

Article

Advancing Urban Flood Risk Mapping: A Hybrid Framework Integrating Interpretable Machine Learning and Uncertainty-Aware Expert Judgment

Shuoyuan Liang  and Tsuyoshi Kinouchi * 

School of Environment and Society, Institute of Science Tokyo, Yokohama 226-8503, Japan;
liang.s.ac@m.titech.ac.jp

* Correspondence: kinouchi.t.ab@m.titech.ac.jp

Abstract

Climate change and urbanization are intensifying urban flood risks worldwide, making flood risk management critically important. While machine learning has been widely applied for flood susceptibility mapping, many studies do not integrate socioeconomic dimensions for comprehensive risk assessment; moreover, traditional subjective evaluation methods provide insufficient quantification of expert judgment uncertainty. This study presents a hybrid framework for urban flood risk mapping, which integrates interpretable machine learning and the Z-number-based Fuzzy Analytic Hierarchy Process (Z-FAHP), applied to Tokyo, Japan, utilizing a high-resolution digital elevation model (DEM) derived from airborne LiDAR and other publicly available geospatial datasets. The framework leverages machine learning efficiency while better accommodating vague linguistic expert judgments by incorporating confidence levels. Flood susceptibility was derived through machine learning with 14 features, with model performance rigorously evaluated using both non-spatial and spatial 4-fold cross-validation, and the CatBoost model demonstrated optimal performance. SHAP (SHapley Additive exPlanations) analysis was further employed to enhance model transparency by quantifying feature contributions. Exposure and vulnerability were quantified from 3 and 5 socioeconomic indicators, respectively, through Z-FAHP. These three criteria were synthesized to produce the flood risk map using simple additive weighting. Results reveal that approximately 32.9% of the study area faces high-to-very-high flood risk, concentrated in eastern lowlands and river corridors. This transferable framework can be applied to other cities globally as an effective tool for urban flood risk management.

Keywords: urban flood risk; interpretable machine learning; SHapley Additive exPlanations; Z-number; fuzzy analytic hierarchy process



Academic Editor: Chang Huang

Received: 15 July 2026

Revised: 25 August 2026

Accepted: 28 August 2026

Published: 30 August 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With ongoing urbanization and a changing climate, the risk of urban flooding continues to rise, posing significant threats to human life and socioeconomic activities. Urban flooding may result from localized extreme precipitation, river overbank flows, coastal storm surges, or their compound interactions, and its occurrence can be exacerbated by factors such as overwhelming of drainage systems, increased impervious surface area, high building density, and flat low-lying terrain [1,2]. Furthermore, continued capital investment in new buildings and their contents may lead to greater losses when these assets are inundated [3].

For flood risk research and management, it is essential to first clarify the definitions of risk and flood risk. Risk is fundamentally defined as the potential for adverse consequences to human or ecological systems; accordingly, flood risk is commonly conceptualized as a function of three elements: flood hazard, exposure of the affected system, and the system's vulnerability [4–7]. Flood hazard is the potential occurrence of flood events that can produce adverse consequences, determined by the physical characteristics and frequency of flood events; exposure refers to people and assets located in flood-prone areas that may be adversely affected, and vulnerability refers to the propensity or predisposition of those people and assets to suffer harm or loss when flooding occurs [6–8].

Urban flood hazard can be analyzed through urban flood modeling, with common approaches including statistical models based on probabilistic techniques, hydrologic–hydrodynamic models, cellular automata, and artificial intelligence [2,9–12]. Artificial intelligence techniques such as machine learning have gained widespread application for over a decade, which can reduce reliance on critical urban drainage data within physical models [2]. Using flood and non-flood locations as the target variable and geographic, climatic, and related features as explanatory variables, machine learning can generate maps that express the study area's susceptibility to flooding; this approach is commonly called flood susceptibility mapping. A major limitation of machine learning is the lack of transparency in interpreting model outputs; the advent of explainable artificial intelligence (XAI) has addressed this challenge [13]. Among the various XAI techniques, SHAP (SHapley Additive exPlanations) has emerged as an important approach that enhances the transparency of black-box model predictions. SHAP, derived from Shapley values in game theory, estimates the mean marginal contribution of each feature across all possible feature subsets [14]; this method has been widely applied in hydrological studies [15].

Care should be taken not to conflate flood susceptibility mapping with flood hazard mapping, as the latter generally incorporates physical characteristics of flooding such as water depth, flow velocity, inundation duration, and sometimes the return period [16]. Due to data constraints, such as a lack of detailed hydrological observations or restricted access to data, deriving physical characteristics of urban flooding via hydrodynamic modeling or estimating event return periods using statistical methods is not always feasible. Consequently, some flood risk studies treat flood susceptibility derived from machine learning as a proxy for flood hazard and incorporate it into the risk assessment framework [17–20]. Flood susceptibility is defined here as the relative predisposition of a location to flooding based mainly on its geographic and climatic characteristics.

Multi-criteria decision analysis (MCDA), also known as multi-criteria decision-making (MCDM), has been widely applied in flood risk research and management. MCDA provides a systematic and explicit framework to rank or weigh the factors influencing flood risk, with the Analytic Hierarchy Process (AHP) being the most popular method; other commonly used methods include the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), Simple Additive Weighting (SAW), the Analytic Network Process (ANP), and hybrid approaches [21–24]. AHP enables the quantification of key flood risk elements such as exposure and vulnerability, and can combine these elements with appropriate weights to produce a composite flood risk result. Examples include using the trapezoidal fuzzy AHP alone to evaluate the flood risk of an urban metro system [25], integrating a hydrodynamic model with AHP to evaluate the flood risk of a small county [26], and integrating automated machine learning with AHP to evaluate the flood risk of an urban agglomeration [27]. Due to the complexity and inherent uncertainty of flood events and their societal impacts, experts' judgments of flood risk determinants are often imprecise. In fuzzy AHP, fuzzy numbers are used in place of exact numbers to accommodate vague linguistic judgment, and the introduction of Z-numbers enables the description of experts' confidence in their own

judgments [28,29]. The Z-number has been applied in certain fields, such as post-disaster damage assessment [30] and groundwater seawater intrusion vulnerability assessment [31], but its use in flood risk assessment remains relatively limited. A more explicit representation of expert judgments can enhance the reliability of weighting under uncertainty, which in turn potentially improves the robustness of the flood risk assessment.

Building on this background, the main objective of this study is to develop a framework for urban flood risk mapping, advancing beyond the scope of conventional susceptibility mapping. The novelty of this study lies in two aspects: (i) it proposes, for the first time, a hybrid framework by incorporating interpretable machine learning and Z-number-based Fuzzy AHP (Z-FAHP) for urban flood risk mapping; and (ii) it delivers the first high-resolution flood risk map for Tokyo Metropolis that explicitly accounts for socioeconomic characteristics, based exclusively on publicly available geospatial datasets. In brief, this urban flood risk assessment framework comprises three criteria: susceptibility (derived from machine learning), exposure, and vulnerability (both computed using Z-FAHP). For each criterion, we computed a normalized index and produced a corresponding map; the three criteria were finally combined via the SAW method to generate the flood risk map. The resulting flood risk map has a spatial resolution of 10 m, enabling representation of risk at neighborhood and street scales; it is particularly relevant for flood risk management and policy decisions in Tokyo. It also has the potential to address a current gap across Japanese cities, where available products are typically limited to flood hazard maps showing inundation depth and duration. This transferable and efficient assessment framework can likewise be applied to other cities worldwide as an effective tool for urban flood risk management. The remainder of this paper is organized as follows. Section 2 provides an overview of the study area. Section 3 details the methodological framework and data sources. Section 4 presents and discusses the results, including model performance evaluation and spatial patterns of flood susceptibility, exposure, vulnerability, and integrated flood risk. Section 5 summarizes the main conclusions.

2. Study Area

Tokyo, Japan's largest city, serves as the political, economic, and cultural center of the nation. It encompasses 23 special wards, multiple cities, towns, villages, and outlying islands. Between 1999 and 2019, water-related disasters in Tokyo were driven mainly by heavy rainfall, including tropical cyclones, localized torrential downpours, and heavy rainfall associated with the Baiu front or the autumn rain front. During this period, records indicate that 25,346 buildings were affected; approximately 87% were associated with urban pluvial flooding, 11% with fluvial flooding, and the remaining 2% with hazards such as debris flows or storm surge [32]. These statistics underscore the urgent need for improved understanding and effective mitigation of flood risk in Tokyo.

This study focuses on the mainland urban region of Tokyo, bounded by the Ken-O Expressway (Metropolitan Inter-City Expressway), a major highway encircling the metropolitan area (Figure 1). This study area, which encompasses the majority of recorded flood events in Tokyo between 1999 and 2019, covers approximately 1229 km², extending eastward from the Ken-O Expressway toward central Tokyo, with an additional 1-km buffer zone on its western side. It comprises 98.4% land and 1.6% water bodies; of the land area, approximately 80% is built up, indicating a highly urbanized region with extensive impervious surfaces [33,34].

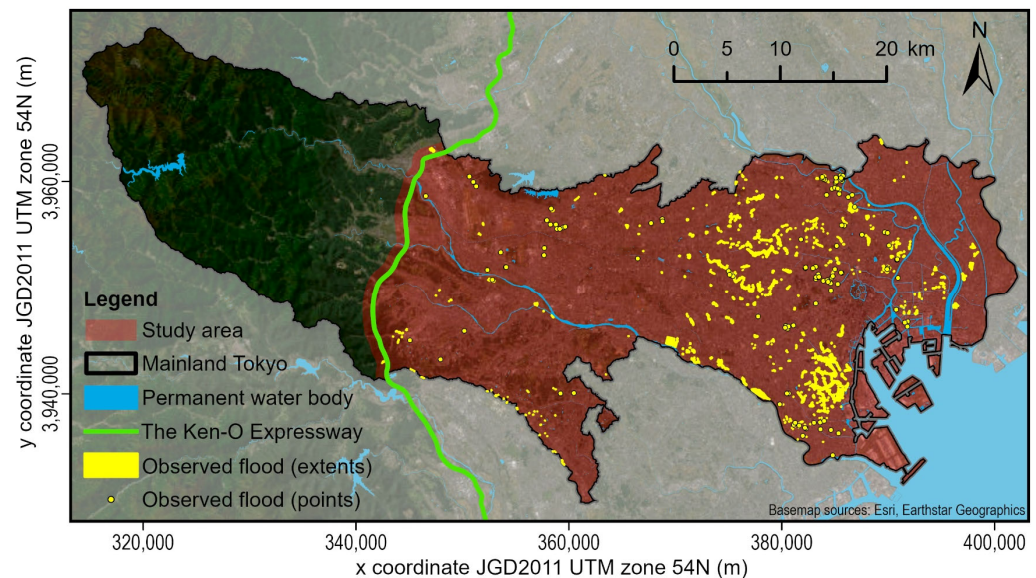


Figure 1. Location map of the study area.

3. Methodology and Materials

3.1. Overview of the Framework

The methodological framework is illustrated in Figure 2. The study proceeds in four main steps: (i) compiling geographic data, meteorological data, and flood records; (ii) comparing and selecting machine learning algorithms for flood susceptibility mapping; (iii) processing socioeconomic indicators and deriving exposure and vulnerability maps through Z-FAHP; and (iv) integrating the susceptibility, exposure, and vulnerability maps to produce the flood risk map.

The data used in this study are shown in Table 1. Specifically, we employed 14 explanatory variables (features) for machine learning, which are commonly used in flood susceptibility mapping studies [12,17,35,36]. These include six DEM-based geographic features, i.e., elevation, slope, tangential curvature, flow direction, terrain ruggedness index (TRI), and topographic wetness index (TWI); a normalized difference vegetation index (NDVI) derived from optical remote sensing imagery; distances to major rivers and distances to roads obtained using open data sources; mean annual precipitation (MAP), maximum daily precipitation (MDP) and maximum hourly precipitation (MHP) obtained from the official meteorological database; and layers of subsurface geology and soil type that characterize surface conditions. Historical urban flood records are essential for supervised machine learning. Based on 21 years of flood observations from 1999 to 2019, a flood inventory was created. The labeled data consist of an equal number of flood samples and non-flood samples. We conducted a systematic comparison of two machine learning algorithms, CatBoost and an artificial neural network (ANN). Model performance was first assessed with multiple metrics under non-spatial 4-fold cross-validation, and then a more rigorous spatial 4-fold cross-validation was implemented to verify generalization capability. The model with the best performance was subsequently selected to generate the flood susceptibility map. The flood susceptibility index ranges from 0 to 1, representing the relative degree of flood susceptibility at each location.

We assembled eight socioeconomic indicators and organized them into two dimensions: exposure and vulnerability. For all indicators, we winsorized the data at the 1st and 99th percentiles to reduce the influence of extreme values and normalized them to [0, 1]. The exposure index was computed from three indicators, that is, population intensity, the proportion of vulnerable population, and the publicly announced land price from

the National Land Information Division, Ministry of Land, Infrastructure, Transport and Tourism (MLIT) of Japan. These normalized exposure indicators were combined using the SAW method with weights determined through Z-FAHP to obtain the exposure map. The vulnerability index was computed from the deficits of five indicators, that is, the densities of designated emergency evacuation sites (DEES), designated evacuation shelters (DES), emergency transportation roads (ETR), fire stations (FS), and disaster base hospital beds (DBHB). Deficits were computed as 1 minus the normalized values, and these deficit layers were similarly combined using the SAW method with weights determined through Z-FAHP to obtain the vulnerability map. Finally, the susceptibility, exposure, and vulnerability maps were combined using the SAW method with equal weights to produce the flood risk map.

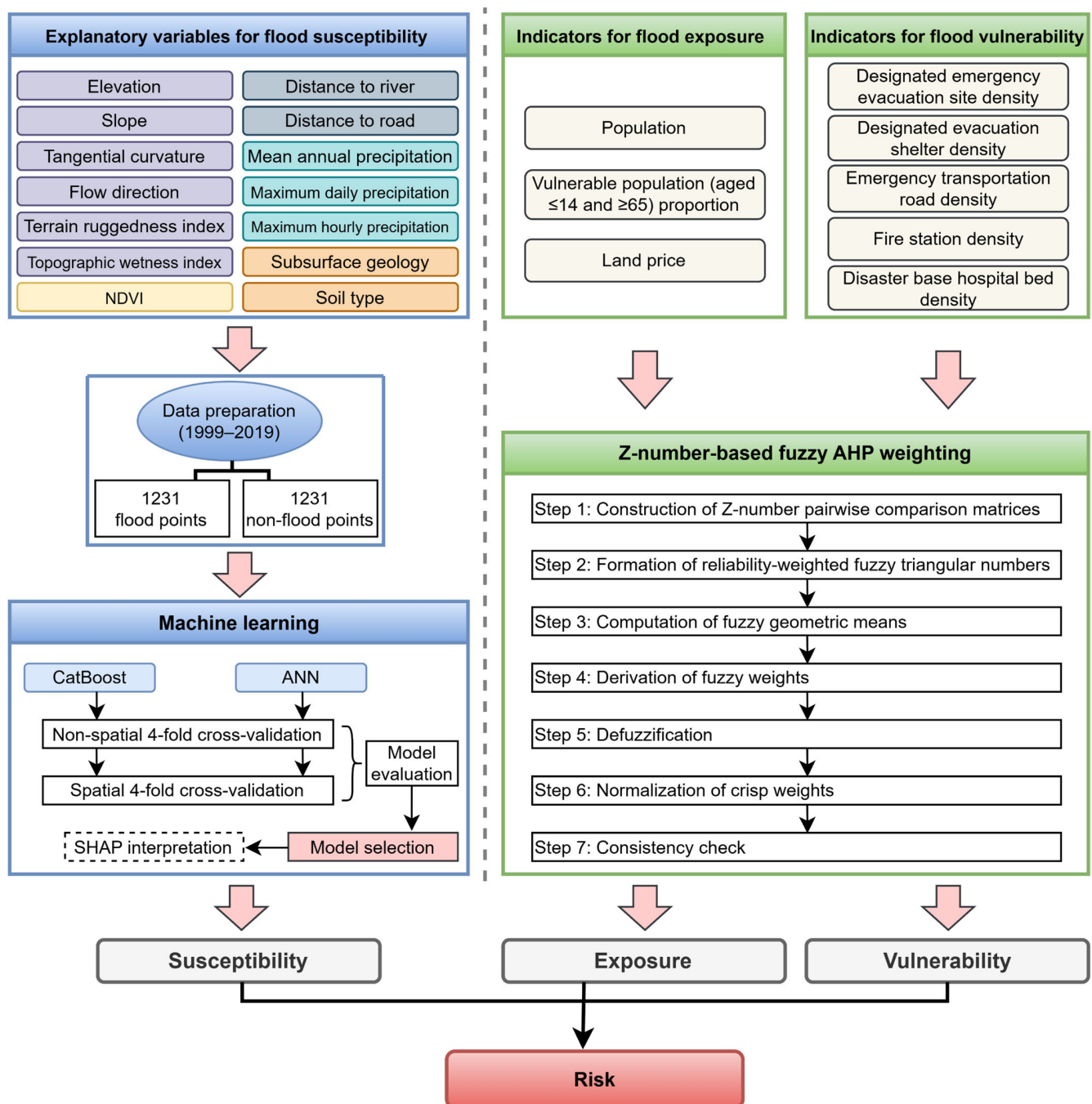


Figure 2. Methodological framework of the study.

Table 1. Summary of the data used in this study.

Category	Item	Dataset	Source	Year	Native Spatial Resolution
Urban flood records	Flood inventory	Flood Risk Information System	Bureau of Construction, Tokyo Metropolitan Government	1999–2019	/
Explanatory variables (features) for flood susceptibility	Elevation	High-Resolution Elevation Dataset	Geospatial Information Authority of Japan (GSI)	2025	5 m
	Slope				
	Tangential curvature				
	Flow direction				
	TRI				
	TWI				
	NDVI	WorldCover product	The European Space Agency (ESA)	2020	10 m
	Distance to river	/	OpenStreetMap	Vector data (water bodies & roads) downloaded in 2025	
	Distance to road				
	MAP	/	Japan Meteorological Agency (JMA) and MLIT	1999–2019	/
	MDP				
	MHP				
Exposure indicators	Subsurface geology	Digital National Land Information	MLIT	1970s	250 m
	Soil type				
	Population				
	Vulnerable population proportion			2020	
Vulnerability indicators	Land price	/	Tokyo Fire Department		/
	ETR density			2020	
	DBHB density			2020	
	FS density			Data last updated in 2025	
	DEES density			Data last updated in 2025	
Others	DES density	/	GSI		
	Tokyo administrative boundary				

3.2. Materials

3.2.1. Urban Flood Records

The flood inventory was compiled from flood records on the Flood Risk Information System website operated by the Bureau of Construction, Tokyo Metropolitan Government,

which provides downloadable geographic information system (GIS)-based historical inundation records from 1974 to 2019 [32]. We considered only records of urban pluvial and fluvial flooding. To reduce the influence of urban change while retaining a sufficient sample for the subsequent machine learning, we limited the inventory to events reported from 1999 to 2019 (21 years). Most downloaded historical inundation records were flood inundation extents that were polygon features, with a minority of 118 records in point form. For polygons, we applied spatial random sampling with a density of approximately one point per square kilometer and a minimum spacing greater than 50 m, which produced 1113 candidate points. All candidate points were checked manually to ensure that no point fell within river channels or large buildings. We then combined these 1113 sampled points with the 118 original point records, yielding 1231 samples that represent inundation (labeled as 1). An equal number of 1231 non-flood samples (labeled as 0) was generated via buffer-controlled random sampling from locations without historical inundation records, while avoiding areas clearly below mean sea level and ensuring a minimum spacing of over 50 m from observed flood inundation extents or points [36,37]. In total, the dataset contains 2462 labeled data points for machine learning.

3.2.2. Explanatory Variables for Flood Susceptibility

A high-resolution digital elevation model (DEM) is essential for urban flood risk mapping, enabling risk identification at neighborhood and street scales. Therefore, the LiDAR-based DEM with a native spatial resolution of approximately 5 m was derived from the GSI High-Resolution Elevation Dataset [38]. The dataset is a bare-earth product in which buildings, bridges, and trees have been removed; therefore, urban land surface changes occurring over several years have a negligible effect on the DEM. The DEM was resampled to a resolution of 10 m using bilinear interpolation to maintain consistency with the NDVI data mentioned below. We then calculated the slope, tangential curvature, flow direction, TRI, and TWI layers based on this DEM. The definitions and equations for TRI and TWI are provided in Appendix A.

The NDVI layer was derived from the ESA WorldCover 2020 v100 product, which is based on the Copernicus Sentinel-1 and Sentinel-2 archives [39]. It represents a median composite of NDVI for the year 2020 at a spatial resolution of approximately 10 m, depicting the vegetation cover condition across the study area.

Precipitation data at 18 meteorological stations within and around the study area for the period 1999–2019 were sourced from JMA [40] and MLIT [41]. From these station records, we derived the MAP, MDP, and MHP, which were then interpolated into raster layers with a spatial resolution of 50 m using the inverse distance weighting (IDW) method.

Vector data for the major rivers and roads (including motorway, trunk, primary, secondary, and tertiary roads) were sourced from OpenStreetMap [42]. In combination with the DEM, we computed surface distance to the nearest river and road along the terrain, producing two raster layers with a spatial resolution of 10 m. Vector data for the subsurface geology and soil type were downloaded from the Digital National Land Information website operated by the National Land Information Division, MLIT [43]. Although both data layers originate from the 1:200,000 Land Classification Basic Survey conducted in the 1970s, subsurface geology and soil type are treated as unchanged over the decades. They were processed into raster layers with a spatial resolution of 10 m.

3.2.3. Flood Exposure Indicators

Our study prioritizes population exposure and economic exposure. Vector data for population and publicly announced land price were downloaded from the Digital National Land Information website [44,45]. The population data from Japan's 2020 national census

were summarized for each 250-m grid cell with age-stratified demographic information. In addition to converting into a raster layer representing population distribution, we produced a raster layer of the vulnerable population proportion by computing the fraction of people aged 14 or younger and 65 or older for each 250-m grid cell, as these dependent age groups are assumed to have relatively lower mobility. The layers were resampled to 50-m resolution via IDW interpolation, which also filled spatial gaps present in the original 250-m census grid where no data were recorded. The resulting surfaces were used as relative indicators of population exposure rather than absolute counts.

Land prices are strongly correlated with local economic vitality and commerce. In a highly urbanized and economically advanced metropolis like Tokyo, Gross Domestic Product (GDP) provides less spatial discrimination of economic exposure than land prices. The land price data we used are those surveyed and published by the Land Appraisal Committee, MLIT, under the Public Notice of Land Prices Act, representing market values per square meter in selected locations as of January 1, 2020; we refer to these data as the publicly announced land price. Because the data are point features, we applied IDW to interpolate over 2600 points to a raster layer with a spatial resolution of 50 m.

All the above indicators were winsorized at the 1st and 99th percentiles and normalized to [0, 1]. The SAW method was then applied to the exposure indicators with the weights determined through Z-FAHP to obtain the flood exposure map.

3.2.4. Flood Vulnerability Indicators

There is ongoing discussion on how to characterize flood vulnerability, and prior studies have proposed hundreds of different indicators [4,46]. In this study, we defined flood vulnerability as deficiencies in the capacities for evacuation, emergency response, and recovery, rather than the broader sense of socioeconomic vulnerability. Specifically, five indicators were selected, i.e., the densities of DEES, DES, ETR, FS, and DBHB.

In Japan, emergency evacuation facilities comprise DEES and DES. The former serve as nearby locations for immediate evacuation during major disasters, while the latter provide temporary accommodation for those unable to return home due to disasters. Vector datasets for these facilities were sourced from GSI as point features [47]. We limited our selection to facilities designated for flood hazards in the GSI dataset, which are applicable to both pluvial and fluvial flooding. Assuming evacuees use available routes to reach nearby DEES, with some requiring temporary accommodation at shelters close to home, we computed kernel density with a 3 km search radius. This produced separate raster layers showing the densities (sites/km²) of DEES and DES at 10-m spatial resolution.

The fire department conducts rescues and transports those in need to hospitals, with emergency response typically originating from the nearest FS. ETR are critical corridors designated by the government to ensure evacuation, rescue, and logistical support during emergencies. Data for FS and ETR were obtained from Tokyo Fire Department [48] and the Digital National Land Information website [49], respectively. Because emergency vehicles can choose appropriate routes during disasters, we computed kernel density with a 3 km search radius and derived raster layers of ETR density (km/km²) and FS density (sites/km²) with a spatial resolution of 10 m.

Although Tokyo has many medical institutions of different tiers, our analysis focused on the distribution of disaster base hospitals, which act as key centers for treating disaster casualties and as medical support bases for local hospitals during major disasters. Bed counts for disaster base hospitals were sourced from the Digital National Land Information website [50]. Because the number of disaster base hospitals is limited but their service areas are relatively extensive, we computed kernel density with a 6 km search radius and derived the raster layer of DBHB density (beds/km²) with a spatial resolution of 10 m.

All the above indicators were winsorized at the 1st and 99th percentiles and normalized to [0, 1]. Deficits were computed as 1 minus the normalized values, and the SAW method was then applied to the five deficit layers with the weights determined through Z-FAHP to obtain the flood vulnerability map.

3.3. Machine Learning

CatBoost is a supervised ensemble machine learning algorithm based on the gradient boosting decision tree framework, which has demonstrated reliable performance in flood susceptibility mapping [27,35,51,52]. We selected the commonly used ANN for comparison. We used the ‘randomized_search’ method from the CatBoost Python library (version 1.2.7) and the ‘RandomizedSearchCV’ method from the scikit-learn Python library (version 1.2.2), implemented in Python (version 3.10.18), to perform hyperparameter optimization for CatBoost and ANN, respectively, using the area under the receiver operating characteristic curve (ROC-AUC) as the optimization criterion, after which we determined the final parameters. Specific parameters for the CatBoost model included a learning rate of 0.05, a maximum tree depth of 4, and 500 iterations. The ANN model featured two hidden layers with 128 and 64 neurons, respectively, a constant learning rate of 0.003, a maximum of 500 epochs, and ReLU activation functions. The detailed hyperparameters and search ranges are summarized in Table 2.

Table 2. Hyperparameters for the machine learning models.

Model	Hyperparameter	Search Range	Selected Value
ANN	Hidden layer sizes (hidden_layer_sizes)	{(128, 64), (128, 64, 32), (64, 32)}	(128, 64)
	Learning rate schedule (learning_rate)	{constant, adaptive}	constant
	Initial learning rate (learning_rate_init)	{0.001, 0.003, 0.005, 0.01, 0.02, 0.05}	0.003
CatBoost	Learning rate (learning_rate)	{0.001, 0.01, 0.03, 0.05, 0.1, 0.2}	0.05
	Tree depth (depth)	{3, 4, 5, 6, 7, 8}	4
	Number of iterations (iterations)	{100, 200, 300, 400, 500}	500

We first snapped the 14 explanatory raster layers to a 10-m grid. Layers with differing resolutions were resampled to this grid. Predictor values were then extracted at the 2462 labeled data points (Figure 3). Both models underwent rigorous validation via non-spatial 4-fold cross-validation and spatial 4-fold cross-validation. In the non-spatial 4-fold cross-validation, the entire set of labeled data was randomly split into four subsets. In each run, one subset was reserved for validation while the other three subsets trained the model. The process was repeated four times, with every subset used once for validation. Model performance was evaluated by averaging results across the four runs.

To account for potential optimistic bias from spatial autocorrelation and to assess model generalization ability, we performed spatial 4-fold cross-validation. The study area was divided into four blocks by a cross, with labeled data in each block forming one subset of roughly balanced size; the intersection point of the cross was determined via a grid search targeting a maximum relative deviation of 10% in the number of labeled data points across blocks. In each run, one subset was used for validation and the other three for training. Subsequently, a 5-km buffer zone was extended outward from each validation block to isolate it from the three training blocks; points within this buffer zone were excluded from both training and validation, slightly reducing the amount of training data. Following

the same procedure as non-spatial validation, the process was repeated four times with performance averaged across the four runs.

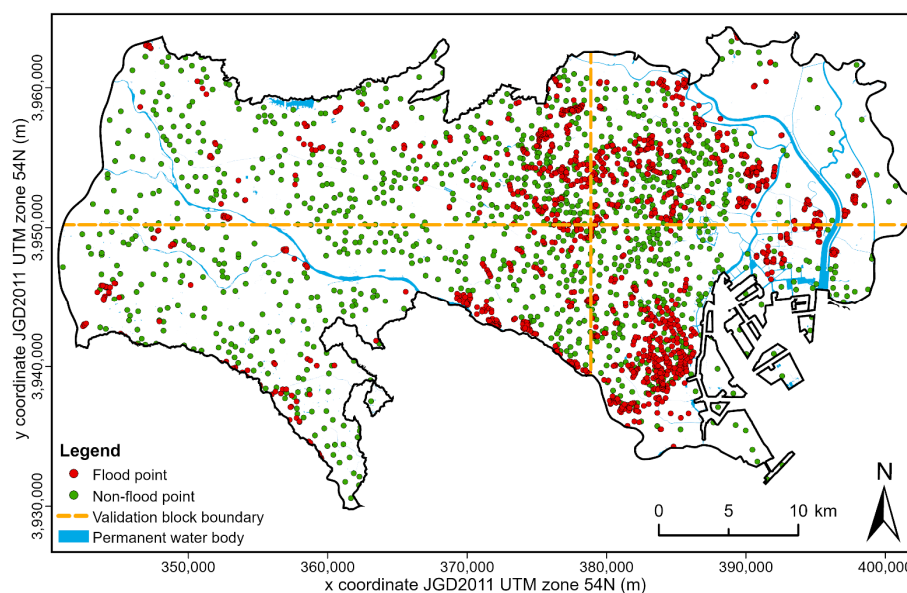


Figure 3. Spatial distribution of labeled data points. Flood samples are denoted in red, and non-flood samples are denoted in green; orange dashed lines indicate the boundaries of the validation blocks used for spatial 4-fold cross-validation.

Model performance was assessed with four threshold-dependent metrics (accuracy, precision, recall, F1-score) and a threshold-free metric (the receiver operating characteristic curve). A threshold of 0.5 is commonly used in well-defined binary classification [53]; given our balanced sample and the assumption of no asymmetric misclassification costs, we adopted this fixed threshold and applied it consistently to both models to ensure a fair comparison. The threshold fixes the operating point for the confusion matrix, from which the four threshold-dependent metrics are derived. Accuracy denotes the proportion of all samples that are classified correctly. Precision denotes the proportion of predicted positives that are truly positive. Recall (also known as sensitivity, hit rate, or true positive rate) is the proportion of true positives among all actual positives. The F1-score (F1) is defined as the harmonic mean of precision and recall. Receiver operating characteristic (ROC) curves were used to assess the overall performance of machine learning models in binary classification across all thresholds by plotting the false positive rate (FPR) on the horizontal axis and the true positive rate on the vertical axis [54]. A curve closer to the upper-left corner indicates better model performance. The area under the curve (AUC) of ROC provides a single quantitative measure of model performance ranging from 0 to 1, where 1 denotes a perfect classifier, and 0.5 denotes random prediction. The equations are given as follows [55,56]:

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

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (4)$$

$$FPR = \frac{FP}{TN + FP} \quad (5)$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

To statistically compare the ROC-AUCs of the two models, we employed the non-parametric DeLong test [57]. This test assesses whether the observed AUC difference between two classifiers is statistically significant; a p -value of less than 0.05 indicates that the performance difference is not due to random chance. By comprehensively evaluating the aforementioned threshold-dependent and threshold-free metrics, alongside the DeLong test results, we could select the optimal model and train it on all labeled data to produce the flood susceptibility map for the study area. Additionally, we employed SHAP to interpret the model's prediction. The SHAP beeswarm plot can visualize these SHAP values as data points grouped by features, with features ranked by importance and colors denoting the corresponding feature magnitudes. It provides an intuitive view of how each feature affects model decisions, thereby enhancing interpretability.

3.4. Z-Number-Based Fuzzy AHP

Fuzzy AHP, which can accommodate the vague linguistic judgment of experts in real-world decision-making scenarios, has been widely applied [28,58]. Combining Z-numbers with fuzzy AHP can better address uncertainties inherent in subjective judgment. A Z-number can be denoted as $(\tilde{A}, \tilde{B})$ when combined with fuzzy numbers, which is an ordered pair of fuzzy numbers [29,59]. The first component $\tilde{A}$ represents the fuzzy restriction on the evaluation value, while the second component $\tilde{B}$ is a measure of the reliability or confidence associated with the first component.

Consider a criterion computed from m ($m \in \mathbb{N}^+, m \geq 2$) indicators, which form a pairwise comparison matrix $Z = [Z_{ij}]_{m \times m}$. The rows and columns of the matrix correspond to the same ordered set of indicators. Each element Z_{ij} ($i, j = 1, 2, \dots, m$) represents a pairwise comparison between indicator i (row) and indicator j (column). In this study, each element Z_{ij} is expressed as a Z-number:

$$Z_{ij} = (\tilde{A}_{ij}, \tilde{B}_{ij}) \quad (6)$$

where $\tilde{A}_{ij}$ is a triangular fuzzy number representing the expert's fuzzy judgment on the relative importance of indicator i with respect to indicator j , and $\tilde{B}_{ij}$ is a triangular fuzzy number that quantifies the expert confidence level (reliability) regarding that judgment.

Let the elements in the lower triangular portion (where $i > j$) be those directly assessed by experts, such that $\tilde{A}_{ij} = (a_{ij}^l, a_{ij}^m, a_{ij}^u)$ where $1 \leq a_{ij}^l \leq a_{ij}^m \leq a_{ij}^u \leq 9$ and $a_{ij}^l, a_{ij}^m, a_{ij}^u \in \{1, 2, \dots, 9\}$, and $\tilde{B}_{ij} = (b_{ij}^l, b_{ij}^m, b_{ij}^u)$ where $0 \leq b_{ij}^l \leq b_{ij}^m \leq b_{ij}^u \leq 1$. The diagonal elements are defined as $\tilde{A}_{ii} = (1, 1, 1)$ and $\tilde{B}_{ii} = (1, 1, 1)$ for all i , as each indicator has equal importance relative to itself with full certainty. For off-diagonal elements (where $i \neq j$), larger values of $\tilde{A}_{ij}$ indicate a higher importance of indicator i over indicator j , and larger values of $\tilde{B}_{ij}$ indicate higher confidence regarding that judgment, and vice versa.

The importance component of the pairwise comparison matrix is reciprocal, while the reliability component is symmetric across the diagonal. Z_{ji} represents the relative importance of indicator j with respect to indicator i and is calculated as:

$$Z_{ji} = (\frac{1}{\tilde{A}_{ij}}, \tilde{B}_{ij}), \text{ where } \frac{1}{\tilde{A}_{ij}} = (\frac{1}{a_{ij}^u}, \frac{1}{a_{ij}^m}, \frac{1}{a_{ij}^l}) \quad (7)$$

For each element in the matrix, convert the Z-number-based fuzzy number to a regular fuzzy number using the square root of the reliability centroid as a scaling factor [59]. For a triangular fuzzy number $\tilde{B}_{ij} = (b_{ij}^l, b_{ij}^m, b_{ij}^u)$, this can be expressed as:

$$\tilde{F}_{ij} = \sqrt{\frac{b_{ij}^l + b_{ij}^m + b_{ij}^u}{3}} \cdot \tilde{A}_{ij} \quad (8)$$

This yields a new regular fuzzy pairwise comparison matrix $F = [\tilde{F}_{ij}]_{m \times m}$. Subsequently, a standard fuzzy AHP procedure is applied to the regular fuzzy pairwise comparison matrix to derive indicator weights, as outlined in the following steps [24,28,58,60]. Firstly, the fuzzy geometric mean of each row in the fuzzy matrix is calculated and normalized to obtain fuzzy weights for each indicator. Secondly, these fuzzy weights are defuzzified using the centroid method to derive crisp weights for each indicator. Finally, a consistency check is performed using the consistency ratio (CR), ensuring the CR is lower than 0.1 for acceptable judgments. The equations are given as follows:

$$\tilde{g}_i = \left(\prod_{j=1}^m \tilde{F}_{ij} \right)^{\frac{1}{m}} \quad (9)$$

$$\tilde{w}_i = \left(\frac{g_i^l}{\sum_{k=1}^m g_k^u}, \frac{g_i^m}{\sum_{k=1}^m g_k^m}, \frac{g_i^u}{\sum_{k=1}^m g_k^l} \right) \quad (10)$$

$$w_i = \frac{w_i^l + w_i^m + w_i^u}{3} \quad (11)$$

$$w_i^* = \frac{w_i}{\sum_{k=1}^m w_k} \quad (12)$$

$$CR = \frac{(\lambda_{max} - m) / (m - 1)}{RI} \quad (13)$$

where $\tilde{g}_i$ is the fuzzy geometric mean for indicator (row) i ; $\tilde{w}_i$ is the fuzzy weight for indicator i ; w_i is the crisp weight for indicator i , defuzzified via the centroid method; w_i^* is the normalized crisp weight for indicator i , which is the weight applied to aggregate the indicators into the criterion if the consistency check is passed; λ_{max} is the maximum eigenvalue of the defuzzified pairwise comparison matrix; and RI is the random index [61].

In this study, to facilitate the expression of fuzzy judgments and ensure judgment consistency, we prepared correspondence tables between linguistic variables and triangular fuzzy numbers for relative importance (Table 3) and confidence levels (Table 4). Using these correspondence tables, we constructed two pairwise comparison matrices for the exposure and vulnerability criteria as an illustrative demonstration of the procedure, based on our domain knowledge of flood risk assessment and our local knowledge of the study area; this is discussed further as a limitation in Section 4.2. We applied the aforementioned Z-FAHP procedure to calculate the weights of indicators under the exposure and vulnerability criteria, and used these weights to synthesize the exposure and vulnerability maps through the SAW method. Finally, we specified equal weights (1/3 each) for the three criteria (susceptibility, exposure, and vulnerability), based on the assumption of equal contribution to flood risk, and integrated them using the SAW method to generate the final flood risk map.

Table 3. Linguistic judgments of relative importance and corresponding triangular fuzzy weights.

Linguistic Judgment	Triangular Fuzzy Number
Equally important	(1, 1, 1)

Table 3. *Cont.*

Linguistic Judgment	Triangular Fuzzy Number
Moderately important	(2, 3, 4)
Important	(4, 5, 6)
Very important	(6, 7, 8)
Extremely important	(9, 9, 9)
Intermediate values	(1, 2, 3)
	(3, 4, 5)
	(5, 6, 7)
	(7, 8, 9)

Table 4. Expert confidence levels and corresponding triangular fuzzy weights.

Confidence Level	Triangular Fuzzy Number
Limited	(0.4, 0.5, 0.6)
Moderate	(0.5, 0.6, 0.7)
Moderate to high	(0.6, 0.7, 0.8)
High	(0.7, 0.8, 0.9)
Very high	(0.8, 0.9, 1.0)

4. Results and Discussion

4.1. Flood Susceptibility Mapping

Figure 4 presents the 14 explanatory variables (features) used for susceptibility assessment through machine learning. The topography of the study area descends from west to east, transitioning progressively from mountains and hills through plateaus to lowlands. Elevation ranges from approximately 559 m in the western mountainous area to below sea level in the eastern lowlands. The topographic characteristics are further reflected in the slope and curvature distributions. The landscape is dissected by numerous rivers and waterways, while the road network exhibits high density and connectivity across the study area. The NDVI distribution shows relatively higher values in the western areas, while scattered high NDVI patches in the middle and eastern parts correspond to urban parks. Precipitation variables exhibit spatial heterogeneity, with MAP and MDP being higher in the western mountainous and hilly area compared to the eastern urban center. In contrast, MHP exhibits higher values in both the western mountainous area and the eastern urban and coastal areas, highlighting the substantial stress that intense short-duration rainfall events exert on drainage systems in these locations. The geological and pedological complexity of the study area is also evident.

To assess multicollinearity among the prepared continuous explanatory variables, Pearson correlation analysis was conducted (Figure 5), with the three categorical explanatory variables (subsurface geology, flow direction, and soil type) excluded from this analysis. Most pairwise correlations among the explanatory variables are relatively low. Two pairs, however, show high correlation, namely slope and TRI, and elevation and MDP. All variables were retained for use in the comparison of machine learning models for the following reasons. First, although slope and TRI are highly correlated, TRI quantifies terrain ruggedness based on local elevation variability, which is physically distinct from slope. Second, elevation and MDP respectively reflect topographic and climatic conditions; their high correlation may arise from spatial coincidence or from a topographic influence on precipitation patterns. Both variables were nonetheless retained, as elevation is a fundamental topographic control on flood susceptibility and MDP conveys climatic information. Third, tree-based models are relatively robust to multicollinearity [62,63]; the L2 regularization and early stopping used in the ANN can also deal with this issue.

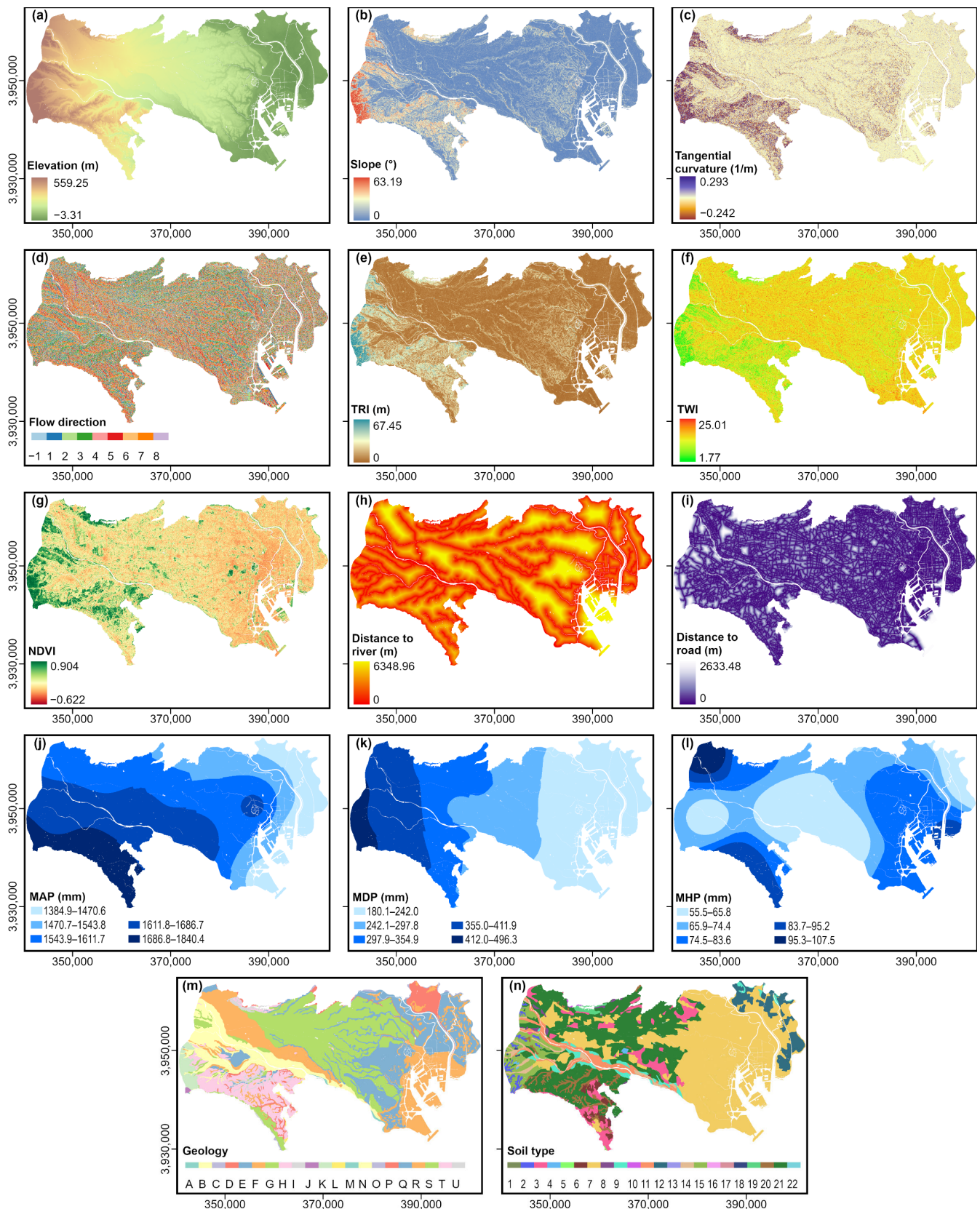


Figure 4. Explanatory variables for flood susceptibility mapping via machine learning: (a) elevation; (b) slope; (c) tangential curvature; (d) flow direction; (e) terrain ruggedness index (TRI); (f) topographic wetness index (TWI); (g) normalized difference vegetation index (NDVI); (h) distance to river; (i) distance to road; (j) mean annual precipitation (MAP); (k) maximum daily precipitation (MDP); (l) maximum hourly precipitation (MHP); (m) subsurface geology; and (n) soil type.

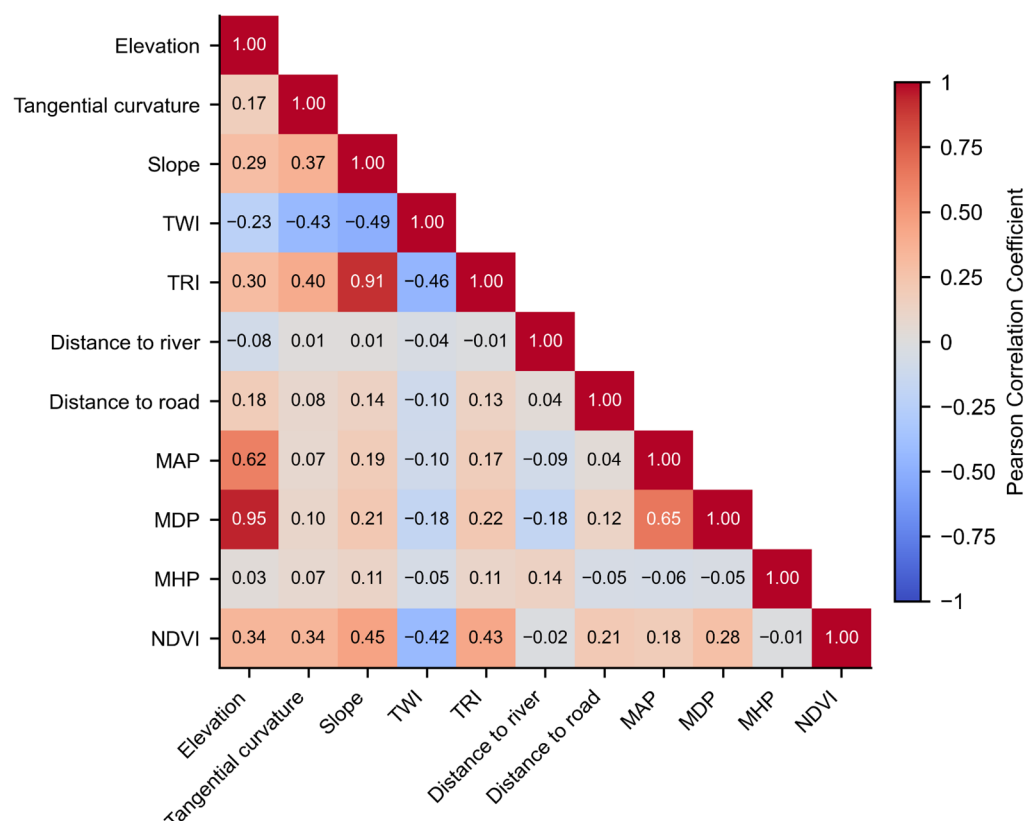


Figure 5. Lower-triangular Pearson correlation matrix for the continuous explanatory variables. Colors indicate the strength and direction of the pairwise linear correlation, ranging from -1 (strong negative correlation, blue) to $+1$ (strong positive correlation, dark red), with values annotated in each cell.

Figure 6 compares the flood susceptibility maps produced by the ANN and CatBoost models. Visually, both models identify the eastern lowlands and river corridors as zones with relatively higher susceptibility. However, the ANN map shows a more fragmented pattern with scattered, discontinuous patches, which is especially evident in the eastern lowlands, where several green patches of low susceptibility are embedded within large red areas of high susceptibility. By comparison, the CatBoost map demonstrates a more continuous and smooth spatial distribution, and also shows a clearer delineation of higher susceptibility zones along small river channels and their surrounding areas. Table 5 and Figure 7 compare the performance of the ANN and CatBoost models under two validation strategies. Under non-spatial 4-fold cross-validation, both models perform well, but CatBoost outperforms ANN across all evaluation metrics. The ROC-AUCs are 0.944 and 0.905 for CatBoost and ANN, respectively, representing a 4.3% improvement (DeLong's test, $p < 0.01$). Spatial 4-fold cross-validation is designed to mitigate the influence of spatial autocorrelation, as validation blocks are isolated from training data via a 5-km buffer zone, providing a more conservative test of model transferability to unseen locations. Under this stricter validation strategy, both models show reduced performance, yet CatBoost still achieves the best overall results, with a ROC-AUC of 0.894, outperforming ANN by 7.6% (DeLong's test, $p < 0.01$). In summary, these results indicate that the CatBoost model has a stronger spatial generalization ability in this study area; therefore, the flood susceptibility map derived by the CatBoost model is adopted for subsequent flood risk mapping.

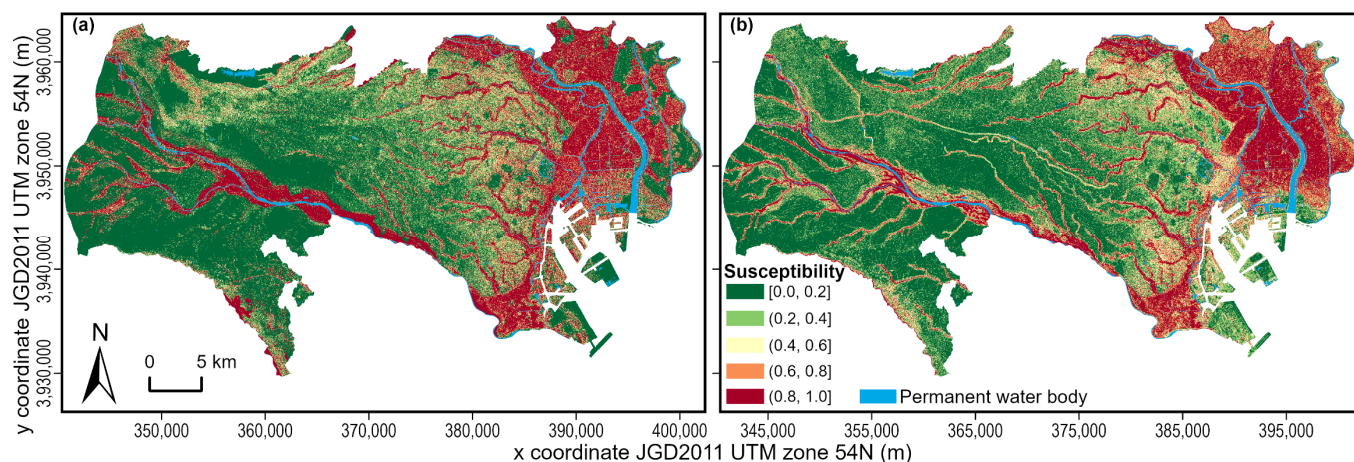


Figure 6. Flood susceptibility maps obtained using the ANN (a) and CatBoost (b) models. To facilitate comparison, the flood susceptibility indices in both maps are divided into the same five equal intervals.

Table 5. Performance of the ANN and CatBoost models.

Model	Validation Type	Accuracy	Precision	Recall	F1-Score	ROC-AUC
ANN	Non-spatial 4-fold	0.811	0.800	0.829	0.814	0.905
CatBoost		0.870	0.874	0.864	0.869	0.944
ANN	Spatial 4-fold	0.743	0.734	0.715	0.711	0.831
CatBoost		0.815	0.819	0.765	0.789	0.894

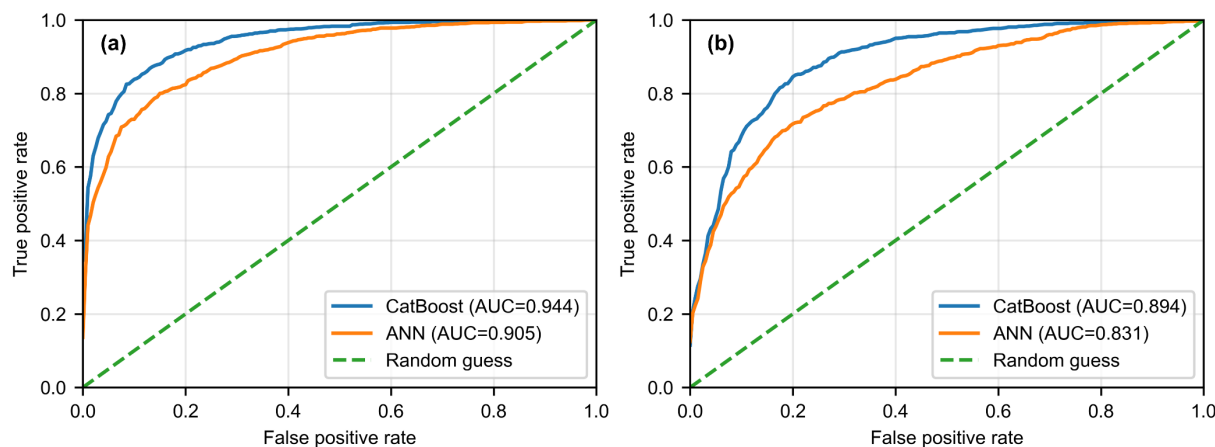


Figure 7. ROC curves obtained using the ANN and CatBoost models under (a) non-spatial 4-fold cross-validation and (b) spatial 4-fold cross-validation. Blue and orange curves represent CatBoost and ANN models, respectively. The green dashed line denotes random classification (AUC = 0.5); performance below this line indicates worse than a random guess.

Figure 8 presents the SHAP beeswarm plot that explains the contribution of each feature to the CatBoost model output. The results indicate that TWI, NDVI, elevation, subsurface geology, and distance to river are the five most important features in determining the model prediction. Higher TWI values, lower NDVI, and lower elevation are associated with increased predicted flood susceptibility. Subsurface geology and distance to river also show strong impacts. The role of river distance is straightforward, as areas close to rivers are more susceptible to flooding, which aligns with common sense. In terms of climatic features, MAP and MHP show moderate contributions, with their spatial patterns exerting

greater influence on the model output than MDP. The remaining features also influence the model output to varying degrees.

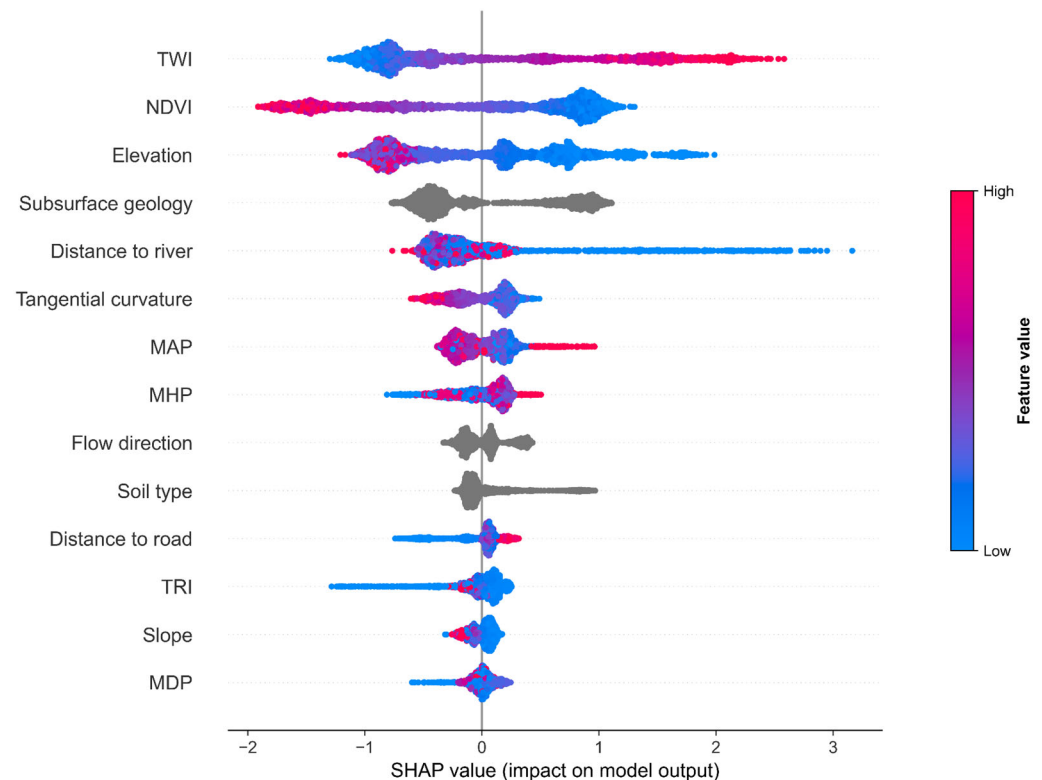


Figure 8. The SHAP beeswarm plot for the CatBoost model. Positive SHAP values indicate that the feature increases flood susceptibility, while negative SHAP values indicate that it decreases flood susceptibility. The vertical axis lists features ranked by their mean absolute SHAP values, which reflect feature importance, from top to bottom. Each data point represents an individual sample, with the color gradient reflecting feature values: red represents high values, and blue represents low values. Subsurface geology, flow direction, and soil type are shown in grey as they are categorical features.

Since subsurface geology is a categorical feature without a meaningful numerical ordering, Figure 9 presents SHAP dependence plots only for the four most important continuous features identified in Figure 8, namely TWI, NDVI, elevation, and distance to river, all of which display clear trends consistent with those described above. It should be noted that SHAP merely interprets the influence of features on the model output; since these effects are not necessarily causal, SHAP cannot be completely relied upon to explain the physical mechanisms underlying flood occurrence [64].

Several sources of uncertainty may exist in the employed machine learning models. First, the completeness of the flood records is unknown. Despite utilizing official flood records, complete coverage of all historical flooding events within the study area cannot be assured. Record incompleteness may result from unobserved events or events with limited impacts that were not officially recorded. Second, the quantity and spatial distribution of samples may introduce sampling bias. While we used a balanced dataset with equal numbers of flood and non-flood samples [36], the sample size and the spatial placement of samples may still influence model outcomes, which deserves further verification in future studies. Third, the flood inventory used for model training covers only urban pluvial and fluvial flooding, so the resulting susceptibility map does not account for coastal flood impacts. Moreover, the model training is based on historical flooding patterns and does not account for future changes in climate, land use, or flood control measures. Recent projections indicate that sub-daily extreme rainfall intensities in Tokyo will likely increase

under global warming, with a greater increase for shorter durations and higher return periods, and the frequency of extreme events is also expected to rise [65]. Future studies could explore integrating such projected changes in extreme rainfall to enable dynamic flood susceptibility mapping that accounts for evolving climatic conditions. Lastly, other factors including the temporal coverage of flood records, machine learning hyperparameters, and feature selection may affect the model results.

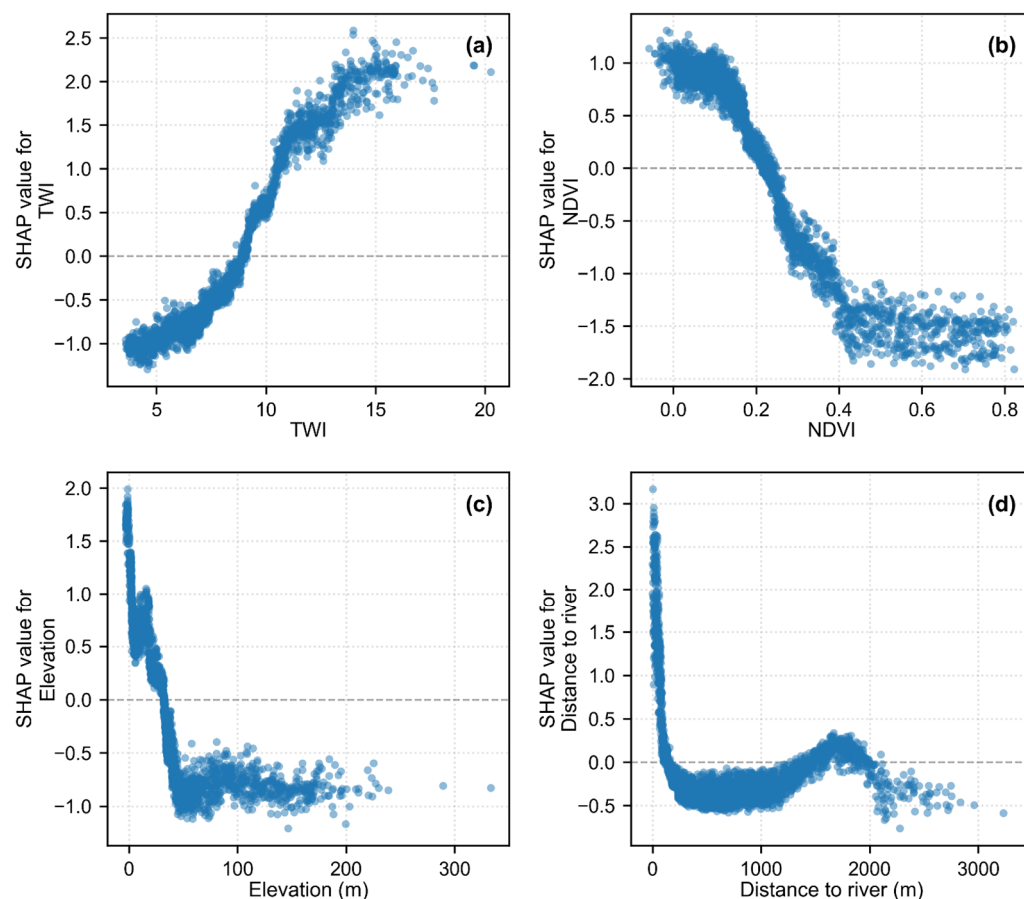


Figure 9. SHAP dependence plots for the four most important continuous features in the CatBoost model: (a) TWI, (b) NDVI, (c) elevation, and (d) distance to river. Each data point represents an individual sample, with the x-axis showing the feature value and the y-axis showing the corresponding SHAP value.

Given the uneven development of cities worldwide and data availability limitations, it is necessary to flexibly select urban flood assessment methods according to specific objectives; where inundation records are available, flood susceptibility assessment based on machine learning represents an efficient option. Flood susceptibility mapping characterizes the tendency of flooding within a study area by addressing the fundamental binary question of whether flooding will occur [66]. This approach generates susceptibility indices rather than quantifying physical characteristics such as water depth or flow velocity. In the contexts of flood risk management and urban planning, susceptibility maps and physically based hazard maps may serve complementary roles; for instance, susceptibility maps facilitate rapid screening of flood-prone areas, whereas physically based hazard maps offer precise quantification of physical characteristics such as inundation depth and duration. We recommend that, in future work, a comparison be made with the Tokyo Flood Hazard Map on the Flood Risk Information System website [32], while further exploring the quantitative relationship between susceptibility index values and predicted water depths.

4.2. Exposure and Vulnerability Mapping

Figure 10 illustrates the spatial distribution of the three indicators used for flood exposure assessment. Population intensity shows a high concentration in the east, with relatively lower density toward the west. The spatial distribution of the vulnerable population proportion is relatively dispersed, with elevated proportions primarily in areas away from the urban center. The publicly announced land price exhibits a similar pattern, with maximum values in the eastern urban center declining toward the west; additionally, some scattered high land price clusters are evident, likely corresponding to subcenter districts and major railway stations. These indicators provide a multidimensional perspective for identifying high-exposure areas. Figure 11 shows the spatial distribution of the five indicators used for flood vulnerability assessment. DEES, serving as immediate escape venues during severe flooding, are typically not located in areas prone to water accumulation or low-lying terrain; this is particularly evident in areas adjacent to rivers in the eastern lowlands. In contrast, DES, which provide temporary accommodation for displaced individuals, display a more even spatial pattern. ETR and FS exhibit density patterns similar to those of the population. DBHB density shows the highest concentration in the urban center.

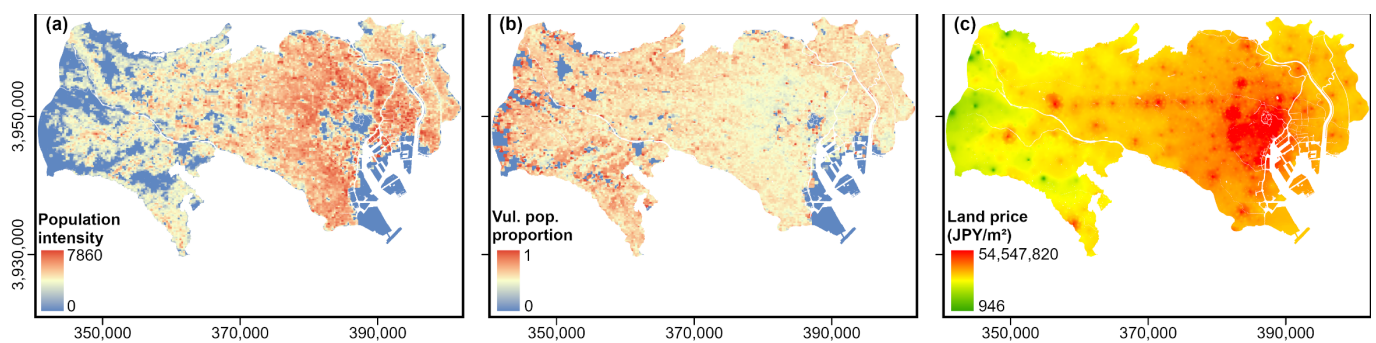


Figure 10. Indicators for flood exposure mapping via Z-FAHP: (a) population intensity, (b) vulnerable population proportion, and (c) the publicly announced land price.

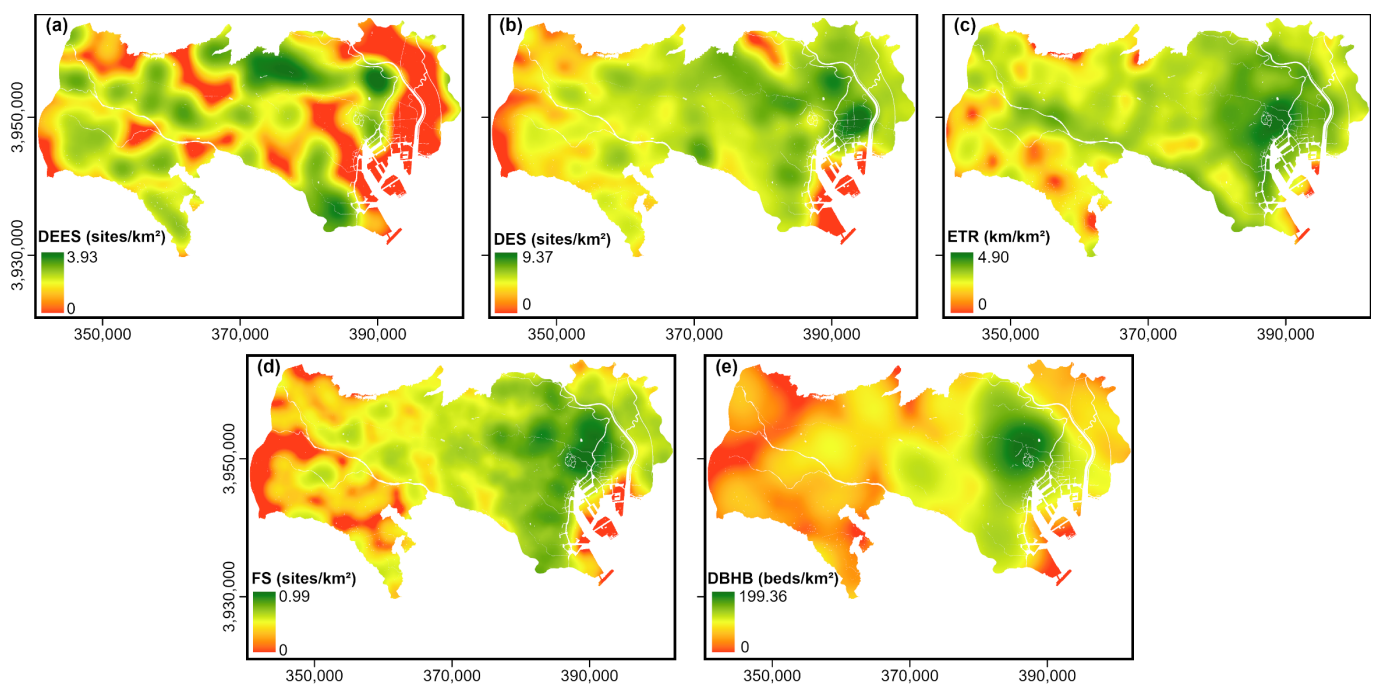


Figure 11. Indicators for flood vulnerability mapping via Z-FAHP: (a) DEES density, (b) DES density, (c) ETR density, (d) FS density, and (e) DBHB density.

The pairwise comparison matrices of indicators for the exposure and vulnerability criteria are presented in Tables 6 and 7, respectively, summarizing the fuzzy judgments on relative importance and the associated confidence. Both matrices passed the consistency check, with CR values below 0.1. The derived weights are given in Tables 8 and 9. For exposure, the vulnerable population proportion receives the largest weight, reflecting the greater need for assistance during emergencies. For vulnerability, DEES density carries the highest weighting, likely because these sites have a direct impact on reducing casualties, while FS density follows because FS provide rescue and emergency medical services. These weights were applied to produce the exposure and vulnerability maps through the SAW method. While exposure was computed directly from the weighted indicators, vulnerability was computed from weighted indicator deficits, maintaining the same weights as their corresponding indicators. The derived flood exposure map (Figure 12a) shows higher exposure predominantly distributed across eastern urban areas, whereas vulnerability in these areas is relatively low (Figure 12b).

Table 6. Pairwise comparison matrix of indicators for exposure criterion.

	Land Price	Population Intensity	Vulnerable Population Proportion
Land Price	(1, 1, 1), (1, 1, 1)	(1, 1, 1), (0.5, 0.6, 0.7)	(1/3, 1/2, 1), (0.6, 0.7, 0.8)
Population intensity	(1, 1, 1), (0.5, 0.6, 0.7)	(1, 1, 1), (1, 1, 1)	(1/3, 1/2, 1), (0.6, 0.7, 0.8)
Vulnerable Population Proportion	(1, 2, 3), (0.6, 0.7, 0.8)	(1, 2, 3), (0.6, 0.7, 0.8)	(1, 1, 1), (1, 1, 1)

Table 7. Pairwise comparison matrix of indicators for vulnerability criterion.

	DES Density	ETR Density	DBHB Density	FS Density	DEES Density
DES Density	(1, 1, 1), (1, 1, 1)	(1, 1, 1), (0.5, 0.6, 0.7)	(1/3, 1/2, 1), (0.5, 0.6, 0.7)	(1/4, 1/3, 1/2), (0.6, 0.7, 0.8)	(1/6, 1/5, 1/4), (0.7, 0.8, 0.9)
ETR Density	(1, 1, 1), (0.5, 0.6, 0.7)	(1, 1, 1), (1, 1, 1)	(1/3, 1/2, 1), (0.5, 0.6, 0.7)	(1/3, 1/2, 1), (0.6, 0.7, 0.8)	(1/4, 1/3, 1/2), (0.7, 0.8, 0.9)
DBHB Density	(1, 2, 3), (0.5, 0.6, 0.7)	(1, 2, 3), (0.5, 0.6, 0.7)	(1, 1, 1), (1, 1, 1)	(1/3, 1/2, 1), (0.6, 0.7, 0.8)	(1/5, 1/4, 1/3), (0.7, 0.8, 0.9)
FS Density	(2, 3, 4), (0.6, 0.7, 0.8)	(1, 2, 3), (0.6, 0.7, 0.8)	(1, 2, 3), (0.6, 0.7, 0.8)	(1, 1, 1), (1, 1, 1)	(1/4, 1/3, 1/2), (0.6, 0.7, 0.8)
DEES Density	(4, 5, 6), (0.7, 0.8, 0.9)	(2, 3, 4), (0.7, 0.8, 0.9)	(3, 4, 5), (0.7, 0.8, 0.9)	(2, 3, 4), (0.6, 0.7, 0.8)	(1, 1, 1), (1, 1, 1)

Table 8. Indicator weights for exposure criterion.

Land Price	Population Intensity	Vulnerable Population Proportion
0.251	0.251	0.497

Table 9. Indicator weights for vulnerability criterion.

DES Density	ETR Density	DBHB Density	FS Density	DEES Density
0.082	0.103	0.143	0.213	0.460

The focus of this study is to develop a framework integrating machine learning and Z-FAHP for urban flood risk assessment, emphasizing the replicable workflow rather than specific indicator selection or weight values. It should be noted that the pairwise comparison matrices applied in this study (Tables 6 and 7) were constructed by the authors as an illustrative demonstration of the Z-FAHP procedure, rather than through a structured expert elicitation process. While this does not affect the validity of the overall

methodological workflow, the specific weights for the indicators should be interpreted with this in mind. Beyond this specific limitation, it should also be recognized that, even when properly elicited, the Z-FAHP method relies on experts' knowledge, which is an inherent characteristic of subjective weighting methods; this introduces a degree of uncertainty, as different expert panels may give varying opinions, and the understanding of flood processes, exposure, and vulnerability is likely to continue evolving over time. In future practical applications, it is recommended that this framework be implemented with a properly constituted panel of experts spanning relevant disciplines, who would participate not only in constructing the pairwise comparison matrices through a documented elicitation and consensus procedure, but also in the indicator screening process, making decisions based on local factors and data availability. In addition, although the Z-FAHP method inherently accounts for expert judgment uncertainty through its reliability component, this study does not include a systematic sensitivity analysis of how variations in the fuzzy judgments on relative importance and the associated confidence levels would propagate to the final flood risk map; such an analysis represents an important direction for future research.

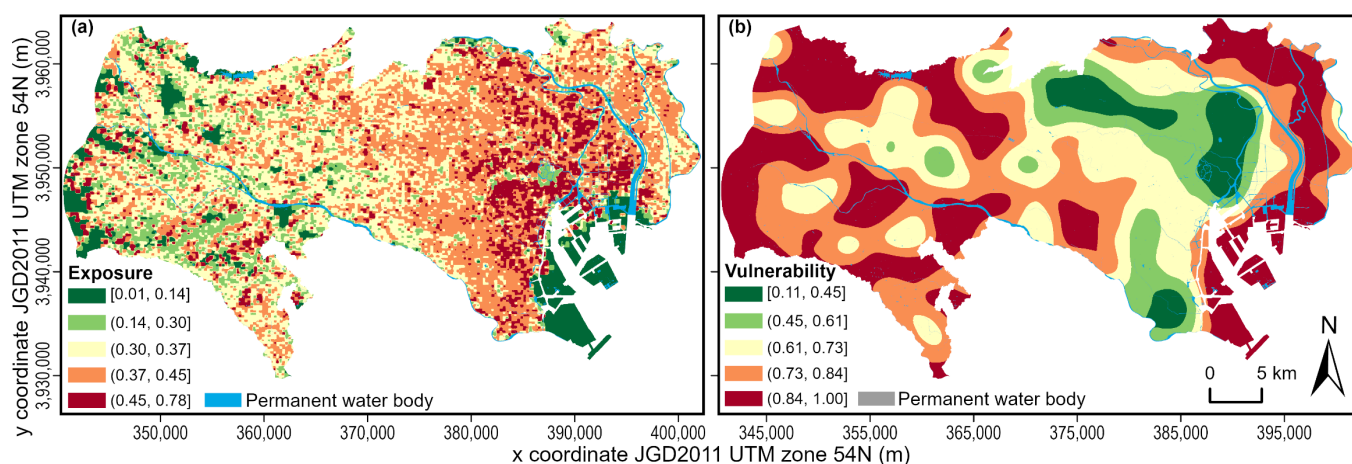


Figure 12. Flood exposure map (a) and flood vulnerability map (b) derived from Z-FAHP. The indices are divided into five intervals based on the natural breaks (Jenks) classification method.

4.3. Flood Risk Mapping

Figure 13 shows the flood risk map synthesizing the three criteria: flood susceptibility, exposure, and vulnerability. It is evident that high- and very-high-risk zones are primarily distributed along river corridors; the eastern lowlands exhibit the highest relative risk, likely due to low-lying terrain and concentrated population and assets. High- and very-high-risk zones comprise approximately 32.9% of the study area (Table 10). Notably, the Tokyo Metropolitan Government has undertaken substantial protective measures for these vulnerable low-lying areas [34,67]. Although major fluvial flooding events have been less frequent in Tokyo due to protective measures, continuous maintenance and upgrading of drainage systems remain essential to mitigate pluvial flood risk. This static flood risk map synthesizes flood susceptibility, exposure, and vulnerability to characterize the relative potential for adverse flood consequences at any given location across the study area. While the Tokyo Metropolitan Government has published predicted inundation depth and duration maps, this flood risk map offers a comprehensive perspective for both the public and decision makers.

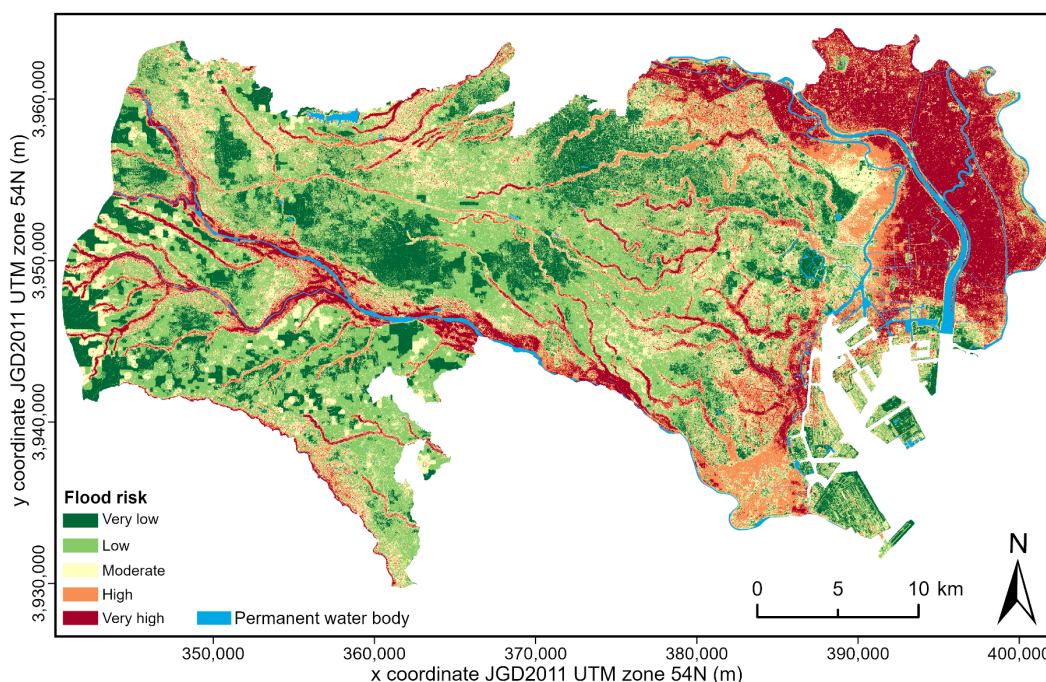


Figure 13. Flood risk map of the study area. The five risk levels are classified via the natural breaks (Jenks) classification method.

Table 10. Risk levels and their corresponding proportions for the study area.

Risk Level	Value Range	Proportion
Very low	(0.097, 0.375]	15.7%
Low	(0.375, 0.449]	30.4%
Moderate	(0.449, 0.538]	20.9%
High	(0.538, 0.641]	16.8%
Very high	(0.641, 0.826]	16.1%

The public can use this map to understand potential flood threats in their surroundings. Residents in flood-prone areas are encouraged to learn necessary life-saving skills for water-related disasters. Particularly, residents living in high- and very-high-risk zones should enhance disaster preparedness awareness by consulting official hazard maps to understand the potential maximum water depth at their residences, and familiarize themselves with evacuation routes and nearby DEES. However, some people are risk-takers while others are risk-averse; flood risk information can at least alert those who are risk-averse, while others may not take it seriously [34].

This flood risk map also provides decision makers with valuable insights for mitigation planning. For instance, areas with greater predicted inundation depth and longer inundation duration do not necessarily have higher flood risk, and vice versa. Future work can also proceed in the following directions. First, given the uncertainties associated with demographic change, urban development, and climate change, periodic reassessment of flood risk remains important. Second, investigating the relationships among flood risk, susceptibility, exposure, vulnerability, and predicted inundation depths and durations may yield important insights for adaptive flood risk management. Third, comparison of the flood risk map against independent flood records, insurance, or damage data could be pursued as such data become accessible. Fourth, while equal weights were assigned to susceptibility, exposure, and vulnerability in this study, applying different weights to these criteria may yield variations in the final risk map. Future studies can attempt to determine weights via expert elicitation and investigate their sensitivity.

5. Conclusions

We propose, for the first time, a transferable and efficient assessment framework integrating machine learning and Z-FAHP for urban flood risk mapping. Taking Tokyo, Japan as a case study, we developed a flood risk map with 10-m spatial resolution that explicitly accounts for socioeconomic characteristics, based entirely on publicly available datasets. This framework utilizes flood susceptibility derived from machine learning as a proxy for flood hazard, enabling efficient assessment. In evaluating exposure and vulnerability, it incorporates not only the expert fuzzy judgment on relative importance but also the confidence level associated with that judgment. The results demonstrate that the CatBoost model outperformed the ANN model across all five evaluation metrics (accuracy, precision, recall, F1-score, and ROC-AUC) for flood susceptibility mapping. The resulting map shows that high- and very-high-risk zones, covering approximately 32.9% of the study area, are primarily distributed along river corridors and in the eastern lowlands. The derived flood risk map can help the public understand local flood risk levels in their residential areas and enhance disaster preparedness, while providing decision makers with a tool to integrate with existing inundation depth and duration maps for more effective mitigation planning. The transferable and efficient assessment framework can likewise be applied to other cities globally as an effective tool for urban flood risk management, allowing flexible adjustment of features and indicators based on local data availability.

Author Contributions: Conceptualization, S.L.; methodology, S.L.; software, S.L.; validation, S.L. and T.K.; formal analysis, S.L.; investigation, S.L. and T.K.; resources, S.L. and T.K.; data curation, S.L.; writing—original draft preparation, S.L.; writing—review and editing, S.L. and T.K.; visualization, S.L.; supervision, T.K.; project administration, S.L. and T.K.; funding acquisition, S.L. and T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JST SPRING, Japan Grant Number JPMJSP2106.

Data Availability Statement: Section 3.2 summarizes all the data used in this study and their sources, all of which are publicly available.

Acknowledgments: This work was supported by JST SPRING, Japan Grant Number JPMJSP2106; the authors gratefully acknowledge this support. We also thank the anonymous reviewers for their valuable time.

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

Abbreviations

The following abbreviations are used in this manuscript:

AHP	Analytic Hierarchy Process
ANN	Artificial Neural Network
ANP	Analytic Network Process
AUC	Area Under the Curve
CR	Consistency Ratio
DBHB	Disaster Base Hospital Beds
DEES	Designated Emergency Evacuation Sites
DEM	Digital Elevation Model
DES	Designated Evacuation Shelters
ESA	European Space Agency
ETR	Emergency Transportation Roads
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FS	Fire Stations

GDP	Gross Domestic Product
GIS	Geographic Information System
GSJ	Geospatial Information Authority of Japan
JMA	Japan Meteorological Agency
MAP	Mean Annual Precipitation
MCDA	Multi-Criteria Decision Analysis
MCDM	Multi-Criteria Decision-Making
MDP	Maximum Daily Precipitation
MHP	Maximum Hourly Precipitation
MLIT	Ministry of Land, Infrastructure, Transport and Tourism
NDVI	Normalized Difference Vegetation Index
RI	Random Index
ROC	Receiver Operating Characteristic
SAW	Simple Additive Weighting
SHAP	SHapley Additive exPlanations
TN	True Negatives
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TP	True Positives
TRI	Terrain Ruggedness Index
TWI	Topographic Wetness Index
XAI	eXplainable Artificial Intelligence
Z-FAHP	Z-number-based Fuzzy Analytic Hierarchy Process

Appendix A

TRI is an indicator that quantifies the amount of elevation difference between a central cell and its eight neighboring cells of a DEM [68]. A higher TRI value indicates more rugged terrain, whereas a lower value represents flatter surfaces. TWI is an indicator that quantifies the influence of topography on soil moisture, based on upstream contributing area and local slope [69,70]. A higher TWI value indicates a greater likelihood of water accumulation. The equations are given as follows:

$$TRI = \sqrt{\frac{1}{8} \sum_{i=1}^8 (Z_i - Z_0)^2} \quad (A1)$$

$$TWI = \ln \left(\frac{\alpha}{\tan(\beta)} \right) \quad (A2)$$

where Z_0 is the elevation of the central cell, and Z_i is the elevation of each of the eight neighboring cells; α is the local upslope contributing area draining through a certain cell per unit contour length, and β is the local slope at that cell.

References

1. Bloch, R.; Jha, A.K.; Lamond, J. *Cities and Flooding: A Guide to Integrated Urban Flood Risk Management for the 21st Century*; World Bank: Washington, DC, USA, 2012.
2. Agonafir, C.; Lakhankar, T.; Khanbilvardi, R.; Krakauer, N.; Radell, D.; Devineni, N. A Review of Recent Advances in Urban Flood Research. *Water Secur.* **2023**, *19*, 100141. [\[CrossRef\]](#)
3. O'Donnell, E.C.; Thorne, C.R. Drivers of Future Urban Flood Risk. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2020**, *378*, 20190216. [\[CrossRef\]](#)
4. Koks, E.E.; Jongman, B.; Husby, T.G.; Botzen, W.J.W. Combining Hazard, Exposure and Social Vulnerability to Provide Lessons for Flood Risk Management. *Environ. Sci. Policy* **2015**, *47*, 42–52. [\[CrossRef\]](#)
5. Kron, W. Flood Risk = Hazard · Values · Vulnerability. *Water Int.* **2005**, *30*, 58–68. [\[CrossRef\]](#)
6. Winsemius, H.C.; Van Beek, L.P.H.; Jongman, B.; Ward, P.J.; Bouwman, A. A Framework for Global River Flood Risk Assessments. *Hydrol. Earth Syst. Sci.* **2013**, *17*, 1871–1892. [\[CrossRef\]](#)

7. Reisinger, A.; Howden, M.; Vera, C.; Garschagen, M.; Hurlbert, M.; Kreibiehl, S.; Mach, K.J.; Mintenbeck, K.; O'Neill, B.; Pathak, M.; et al. *The Concept of Risk in the IPCC Sixth Assessment Report: A Summary of Cross-Working Group Discussions*; Intergovernmental Panel on Climate Change: Geneva, Switzerland, 2020.
8. Pörtner, H.-O.; Roberts, D.C.; Poloczanska, E.S.; Mintenbeck, K.; Tignor, M.; Alegría, A.; Craig, M.; Langsdorf, S.; Löschke, S.; Möller, V.; et al. (Eds.) IPCC Summary for Policymakers. In *Climate Change 2022—Impacts, Adaptation and Vulnerability. Contribution of Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2022; pp. 3–33.
9. Liu, W.; Feng, Q.; Engel, B.A.; Yu, T.; Zhang, X.; Qian, Y. A Probabilistic Assessment of Urban Flood Risk and Impacts of Future Climate Change. *J. Hydrol.* **2023**, *618*, 129267. [\[CrossRef\]](#)
10. Liu, Y.; Bates, P.; Neal, J.C.; Slater, L.; Zhao, J.; Ben-Haim, Z. A Bare-Earth GoogleDEM to Simulate Flooding in New Delhi, India. *Water Resour. Res.* **2025**, *61*, e2024WR038577. [\[CrossRef\]](#)
11. Jamali, B.; Bach, P.M.; Cunningham, L.; Deletic, A. A Cellular Automata Fast Flood Evaluation (CA-ffé) Model. *Water Resour. Res.* **2019**, *55*, 4936–4953. [\[CrossRef\]](#)
12. Zhao, G.; Pang, B.; Xu, Z.; Peng, D.; Xu, L. Assessment of Urban Flood Susceptibility Using Semi-Supervised Machine Learning Model. *Sci. Total Environ.* **2019**, *659*, 940–949. [\[CrossRef\]](#)
13. Choubin, B.; Jaafari, A.; Henareh, J.; Karimi, O.; Sajedi Hosseini, F. Explainable Artificial Intelligence (XAI) for Interpreting Predictive Models and Key Variables in Flood Susceptibility. *Results Eng.* **2025**, *27*, 105976. [\[CrossRef\]](#)
14. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
15. Zounemat-Kermani, M.; Kheimi, M. Explainable Artificial Intelligence in Hydrology: A Review. *Water Resour. Manag.* **2026**, *40*, 106. [\[CrossRef\]](#)
16. Pugliese Vilorio, A.d.J.; Folini, A.; Carrion, D.; Brovelli, M.A. Hazard Susceptibility Mapping with Machine and Deep Learning: A Literature Review. *Remote Sens.* **2024**, *16*, 3374. [\[CrossRef\]](#)
17. Qin, X.; Wang, S.; Meng, M.; Long, H.; Zhang, H.; Shi, H. Enhancing Urban Resilience through Machine Learning-Supported Flood Risk Assessment: Integrating Flood Susceptibility with Building Function Vulnerability. *npj Urban Sustain.* **2025**, *5*, 19. [\[CrossRef\]](#)
18. Dey, H.; Shao, W.; Haque, M.M.; VanDyke, M. Enhancing Flood Risk Analysis in Harris County: Integrating Flood Susceptibility and Social Vulnerability Mapping. *J. Geovis. Spat. Anal.* **2024**, *8*, 19. [\[CrossRef\]](#)
19. Pham, B.T.; Luu, C.; Phong, T.V.; Nguyen, H.D.; Le, H.V.; Tran, T.Q.; Ta, H.T.; Prakash, I. Flood Risk Assessment Using Hybrid Artificial Intelligence Models Integrated with Multi-Criteria Decision Analysis in Quang Nam Province, Vietnam. *J. Hydrol.* **2021**, *592*, 125815. [\[CrossRef\]](#)
20. Liu, Y.; Liu, L.; Sun, H.; Chen, B.; Ma, X.; Ning, Y.; Qi, S. Flood Risk Assessment Combining Machine Learning with Multi-Criteria Decision Analysis in Jiangxi Province, China. *Int. J. Disaster Risk Sci.* **2025**, *16*, 858–869. [\[CrossRef\]](#)
21. Abdullah, M.F.; Siraj, S.; Hodgett, R.E. An Overview of Multi-Criteria Decision Analysis (MCDA) Application in Managing Water-Related Disaster Events: Analyzing 20 Years of Literature for Flood and Drought Events. *Water* **2021**, *13*, 1358. [\[CrossRef\]](#)
22. De Brito, M.M.; Evers, M. Multi-Criteria Decision-Making for Flood Risk Management: A Survey of the Current State of the Art. *Nat. Hazards Earth Syst. Sci.* **2016**, *16*, 1019–1033. [\[CrossRef\]](#)
23. Liu, K.; Kinouchi, T.; Tan, R.; Heng, S.; Chhuon, K.; Zhao, W. Unraveling Urban Hydro-Environmental Response to Climate Change and MCDA-Based Area Prioritization in a Data-Scarce Developing City. *Sci. Total Environ.* **2024**, *948*, 174389. [\[CrossRef\]](#)
24. Saaty, R.W. The Analytic Hierarchy Process—What It Is and How It Is Used. *Math. Model.* **1987**, *9*, 161–176. [\[CrossRef\]](#)
25. Wang, G.; Liu, L.; Shi, P.; Zhang, G.; Liu, J. Flood Risk Assessment of Metro System Using Improved Trapezoidal Fuzzy AHP: A Case Study of Guangzhou. *Remote Sens.* **2021**, *13*, 5154. [\[CrossRef\]](#)
26. Cai, T.; Li, X.; Ding, X.; Wang, J.; Zhan, J. Flood Risk Assessment Based on Hydrodynamic Model and Fuzzy Comprehensive Evaluation with GIS Technique. *Int. J. Disaster Risk Reduct.* **2019**, *35*, 101077. [\[CrossRef\]](#)
27. Gao, Y.; Lu, H.; Zhang, Y.; Jin, H.; Wu, S.; Gao, Y.; Zhang, S. Evaluating Yangtze River Delta Urban Agglomeration Flood Risk Using a Hybrid Method of Automated Machine Learning and Analytic Hierarchy Process. *Nat. Hazards Earth Syst. Sci.* **2025**, *25*, 3087–3108. [\[CrossRef\]](#)
28. Liu, Y.; Eckert, C.M.; Earl, C. A Review of Fuzzy AHP Methods for Decision-Making with Subjective Judgements. *Expert Syst. Appl.* **2020**, *161*, 113738. [\[CrossRef\]](#)
29. Zadeh, L.A. A Note on Z-Numbers. *Inf. Sci.* **2011**, *181*, 2923–2932. [\[CrossRef\]](#)
30. Anbari Moghadam, M.; Bagherpour, M. A Framework to Assess and Report Social, Environmental, and Economic Post-Disaster Damages Based on Z-Numbers and the Delphi Method. *Nat. Hazards* **2024**, *120*, 14459–14494. [\[CrossRef\]](#)
31. Nourani, V.; Najafi, H.; Maleki, S.; Jabbarian Paknezad, N.; Jeanne Huang, J.; Zhang, P.; Mohammadisepasi, S. Z-Number Based Assessment of Groundwater Vulnerability to Seawater Intrusion. *J. Hydrol.* **2024**, *632*, 130859. [\[CrossRef\]](#)

32. Tokyo Bureau of Construction. Flood Risk Information System. Available online: <https://www.suigai-risk.metro.tokyo.lg.jp/> (accessed on 28 August 2025).
33. JAXA EORC. High-Resolution Land-Use and Land-Cover Map of Japan [2024-2022-2020] (Released in April 2025/Version 25.04). Available online: https://www.eorc.jaxa.jp/ALOS/en/dataset/lulc/lulc_v2504_e.htm (accessed on 20 August 2025).
34. Sado-Inamura, Y.; Fukushima, K. Empirical Analysis of Flood Risk Perception Using Historical Data in Tokyo. *Land Use Policy* **2019**, *82*, 13–29. [CrossRef]
35. Zerouali, B.; Almaliki, A.H.; Santos, C.A.G. Flood Susceptibility Mapping in Arid Urban Areas Using SHAP-Enhanced Stacked Ensemble Learning: A Case Study of Jeddah. *J. Environ. Manag.* **2025**, *393*, 127128. [CrossRef]
36. Ren, H.; Pang, B.; Bai, P.; Zhao, G.; Liu, S.; Liu, Y.; Li, M. Flood Susceptibility Assessment with Random Sampling Strategy in Ensemble Learning (RF and XGBoost). *Remote Sens.* **2024**, *16*, 320. [CrossRef]
37. Wei, Q.; Zhang, H.; Chen, Y.; Xie, Y.; Yin, H.; Xu, Z. City Scale Urban Flooding Risk Assessment Using Multi-Source Data and Machine Learning Approach. *J. Hydrol.* **2025**, *651*, 132626. [CrossRef]
38. GSI. High-Resolution Elevation Dataset. Available online: <https://service.gsi.go.jp/kiban/> (accessed on 18 August 2025).
39. Zanaga, D.; Van De Kerchove, R.; De Keersmaecker, W.; Souverijns, N.; Brockmann, C.; Quast, R.; Wevers, J.; Grosu, A.; Paccini, A.; Vergnaud, S.; et al. *ESA WorldCover 10 m 2020 V100*; Zenodo: Geneva, Switzerland, 2021. [CrossRef]
40. JMA. Historical Meteorological Data. Available online: <https://www.data.jma.go.jp/stats/etrn/index.php> (accessed on 19 August 2025).
41. MLIT. Water Information System. Available online: <https://www1.river.go.jp/>.
42. OpenStreetMap Contributors. OpenStreetMap. Available online: <https://planet.openstreetmap.org> (accessed on 24 July 2025).
43. MLIT. The 1:200,000 Land Classification Basic Survey. Available online: <https://nlftp.mlit.go.jp/kokjo/inspect/landclassification/download.html> (accessed on 30 July 2025).
44. MLIT. Future Projected Population Data for 250-Meter Grid Cells. Available online: <https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-mesh250r6.html> (accessed on 13 September 2025).
45. MLIT. Land Price Public Notice: 2020 (Reiwa 2) Edition. Available online: <https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-L01-2020.html> (accessed on 13 September 2025).
46. Moreira, L.L.; de Brito, M.M.; Kobiyama, M. Review Article: A Systematic Review and Future Prospects of Flood Vulnerability Indices. *Nat. Hazards Earth Syst. Sci.* **2021**, *21*, 1513–1530. [CrossRef]
47. GSI. Dataset of Designated Emergency Evacuation Sites and Evacuation Shelters. Available online: <https://www.gsi.go.jp/bousaichiri/hinanbasho.html> (accessed on 3 September 2025).
48. Tokyo Fire Department. List of Tokyo Fire Department Headquarters and Fire Station Locations. Available online: <https://catalog.data.metro.tokyo.lg.jp/en/dataset/t000017d00000000001> (accessed on 13 September 2025).
49. MLIT. Emergency Transportation Road Data: 2020 (Reiwa 2) Edition. Available online: <https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N10-2020.html> (accessed on 10 September 2025).
50. MLIT. Medical Institution Data: 2020 (Reiwa 2) Edition. Available online: <https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P04-2020.html> (accessed on 13 September 2025).
51. Kaiser, M.; Günnemann, S.; Disse, M. Regional-Scale Prediction of Pluvial and Flash Flood Susceptible Areas Using Tree-Based Classifiers. *J. Hydrol.* **2022**, *612*, 128088. [CrossRef]
52. Saber, M.; Boulmaiz, T.; Guermoui, M.; Abdrabo, K.I.; Kantoush, S.A.; Sumi, T.; Boutaghane, H.; Hori, T.; Binh, D.V.; Nguyen, B.Q.; et al. Enhancing Flood Risk Assessment through Integration of Ensemble Learning Approaches and Physical-Based Hydrological Modeling. *Geomat. Nat. Hazards Risk* **2023**, *14*, 2203798. [CrossRef]
53. Portalés-Julià, E.; Mateo-García, G.; Gómez-Chova, L. Understanding Flood Detection Models across Sentinel-1 and Sentinel-2 Modalities and Benchmark Datasets. *Remote Sens. Environ.* **2025**, *328*, 114882. [CrossRef]
54. Schumann, G.J.-P.; Vernieuwe, H.; De Baets, B.; Verhoest, N.E.C. ROC-based Calibration of Flood Inundation Models. *Hydrol. Process.* **2014**, *28*, 5495–5502. [CrossRef]
55. Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
56. Bennett, N.D.; Croke, B.F.W.; Guariso, G.; Guillaume, J.H.A.; Hamilton, S.H.; Jakeman, A.J.; Marsili-Libelli, S.; Newham, L.T.H.; Norton, J.P.; Perrin, C.; et al. Characterising Performance of Environmental Models. *Environ. Model. Softw.* **2013**, *40*, 1–20. [CrossRef]
57. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837. [CrossRef]
58. Buckley, J.J. Fuzzy Hierarchical Analysis. *Fuzzy Sets Syst.* **1985**, *17*, 233–247. [CrossRef]
59. Kang, B.; Wei, D.; Li, Y.; Deng, Y. A Method of Converting Z-Number to Classical Fuzzy Number. *J. Inf. Comput. Sci.* **2012**, *9*, 703–709.
60. Saaty, T.L. A Scaling Method for Priorities in Hierarchical Structures. *J. Math. Psychol.* **1977**, *15*, 234–281. [CrossRef]
61. Saaty, T.L. *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*; McGraw-Hill: New York, NY, USA, 1980.

62. Han, Y.; Semnani, S.J. Important Considerations in Machine Learning-Based Landslide Susceptibility Assessment under Future Climate Conditions. *Acta Geotech.* **2025**, *20*, 475–500. [[CrossRef](#)]
63. Vemula, S.; Gatti, F.; Jehel, P. Graph Transformer-Based Flood Susceptibility Mapping: Application to the French Riviera and Railway Infrastructure under Climate Change. *J. Hydrol. Reg. Stud.* **2026**, *66*, 103644. [[CrossRef](#)]
64. Husic, A.; Hammond, J.; Price, A.N.; Roundy, J.K. Interrogating Process Deficiencies in Large-Scale Hydrologic Models with Interpretable Machine Learning. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 4457–4472. [[CrossRef](#)]
65. Zhao, W.; Abhishek; Takhellambam, B.S.; Zhang, J.; Zhao, Y.; Kinouchi, T. Spatiotemporal Variability of Current and Future Sub-Daily Rainfall in Japan Using State-Of-The-Art High-Quality Data Sets. *Water Resour. Res.* **2023**, *59*, e2022WR034305. [[CrossRef](#)]
66. Bentivoglio, R.; Isufi, E.; Jonkman, S.N.; Taormina, R. Deep Learning Methods for Flood Mapping: A Review of Existing Applications and Future Research Directions. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 4345–4378. [[CrossRef](#)]
67. Cao, A.; Esteban, M.; Onuki, M. Public Support for Flood Adaptation Policy in Tokyo Lowland Areas. *Clim. Policy* **2024**, *24*, 1275–1292. [[CrossRef](#)]
68. Riley, S.J.; DeGloria, S.D.; Elliot, R. A Terrain Ruggedness Index That Quantifies Topographic Heterogeneity. *Intermt. J. Sci.* **1999**, *5*, 23–27.
69. Buchanan, B.P.; Fleming, M.; Schneider, R.L.; Richards, B.K.; Archibald, J.; Qiu, Z.; Walter, M.T. Evaluating Topographic Wetness Indices across Central New York Agricultural Landscapes. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 3279–3299. [[CrossRef](#)]
70. Beven, K.J.; Kirkby, M.J. A Physically Based, Variable Contributing Area Model of Basin Hydrology/Un Modèle à Base Physique de Zone d'appel Variable de l'hydrologie Du Bassin Versant. *Hydrol. Sci. Bull.* **1979**, *24*, 43–69. [[CrossRef](#)]

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.