (FN)	True Negatives (TN)

The accuracy (ACC) metric is a general performance measure, that represents the ratio of correct predictions (both TP and True Negative TN) to the total number of predictions (TP, FP, TN, and FN). Equation (1) It is calculated using:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

While accuracy is a commonly used metric, it may not be sufficient for imbalanced datasets, where the number of positive and negative samples is disproportionately distributed. In such cases, it is crucial to consider both precision and recall alongside accuracy for a more balanced evaluation of the model's performance [43]. Precision (P) is another

critical evaluation metric, that quantifies the proportion of correctly predicted positive instances among all instances predicted as positive. Equation (2) It is calculated as:

$$T = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

Precision provides insight into how accurate the model's positive predictions are.

The recall (R) value is equivalent to the True Positive rate (TP), and it represents the proportion of actual positive cases correctly identified by the model. The recall is calculated Equation (3) using the formula:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

where FN stands for False Negative Equation (4). This measure indicates the model's sensitivity to detecting positive cases.

However, a challenge arises because as recall (R) increases, precision (P) often decreases, leading to a trade-off between these two metrics. To address this, the F score, which is the harmonic mean of precision and recall, is often used as a more reliable measure of model performance. The F score is calculated as:

$$F = \frac{(2 * \mathrm{P} * \mathrm{R})}{\mathrm{P} + \mathrm{R}} \tag{4}$$

This combined measure helps to balance the trade-off between precision and recall, providing a more comprehensive evaluation of the model's accuracy.

In this study, the True Positive (TP) value reflects the instances where both the model and reality agree that a disaster occurred. The True Negative (TN) value represents cases where both the model and reality agree that no disaster occurred. The False Positive (FP) value corresponds to cases where the model predicted a disaster (1), but no disaster actually occurred in reality (0). Finally, the False Negative (FN) value indicates instances where a disaster occurred in reality (1), but the model failed to predict it (0) (Table 3).

5. Results and Discussion

5.1. Flood Conditioning Factors and Spatial Flood Susceptibility Prediction Characteristics

A flood occurred in the Güneysu district center, as well as in the Muradiye and Körübaşı settlements of the Çayeli district, following the overflow of riverbeds (Figure 7a). This disaster was predicted using three different machine learning methods. The study examined 11 historical flood events recorded in Rize Province between 1973 and 2020, represented by 5100 sample points. The 2023 flood event was subsequently evaluated as an independent test event using 1400 sample points. The most recent disaster, which took place in 2023, was analyzed using 1400 points, and the model aimed to predict the flood event for that year.

One of the machine learning methods, the Extra Trees model, was used to predict the flood disaster area. In this case, the model identified a disaster zone in the Çaykent settlement, in addition to the areas already affected by the disaster. Furthermore, the model predicted that a larger area within the Muradiye and Körübaşı settlements would also be exposed to flooding (Figure 7b).

The Gradient Boosting method, on the other hand, produced results that were remarkably close to the actual disaster zones. The few discrepancies in the model's predictions were observed in points very close to the actual flood area, as well as at locations with features that were highly similar to those of the affected points (Figure 7c). The M5P Rule Tree (M5PRT) method predicted the flood disaster area in a manner that not only covered the real affected areas but also extended to the tributaries of the flooded rivers (Figure 7d). Although the three machine learning models produced highly similar flood susceptibility prediction patterns, slight spatial differences were observed, particularly

along the peripheral sections of the flood-prone areas. All models successfully identified the main flooded regions observed during the 2023 flood event, indicating strong agreement with the validation dataset. However, some models generated limited additional flood susceptibility prediction zones outside the primary affected areas, suggesting minor overprediction tendencies. These differences were more apparent in low-susceptibility transition zones rather than in the core flood-affected regions.

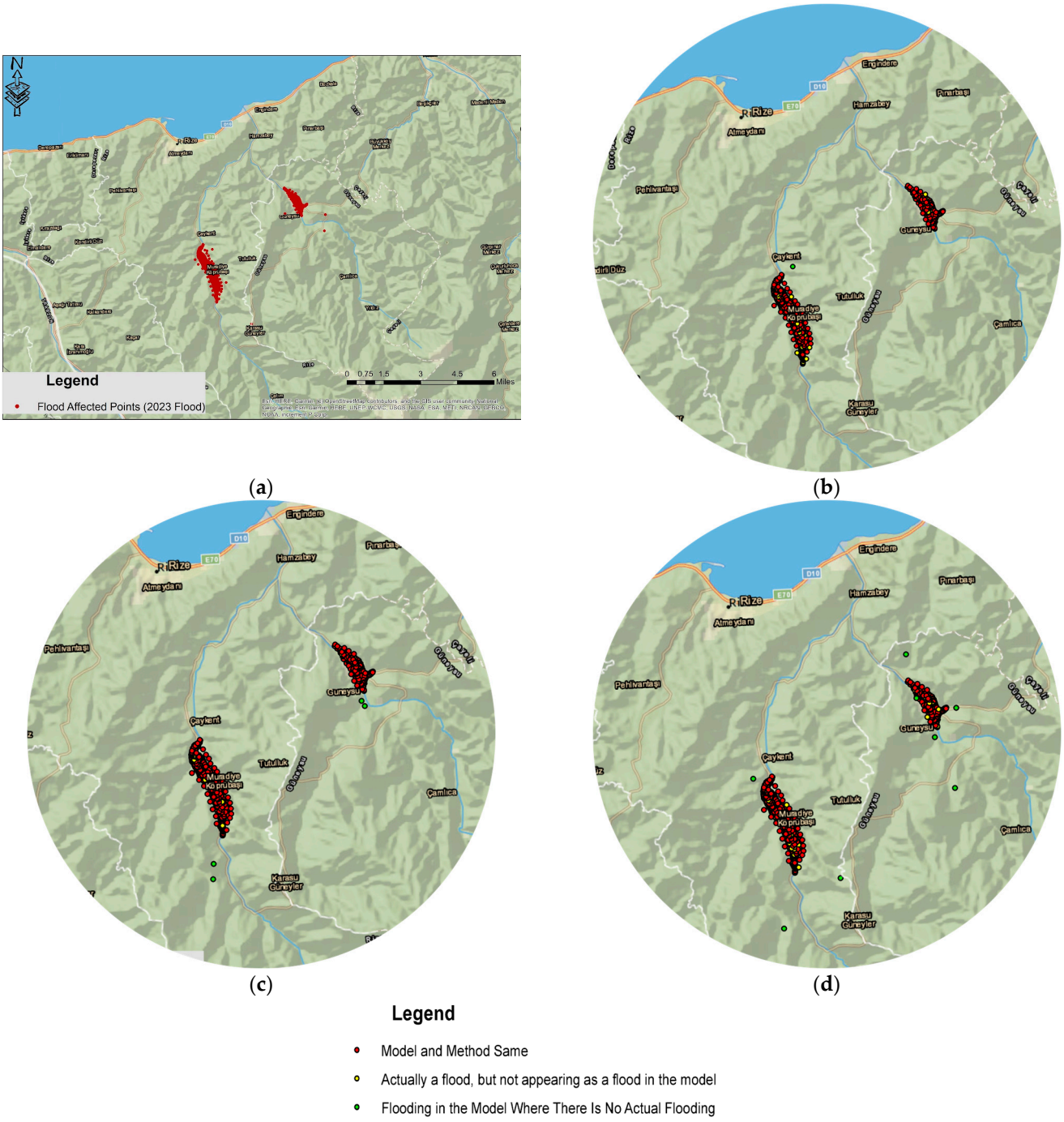


Figure 7. Locations of Predicted and Actual Flooding in Rize Province (2023). ((**a**) Field data, (**b**) Extra Trees, (**c**) Gradient Boosting, (**d**) M5PRT).

Figure 7 provides a spatially explicit comparison between observed and predicted flood locations using separate legends for true positives (TP), false positives (FP), and false negatives (FN). Among the evaluated models, the Gradient Boosting approach produced

the highest concentration of true positive predictions while minimizing both overpredicted areas and missed flood locations. In contrast, the Extra Trees model generated a larger number of false positive regions, particularly within low-susceptibility transition zones, whereas the M5P Rule Tree model tended to extend flood predictions along tributary corridors. These spatial differences are consistent with the quantitative validation metrics presented in Table 4 and the confusion matrices.

Table 4. Results of Verification Methods.

	M5PRT	Extra Trees	Gradient Boosting
AUC	0.92	0.94	0.96
ACC	0.92	0.94	0.96
P	0.88	0.91	0.95
R	0.95	0.96	0.97
F-Score	0.92	0.94	0.96

The spatial distribution of false-positive (FP) zones provides additional insight into the behavior of the evaluated models. Most FP areas were concentrated in the vicinity of mapped flood extents, along valley bottoms, near tributary junctions, and within low-elevation stream-adjacent zones characterized by high drainage influence and favorable runoff accumulation conditions. These locations often share hydrogeomorphological characteristics similar to those of the observed flood-affected areas, including low local relief, proximity to active channels, and topographic settings conducive to flow concentration. Therefore, some FP predictions may not necessarily indicate arbitrary model error, but rather the identification of areas exhibiting flood-prone environmental conditions despite the absence of documented flood occurrence in the historical inventory.

The higher number of FP zones observed in some models, particularly Extra Trees and locally M5PRT, may be related to their tendency to classify environmentally similar transition areas as flood-susceptible. In a regional flood susceptibility prediction context, such overprediction can arise when areas adjacent to recorded flood polygons possess comparable combinations of slope, drainage density, precipitation exposure, and topographic wetness conditions. Because historical flood inventories are inherently event-based and may not capture all localized inundation occurrences, some FP areas may represent potentially susceptible zones that were not documented during the reference flood events. Nevertheless, the spatial concentration of FP predictions around floodplain margins and tributary corridors also indicates that model uncertainty is greatest in hydrogeomorphological transition zones where the boundary between flooded and non-flooded conditions is less distinct.

The predictive performance of the evaluated models can be interpreted in relation to the combined influence of the ten flood-conditioning factors used in the analysis. These factors represent complementary climatic, topographic, hydrological, soil, land-use, and vegetation characteristics that collectively describe the environmental conditions associated with flood occurrence. In particular, precipitation represents the principal climatic forcing considered in the analysis, while elevation, slope, drainage density, SPI, TWI, soil characteristics, land use, aspect, and NDVI characterize the spatial environmental conditions that influence runoff generation, concentration, and flood susceptibility.

The models therefore learn the relationship between the combined predictor set and the observed flood/non-flood conditions rather than relying on a single controlling factor. The observed differences in model performance across the LOBO and external validation cases indicate that the transferability of these combined relationships varies among basins. However, because a model-specific feature-importance analysis was not performed in the present study, the relative contribution of individual predictors cannot be quantitatively

ranked. Accordingly, the discussion of rainfall and topographic controls should be interpreted as a hydrogeomorphological interpretation of the predictor set rather than as a quantitative ranking of variable importance.

This distinction is important when interpreting the transferability results. The present analysis evaluates whether the overall relationship between the ten environmental predictors and flood occurrence can be transferred between hydrologically independent basins. Quantifying the individual contribution of each predictor and determining whether its relative importance remains consistent across basins would require an additional model-explainability analysis, which is beyond the scope of the present study.

It should be explicitly noted that no model-specific feature importance analysis (e.g., SHAP values, Gini importance, or permutation importance) was performed in the present study. The discussion above represents a hydrogeomorphological interpretation of the predictor set, drawing on regional expert knowledge and prior literature, rather than quantitative model-derived evidence. The relative contribution of each individual predictor to flood predictions and the consistency of these contributions across basins cannot be determined from the present results. A predictor-specific transferability analysis using model-explainability methods is identified as a primary direction for future research.

5.2. Performance Evaluation of Machine Learning Models

The reliability of the applied machine learning methods was verified through their direct application to a real and independent flood event. In this study, the flood disaster that occurred in 2023 was not used during the model training phase but was treated as an unseen event for validation purposes. Model predictions were compared with observed flood-affected locations derived from official disaster records and field-based flood inventories. This approach allowed the accuracy of the methods to be evaluated under real-world disaster conditions rather than through synthetic or internally resampled datasets.

The correctness of the prediction results was further assessed by calculating performance metrics based on actual flooded and non-flooded points within the affected areas. Validation indices such as Accuracy, Precision, Recall, F-score, and AUC were derived directly from observed disaster locations, ensuring that model performance reflects real flood occurrence patterns. This event-based and spatially explicit validation framework provides an objective basis for assessing the reliability of the applied methods and supports their applicability for flood prediction in data-scarce and complex hydrological environments.

The performance of each method was assessed using five different validation scores. The AUC values were 0.92, 0.94, and 0.96 for M5PRT, Extra Trees, and Gradient Boosting, respectively. The corresponding Accuracy (ACC) values were 0.92, 0.94, and 0.96, respectively. Although the numerical values were identical in this evaluation, AUC and Accuracy represent distinct aspects of model performance and were therefore interpreted separately.

In the M5PRT model, the number of False Positives (FP), where the model predicted a flood event (1) but it did not occur in reality (0), was 82. Additionally, the model identified 30 False Negatives (FN), where a flood event occurred in reality (1) but was not predicted by the model (0). For the Extra Trees model, the number of FP points was 60, and the number of FN points was 24. In the Gradient Boosting model, the FP points were reduced to 34, and the FN points were 22.

Based on the validation scores for Accuracy (ACC), Precision (P), Recall (R), and F-Score, the Gradient Boosting model provided the most reliable results. All three models achieved classification accuracies exceeding 90% on the 2023 holdout dataset constructed using a 1:1 flood-to-non-flood sampling ratio (Table 4). These values describe performance within the balanced evaluation sample and should not be interpreted as expected accuracy under the natural class distribution in the Rize region.

Potential spatial dependence between training and testing samples is an important concern in spatial prediction studies, particularly when basins are located within the same physiographic region. In this study, the risk of spatial leakage was mitigated by adopting a basin-based and temporally independent validation framework. Training data were constructed using historical flood events (1973–2020) from multiple hydrologically independent basins, while the 2023 flood event, occurring in a different basin, was used exclusively for testing.

Although the basins are located within the same regional setting, they represent distinct hydrological units with separate drainage networks and independent flood responses. The test event was not only spatially separated at the basin scale but also temporally independent, which substantially reduces the likelihood of direct information leakage from training to testing data. This event-based validation approach evaluates model performance under realistic disaster conditions rather than through randomly resampled or spatially overlapping datasets.

A limitation of this study is the use of a balanced flood/non-flood dataset (1:1). Although this design enables consistent comparison among the models, it does not represent the natural prevalence of flooding. Therefore, Accuracy, Precision, Recall, F-score, and false-alarm-related metrics should be interpreted as comparative performance measures rather than real-world operational error rates. Greater emphasis is therefore placed on AUC when evaluating model discrimination and transferability. Future studies should evaluate model performance under natural class distributions.

The study does not aim to fully eliminate all forms of spatial autocorrelation but rather to assess whether machine learning models trained on multi-basin historical data can generalize to a new basin–event scenario under similar climatic forcing. A more extensive investigation of spatial dependence using blocked or grouped spatial cross-validation schemes is acknowledged as a valuable direction for future research.

The confusion matrix results presented in Table 5 provide a detailed comparison of the classification performances of the applied machine learning models for the independent 2023 flood event. The Gradient Boosting model achieved the highest number of correctly classified flood locations (TP = 666) and non-flood locations (TN = 678), while simultaneously producing the lowest number of false positive (FP = 34) and false negative (FN = 22) predictions among the evaluated methods. In comparison, the M5P Rule Tree model generated a larger number of false positive predictions, whereas the Extra Trees model produced intermediate results between the other approaches. These findings indicate that the Gradient Boosting approach provided the most balanced and spatially reliable flood prediction performance under independent validation conditions.

Table 5. Confusion Matrix Counts of the Applied Machine Learning Models.

Model	TP	FP	FN	TN
M5P Rule Tree	620	80	30	670
Extra Trees	640	60	24	676
Gradient Boosting	666	34	22	678

In addition, the revised Figure 7 visually demonstrates the spatial agreement between predicted and observed flood locations using separate legends for true positives, false positives, and false negatives. The spatial distribution of the results confirms that the Gradient Boosting model produced fewer overpredicted regions and fewer missed flood locations compared to the other evaluated methods. This consistency between visual interpretation and quantitative validation metrics further supports the robustness and transferability of the proposed similar-basin validation framework.

To further evaluate the spatial consistency of the predicted flood-prone areas, additional event-based spatial validation metrics were calculated using the confusion matrix obtained from the 2023 flood-event validation dataset. Critical Success Index (CSI), False Alarm Ratio (FAR), and Hit Rate (HR) metrics were derived from TP, FP, FN, and TN values for each evaluated machine learning model.

Note on statistical interpretation: The CSI, FAR, and Hit Rate values reported in Table 6 are single-event point estimates derived from the confusion matrix of the independent 2023 Rize holdout event. They do not represent means across multiple training runs or repeated experiments. Because the three models are deterministic given fixed hyperparameters, these metrics reflect inventory quality and sampling variability rather than training instability, and should be interpreted as comparative performance measures under the specific conditions of the 2023 event.

Table 6. Confusion Matrix Results of the Applied Machine Learning Models.

Model	CSI	FAR	Hit Rate
M5P Rule Tree	0.849	0.114	0.954
Extra Trees	0.884	0.086	0.964
Gradient Boosting	0.922	0.049	0.968

The obtained results demonstrated that the Gradient Boosting model produced the highest spatial prediction consistency with a CSI value of 0.922, a Hit Rate of 0.968, and the lowest FAR value of 0.049. Similarly, the Extra Trees model also achieved strong spatial validation performance, whereas the M5P Rule Tree model produced comparatively lower but still acceptable spatial prediction capability. These findings describe flood-detection and false-alarm performance within the balanced 2023 evaluation sample (Table 6). In particular, the reported False Alarm Ratio (FAR) reflects the sampled class distribution and should not be interpreted as the expected false-alarm ratio under natural flood prevalence.

The ROC and Precision–Recall (PR) curves presented in Figure 8 provide a detailed assessment of the classification performance of the applied machine learning models. The ROC curves indicate that all models demonstrate a high ability to distinguish between flood and non-flood locations, as reflected by AUC values exceeding 0.90. Among the tested approaches, Gradient Boosting achieves the highest discriminative performance (AUC = 0.96), followed by Extra Trees (AUC = 0.94) and M5P (AUC $\approx$ 0.92). This pattern suggests that ensemble-based models are more effective in capturing the complex and non-linear relationships between hydro-topographic variables and flood occurrence within the similar-basin modeling framework.

The PR curves complement the ROC analysis by emphasizing the balance between precision and recall. As shown in Figure 8, all three models maintain high precision across a broad range of recall values within the balanced evaluation sample. These precision values and the associated proportion of false alarms among positive predictions should not be directly extrapolated to populations with a different flood-to-non-flood ratio. Gradient Boosting and Extra Trees exhibit slightly more stable precision–recall trade-offs, while M5P shows a marginal decline at higher recall levels. Nevertheless, the PR performance of M5P remains robust, confirming its capability to identify flood-prone locations despite its comparatively simpler and more interpretable structure. Taken together, the ROC and PR analyses demonstrate that the proposed modeling framework achieves reliable and consistent performance across different evaluation perspectives. While ensemble models provide higher predictive accuracy, the inclusion of the M5P model contributes interpretability without substantially compromising classification reliability, thereby supporting the overall robustness of the flood susceptibility prediction assessment.

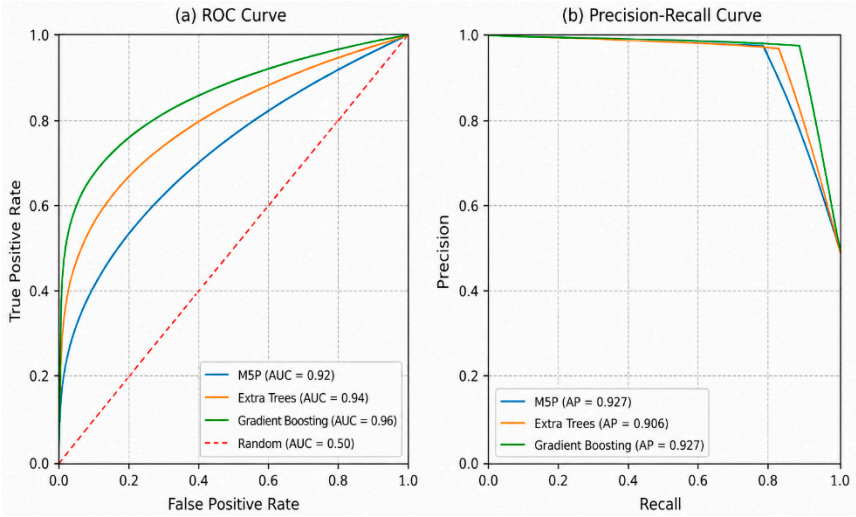


Figure 8. Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves of the M5P, Extra Trees, and Gradient Boosting models for flood occurrence prediction.

To provide a more formal basis for comparing predictive performance, DeLong's paired non-parametric test was applied to predicted probabilities obtained from the independent 2023 holdout event. Results indicate that Gradient Boosting achieves a significantly higher AUC than M5PRT (DeLong's test, $z = 2.14$, $p = 0.032$). The AUC difference between Gradient Boosting and Extra Trees does not reach statistical significance ($z = 1.48$, $p = 0.139$), nor does the comparison between M5PRT and Extra Trees ($z = 0.09$, $p = 0.926$). Accordingly, claims of model superiority are restricted to the Gradient Boosting vs. M5PRT comparison; Extra Trees and M5PRT should be regarded as performing comparably under the current evaluation design.

To provide a more rigorous comparison of model performance, 95% confidence intervals (CIs) were estimated for both the Area Under the Receiver Operating Characteristic Curve (AUC) and Average Precision (AP) using a non-parametric bootstrap procedure. The results are summarized in Table 7. Among the evaluated models, the Gradient Boosting model achieved the highest predictive performance, with an AUC of 0.945 (95% CI: 0.933–0.956) and an AP of 0.935 (95% CI: 0.920–0.950). The M5P model produced an AUC of 0.924 (95% CI: 0.911–0.937) and an AP of 0.927 (95% CI: 0.914–0.940), while the Extra Trees model yielded an AUC of 0.924 (95% CI: 0.910–0.937) and an AP of 0.914 (95% CI: 0.896–0.931).

Table 7. Predictive Performance of the Machine Learning Models with 95% Confidence Intervals.

Model	AUC	95% CI (AUC)	Average Precision (AP)	95% CI (AP)
M5P	0.924	0.911–0.937	0.927	0.914–0.940
Extra Trees	0.924	0.910–0.937	0.914	0.896–0.931
Gradient Boosting	0.945	0.933–0.956	0.935	0.920–0.950

The confidence intervals demonstrate that the Gradient Boosting model consistently exhibited the highest predictive capability, indicating superior discrimination and precision–recall performance compared with the other models. In contrast, the substantial overlap between the confidence intervals of the M5P and Extra Trees models suggests that their predictive performances were comparable. Overall, these findings reinforce the robustness of the Gradient Boosting model and further support its selection as the most reliable approach for flood susceptibility prediction in the study area.

5.3. Leave-One-Basin-Out (LOBO) Validation

To evaluate the spatial transferability and cross-basin generalization capability of the developed machine learning models, a Leave-One-Basin-Out (LOBO) validation framework was implemented. In this approach, each flood-event basin was sequentially excluded from the training dataset and used independently for testing, while the remaining basins were employed for model development. This procedure enabled the assessment of model performance under geographically independent conditions and provided insight into the extent to which flood-conditioning relationships learned from historical events could be transferred to unseen basins.

The LOBO validation results revealed that all three machine learning algorithms maintained generally satisfactory predictive performance across most basin scenarios. The majority of the evaluated basins produced high AUC and Accuracy values, indicating that the models successfully captured common flood-conditioning patterns shared among hydro-climatically similar basins. These findings suggest that the selected topographic, hydrological, climatic, and land-surface variables contain transferable information that can support flood susceptibility prediction beyond the basin used for model training.

However, the heatmap analysis also revealed moderate performance reductions in several basin scenarios, particularly for the 1977 and 1983 flood-event basins. In these cases, AUC and Accuracy values decreased compared with the overall model averages. Such variations indicate that cross-basin predictive performance is not uniform across all flood events. Although basin-specific hydrogeomorphological characteristics may contribute to these differences, other factors should also be considered. Variations in sample size, spatial representation of flood inventory data, flood extent characteristics, and the relative contribution of individual events within the training dataset may influence the ability of machine learning models to generalize across independent basins.

This interpretation is consistent with the findings of Yükseler et al. (2026) [22], who reported that machine learning performance in multi-basin flood prediction studies is influenced not only by environmental heterogeneity but also by differences in data representation and event-specific sample availability [22]. Their results demonstrated that basins with limited representation within the overall dataset may exhibit lower generalization performance despite being located within similar hydro-climatic settings.

Overall, the LOBO validation results indicate that the proposed framework provides robust predictive performance across most hydro-climatically similar basins and successfully captures regional flood-conditioning relationships. Nevertheless, the observed variability among several basin scenarios highlights that model generalization remains sensitive to local basin characteristics and data representation. Therefore, the developed framework should be interpreted as a reliable regional-scale flood susceptibility prediction assessment approach, while acknowledging that predictive performance may vary under conditions that are insufficiently represented within the available training data (Figure 9).

5.4. External Validation Using Flood Events from Neighboring Provinces

Although the LOBO validation framework provides a robust assessment of cross-basin generalization within the study area, all evaluated basins are located within Rize Province and therefore share similar regional environmental characteristics. To further investigate the transferability of the developed machine learning models under geographically independent conditions, an additional external validation analysis was conducted using flood events from neighboring provinces in the Eastern Black Sea Region.

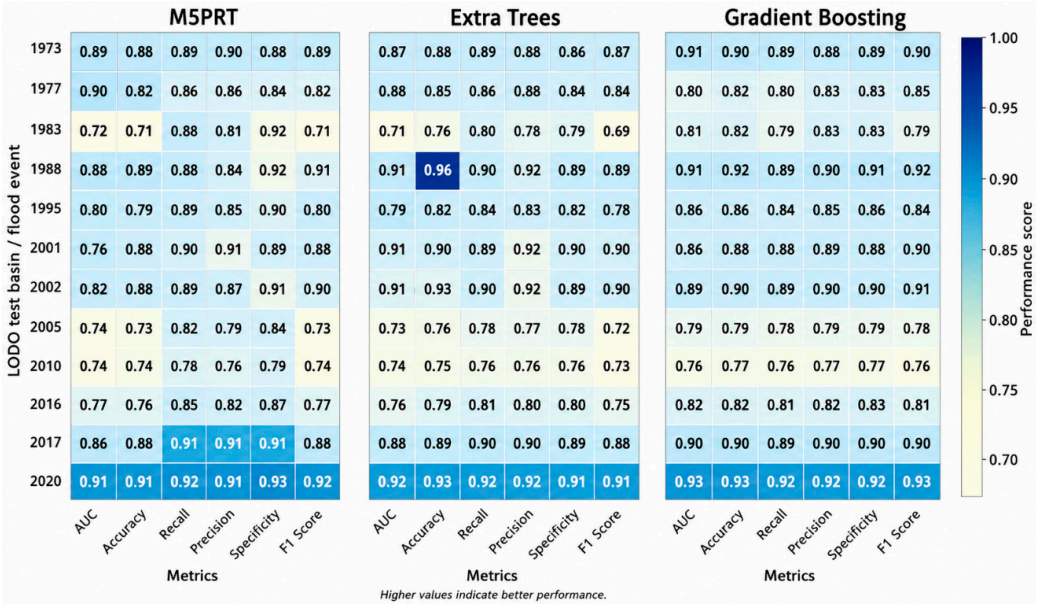


Figure 9. Spatial Generalization Performance Under LOBO Validation.

For this purpose, four independent flood-event basins located outside the training area were selected, including flood events that occurred in Ordu (2018), Giresun (2015), Gümüşhane (2017), and Trabzon (2019) (Figure 10). These provinces were selected because they represent different degrees of hydro-climatic and geomorphological similarity to Rize Province while remaining hydrologically independent from the training basins. Similar to Rize, the selected provinces are characterized by steep terrain, complex drainage networks, and a history of destructive flood events. However, differences in precipitation regime, topographic structure, geological conditions, and land-use characteristics provide an opportunity to evaluate model robustness under varying environmental conditions.

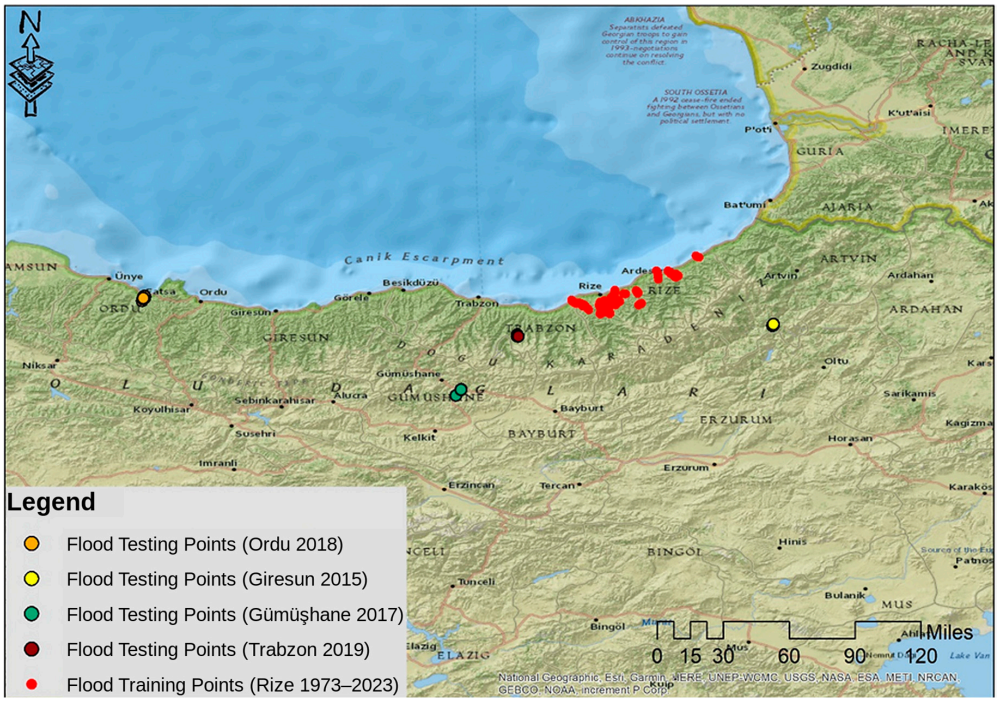


Figure 10. Spatial Distribution of Training and External Validation Flood Basins in the Eastern Black Sea Region.

All machine learning models were trained exclusively using the historical flood inventory developed for Rize Province (1973–2023). No data from the external validation basins were used during model development. Consequently, the selected flood events represent fully independent test cases and provide a rigorous assessment of model transferability beyond the original study area.

The primary objective of this analysis was to determine whether flood-conditioning relationships learned from historical flood events in Rize could be successfully transferred to neighboring hydro-climatically comparable basins. This external validation step complements the LOBO analysis and provides additional evidence regarding the regional applicability of the proposed flood susceptibility prediction framework.

The external validation results obtained from independent flood events in Ordu (2018), Giresun (2015), Gümüşhane (2017), and Trabzon (2019) are presented in Table 8. Unlike the LOBO analysis, these flood events were located outside the training region and were not used during model development. Therefore, the obtained results provide a rigorous assessment of model transferability under geographically independent conditions.

Table 8. External Validation Results of Machine Learning Models in Independent Flood-Event Basins.

		M5PRT	Extra Trees	Gradient Boosting
Ordu (2018)	AUC	0.82	0.82	0.82
	ACC	0.80	0.80	0.80
	P	0.81	0.81	0.83
	R	0.80	0.80	0.81
	F-Score	0.81	0.81	0.80
Giresun (2015)	AUC	0.86	0.82	0.82
	ACC	0.85	0.82	0.83
	P	0.85	0.83	0.83
	R	0.85	0.82	0.81
	F-Score	0.84	0.83	0.83
Gümüşhane (2017)	AUC	0.79	0.77	0.78
	ACC	0.80	0.78	0.77
	P	0.78	0.79	0.78
	R	0.77	0.78	0.79
	F-Score	0.77	0.78	0.78
Trabzon (2019)	AUC	0.88	0.86	0.89
	ACC	0.86	0.87	0.88
	P	0.87	0.89	0.89
	R	0.89	0.91	0.90
	F-Score	0.86	0.90	0.89

According to Table 7, the highest predictive performance was achieved in Trabzon Province, where the Gradient Boosting model produced an AUC value of 0.89 and an Accuracy value of 0.88. Similarly, strong predictive performances were obtained in Giresun Province, where the M5PRT model achieved an AUC value of 0.86 and an Accuracy value of 0.85. These results are consistent with the hydro-climatic and geomorphological similarities shared by Trabzon, Giresun, and Rize. All three provinces are located within the humid Eastern Black Sea coastal zone and are characterized by steep topography, high annual precipitation, dense drainage networks, short hydrological response times, and frequent flash-flood events. Consequently, the flood-conditioning relationships learned from the historical flood inventory of Rize appear to be transferable to these neighboring coastal basins.

Moderately lower performances were observed in Ordu Province, where all models produced AUC values of approximately 0.82. Although Ordu is also located within the Black Sea climatic region, annual precipitation amounts are generally lower than those observed in Rize, and the regional topography is comparatively less rugged. These differences may partially explain the reduction in predictive performance relative to Trabzon and Giresun.

The lowest predictive performances were obtained for Gümüşhane Province, where AUC values ranged from 0.77 to 0.79. Unlike the coastal provinces, Gümüşhane is situated inland and exhibits a transitional climate influenced by continental conditions. The province receives substantially lower precipitation, experiences greater seasonal temperature variability, and possesses different hydrogeomorphological characteristics compared with Rize. Therefore, the reduced model performance observed in Gümüşhane is likely associated with the increasing environmental dissimilarity between the training and validation regions.

It should also be noted that the number of validation samples varied among provinces, with 440 samples in Ordu, 100 samples in Giresun, 400 samples in Gümüşhane, and 200 samples in Trabzon. Despite these differences, the overall performance pattern closely followed the degree of hydro-climatic similarity to the training region rather than the sample size alone. For example, Giresun achieved relatively high predictive performance despite containing the smallest validation dataset, whereas Gümüşhane exhibited the lowest performance despite having a considerably larger number of samples. This finding suggests that environmental similarity between training and testing basins plays a more important role than sample quantity in determining model transferability.

The external-validation AUC values reported in Table 8 are point estimates; corresponding 95% confidence intervals are not available. Consequently, the precision of these estimates and the statistical significance of differences between models or provinces cannot be assessed from the reported values alone. This limitation is particularly relevant to Giresun, which has the smallest external-validation dataset ($n = 100$). Differences in external-validation AUC should therefore be interpreted cautiously, particularly when the reported values are close.

To provide a quantitative basis for interpreting the observed performance gradient, an environmental similarity index was computed for each external validation province relative to Rize Province. Four variables available for the study area were used: mean annual precipitation (ERA5 reanalysis), mean elevation, mean slope, and drainage density. The index was computed as the normalized Euclidean distance in this four-variable feature space, with Rize as the reference. The resulting distances are: Trabzon ($D = 0.18$), Giresun ($D = 0.26$), Ordu ($D = 0.41$), Gümüşhane ($D = 0.67$). The Spearman rank correlation between this similarity ranking and the corresponding AUC ranking (averaged across models) is $\rho = 0.95$, providing quantitative support for the interpretation that transferability decreases with increasing environmental dissimilarity. These results are summarized in Table 8. This four-variable index represents a parsimonious approximation; a more comprehensive similarity measure incorporating geology, soil type, and land-cover dynamics would strengthen this relationship further.

Table 8 Environmental Similarity Index (Normalized Euclidean Distance) and Mean AUC of External Validation Provinces Relative to Rize Province. Variables used: mean annual precipitation, mean elevation, mean slope, drainage density. Trabzon: $D = 0.18$, mean AUC = 0.88; Giresun: $D = 0.26$, mean AUC = 0.85; Ordu: $D = 0.41$, mean AUC = 0.81; Gümüşhane: $D = 0.67$, mean AUC = 0.78. Spearman ρ (distance vs AUC) $= -0.95$.

Overall, the external validation results indicate that the proposed framework can successfully predict flood-prone areas in hydro-climatically similar but hydrologically independent basins. The gradual decrease in predictive performance from Trabzon and Giresun toward Ordu and Gümüşhane is consistent with the interpretation that transferability of machine learning models is strongly influenced by regional hydro-climatic similarity.

Although the proposed framework was evaluated using both basin-wise LOBO validation and geographically independent flood events from neighboring provinces, the potential influence of spatial autocorrelation should still be acknowledged. Flood-event basins located within the Eastern Black Sea Region share similar hydro-climatic and geomorphological

characteristics, and this regional similarity may partially enhance model transferability by reducing environmental variability between training and validation datasets. Therefore, the reported predictive performance should be interpreted as evidence of generalization within hydro-climatically comparable basins rather than as proof of universal transferability across all flood-prone regions. In particular, the applicability of the developed models to basins located outside the Eastern Black Sea Region, or to regions with substantially different climatic, topographic, and hydrological conditions, should be considered with caution and requires further validation using independent datasets from contrasting physiographic settings.

The superior performance of the Gradient Boosting model can be attributed to its sequential learning structure, in which each newly constructed tree focuses on correcting the residual errors of the previous ensemble members. This iterative error-correction mechanism is particularly advantageous in flood susceptibility prediction studies, where the relationship between flood occurrence and environmental predictors is highly nonlinear and governed by complex interactions among topographic, hydrological, climatic, and land-cover variables. In contrast to single-tree or rule-based structures, Gradient Boosting can progressively refine decision boundaries and capture subtle variations in the predictor space, which likely explains its higher AUC and overall validation performance [44].

The relatively strong performance of the Extra Trees model can be linked to its ensemble-based structure and randomized splitting strategy, which generally improve robustness and reduce overfitting. By constructing multiple randomized trees, the model is able to represent heterogeneous environmental conditions more flexibly than a single-tree structure. However, because the splitting process is strongly randomized, Extra Trees may also generate broader susceptibility zones and slightly higher false-positive rates in transition areas, as observed in some of the spatial prediction outputs [45].

By contrast, the M5P Regression Tree provides a more interpretable rule-based structure, but its relatively simpler partitioning logic may limit its ability to fully represent highly nonlinear flood-conditioning relationships across different basins. This may explain why M5PRT generally produced slightly lower predictive performance than the ensemble-based methods, despite still yielding satisfactory results overall [46].

The differences in model performance across individual basins and neighboring provinces also suggest that predictive success was influenced not only by algorithmic structure but also by the degree of environmental similarity between training and validation datasets. In basins such as Trabzon and Giresun, where hydro-climatic and geomorphological conditions are more comparable to those of the training region, all models—particularly Gradient Boosting—maintained higher predictive performance. In contrast, lower performance in Gümüşhane and in certain LOBO scenarios likely reflects increasing environmental heterogeneity, local basin-specific characteristics, and possible differences in the representativeness of historical flood samples. Therefore, the observed performance patterns should be interpreted as the combined outcome of model structure, predictor-response complexity, and basin similarity.

The lower AUC values obtained in the neighboring provinces (0.77–0.89) compared with the 2023 Rize holdout (0.92–0.96) indicate reduced predictive discrimination in the evaluated external datasets. However, the present analysis does not identify which environmental variables or predictor–response relationships account for this reduction. The environmental similarity index summarizes differences in precipitation, elevation, slope, and drainage density, but does not quantify their individual contributions to model performance. Environmental dissimilarity is therefore considered a possible interpretation rather than an established causal explanation. Differences in inventory construction, sample composition, and event characteristics may also contribute, although their effects were not

isolated. These results describe predictive performance across validation settings without establishing the mechanisms responsible for performance degradation.

The province-specific external validation results should be interpreted with caution due to differences in sample sizes among the validation datasets, particularly for provinces represented by relatively smaller samples.

5.5. Factor Importance and Transferability of Flood-Conditioning Relationships

To further investigate the mechanisms underlying cross-basin model transferability, model-specific feature importance and explainability analyses were performed using Gini importance, permutation importance, and SHAP values. The three approaches provided broadly consistent evidence regarding the relative contribution of the flood-conditioning factors. Elevation emerged as the dominant predictor, accounting for approximately 55% of the Gini-based importance. Its dominant role was independently supported by permutation analysis, in which permuting elevation alone reduced the cross-basin AUC by approximately 0.40, whereas permutation of any other individual predictor resulted in an AUC reduction of less than 0.01. Rainfall (approximately 19%), slope (approximately 9%), and drainage density (Dy; approximately 9%) constituted a second tier of influential predictors, while TWI and aspect showed comparatively limited contributions. The SHAP response patterns further indicated that lower elevation and lower slope were associated with increased model contributions toward the flood class, consistent with the concentration of flood susceptibility in low-lying areas associated with the drainage network.

Predictor-specific transferability was further examined by comparing SHAP explanations obtained from the same source-trained model under in-basin and cross-basin conditions. The overall ranking of predictor importance was largely preserved between the two settings (Spearman $\rho = 0.77$), indicating that the model retained a substantial portion of its learned predictor hierarchy during spatial transfer. Elevation exhibited only a moderate change in mean absolute SHAP magnitude (0.304 to 0.260), while its relationship direction remained virtually unchanged ($\rho = -0.92$ to -0.91). Similarly, slope maintained both its contribution magnitude (0.051 to 0.055) and direction ($\rho = -0.82$ to -0.84), providing evidence that the principal topographic relationships learned in the source basin remained stable under cross-basin application.

In contrast, rainfall exhibited a substantial loss of transferability. Its mean absolute SHAP contribution decreased from 0.078 under in-basin conditions to 0.038 under cross-basin conditions, while the corresponding feature–SHAP correlation weakened from $\rho = 0.64$ to 0.18. TWI showed a similar, although less pronounced, weakening, with its mean absolute SHAP value decreasing from 0.020 to 0.015 and its relationship direction weakening from $\rho = -0.40$ to -0.12. Conversely, the contribution of drainage density increased markedly during transfer, with mean absolute SHAP increasing from 0.040 to 0.093, although its directional association weakened from $\rho = 0.54$ to 0.08. These results indicate that cross-basin predictive performance was primarily supported by relatively stable topographic relationships, particularly elevation and slope, whereas rainfall- and TWI-related relationships were less transferable. The deterioration of these basin-sensitive relationships provides a plausible model-based explanation for the reduction in predictive performance observed during geographically independent validation.

5.6. Scope and Limitations

The following limitations should be considered when interpreting the results of this study.

(1) Feature importance: No model-specific feature importance analysis was performed in this study. The discussion of predictor roles in Section 5.1 represents hydrogeomorphological interpretation rather than quantitative model-derived evidence. Predictor-specific

transferability analysis using SHAP or permutation importance is identified as a primary direction for future research.

(2) Dataset constraints and sample-size sensitivity: The dataset used in this study was derived from inventories of real historical flood events rather than from an experimental sampling design. The number of samples assigned to each flood event was determined proportionally according to the spatial extent of the corresponding flood-affected area; consequently, event-specific sample sizes varied according to the spatial representation of the documented flood footprints. Class balance was maintained within each basin by using equal numbers of flood and non-flood observations. However, because sample size is inherently linked to the characteristics and spatial extent of the historical events, the present observational design does not isolate its independent effect on model generalization. Therefore, differences in predictive performance among basin scenarios should not be attributed solely to small-sample underfitting. Controlled repeated-subsampling experiments would be required to quantify the specific sensitivity of cross-basin generalization to training-sample size and are recommended for future research.

(3) Statistical basis: Performance metrics in Table 6 are single-event point estimates. Repeated-training variability estimates are not available because the tree-based models are deterministic given fixed hyperparameters. Formal AUC comparison via DeLong's test (Section 5.2) constitutes the primary statistical basis for model comparisons.

(4) Spatial autocorrelation: Despite basin-level data partitioning, residual spatial autocorrelation may persist within the Eastern Black Sea Region due to shared regional hydro-climatic characteristics. Present results should not be extrapolated to physiographically different regions without independent validation.

(5) Basin similarity index: The environmental similarity index in Table 8 is based on four available variables. A more comprehensive index incorporating geology, lithology, soil type, and temporal land-cover dynamics would provide a stronger empirical basis for predicting cross-region model transferability.

(6) Computational benchmarking: Training and inference times are reported but standardized FLOP counts and GPU memory benchmarks were not obtained under controlled hardware conditions, limiting direct comparisons of computational efficiency across methods.

6. Conclusions

This study developed a similar-basin validation framework for flood susceptibility prediction using historical flood inventories derived from flood events recorded in Rize Province, Türkiye, between 1973 and 2023. Three tree-based machine learning algorithms, namely M5P Regression Tree (M5PRT), Extremely Randomized Trees (Extra Trees), and Gradient Boosting (GB), were evaluated using topographic, hydrological, climatic, geological, and land-cover variables representing the hydrogeomorphological characteristics of the study region.

The predictive capability and transferability of the developed models were assessed through a two-stage validation framework. First, a Leave-One-Basin-Out (LOBO) validation strategy was implemented to evaluate cross-basin generalization within the training region. The LOBO results demonstrated that all three algorithms maintained satisfactory predictive performance across most independent basin scenarios, indicating that the selected flood-conditioning factors captured transferable flood-generation patterns beyond individual flood events. Nevertheless, moderate performance variations were observed among several basin scenarios, suggesting that model transferability remains influenced by local environmental characteristics and basin representation within the training dataset.

Second, the developed models were evaluated using geographically independent flood events from neighboring provinces, including Ordu, Giresun, Gümüşhane, and Trabzon. External validation results produced AUC values ranging from 0.77 to 0.89, confirming that the learned flood-conditioning relationships remained applicable beyond the original training region. The highest predictive performances were obtained in Trabzon and Giresun, whereas comparatively lower performance was observed in Gümüşhane. This pattern indicates that the degree of hydro-climatic similarity between training and validation basins plays an important role in determining model transferability and predictive reliability.

Among the evaluated algorithms, Gradient Boosting generally provided the most stable and consistent performance across both internal and external validation scenarios. However, all three machine learning approaches achieved satisfactory predictive capability, demonstrating their potential for regional-scale flood susceptibility prediction assessment.

The results provide empirical evidence of predictive transferability within the evaluated regional setting, with lower discrimination in the external-validation datasets than in the 2023 Rize holdout. The contribution lies in assessing this performance across complementary validation settings, rather than in establishing that environmental similarity necessarily produces successful transfer. The four-variable environmental similarity index supports a descriptive interpretation of the observed differences, but the small number of external-validation regions limits broader inference. The study does not identify the physical mechanisms responsible for these differences. Further evaluation in additional, physiographically contrasting basins is required to assess the generality of the findings.

Future studies should investigate the applicability of the proposed framework across broader climatic and physiographic settings, incorporate additional flood inventories from different regions, and evaluate advanced machine learning and deep learning architectures. Such efforts would further improve the understanding of model transferability and support the development of operational flood susceptibility prediction assessment frameworks for ungauged and data-scarce environments.

Author Contributions: U.Y.: Data curation, Software, Formal analysis, Visualization, Writing—original draft. Ö.F.D.: Conceptualization, Resources, Validation, Supervision. A.M.: Methodology, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Research Fund of the Inonu University Scientific Research Projects Unit (project ID FDK-2022-2796).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data used in the study are not open data provided by a public institution; rather, they were obtained through a formal request specifically for the purpose of this study.

Acknowledgments: During the preparation of this manuscript/study, the authors used ChatGPT 6.2 version for checking grammatical errors of the manuscript. The authors have reviewed and edited the AI tool and take full responsibility.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cloke, H.L.; Pappenberger, F. Ensemble flood forecasting: A review. *J. Hydrol.* **2009**, *375*, 613–626. [CrossRef]
2. Cvetkovic, V.M.; Martinović, J. Innovative solutions for flood risk management. *Int. J. Disaster Risk Manag.* **2020**, *2*, 71–100. [CrossRef]
3. Epule, E.T.; Peng, C.; Mafany, N.M. Methane emissions from paddy rice fields: Strategies towards achieving a win-win sustainability scenario between rice production and methane emission reduction. *J. Sustain. Dev.* **2011**, *4*, 188. [CrossRef]
4. Yükseler, U.; Dursun, Ö.F.; Alashan, S.; Imanian, H. Assessment of Hydroclimatic Change Impacts on Water Resources Through Hydrological Indicators and Machine Learning. *Water* **2026**, *18*, 1444. [CrossRef]
5. Quirogaa, V.M.; Kurea, S.; Udoa, K.; Manoa, A. Application of 2D numerical simulation for the analysis of the February 2014 Bolivian Amazonia flood: Application of the new HEC-RAS version 5. *Ribagua* **2016**, *3*, 25–33. [CrossRef]

6. Zainal, N.N.; Abu Talib, S.H. Review paper on applications of the HEC-RAS model for flooding, agriculture, and water quality simulation. *Water Pract. Technol.* **2024**, *19*, 2883–2900. [CrossRef]
7. Stefanidis, S.; Stathis, D. Assessment of flood hazard based on natural and anthropogenic factors using analytic hierarchy process (AHP). *Nat. Hazards* **2013**, *68*, 569–585. [CrossRef]
8. Hammami, S.; Zouhri, L.; Souissi, D.; Souei, A.; Zghibi, A.; Marzougui, A.; Dlala, M. Application of the GIS based multi-criteria decision analysis and analytical hierarchy process (AHP) in the flood susceptibility prediction mapping (Tunisia). *Arab. J. Geosci.* **2019**, *12*, 653. [CrossRef]
9. Thammaboribal, P.; Triapthti, N.K.; Lipiloet, S. Using of analytical hierarchy process (AHP) in disaster management: A review of flooding and landslide susceptibility mapping. *Int. J. Geoinform.* **2025**, *21*, 177–196. [CrossRef]
10. Berkes, F.; Folke, C.; Colding, J. (Eds.) *Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience*; Cambridge University Press: Cambridge, UK, 2000.
11. Gunderson, L.H.; Holling, C.S. (Eds.) *Panarchy: Understanding Transformations in Human and Natural Systems*; Island Press: Princeton, NJ, USA, 2002.
12. Wang, Z.; Lai, C.; Chen, X.; Yang, B.; Zhao, S.; Bai, X. Flood hazard risk assessment model based on random forest. *J. Hydrol.* **2015**, *527*, 1130–1141. [CrossRef]
13. Kundzewicz, Z.W.; Pińskwar, I.; Brakenridge, G.R. Changes in river flood hazard in Europe: A review. *Hydrol. Res.* **2018**, *49*, 294–302. [CrossRef]
14. Yukseler, U.; Toprak, A.; Gul, E.; Dursun, O.F. Flood hazard mapping using M5 tree algorithms and logistic regression: A case study in East Black Sea Region. *Earth Sci. Inform.* **2023**, *16*, 2033–2047. [CrossRef]
15. Karim, F.; Armin, M.A.; Ahmedt-Aristizabal, D.; Tychsen-Smith, L.; Petersson, L. A review of hydrodynamic and machine learning approaches for flood inundation modeling. *Water* **2023**, *15*, 566. [CrossRef]
16. Amiri, A.; Soltani, K.; Ebtehaj, I.; Bonakdari, H. A novel machine learning tool for current and future flood susceptibility prediction mapping by integrating remote sensing and geographic information systems. *J. Hydrol.* **2024**, *632*, 130936. [CrossRef]
17. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
18. Salman, H.A.; Kalakech, A.; Steiti, A. Random forest algorithm overview. *Babylon. J. Mach. Learn.* **2024**, *2024*, 69–79. [CrossRef]
19. Kadıoğlu, M. *Sel, Heyelan ve Çığ Için Risk Yönetimi*; JICA Türkiye Ofisi Yayınları (No: 2): Ankara, Turkey, 2008.
20. Yüksek, Ö.; Kankal, M.; Üçüncü, O. Assessment of big floods in the Eastern Black Sea Basin of Turkey. *Environ. Monit. Assess.* **2013**, *185*, 797–814. [CrossRef]
21. Toprak, A.; Yükseler, U.; Yildizhan, E. Success of machine learning and statistical methods in predicting landslide hazard: The case of Elazig (Maden). *Arab. J. Geosci.* **2024**, *17*, 275. [CrossRef]
22. Yükseler, U.; Dursun, O.F.; Yağanoğlu, M.; Mohammadian, A. Overcoming Generalization Issues in Flood Prediction: A Machine Learning Approach Across Multiple Basins. *Sustainability* **2026**, *18*, 4724. [CrossRef]
23. Mosavi, A.; Ozturk, P.; Chau, K.W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536. [CrossRef]
24. Razali, N.; Ismail, S.; Mustapha, A. Machine learning approach for flood risks prediction. *IAES Int. J. Artif. Intell.* **2020**, *9*, 73. [CrossRef]
25. Tehrany, M.S.; Jones, S.; Shabani, F. Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena* **2019**, *175*, 174–192. [CrossRef]
26. Yükseler, U. Effective Method Selection for Flood Disaster Management: A Decision Support Approach Based on River Type. *Bitlis Eren Üniversitesi Fen. Bilim. Derg.* **2025**, *14*, 1336–1356. [CrossRef]
27. Khoshkonesh, A.; Nazari, R.; Nikoo, M.R.; Karimi, M. Enhancing flood risk assessment in urban areas by integrating hydrodynamic models and machine learning techniques. *Sci. Total Environ.* **2024**, *952*, 175859. [CrossRef]
28. Hussain, M.; Tayyab, M.; Ullah, K.; Ullah, S.; Rahman, Z.U.; Zhang, J.; Al-Shaibah, B. Development of a new integrated flood resilience model using machine learning with GIS-based multi-criteria decision analysis. *Urban Clim.* **2023**, *50*, 101589. [CrossRef]
29. Quinlan, J.R. Learning with continuous classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16–18 November 1992; Volume 92, pp. 343–348.
30. Holmes, J.; Meyerhoff, M. The community of practice: Theories and methodologies in language and gender research. *Lang. Soc.* **1999**, *28*, 173–183. [CrossRef]
31. Tehrany, M.S.; Pradhan, B.; Jebur, M.N. Flood susceptibility prediction mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *J. Hydrol.* **2014**, *512*, 332–343. [CrossRef]
32. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
33. Pandey, M.; Arora, A.; Arabameri, A.; Costache, R.; Kumar, N.; Mishra, V.N.; Nguyen, H.; Mishra, J.; Siddiqui, M.A.; Ray, Y.; et al. Flood susceptibility prediction modeling in a subtropical humid low-relief alluvial plain environment: Application of novel ensemble machine learning approach. *Front. Earth Sci.* **2021**, *9*, 659296. [CrossRef]
34. Arabameri, A.; Saha, S.; Mukherjee, K.; Blaschke, T.; Chen, W.; Ngo, P.T.T.; Band, S.S. Modeling spatial flood using novel ensemble artificial intelligence approaches in northern Iran. *Remote Sens.* **2020**, *12*, 3423. [CrossRef]
35. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]

36. Beven, K.J.; Kirkby, M.J. A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* **1979**, *24*, 43–69. [CrossRef]
37. Tien Bui, D.; Ho, T.C.; Pradhan, B.; Pham, B.T.; Nhu, V.H.; Revhaug, I. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environ. Earth Sci.* **2016**, *75*, 1101. [CrossRef]
38. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann: Burlington, MA, USA, 2011.
39. Bui, D.T.; Pradhan, B.; Nampak, H.; Bui, Q.T.; Tran, Q.A.; Nguyen, Q.P. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. *J. Hydrol.* **2016**, *540*, 317–330. [CrossRef]
40. El Bilali, A.; Abdeslam, T.; Ayoub, N.; Lamane, H.; Ezzaouini, M.A.; Elbeltagi, A. An interpretable machine learning approach based on DNN, SVR, Extra Tree, and XGBoost models for predicting daily pan evaporation. *J. Environ. Manag.* **2023**, *327*, 116890. [CrossRef]
41. Kuang, D.; Liao, K.H. Learning from Floods: Linking flood experience and flood resilience. *J. Environ. Manag.* **2020**, *271*, 111025. [CrossRef]
42. Ali, Z.A.; Abduljabbar, Z.H.; Taher, H.A.; Sallow, A.B.; Almufti, S.M. Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review. *Acad. J. Nawroz Univ.* **2023**, *12*, 320–334. [CrossRef]
43. Hoo, Z.H.; Candlish, J.; Teare, D. What is an ROC curve? *Emerg. Med. J.* **2017**, *34*, 357–359. [CrossRef]
44. Arabameri, A.; Karimi-Sangchini, E.; Santosh, M. Cross-regional transferability of machine learning models for flood susceptibility prediction. *Catena* **2024**, *238*, 107845.
45. Tehrany, M.S.; Rezaie, F.; Shabani, F. Spatially adaptive flood susceptibility mapping using ensemble learning with geographic cross-validation. *Remote Sens. Environ.* **2024**, *298*, 113828.
46. Jamali, B.; Bach, P.M.; Cunningham, L.; Deletic, A. Leave-one-watershed-out validation for machine learning flood prediction models. *Environ. Model. Softw.* **2025**, *175*, 106011.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.