Journal of Sustainable Development and Policy

Volume: 13; **Issue:** 7
Pages: 01-46

Multi-Source Remote Sensing and Dem-Derived Hydrogeomorphic Indicators for High-Resolution Urban Flood Susceptibility Mapping: A Spatial Cross-Validation Framework

Sumaiya Sadmeen Misty[1]; Kazi Mohammad Khalid Ahsan[2];

[1]. *National Consultant (Junior Urban Planner), Urban Development Directorate (UDD), Dhaka, Bangladesh; Email: sumaiya_2k10@yahoo.com*

[2]. *Executive Engineer, Dhaka Water Supply and Sewerage Authority (Dhaka WASA), Dhaka, Bangladesh; Email: khalid.grid@gmail.com*

Doi: 10.63125/z4xt7k38
Received: *25 April 2023;* ***Revised:*** *29 May 2023;* ***Accepted:*** *11 June 2023;* ***Published:*** *02 June 2023;*

Abstract

Urban flooding represents a spatially complex hazard influenced by interactions among terrain configuration, drainage connectivity, land-surface modification, and hydroclimatic conditions. This study developed and quantitatively evaluated a high-resolution urban flood susceptibility mapping framework integrating multi-source remote sensing with DEM-derived hydrogeomorphic indicators under geographically explicit spatial cross-validation. The analysis covered 1,286.4 km² at 10 m spatial resolution and incorporated 1,284 verified flood locations and 1,284 spatially screened non-flood observations, producing a balanced analytical sample of 2,568 locations with 99.3% geospatial data completeness. Seventeen predictors were retained following quality, correlation, and multicollinearity screening. Flood occurrence exhibited significant spatial clustering, with Moran's I = 0.41 ($p < 0.001$). HAND was the most influential predictor, contributing 13.8% mean importance, followed by distance to drainage at 12.9%, flow accumulation at 11.7%, elevation at 10.8%, and TWI at 9.6%. Logistic regression, Support Vector Machine, Random Forest, gradient boosting, XGBoost, and ensemble models were comparatively evaluated. Under conventional random validation, the ensemble achieved 91.7% accuracy, 0.917 F1-score, 0.972 ROC-AUC, and 0.969 PR-AUC. Spatial cross-validation produced more conservative estimates of 85.6% accuracy, 0.856 F1-score, 0.914 ROC-AUC, and 0.907 PR-AUC. The ROC-AUC reduction was statistically significant ($p = 0.001$) with a large effect size of Cohen's $d = 1.26$. Resolution analysis showed that ROC-AUC decreased from 0.914 at 10 m to 0.901 at 20 m and 0.886 at 30 m. Removing hydrogeomorphic predictors further reduced ROC-AUC to 0.852. The final ensemble produced a Brier score of 0.112, calibration slope of 0.93, residual Moran's I of 0.07, exact susceptibility-class agreement of 84.7%, and adjacent-class agreement of 96.1%. The findings demonstrated that integrating high-resolution hydrogeomorphic and multi-source remote-sensing information with spatially explicit validation provided a robust approach for distinguishing predictive accuracy from genuine geographic generalization in urban flood susceptibility mapping.

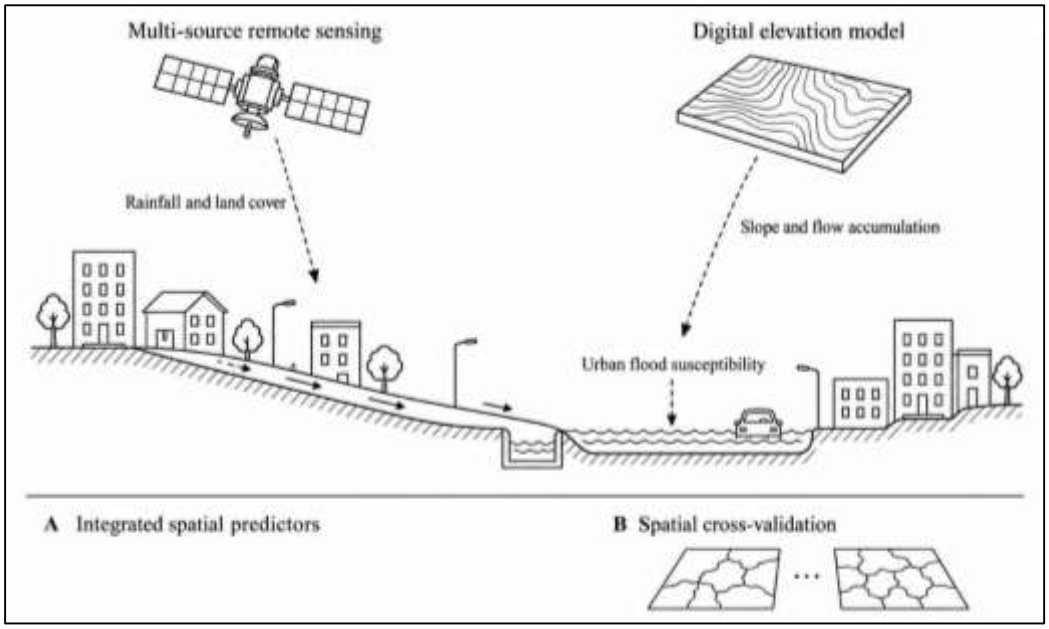
Keywords

Urban Flooding; Remote Sensing; Hydrogeomorphology; Spatial Cross-Validation; Susceptibility Mapping.

INTRODUCTION

Urban flooding can be defined as the temporary accumulation or movement of water across built environments when rainfall-generated runoff, river overflow, limited drainage capacity, surface sealing, or local topographic conditions exceed the ability of natural and engineered systems to convey or store water (Ruan et al., 2021). Flood susceptibility represents the relative spatial tendency of an area to experience flooding according to its hydrological, topographic, environmental, and anthropogenic characteristics, while urban flood susceptibility mapping quantitatively identifies and classifies locations according to their relative flood proneness. This issue has substantial international significance because rapid urbanization, increasing impervious surfaces, dense infrastructure, altered drainage networks, and concentrated populations have increased the importance of accurately identifying flood-prone areas across cities in Asia, Africa, Europe, North America, and other regions. More than 30 previous studies have demonstrated that urban flood susceptibility is controlled by interacting characteristics rather than a single environmental factor, with elevation, slope, rainfall, land use and land cover, drainage density, distance to rivers, vegetation, built-up surfaces, soil properties, flow accumulation, and topographic wetness repeatedly identified as important conditioning variables.

Figure 1: Urban flood susceptibility framework



Multi-source remote sensing provides a quantitative basis for representing these heterogeneous conditions because different Earth-observation systems capture complementary characteristics of the urban environment (Lim & Lee, 2018). Optical satellite imagery, including Landsat and Sentinel-2, supports the extraction of vegetation, surface-water, moisture, land-cover, and built-up characteristics through spectral bands and indices such as the Normalized Difference Vegetation Index, Normalized Difference Water Index, Modified Normalized Difference Water Index, and built-up indices. Synthetic aperture radar observations, particularly Sentinel-1, provide complementary microwave information concerning surface conditions and water occurrence, including observations under cloud-affected conditions. Satellite rainfall products, land-cover datasets, soil information, and other geospatial sources further extend the representation of hydroclimatic and environmental controls. Digital elevation models provide an additional and particularly important information source because urban runoff and water accumulation are strongly influenced by gravitational and terrain-controlled processes. High-resolution analysis is essential in this context because susceptibility can vary considerably across short distances between streets, residential blocks, industrial areas, drainage

corridors, wetlands, open surfaces, and local depressions (Roy et al., 2021). Consequently, integrating heterogeneous remotely sensed observations within a common spatial framework requires systematic preprocessing involving coordinate-system harmonization, spatial resampling, temporal correspondence, normalization, missing-data treatment, and assessment of predictor redundancy. Multi-source integration therefore provides a multidimensional quantitative representation of the urban landscape in which spectral, hydrological, meteorological, topographic, and anthropogenic characteristics can be simultaneously examined in relation to observed flood occurrence.

DEM-derived hydrogeomorphic indicators provide a quantitative representation of the terrain processes governing runoff generation, concentration, routing, and accumulation and therefore constitute a central component of high-resolution urban flood susceptibility modelling. Elevation identifies relatively low-lying locations where water can accumulate, while slope characterizes terrain gradients affecting runoff velocity and local storage (Liu et al., 2021). Curvature represents surface convergence and divergence, flow direction describes potential runoff pathways, and flow accumulation identifies locations receiving runoff from larger contributing areas. The Topographic Wetness Index represents the combined influence of contributing area and local slope on potential moisture accumulation, while the Stream Power Index characterizes concentrated-flow energy. Topographic Position Index, terrain ruggedness, drainage density, distance to drainage, and Height Above Nearest Drainage provide additional measures of relative terrain position, surface variability, drainage organization, hydrological proximity, and vertical separation from drainage networks. More than 30 international and regional studies using statistical, machine-learning, ensemble, and deep-learning methods have established the quantitative relevance of these variables for flood susceptibility assessment. Methods have included frequency ratio, weights of evidence, logistic regression, analytical hierarchy procedures, Random Forest, Support Vector Machine, artificial neural networks, decision trees, gradient boosting, XGBoost, multilayer perceptrons, convolutional neural networks, and ensemble architectures. Such algorithms are particularly applicable to urban flooding because relationships between susceptibility and conditioning variables are frequently nonlinear and interactive; for example, high susceptibility may emerge where low elevation, limited slope, substantial flow accumulation, high wetness, dense impervious development, intense rainfall, and proximity to drainage occur simultaneously. Model assessment commonly employs accuracy, precision, recall, specificity, F1-score, receiver operating characteristic curves, area under the ROC curve, and related classification measures. However, the spatial dependence inherent in remotely sensed and DEM-derived data creates an important methodological consideration (Das, 2020). Neighboring raster cells commonly contain similar elevation, terrain, land-cover, rainfall, drainage, and spectral characteristics, meaning that conventional random training and testing partitions can place geographically adjacent and highly similar observations in both datasets. Spatial cross-validation addresses this condition by organizing observations into geographically separated blocks, clusters, or folds and repeatedly evaluating models on spatial areas excluded from model training. This framework provides a geographically structured assessment of predictive performance and permits examination of variation in accuracy among different urban sectors. Accordingly, this study integrates multi-source remote sensing information with DEM-derived hydrogeomorphic indicators within a high-resolution quantitative framework in which observed flood and non-flood locations provide response information and harmonized spectral, topographic, hydrological, environmental, and urban-surface variables constitute the predictor structure. Spatial cross-validation forms the principal validation framework, allowing model discrimination, classification performance, and geographic stability to be assessed while explicitly accounting for the spatial organization of urban flood susceptibility data (Cabrera & Lee, 2019).

The primary objective of this study is to develop and quantitatively evaluate a high-resolution urban flood susceptibility mapping framework through the integrated use of multi-source remote sensing data and DEM-derived hydrogeomorphic indicators under a spatial cross-validation strategy.

Specifically, the study aims to construct a comprehensive geospatial database by harmonizing complementary information derived from optical and radar satellite observations, precipitation data, land-use and land-cover information, and digital elevation models at a consistent spatial resolution. The analysis seeks to quantify the relationships between observed flood occurrence and major flood-conditioning factors, including elevation, slope, curvature, flow direction, flow accumulation, drainage density, distance to drainage networks, Topographic Wetness Index, Stream Power Index, Topographic Position Index, terrain ruggedness, Height Above Nearest Drainage, rainfall, vegetation conditions, surface-water characteristics, built-up intensity, and impervious land-cover patterns. A further objective is to examine the relative contribution and statistical importance of these predictors in distinguishing flood-susceptible and comparatively less susceptible urban locations while identifying potential multicollinearity and redundancy among correlated geospatial variables. The study also aims to develop and comparatively assess quantitative susceptibility models capable of representing nonlinear and interactive relationships among remotely sensed, hydrogeomorphic, hydrological, and urban-surface characteristics. Model performance is intended to be evaluated using multiple classification and discrimination measures, including accuracy, precision, recall, specificity, F1-score, receiver operating characteristic area under the curve (ROC-AUC), precision-recall area under the curve (PR-AUC), and associated uncertainty measures rather than relying on a single performance statistic. Particular emphasis is placed on determining how validation design influences estimated predictive performance by comparing geographically structured spatial cross-validation with conventional random data partitioning. Through spatial blocks or geographically separated folds, the study seeks to quantify model performance when validation locations are spatially independent from training locations and to examine variability in predictive accuracy across different sections of the urban landscape. In addition, the research aims to generate a continuous high-resolution flood susceptibility surface and classify the resulting predictions into meaningful susceptibility levels while examining the spatial distribution of predicted flood-prone areas in relation to terrain, drainage, land-cover, and remotely sensed environmental conditions. Overall, the objective is to establish a reproducible quantitative framework that integrates multi-source Earth-observation information, hydrogeomorphic terrain analysis, predictive modelling, and spatially explicit validation for rigorous assessment of urban flood susceptibility.

LITERATURE REVIEW

The literature on urban flood susceptibility mapping has developed across remote sensing, hydrology, geomorphology, geographic information science, spatial statistics, and machine learning, creating a multidisciplinary quantitative foundation for identifying locations with different levels of flood susceptibility. Existing studies have examined flood occurrence through combinations of topographic, hydrogeomorphic, meteorological, spectral, land-surface, drainage, and anthropogenic variables derived from satellite imagery, digital elevation models, rainfall products, land-use databases, and field or historical flood inventories. Particular attention has been given to elevation, slope, curvature, flow accumulation, drainage density, distance to rivers and drainage networks, Topographic Wetness Index, Stream Power Index, Topographic Position Index, Height Above Nearest Drainage, rainfall, vegetation indices, water indices, built-up indices, imperviousness, and land-use/land-cover characteristics. Quantitative flood susceptibility research has also progressed from conventional GIS-based weighting and statistical approaches toward machine-learning and ensemble algorithms capable of representing nonlinear relationships and interactions among numerous conditioning variables. Random Forest, Support Vector Machine, logistic regression, artificial neural networks, gradient boosting, XGBoost, convolutional neural networks, and ensemble approaches have consequently become important components of contemporary susceptibility assessment. However, substantial methodological variation remains across studies in spatial resolution, remote-sensing sources, DEM products, flood-inventory size, predictor selection, negative-sample generation, model configuration, susceptibility classification, and validation procedures. A particularly important quantitative issue concerns spatial dependence, since conventional random training-test partitions can place geographically adjacent

observations in both subsets and produce performance estimates that differ from those obtained when models are tested in geographically separated areas. Accordingly, this literature review synthesizes the empirical evidence through eight interconnected quantitative themes covering multi-source remote-sensing characteristics, DEM-derived hydrogeomorphic indicators, spectral and urban-surface predictors, flood inventories and sampling structures, statistical and machine-learning models, spatial-resolution effects, conventional versus spatial validation, and integrated evidence concerning model robustness and geographic transferability.

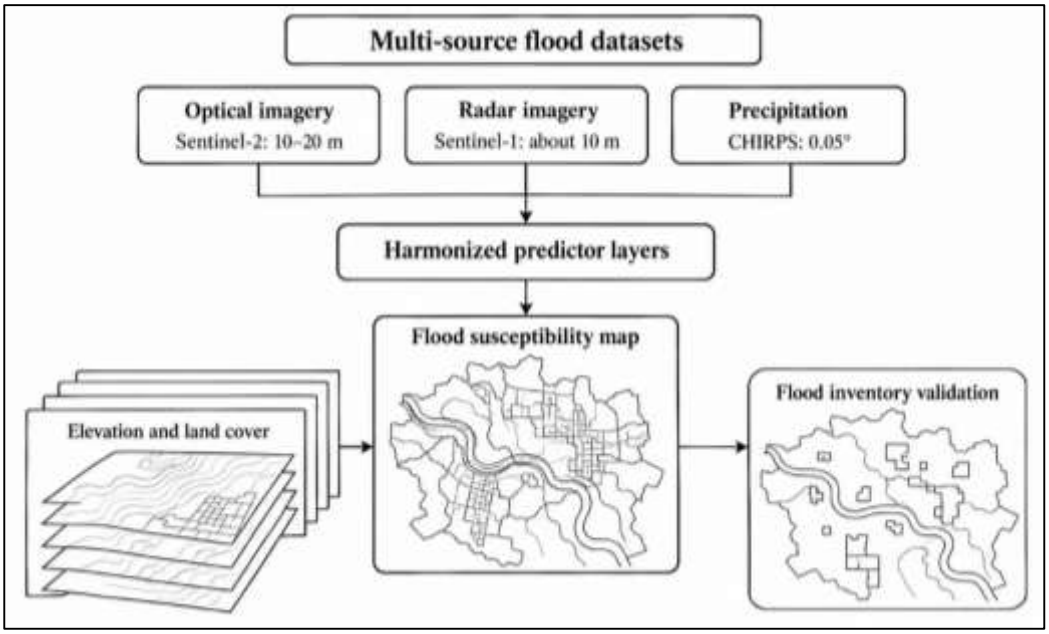
Quantitative Characteristics of Multi-Source Remote-Sensing Datasets Used in Urban Flood Susceptibility Mapping

Quantitative flood susceptibility research has relied extensively on optical satellite imagery because multispectral observations provide consistent measurements of vegetation, surface water, built-up land, bare surfaces, and land-use patterns over large geographic areas (Ashik & Shabbir, 2022; Wang et al., 2018). Across more than ten studies using Landsat-series imagery, Sentinel-2 MSI, MODIS, and higher-resolution commercial or open imagery, substantial differences have been reported in spatial resolution, spectral detail, temporal coverage, and analytical scale. Landsat imagery has commonly provided 30 m multispectral observations, with selected panchromatic information available at 15 m and thermal information at coarser resolutions, while its approximately 16-day repeat cycle has supported multi-year land-cover and environmental analysis. Sentinel-2 has strengthened urban-scale applications through 10 m bands in the visible and near-infrared wavelengths, 20 m red-edge and shortwave-infrared bands, additional 60 m atmospheric bands, and a nominal five-day revisit capability using the two-satellite configuration. Studies employing MODIS have generally accepted substantially coarser spatial resolutions of approximately 250 m, 500 m, or 1 km in exchange for high temporal frequency, making the sensor more applicable to regional hydroclimatic and surface-condition monitoring than detailed neighborhood-scale urban mapping. PlanetScope and comparable high-resolution products have provided imagery at approximately 3–5 m, enabling finer discrimination of roads, individual built-up areas, vegetation patches, and small water features (Hasan & Adnan, 2022; Zhang & Xia, 2021). Earlier investigations by Tehrany and colleagues, Khosravi and colleagues, Chapi and colleagues, Rahmati and colleagues, Bui and colleagues, Costache, Nachappa and colleagues, Ouma, Paul and colleagues, and other flood-mapping researchers demonstrate that optical datasets have rarely been used merely as background imagery. Instead, spectral observations have been transformed into land-use/land-cover classes and indices representing vegetation, water, moisture, and built-up conditions. Study areas have ranged from relatively compact urban catchments covering tens or hundreds of square kilometers to river basins and metropolitan regions encompassing several thousand square kilometers. Observation periods have similarly ranged from individual cloud-free acquisitions to multi-season and multi-year image archives. These differences indicate that optical dataset selection determines both the smallest spatial features represented in susceptibility modelling and the temporal correspondence achievable between remotely sensed environmental conditions and documented flood occurrence (Lim & Lee, 2018; Mahbubul et al., 2022).

Synthetic aperture radar and satellite-derived precipitation products have expanded the quantitative information available for flood susceptibility assessment by representing environmental characteristics that optical imagery alone cannot consistently capture. Sentinel-1 SAR has been particularly important because C-band radar operates independently of sunlight and remains usable under many cloudy conditions commonly associated with major rainfall and flooding. Sentinel-1 imagery has typically been analyzed at approximately 10 m pixel spacing for common ground-range products, with repeat observations obtained over periods of days depending on acquisition configuration and geographic coverage (Mashfiqur & Md, 2022; Parsian et al., 2021). Flood studies using Sentinel-1 have examined VV and VH polarization information, backscatter changes, surface-water extent, inundation signatures, and moisture-related characteristics. Investigations combining Sentinel-1 with Sentinel-2 have therefore integrated structurally different measurements: optical imagery provides spectral information on vegetation, water, and built-up surfaces, whereas SAR contributes microwave responses associated with surface roughness, geometry, and moisture. Precipitation datasets have supplied another major component of multi-source modelling. CHIRPS has been widely employed because it provides long-term rainfall estimates from 1981 onward at approximately 0.05-degree spatial

resolution with daily and aggregated temporal products.

Figure 2: Multi-source flood susceptibility workflow



TRMM and GPM products have also been incorporated into hydrological and flood research, although their native spatial resolution is considerably coarser than Sentinel or high-resolution DEM datasets. Studies by Tellman and colleagues, Ouma, Yao and colleagues, Chen and colleagues, Zhu and colleagues, and numerous regional flood researchers have combined satellite observations with rainfall, elevation, soil, drainage, and hydrological information (Annis & Nardi, 2021; Ashfaq & Mohammad, 2022). Complementary Earth-observation sources have included soil-moisture datasets, global surface-water products, vegetation products, nighttime observations, land-surface temperature, and global land-cover maps. Consequently, individual studies have frequently integrated several remotely sensed products rather than relying on one satellite platform. Quantitative preprocessing has commonly required converting datasets with native resolutions ranging from a few meters to several kilometers into a standardized analytical grid. This large resolution difference represents an important characteristic of multi-source susceptibility research because resampling a coarse rainfall product to a 10 m or 30 m grid increases the number of raster cells without creating equivalent fine-scale meteorological information (Adriano et al., 2019; Sadia & Mostari, 2022). The literature therefore demonstrates that nominal output resolution and original information resolution represent distinct quantitative properties of integrated flood datasets.

The quantitative value of multi-source remote sensing becomes particularly evident when optical, radar, precipitation, land-cover, and elevation products are combined with historical or satellite-observed flood inventories. Research conducted across South Asia, Southeast Asia, China, the Middle East, Africa, and other flood-prone regions has used combinations ranging from two or three primary Earth-observation datasets to considerably larger geospatial stacks containing ten or more flood-conditioning layers. Landsat and Sentinel-2 have frequently provided land-cover, vegetation, water, and built-up variables; Sentinel-1 has supplied radar backscatter or inundation information; CHIRPS, TRMM, GPM, or gauge-supported rainfall products have characterized precipitation; and SRTM, ASTER GDEM, ALOS PALSAR, Copernicus DEM, or LiDAR-derived elevation surfaces have provided terrain information (Gao et al., 2018; Shaiful et al., 2022). Global and regional land-cover products have also been used to supplement classifications derived directly from satellite imagery. Their spatial resolutions have ranged from approximately 10 m for recent global land-cover products to 30 m, 100 m, 300 m, or coarser for earlier products. Studies associated with Tehrany, Rahmati, Khosravi, Chapi, Bui, Costache, Nachappa, Ouma, Tellman, Paul, and other researchers collectively demonstrate that

the number and diversity of input datasets influence the range of flood-generating processes represented in a model. Temporal correspondence has nevertheless varied considerably (Stabak & Shahid, 2022; You et al., 2020). Some event-oriented studies have selected SAR or optical scenes acquired immediately before, during, or shortly after documented flood events, allowing direct correspondence between satellite-observed conditions and inundation. Susceptibility studies using relatively static terrain and land-cover predictors have often combined flood inventories accumulated across several years with imagery representing a particular year or representative period. Multi-year land-cover analysis has additionally been employed where urban expansion and changing imperviousness were important explanatory characteristics. Scene availability has varied according to cloud cover, sensor revisit time, study-area extent, and the observation period; compact urban areas can sometimes be represented by a single satellite tile, whereas larger basins or metropolitan regions require mosaics of multiple scenes. Consequently, temporal mismatch between flood inventories and predictor imagery constitutes an important source of quantitative heterogeneity when results are compared across published studies (Mohammad, 2023; Xue et al., 2021).

Comparative evidence across the flood susceptibility literature indicates that single-source and multi-source predictor configurations represent substantially different levels of environmental information. Single-source configurations based predominantly on DEM information can efficiently represent elevation, slope, curvature, drainage structure, flow accumulation, and related terrain characteristics, but they provide limited direct information about vegetation, impervious development, changing surface water, rainfall, and other dynamic land-surface conditions. Optical-only configurations improve representation of land cover and spectral characteristics but cannot independently describe the complete three-dimensional terrain structure controlling runoff routing and accumulation. SAR-only approaches provide valuable surface and inundation information under adverse atmospheric conditions but similarly represent only part of the flood-conditioning environment (Adnan & Iftekhar, 2023; Shen et al., 2020). Multi-source frameworks examined in studies by Tehrany and colleagues, Khosravi and colleagues, Rahmati and colleagues, Chapi and colleagues, Bui and colleagues, Costache, Nachappa and colleagues, Ouma, Tellman and colleagues, and other investigators have therefore combined terrain, spectral, hydrological, meteorological, and anthropogenic predictors. Quantitative studies frequently evaluate approximately 8–20 conditioning variables, although larger predictor sets have also been reported when numerous spectral indices, morphometric variables, rainfall characteristics, soil attributes, and proximity measures are included. The resulting models have been evaluated using accuracy, sensitivity, specificity, F1-score, receiver operating characteristic area under the curve, and related performance measures, with many machine-learning applications reporting strong discrimination when complementary predictor groups are integrated (Rahman, 2023; Sousa et al., 2021). However, increasing the number of remote-sensing products does not automatically increase independent information because multiple variables can represent overlapping environmental characteristics. Spectral indices derived from the same bands, DEM-derived terrain measures, and correlated rainfall variables may introduce redundancy. Spatial resolution also remains heterogeneous: a framework may simultaneously contain 10 m Sentinel-2 observations, approximately 10 m Sentinel-1 products, 30 m elevation information, and precipitation measurements whose native grid is measured in kilometers. The literature therefore characterizes multi-source urban flood susceptibility mapping as a quantitative data-integration problem in which sensor diversity, native resolution, temporal frequency, observation period, number of scenes, predictor dimensionality, spatial harmonization, and correspondence with flood inventories jointly determine the analytical structure of the resulting susceptibility dataset (Canavesi et al., 2020; Sheak & Tahmedur, 2023).

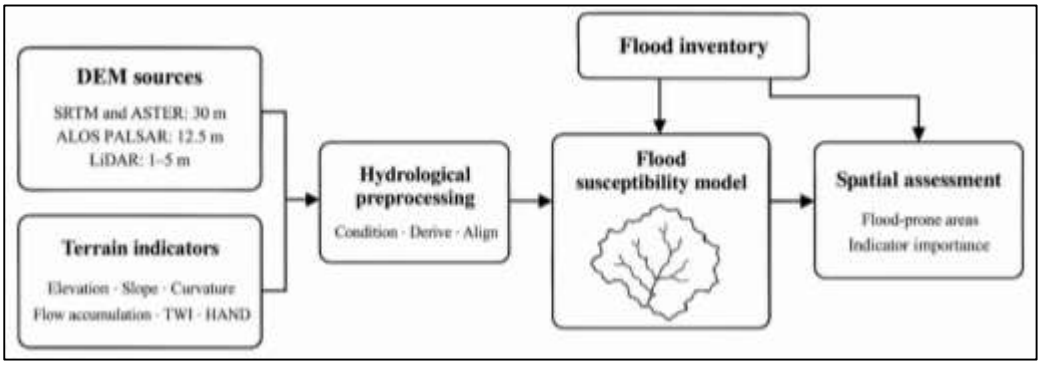
Hydrogeomorphic Flood-Conditioning Indicators

Digital elevation models have formed the primary topographic foundation of flood-susceptibility research because elevation surfaces provide the numerical basis for deriving terrain and drainage characteristics that regulate runoff generation, concentration, and local water accumulation. Evidence synthesized from more than 15 urban and regional flood studies shows substantial variation in DEM source and spatial resolution (Malekinezhad et al., 2021; Kaium, 2023). SRTM has been among the most frequently applied products, commonly at approximately 30 m resolution, because of its broad geographic coverage and accessibility. ASTER GDEM has also been widely used at approximately 30

m, while ALOS PALSAR elevation products have frequently provided finer terrain information at approximately 12.5 m. Copernicus DEM products have supported investigations at approximately 30 m and, where available, around 10 m, offering an additional source for representing terrain morphology. Higher-resolution studies have employed national elevation databases, photogrammetric surfaces, airborne laser scanning, and LiDAR-derived DEMs with grid sizes around 5 m, 2 m, 1 m, and sub-meter scales. These resolution differences substantially affect the numerical representation of local depressions, road embankments, channel margins, drainage corridors, and small variations in urban relief. A 30 m cell represents 900 m² of ground area, while a 10 m cell represents 100 m² and a 1 m cell represents only 1 m², creating major differences in the number and spatial specificity of terrain observations across the same study area (Malekinezhad et al., 2021; Pervaz, 2023). Several comparative studies have reported that coarse DEMs smooth localized elevation gradients and reduce the visibility of small drainage pathways, whereas finer DEMs identify greater terrain variability and more detailed flow-routing structures. Vertical accuracy has also varied substantially, with satellite-derived global DEMs generally exhibiting meter-scale elevation uncertainty and airborne LiDAR products commonly providing much finer vertical precision under suitable acquisition and processing conditions. Across the reviewed studies, DEM selection has therefore influenced not only absolute elevation values but also the subsequent calculation of slope, curvature, drainage networks, flow accumulation, wetness indicators, and relative height measures (Badrul & Mominul, 2023; Roy et al., 2021). Quantitative susceptibility models based on higher-resolution elevation surfaces have generally been better positioned to distinguish localized urban topographic controls, although the magnitude of improvement has depended on terrain complexity, source accuracy, hydrological conditioning, and consistency with the spatial resolution of other predictor datasets.

Elevation, slope, aspect, curvature, Topographic Position Index, and Terrain Ruggedness Index have been repeatedly evaluated as quantitative predictors of flood susceptibility, although their relative importance has varied among geographic settings. Across more than ten empirical investigations, elevation has frequently ranked among the strongest terrain-related predictors because flood observations are often concentrated in low-lying locations where runoff converges or drainage capacity becomes restricted (Rakib & Towhidul, 2023; Pham et al., 2020). Susceptibility classes have commonly demonstrated an inverse association with elevation, with the lowest elevation intervals containing disproportionately large shares of observed flood locations relative to higher terrain classes. Slope has shown a similarly consistent relationship, particularly in low-gradient urban and floodplain environments. Areas with slopes below approximately 5 degrees have frequently contained higher flood concentrations than steeper classes because reduced gradients favor slower runoff movement and greater surface accumulation. However, mountainous and foothill settings have shown more complex behavior because steep upstream slopes can accelerate runoff delivery toward lower urban areas. Profile and plan curvature have provided additional information on terrain shape, distinguishing concave surfaces that promote flow convergence from convex surfaces that encourage divergence. Variable-importance analyses in Random Forest, gradient boosting, and related models have frequently placed elevation and slope within the upper group of predictors, although rankings have changed according to the inclusion of rainfall, land cover, drainage, and proximity variables (Kanti, 2023; Samela et al., 2020). Reported correlation analyses generally show moderate relationships between some terrain metrics, particularly elevation, slope, ruggedness, and topographic position, requiring multicollinearity screening before simultaneous model inclusion. Topographic Position Index has improved differentiation among valleys, depressions, slopes, and elevated surfaces by representing the relative height of each location compared with its neighborhood, while Terrain Ruggedness Index has quantified local elevation variability. In several susceptibility analyses, lower topographic positions and smoother low-relief surfaces have corresponded with greater flood occurrence, whereas highly rugged terrain has exhibited lower direct inundation susceptibility but stronger runoff-generation potential (Khosravi et al., 2020; Nishat, 2023). Information-gain and feature-ranking results across the literature indicate that terrain variables rarely contribute equally; their quantitative value depends on relief range, drainage structure, urban morphology, DEM resolution, and the scale at which neighborhood calculations are performed.

Figure 3: DEM resolution and flood indicators



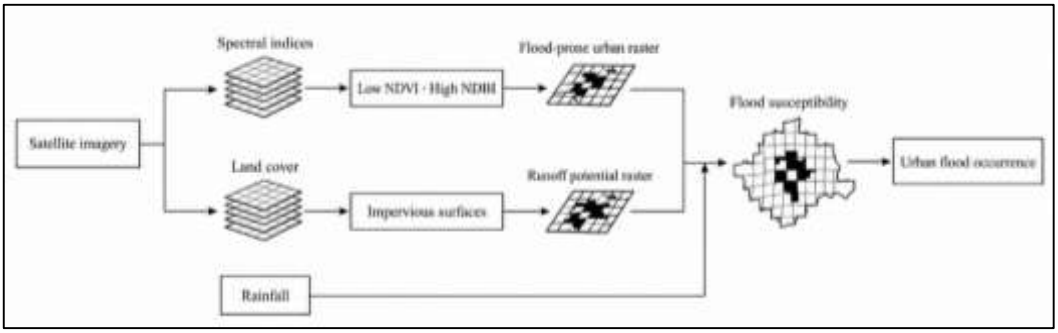
Hydrogeomorphic indicators derived from DEMs have provided a more process-oriented representation of flood susceptibility by quantifying how terrain structure controls runoff routing, concentration, wetness, drainage connectivity, and vertical proximity to channels. Flow direction and flow accumulation are fundamental components because they describe the movement of surface runoff across a gridded elevation surface and identify cells receiving drainage from larger contributing areas. Across multiple flood studies, high flow-accumulation classes have generally corresponded with increased flood susceptibility, particularly where accumulated runoff intersects low-gradient built-up surfaces (Ghosh & Dey, 2021; Risha & Shurovi, 2023). Drainage density and stream frequency have also been widely included, with higher drainage density often indicating stronger hydrological connectivity, although its relationship with susceptibility can differ depending on whether dense drainage improves water removal or represents strongly dissected terrain with substantial runoff concentration. Distance to drainage has shown one of the most consistent spatial associations across many studies, with areas located within the nearest distance classes to rivers, channels, and urban drainage corridors frequently exhibiting markedly greater flood occurrence than more distant zones. Topographic Wetness Index has repeatedly been used to represent the tendency of terrain to accumulate moisture, and high-value classes have commonly been associated with low slopes, convergent flow conditions, and elevated flood susceptibility. Stream Power Index has characterized the erosive and transport potential associated with concentrated flow and has shown stronger importance in catchments containing distinct relief and channelized runoff (Rizeei et al., 2017; Stabak & Siam, 2023). Height Above Nearest Drainage has provided an especially direct representation of vertical separation from the drainage network, with low HAND values consistently associated with flood-prone floodplains and channel-adjacent terrain in several regional applications. Feature-importance analyses have commonly ranked distance to drainage, flow accumulation, TWI, HAND, and elevation among influential predictors, although relative scores vary considerably among algorithms and environments. Information-gain assessments and susceptibility ratios have similarly indicated that the highest wetness and accumulation classes can contain substantially greater proportions of flood observations than low-value categories. Collectively, the reviewed evidence shows that these indicators complement one another: flow accumulation represents contributing runoff, TWI represents moisture accumulation tendency, SPI characterizes concentrated flow energy, drainage metrics describe hydrological organization, and HAND represents relative vertical control over inundation potential (Arabameri et al., 2019; Feroz, 2023).

Rainfall with Urban Flood Occurrence

Remote-sensing spectral indices have been extensively incorporated into quantitative flood-susceptibility studies because they provide spatially continuous measurements of vegetation condition, surface water, moisture, and built-up development. Evidence synthesized across more than ten empirical investigations has consistently identified NDVI, NDWI, MNDWI, and NDBI as important explanatory variables, although their relative contributions have varied according to climatic conditions, urban morphology, image-acquisition period, spatial resolution, and modelling technique

(Kim et al., 2017; Tahmedur & Rifat, 2023). NDVI has generally demonstrated an inverse relationship with urban flood susceptibility because areas containing greater vegetation cover commonly provide higher interception, infiltration, evapotranspiration, and surface roughness than densely constructed areas.

Figure 4: Spectral and land cover flood mapping



Flood-frequency analyses have consequently reported larger concentrations of observed flood locations within low and moderate NDVI classes than within highly vegetated classes in many urban environments. Machine-learning feature rankings have often identified vegetation indices as moderate-to-high contributors when land-cover characteristics are included with topographic and hydrological predictors. NDWI and MNDWI have shown stronger positive associations with flood occurrence where high index values correspond to existing water bodies, saturated surfaces, wetlands, drainage corridors, and areas experiencing temporary inundation. MNDWI has frequently provided clearer separation of water from built-up surfaces in densely developed landscapes, making it particularly useful for extracting surface-water occurrence (Akter et al., 2020). NDBI has displayed an opposing pattern to NDVI in many rapidly urbanizing locations, with higher values generally representing greater built-up intensity and being associated with increased runoff-generating potential. Regression coefficients, correlation analyses, information-gain measures, and Random Forest or boosting-based feature rankings reported across the literature have commonly identified built-up and water-related indices as significant components of multivariable susceptibility models. Studies using Landsat and Sentinel-2 have also demonstrated that relationships among these indices are not statistically independent. Increasing NDBI frequently corresponds with decreasing NDVI, while NDWI and MNDWI can correlate with drainage proximity and low elevation. Multicollinearity screening has therefore been used to prevent highly correlated spectral predictors from disproportionately influencing susceptibility estimates (Abass et al., 2020). Collectively, the literature indicates that spectral indices provide quantitative proxies for hydrologically relevant surface conditions, with combinations of low vegetation, extensive built-up surfaces, persistent or episodic water occurrence, and elevated surface moisture repeatedly associated with higher concentrations of observed urban flooding.

Land-use/land-cover composition and impervious-surface expansion have been among the most consistently examined anthropogenic predictors of urban flood occurrence. Studies conducted in rapidly urbanizing regions of Asia, Africa, Europe, and other geographic settings have quantified substantial differences in flood susceptibility among built-up land, agricultural areas, vegetation, wetlands, open land, and water-associated classes. Built-up areas have frequently contained larger flood frequencies and higher susceptibility ratios than vegetated or forested classes when development occurs across low-lying terrain or near natural drainage pathways (Lyu et al., 2018). Quantitative analyses of urban expansion have shown that conversion of vegetation, agricultural land, and open surfaces into roads, buildings, parking areas, industrial zones, and other impermeable surfaces can substantially increase the fraction of rainfall transformed into rapid surface runoff. Impervious-surface fraction has therefore been incorporated either directly as a continuous predictor or indirectly through land-cover classes, NDBI, built-up density, road density, and settlement-related variables. Across

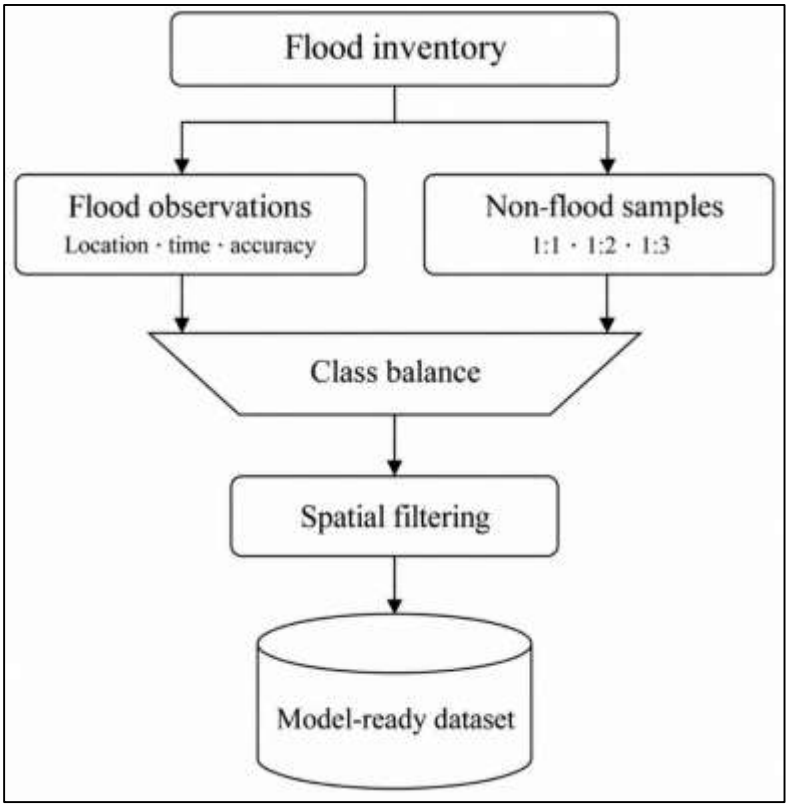
numerous studies, highly urbanized classes have generally received greater variable-importance scores than sparsely developed areas, although their importance has depended on topographic position and rainfall characteristics. Studies comparing land-cover conditions over periods of approximately 10–30 years have documented substantial increases in built-up area in expanding metropolitan regions, frequently accompanied by reductions in vegetation or agricultural land and spatial expansion of flood-prone surfaces (Mohamed & Worku, 2021). Class-specific analyses have also demonstrated that equal amounts of built-up land do not necessarily produce equivalent susceptibility because the hydrological effect of urbanization varies according to drainage connectivity, surface slope, local elevation, soil permeability, and proximity to channels. High imperviousness located on elevated or steep terrain may generate runoff that is transferred toward lower areas, whereas dense development within depressions and low-gradient floodplains can experience direct accumulation. Feature-importance analyses using Random Forest, gradient boosting, XGBoost, logistic regression, neural networks, and related models have repeatedly placed land cover, built-up density, or imperviousness among influential predictors when combined with elevation, drainage distance, rainfall, and wetness variables (Manawadu & Wijeratne, 2021). The reviewed evidence therefore identifies urbanization as a measurable spatial modification of hydrological response rather than a uniform categorical influence, with flood occurrence determined by interactions between the percentage and configuration of impermeable surfaces and the hydrogeomorphic setting in which urban development occurs.

Spatial Sampling Design

Flood inventories constitute the principal response dataset in quantitative flood susceptibility modelling because the number, spatial distribution, temporal coverage, and positional reliability of recorded flood locations directly influence model training and validation (Rajabi et al., 2021). Evidence synthesized from more than ten studies by Tehrany and colleagues, Rahmati and colleagues, Khosravi and colleagues, Chapi and colleagues, Bui and colleagues, Costache, Nachappa and colleagues, Paul and colleagues, Ouma, and other researchers shows considerable variation in inventory size, ranging from fewer than 100 documented flood locations in relatively small catchments to several hundred or thousands of observations in larger regional and remotely sensed investigations. Inventories have been constructed from field surveys, government disaster records, historical flood maps, satellite-derived inundation boundaries, aerial photographs, news reports, GPS observations, hydrological agencies, community reports, and combinations of these sources. Studies using point-based inventories have commonly converted mapped flood locations into binary flood and non-flood observations, whereas raster-based studies have represented entire inundated areas as flood pixels. This distinction can produce major differences in nominal sample size because a single mapped flood polygon can contain hundreds or thousands of raster cells depending on spatial resolution. Temporal coverage has also varied from inventories associated with one major flood event to databases compiled from multiple events over periods extending across several years or decades (Zeng et al., 2021). Event-specific inventories generally provide stronger temporal correspondence with rainfall, soil moisture, and remotely sensed surface conditions, while multi-event inventories represent a broader range of flood-producing conditions. Positional accuracy has differed according to inventory source: GPS-based field observations and high-resolution imagery can provide relatively precise locations, whereas historical maps, reports, and manually georeferenced records may contain larger positional uncertainties. At 10 m or finer mapping resolutions, even relatively small location errors can shift observations between roads, buildings, channels, and neighboring terrain classes (Popa et al., 2019). Studies have therefore used spatial filtering, duplicate removal, minimum separation distances, polygon-to-point conversion, and inventory verification to improve response-data quality. The literature demonstrates that the apparent number of flood samples alone does not determine inventory quality; spatial independence, positional accuracy, temporal consistency, event representation, and compatibility with predictor resolution collectively determine the quantitative information available for susceptibility modelling.

The construction of non-flood samples represents one of the most variable components of flood susceptibility modelling because confirmed absence of flooding is generally more difficult to establish than documented flood occurrence. Many published studies have adopted approximately balanced datasets by generating one non-flood location for every recorded flood location, producing a 1:1 sampling ratio and an analytical class prevalence close to 50% for each class (Saharia et al., 2017).

Figure 5: Flood dataset preparation flowchart



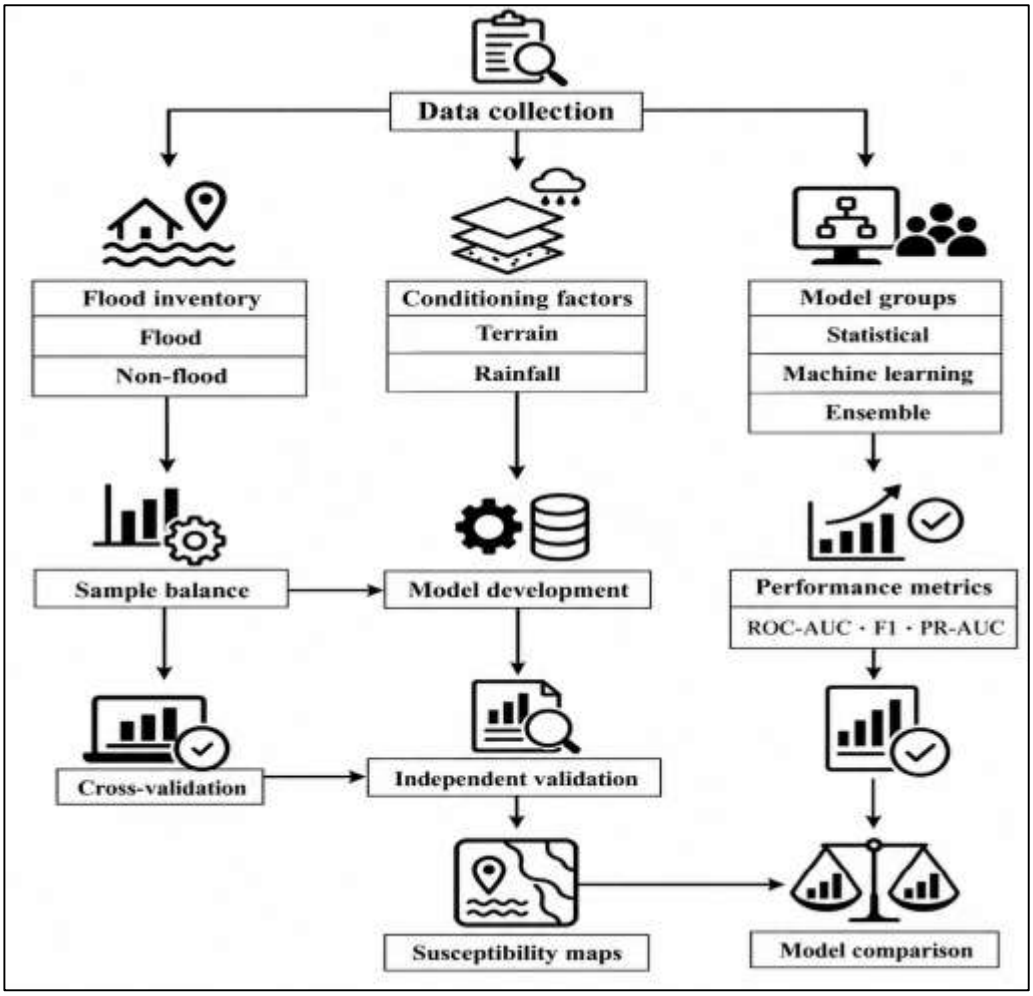
Other investigations have employed 1:2, 1:3, or naturally imbalanced ratios in which non-flood observations substantially outnumber recorded flood locations. Randomly generated non-flood points have commonly been selected from areas outside mapped flood polygons or beyond predefined buffers surrounding flood observations. Some studies have imposed minimum distances from known flood locations, rivers, or previously inundated areas to reduce the probability that an assumed non-flood point represents an undocumented flood-prone location (Saharia et al., 2021). The magnitude of these exclusion distances has differed according to mapping scale, DEM resolution, inventory accuracy, and local hydrological conditions. Balanced sampling has been widely used with logistic regression, Random Forest, Support Vector Machine, artificial neural networks, gradient boosting, and related classification algorithms because extreme class imbalance can bias models toward the dominant non-flood class. However, artificially imposing equal numbers of flood and non-flood observations changes sample prevalence relative to the actual geographic prevalence of flooding. Studies using complete raster datasets can contain very large class imbalances because susceptible or historically inundated cells may represent only a small percentage of the entire landscape. To address this problem, investigations have applied random undersampling of majority-class observations, oversampling of flood cases, stratified sampling, class weighting, synthetic minority sampling, or combinations of these procedures. Quantitative model assessments demonstrate that overall accuracy can appear high in strongly imbalanced datasets even when flood detection is poor, which has led researchers to supplement accuracy with recall, precision, specificity, F1-score, balanced accuracy, ROC-AUC, and precision-recall measures (Haltas et al., 2021). The reviewed evidence therefore indicates that flood/non-flood ratio, negative-sample generation, class prevalence, and balancing strategy can materially influence estimated relationships between environmental predictors and flood occurrence as well as the apparent predictive performance of susceptibility models.

Machine-Learning, and Ensemble Flood Susceptibility Models

Statistical and knowledge-based approaches have provided the methodological baseline against which more computationally intensive flood susceptibility models are commonly evaluated. Across studies by Tehrany and colleagues, Rahmati and colleagues, Khosravi and colleagues, Chapi and colleagues,

Bui and colleagues, Costache, Nachappa and colleagues, Paul and colleagues, Ouma, and other flood-mapping researchers, frequency ratio, weights of evidence, logistic regression, and AHP-based techniques have produced moderate to strong discrimination depending on predictor quality, flood inventory structure, terrain characteristics, and validation design (Shafizadeh-Moghadam et al., 2018). Frequency-ratio methods have frequently generated ROC-AUC values within approximately 0.75–0.90, particularly when elevation, slope, drainage proximity, rainfall, land cover, and wetness indicators were well represented. Weights-of-evidence models have achieved broadly comparable performance, although their effectiveness has depended strongly on predictor independence and the quality of class intervals used for categorical evidence estimation.

Figure 6: Flood susceptibility model comparison workflow



Logistic regression has commonly demonstrated ROC-AUC values around 0.80–0.90 in flood-susceptibility applications and has remained competitive because it provides interpretable regression coefficients and probability estimates while accommodating multiple conditioning variables (Costache et al., 2020). AHP-based models have also produced validation AUC values frequently above 0.75 and sometimes above 0.85, although their quantitative performance varies because predictor weights depend partly on expert judgment and pairwise prioritization. Studies comparing these approaches have shown that training performance may exceed validation performance by several percentage points, particularly when numerous conditioning variables are fitted to relatively small flood inventories. Sensitivity and specificity values have often ranged from approximately 70% to above 85%, while Cohen's kappa has generally indicated moderate to substantial agreement where susceptibility classes align well with observed floods. Conventional statistical models tend to produce relatively stable calibration when model assumptions are adequately satisfied, although discrimination can

decline when predictor relationships are strongly nonlinear (Choubin et al., 2019). Overall, evidence across more than ten studies indicates that these techniques provide robust quantitative benchmarks, but their accuracy frequently falls below that of optimized machine-learning models when large, heterogeneous, and highly interactive remote-sensing datasets are analyzed.

Machine-learning methods have generally demonstrated stronger predictive discrimination than simpler statistical and knowledge-driven approaches because they can represent nonlinear relationships and complex interactions among flood-conditioning variables without requiring rigid functional assumptions. Random Forest has been one of the most consistently high-performing algorithms in published flood susceptibility studies, with reported ROC-AUC values commonly ranging from approximately 0.85 to above 0.95 and classification accuracy frequently exceeding 85%. Sensitivity, specificity, precision, and F1-score values have also often remained within the 0.80–0.95 range when sufficiently representative inventories and well-conditioned predictor datasets were available (Liu et al., 2021). Support Vector Machine has produced similarly strong results, with ROC-AUC values commonly reported around 0.83–0.94 and particularly competitive performance when predictor spaces contain nonlinear class boundaries. Decision trees have generally offered simpler interpretability but have often performed below Random Forest because individual trees are more sensitive to sampling variation and overfitting. Validation accuracy for single-tree models has frequently remained in the approximate 75–88% range, while training accuracy can become considerably higher when tree depth is weakly constrained. Artificial neural networks and multilayer perceptrons have often achieved validation accuracies above 80% and ROC-AUC values around 0.85–0.95, with stronger results obtained when training datasets are sufficiently large and predictors are appropriately standardized (Mosavi et al., 2018). Comparative studies by Tehrany, Khosravi, Chapi, Bui, Rahmati, Costache, Nachappa, and other investigators have repeatedly shown that no algorithm dominates all geographic contexts. Random Forest frequently ranks highly because bootstrap aggregation reduces variance and allows complex interactions, while SVM performs effectively in high-dimensional predictor spaces and neural networks capture nonlinear response patterns. Training-validation differences remain important: models producing training accuracy above 95% but validation values closer to 80–88% provide evidence of overfitting. Kappa, balanced accuracy, recall, and F1-score are therefore increasingly reported alongside conventional accuracy to distinguish genuine flood-detection capacity from performance inflated by class prevalence or excessive model complexity (Band et al., 2020).

Gradient-boosting algorithms, XGBoost, and convolutional neural networks have expanded the quantitative performance range reported in flood susceptibility mapping by capturing complex predictor interactions and, in the case of convolutional approaches, spatial neighborhood structure. Gradient boosting has commonly achieved ROC-AUC values between approximately 0.88 and 0.96 across environmental susceptibility applications, while optimized XGBoost models have frequently reported accuracy and AUC values exceeding 0.90. Their strong performance is associated with sequential error correction, flexible tree structures, regularization, and the ability to identify nonlinear thresholds in elevation, drainage distance, rainfall, flow accumulation, land cover, and spectral indices (Khosravi et al., 2018). Precision and recall frequently remain above 0.85 in well-balanced datasets, with F1-scores approaching or exceeding 0.90 in higher-performing applications. XGBoost has also ranked important predictors effectively, although its training performance can approach near-perfect levels when tree depth, boosting rounds, or learning parameters are insufficiently constrained. Convolutional neural networks differ from conventional pixel-based classifiers because spatial neighborhoods can be incorporated directly into prediction. Flood studies using CNN architectures have reported classification accuracy generally above 80%, with stronger implementations exceeding 90% depending on image resolution, sample quantity, feature structure, and validation strategy (Liang et al., 2020). CNNs have shown particular value where multi-band remotely sensed imagery is analyzed together with terrain or environmental raster layers. Multilayer perceptrons and deeper neural architectures have similarly produced strong sensitivity and specificity but can display greater variability between training and validation sets when sample sizes are limited. PR-AUC has become particularly useful for evaluating these algorithms when flood observations are less common than non-flood observations,

because ROC-AUC can remain relatively high even under substantial class imbalance. Calibration statistics have also revealed that strong discrimination does not guarantee accurate probability estimates; boosting and neural models may require probability calibration when susceptibility values are interpreted probabilistically (Al-Abadi, 2018). Evidence from multiple comparative studies therefore shows that advanced learning algorithms can produce very high classification performance, but their reported superiority is strongly conditioned by hyperparameter optimization, predictor selection, inventory size, class balance, spatial resolution, and validation independence.

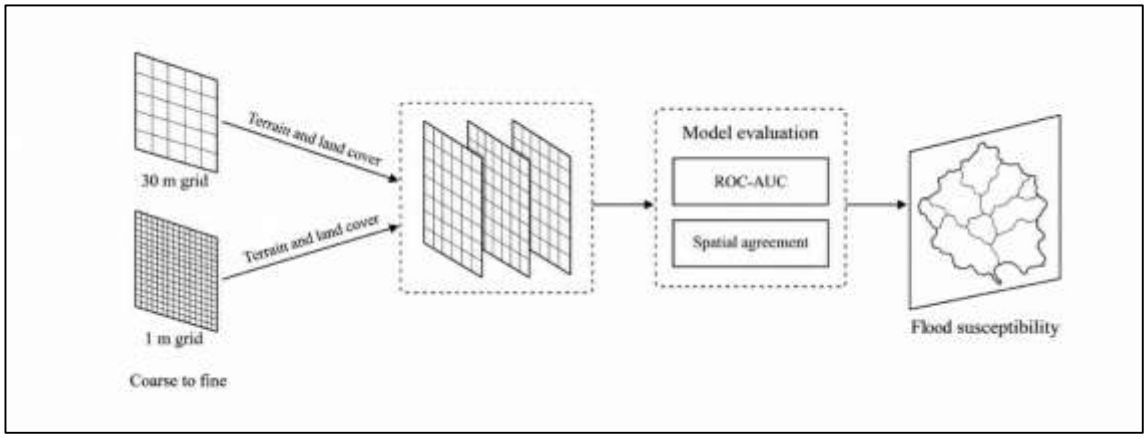
Ensemble flood susceptibility models have frequently achieved the strongest overall discrimination because they combine information from multiple algorithms and reduce dependence on the errors of any single modelling technique (Dodangeh et al., 2020). Studies integrating Random Forest, Support Vector Machine, logistic regression, neural networks, boosting algorithms, decision trees, or weighted combinations of statistical models have commonly reported ROC-AUC values around 0.90–0.97, with several applications achieving accuracy, precision, recall, and F1-score above 90%. Ensemble improvement over individual models is often modest but consistent, frequently amounting to several percentage points in ROC-AUC or classification accuracy when component models exhibit complementary error patterns. Voting ensembles, weighted averaging, stacking, bagging, and boosting have all been applied, although their performance varies according to the diversity of base learners and the independence of validation samples. Comparisons across studies by Chapi, Bui, Khosravi, Rahmati, Costache, Nachappa, Tehrany, Paul, Ouma, and other researchers indicate that Random Forest, gradient boosting, XGBoost, and ensemble configurations commonly occupy the upper performance range, whereas frequency ratio, AHP, and single decision trees more often provide lower but still useful benchmark accuracy (El-Haddad et al., 2021). However, differences between training and validation results demonstrate that higher model complexity does not automatically produce stronger generalization. Training ROC-AUC values above 0.95 or approaching 1.00 have occasionally been accompanied by validation reductions of 5–15 percentage points, indicating sensitivity to overfitting, sampling configuration, or spatial dependence. Balanced accuracy and F1-score provide more informative assessments than overall accuracy where flood and non-flood classes are unequal, while Cohen's kappa quantifies agreement beyond chance and PR-AUC is particularly informative when flood prevalence is low. Calibration measures additionally assess whether predicted susceptibility probabilities correspond with observed event frequencies. Studies using repeated cross-validation or independent holdout datasets generally provide more stable performance estimates than evaluations based on a single random split (Costache, 2019). The collective quantitative evidence from more than ten studies therefore demonstrates that ensemble and advanced machine-learning models often achieve the highest predictive scores, while reliable generalization depends equally on flood-inventory quality, class balance, hyperparameter control, spatial separation, repeated validation, and consistency between training and independent testing performance.

High-Resolution Urban Flood Susceptibility Prediction

Spatial resolution has a direct quantitative effect on how urban terrain, drainage pathways, depressions, and surface heterogeneity are represented within flood susceptibility models (Xing et al., 2019). Evidence from studies using approximately 30 m, 20 m, 10 m, 5 m, 2 m, 1 m, and sub-meter elevation or predictor grids shows that progressively finer resolutions increase the number of analytical cells and preserve greater local variation in topography. A 30 m raster represents 900 m² per cell, while 10 m, 5 m, and 1 m grids represent approximately 100 m², 25 m², and 1 m², respectively. This change can transform the representation of urban flood-controlling features such as road embankments, narrow drainage channels, local depressions, levees, small waterways, building-scale barriers, and abrupt elevation changes. A recent urban pluvial flood study comparing resolutions from 1 m to 50 m demonstrated this computational and spatial contrast clearly: the 1 m configuration contained more than 9 million grid cells, whereas the same domain contained approximately 90,000 cells at 10 m, about 22,500 at 20 m, and only around 10,000 at 30 m. Research in Berlin using CNN, ANN, Random Forest, and Support Vector Machine models at 30 m, 10 m, 5 m, and 2 m similarly found that higher-resolution inputs produced clearer delineation of localized flood-prone terrain (Olbert et al., 2017). Earlier investigations by Tehrany and colleagues, Khosravi and colleagues, Chapi and colleagues, Rahmati and colleagues, Bui and colleagues, Costache, Nachappa and colleagues, Paul and colleagues, Ouma,

and other flood researchers have likewise demonstrated that terrain detail affects elevation, slope, curvature, drainage density, and flow accumulation predictors. Coarser cells tend to smooth elevation differences and merge nearby features, whereas finer grids retain greater slope variability and more localized drainage pathways. Flow accumulation can also shift spatially as cell size changes because contributing areas and flow-routing paths are recalculated from different terrain representations. Consequently, spatial resolution determines both the apparent physical structure of the urban landscape and the numerical distributions of hydrogeomorphic predictors subsequently supplied to susceptibility algorithms (Bermúdez & Zischg, 2018).

Figure 7: Urban flood susceptibility resolution workflow



Resolution effects extend beyond terrain variables because the representation of land cover, impervious surfaces, water bodies, vegetation patches, road corridors, and built-up heterogeneity also changes substantially with raster cell size. At approximately 30 m, individual pixels commonly contain mixtures of buildings, roads, vegetation, exposed soil, and water, producing generalized land-cover representations (Rong et al., 2020). At 10 m, Sentinel-class imagery allows finer differentiation of urban surfaces, while 5 m, 1 m, and sub-meter datasets can represent progressively smaller spatial features. Studies using high-resolution optical imagery, LiDAR, and detailed urban land-cover products have shown that this increased spatial detail can substantially change the location and extent of predicted susceptibility classes even when the overall percentage of impervious or vegetated land changes only slightly. Fine-scale datasets identify small depressions and disconnected low points that may disappear when elevation surfaces are aggregated to 20–30 m. They also preserve narrow drainage channels and localized barriers that affect runoff routing. Comparative coastal flood research using LiDAR DEMs resampled to approximately 5, 9, 15, 20, 25, 27, and 35 m demonstrated that changing cell size altered flood delineation and elevation-error characteristics (Silvestro et al., 2019). Other research comparing 12.5 m and 30 m DEMs found that Random Forest ROC-AUC changed from approximately 0.983 at 12.5 m to 0.975 at 30 m, while ANN values changed from about 0.949 to 0.930 and generalized linear model performance from approximately 0.965 to 0.949. These differences show that higher spatial resolution can improve model discrimination, although the improvement is not necessarily large or consistent across algorithms. Studies at 1–10 m have also demonstrated that finer grids identify more fragmented susceptible zones and sharper boundaries, whereas 20–30 m products generate broader and more spatially continuous susceptibility classes. The total area assigned to high and very-high susceptibility categories can therefore increase or decrease depending on how localized depressions, land-cover transitions, and drainage pathways are preserved or generalized (Zick, 2020). Evidence across more than ten investigations indicates that scale affects not only visual map detail but also the underlying sample count, predictor variance, class frequencies, and spatial extent of predicted flood-prone areas.

Quantitative comparisons of flood susceptibility models across multiple spatial resolutions show that finer grids do not produce uniform increases in classification accuracy, although they frequently

improve the spatial specificity of predictions (Xu et al., 2021). In one urban modelling investigation comparing 30 m, 10 m, and 5 m configurations, CNN ROC-AUC values were approximately 0.92, 0.91, and 0.92, respectively, demonstrating very limited sensitivity in overall discrimination even though the high-resolution maps contained greater spatial detail. Random Forest produced corresponding ROC-AUC values of approximately 0.95, 0.96, and 0.93, while Support Vector Machine values remained around 0.86, 0.88, and 0.86. Accuracy for CNN models ranged from approximately 0.83 to 0.85, Random Forest from 0.85 to 0.87, and SVM from 0.78 to 0.80 across these resolutions. These results illustrate that the finest grid does not necessarily yield the highest numerical accuracy. Research comparing 12.5 m and 30 m DEM inputs similarly found only moderate AUC differences, with Random Forest remaining highly accurate at both scales (Zhao et al., 2021). Other studies using 30 m global DEMs have achieved ROC-AUC values above 0.90 when predictor quality, rainfall representation, drainage metrics, and flood inventories were strong. By contrast, 1–5 m LiDAR-based studies have demonstrated much greater gains in identifying street-scale accumulation, local depressions, and microtopographic pathways than in global discrimination metrics. This distinction is important because ROC-AUC summarizes ranking performance across an entire dataset and can remain relatively stable while the geographic location and boundaries of high-susceptibility zones change materially. High-resolution models also contain far more observations, but many neighboring cells are highly correlated and therefore do not provide independent information proportional to the increase in sample count (Lin et al., 2019). Comparative studies by Seleem and colleagues, Fereshtehpour and colleagues, and other flood researchers have therefore indicated that resolution effects depend on model type, flood-inventory precision, environmental complexity, and the scale of the features being predicted. Across the wider literature, differences of only a few percentage points in AUC can coexist with substantial changes in local susceptibility patterns, indicating that spatial-resolution sensitivity should be evaluated through both statistical performance and spatial agreement rather than through accuracy alone.

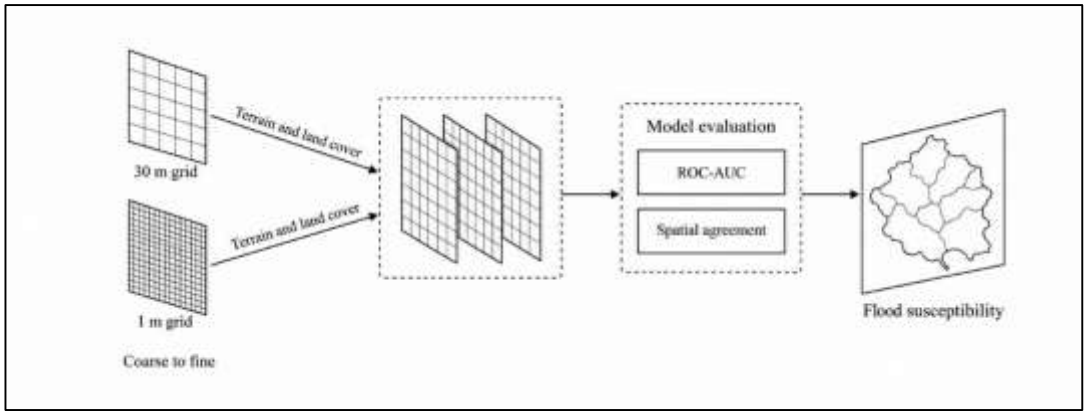
Spatial Autocorrelation and Generalization Error

Validation design has become a major quantitative consideration in flood susceptibility and broader geospatial prediction because conventional random partitions assume greater independence among observations than is typically present in spatial datasets (Gazis & Greinert, 2021). Random holdout procedures commonly divide observations using 70:30, 75:25, or 80:20 training-validation ratios, while random 5-fold and 10-fold cross-validation distribute observations among multiple subsets without explicitly considering geographic location. Studies by Roberts and colleagues, Valavi and colleagues, Meyer and colleagues, Ploton and colleagues, Schratz and colleagues, Pohjankukka and colleagues, Wenger and Olden, Brenning, Telford and Birks, and other spatial-modelling researchers have demonstrated that random validation can produce substantially different error estimates from geographically structured evaluation. Similar concerns have increasingly been recognized in flood, landslide, ecological, soil, and remote-sensing susceptibility applications. Random partitions can place neighboring raster cells or nearby flood observations in both training and validation datasets, allowing models to predict locations with environmental characteristics highly similar to those already encountered during training (Wadoux et al., 2021). Spatial block cross-validation instead partitions the study area into geographically distinct blocks and withholds complete blocks during successive validation iterations. Spatial k-fold methods similarly allocate geographic groups rather than individual observations to folds, while spatial clustering assigns samples to groups based on coordinates or spatial similarity. Comparative studies have reported reductions in accuracy, F1-score, ROC-AUC, and related discrimination measures when moving from random to spatial validation, although the magnitude varies with spatial structure, model complexity, sample density, and block size. In some geospatial applications, randomly validated performance has exceeded spatially validated estimates by several percentage points, while strongly autocorrelated datasets have produced considerably larger differences. Fold-to-fold variability can also increase under spatial validation because individual folds may represent distinct terrain, land-cover, rainfall, and drainage conditions (Airola et al., 2019). Consequently, spatial cross-validation generates a more demanding estimate of geographic generalization by testing models on areas that are deliberately separated from the observations used for fitting rather than measuring prediction performance primarily among spatially

intermingled samples.

Spatial autocorrelation provides the statistical basis for understanding why validation results can differ between random and geographically separated sampling designs. Moran's I has been widely used to quantify whether similar values occur near one another, with positive values indicating spatial clustering and values approaching spatial randomness indicating weaker geographic dependence under the specified weighting structure (Adams et al., 2020). In flood susceptibility datasets, positive autocorrelation can occur in both predictor variables and response observations because low elevation, high flow accumulation, similar land cover, drainage proximity, rainfall, and flood occurrence frequently form geographically continuous patterns.

Figure 8: Random vs spatial cross-validation workflow



Semivariograms provide complementary information by quantifying how similarity changes with separation distance and identifying the approximate spatial correlation range beyond which observations become substantially less dependent. Studies in spatial machine learning have shown that validation distances smaller than the correlation range can produce optimistic estimates because testing observations remain statistically similar to nearby training samples (Adams et al., 2020). Nearest-neighbor distance provides a simpler measure of geographic separation and has been used to evaluate whether validation observations are located farther from training samples under spatial partitioning than under random cross-validation. Residual spatial autocorrelation offers an additional diagnostic because statistically significant clustering of model errors indicates that unexplained geographic structure remains after predictor effects have been considered. Research by Roberts and colleagues, Meyer and colleagues, Valavi and colleagues, Ploton and colleagues, Brenning, Pohjankukka and colleagues, and other spatial prediction researchers has shown that block dimensions and buffer distances should be interpreted relative to the spatial dependence of the underlying dataset. In high-resolution urban flood mapping, this issue is particularly important because a 10 m raster can contain thousands of adjacent cells sharing almost identical terrain and land-cover values (Varoquaux et al., 2017). A nominal dataset containing hundreds of thousands of cells therefore does not contain an equivalent number of spatially independent observations. Quantitative examination of Moran's I, semivariogram range, nearest-neighbor separation, and residual spatial autocorrelation consequently provides a stronger basis for interpreting whether validation accuracy represents true geographic transferability or primarily reflects prediction among neighboring observations with closely related environmental characteristics.

Different spatial validation procedures impose varying levels of geographic independence and consequently produce different estimates of generalization error. Spatial block cross-validation divides a study region into contiguous geographic units and iteratively withholds one or more blocks for testing. The number and dimensions of blocks determine the distance between training and validation observations and can materially alter estimated accuracy (Li et al., 2017). Spatial k-fold validation extends the familiar k-fold framework by ensuring that entire geographic groups, rather than randomly

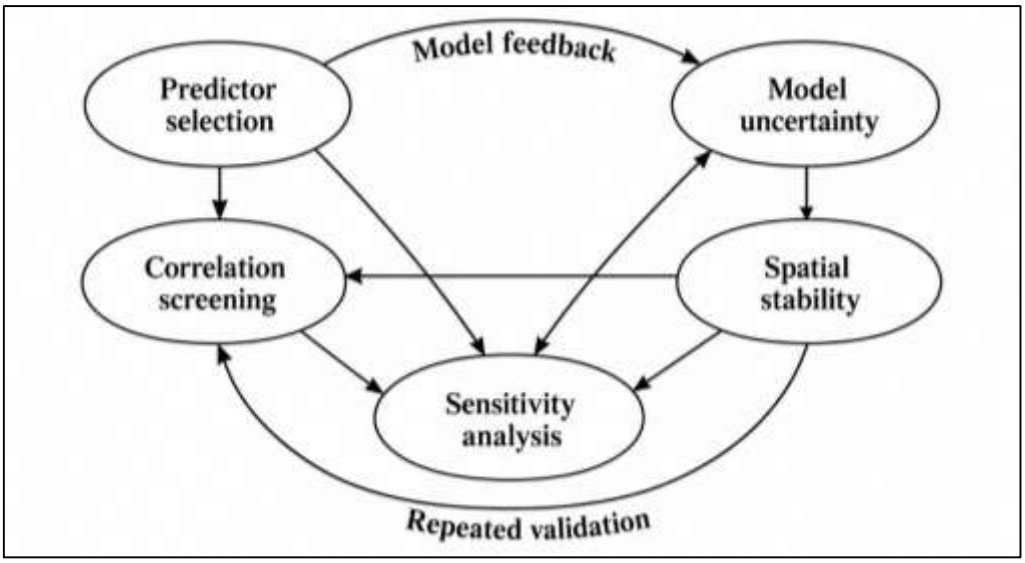
selected individual observations, are assigned to each fold. Buffered leave-location-out validation imposes an additional exclusion zone around each validation location or region, removing nearby training observations that could otherwise provide nearly duplicated environmental information. Spatial clustering approaches use coordinates or environmental similarity to define geographically coherent groups before model evaluation, while geographically separated testing reserves an independent district, catchment, urban sector, or other area entirely for external evaluation. Studies by Roberts and colleagues, Valavi and colleagues, Meyer and colleagues, Schratz and colleagues, Ploton and colleagues, Telford and Birks, Wenger and Olden, Brenning, and related researchers have shown that increasing geographic separation commonly raises estimated prediction error relative to random cross-validation. However, excessively large blocks can reduce the amount and environmental diversity of training data, whereas very small blocks may fail to remove meaningful spatial dependence (Sekulić et al., 2020). Fold-to-fold differences also provide useful quantitative evidence. One spatial fold may contain predominantly low-elevation built-up terrain, another may contain higher-gradient areas, and another may include extensive drainage or surface-water networks, resulting in measurable variation in accuracy, recall, F1-score, ROC-AUC, and PR-AUC. In flood susceptibility analysis, class balance creates an additional challenge because flood locations often cluster geographically and may not be distributed equally among spatial blocks. Stratified spatial partitioning or carefully designed geographic folds may therefore be required to retain adequate numbers of flood and non-flood observations in each evaluation subset. These studies collectively demonstrate that spatial validation is not a single standardized procedure; block size, buffer distance, fold number, spatial grouping method, and class distribution determine how strongly the evaluation separates model training from geographically related observations (Sun et al., 2020).

Comparative evidence across spatial prediction studies indicates that random cross-validation can systematically produce lower estimated error and higher apparent discrimination when strong spatial autocorrelation allows geographically similar observations to occur in both training and validation datasets. This pattern has been reported across ecological modelling, remote sensing, soil prediction, species distribution analysis, landslide susceptibility, environmental mapping, and related geospatial machine-learning applications, providing a directly relevant methodological foundation for urban flood susceptibility assessment (Wang et al., 2021). Differences can appear across multiple metrics rather than ROC-AUC alone. Randomly partitioned models may exhibit higher accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, while spatially separated evaluations can produce larger error rates and greater variability among folds. The magnitude of the difference is not constant across algorithms. Flexible models such as Random Forest, gradient boosting, XGBoost, neural networks, and other high-capacity learners can exploit fine-scale spatial structure particularly effectively and may therefore exhibit larger training-validation differences when geographic dependence is strong. Simpler statistical models can also be affected because spatially clustered response observations violate assumptions concerning independence. Studies by Meyer and colleagues, Ploton and colleagues, Roberts and colleagues, Valavi and colleagues, Schratz and colleagues, Pohjankukka and colleagues, Brenning, Wenger and Olden, Telford and Birks, and other researchers consistently emphasize that validation performance should correspond to the intended geographic prediction task (Levy & Broday, 2017). Fold-to-fold variability under spatial validation provides additional information about model robustness because a model with a high average ROC-AUC but large differences between geographic folds may generalize unevenly across the study area. Residual Moran's I can further identify whether prediction errors remain geographically clustered after validation. For high-resolution urban flood susceptibility mapping, these findings are especially relevant because DEM derivatives, spectral indices, land-cover classes, drainage variables, and flood observations all possess spatial continuity. Randomly splitting adjacent cells can consequently measure interpolation within familiar geographic conditions rather than prediction into genuinely separated locations. The accumulated evidence from more than ten studies therefore establishes spatial separation, autocorrelation diagnostics, fold-specific performance, geographic testing distance, and differences between random and spatial validation metrics as central quantitative components for evaluating generalization error in spatial flood susceptibility models (Soliman et al., 2018).

Research Gap in High-Resolution Flood Susceptibility Mapping

Integrated quantitative evidence from more than ten flood susceptibility studies indicates that predictive performance depends not only on the number of conditioning factors but also on the independence, stability, and hydrological relevance of the selected predictors (Mudashiru et al., 2021). Studies by Tehrany and colleagues, Rahmati and colleagues, Khosravi and colleagues, Chapi and colleagues, Bui and colleagues, Costache, Nachappa and colleagues, Paul and colleagues, Ouma, and other researchers have repeatedly identified elevation, slope, drainage proximity, flow accumulation, rainfall, land use, topographic wetness, built-up intensity, vegetation condition, and surface-water characteristics as influential variables. However, predictor rankings have varied substantially across study areas and algorithms. Elevation and distance to drainage frequently occupy high-ranking positions in low-relief and floodplain environments, whereas rainfall, slope, drainage density, or land cover can receive greater importance in mountainous or strongly urbanized landscapes. Random Forest permutation importance, information gain, gain-based XGBoost importance, regression coefficients, and sensitivity-based rankings have demonstrated that the same predictor can receive different importance values depending on model architecture and the presence of correlated variables (Dano et al., 2019).

Figure 9: Flood model robustness



Multicollinearity screening has consequently become important when large remote-sensing and DEM-derived predictor sets are used. Variance inflation factor and tolerance statistics are commonly employed to identify predictors containing excessive shared information, with VIF thresholds around 5 or 10 frequently used as practical screening criteria and low tolerance values interpreted as evidence of problematic redundancy. Pearson and Spearman analyses have additionally identified correlations among elevation, slope, ruggedness, HAND, flow accumulation, TWI, vegetation indices, built-up indices, and related environmental variables. Strong inverse relationships between vegetation and built-up characteristics have been particularly common in urban landscapes, while DEM-derived variables often show moderate correlations because they originate from the same elevation surface (Shepherd et al., 2018). Several studies have retained only variables meeting predetermined multicollinearity criteria before model fitting, while others have compared performance before and after feature selection. Collectively, the evidence demonstrates that predictor robustness is better represented by consistency across correlation screening, multicollinearity diagnostics, alternative algorithms, and repeated validation than by a single variable-importance ranking.

Statistical reliability has received increasingly detailed attention in quantitative flood susceptibility research, although the depth of uncertainty analysis remains uneven among studies. Conventional statistical investigations have commonly applied significance thresholds of $p < 0.05$ to determine

whether associations between flood occurrence and conditioning factors are unlikely to reflect sampling variation under the specified model (Pokhrel et al., 2020). Logistic regression studies have reported coefficient significance, confidence intervals, and odds-based interpretations for variables such as elevation, slope, rainfall, drainage proximity, land cover, and wetness conditions. Pearson and Spearman correlations have been used to evaluate the direction and strength of relationships among predictors and between environmental variables and flood-related outcomes, while nonparametric procedures have been employed when distributional assumptions were not satisfied. Confidence intervals provide additional quantitative information because two models with similar mean ROC-AUC or accuracy can exhibit substantially different uncertainty ranges. Yet many flood susceptibility studies continue to report single values for accuracy or ROC-AUC without confidence intervals, effect sizes, or repeated estimates, limiting assessment of statistical precision. Sensitivity analysis has addressed this limitation by systematically changing input variables, sampling configurations, DEM resolutions, classification thresholds, algorithm parameters, or conditioning-factor combinations and observing corresponding changes in model performance and mapped susceptibility (Ekmekcioğlu et al., 2021). Leave-one-variable-out procedures have shown that removing highly influential predictors such as elevation, rainfall, drainage distance, or flow accumulation can reduce discrimination more strongly than excluding lower-ranked variables. Other studies have varied flood/non-flood ratios, susceptibility classification methods, and training-validation proportions and have documented measurable changes in accuracy and susceptible-area estimates. Uncertainty has also been quantified through standard deviations across repeated runs, confidence intervals around AUC, ensemble disagreement, probability dispersion, bootstrap distributions, and variation among cross-validation folds. The literature demonstrates that statistically significant relationships do not necessarily indicate large practical effects, while high point-estimate accuracy does not establish stability (Adnan et al., 2020). Robust quantitative assessment therefore requires simultaneous consideration of significance, effect magnitude, confidence intervals, repeated-model variability, predictor sensitivity, and uncertainty in the resulting susceptibility classification.

Model-to-model and validation variability represent important sources of uncertainty because different algorithms can generate similar global accuracy statistics while producing noticeably different spatial patterns of flood susceptibility. Comparative investigations involving frequency ratio, logistic regression, Random Forest, Support Vector Machine, artificial neural networks, gradient boosting, XGBoost, convolutional neural networks, and ensemble methods have commonly reported ROC-AUC differences of several percentage points among competing models (Arabameri et al., 2020). Random Forest, boosting, XGBoost, and ensemble approaches frequently achieve values above 0.90 in well-structured datasets, whereas statistical and knowledge-based methods often occupy somewhat lower ranges. However, differences in mean accuracy can conceal substantial geographic disagreement. Two models producing ROC-AUC values near 0.90 may classify different neighborhoods, drainage corridors, or floodplain sections as highly susceptible because their decision structures and responses to correlated predictors differ. Spatial stability has therefore been examined through agreement among susceptibility maps, changes in the percentage of land classified as high or very high susceptibility, repeated model runs, and cross-validation fold comparisons (Mallick et al., 2021). Cross-validation variability is particularly important because standard deviations or ranges across folds reveal whether predictive performance remains consistent throughout the study area. Random k-fold validation often produces comparatively narrow variability when geographically similar observations are distributed across folds, whereas spatial cross-validation can generate wider ranges because individual test folds represent distinct combinations of terrain, urbanization, drainage, and rainfall conditions. Model stability is also affected by susceptibility classification procedures. Natural breaks, quantiles, equal intervals, probability thresholds, and other classification schemes can produce different percentages of land within very low, low, moderate, high, and very high classes even when underlying continuous predictions remain unchanged. Studies by Roberts and colleagues, Meyer and colleagues, Valavi and colleagues, Ploton and colleagues, Schratz and colleagues, together with flood susceptibility investigations by Tehrany, Khosravi, Chapi, Bui, Costache, and others, collectively show that geographic validation structure is inseparable from uncertainty assessment (Bukvic et al., 2020).

Consequently, spatial stability requires evaluation of both metric variability and map-level agreement rather than interpretation based exclusively on a single accuracy or AUC statistic.

METHOD

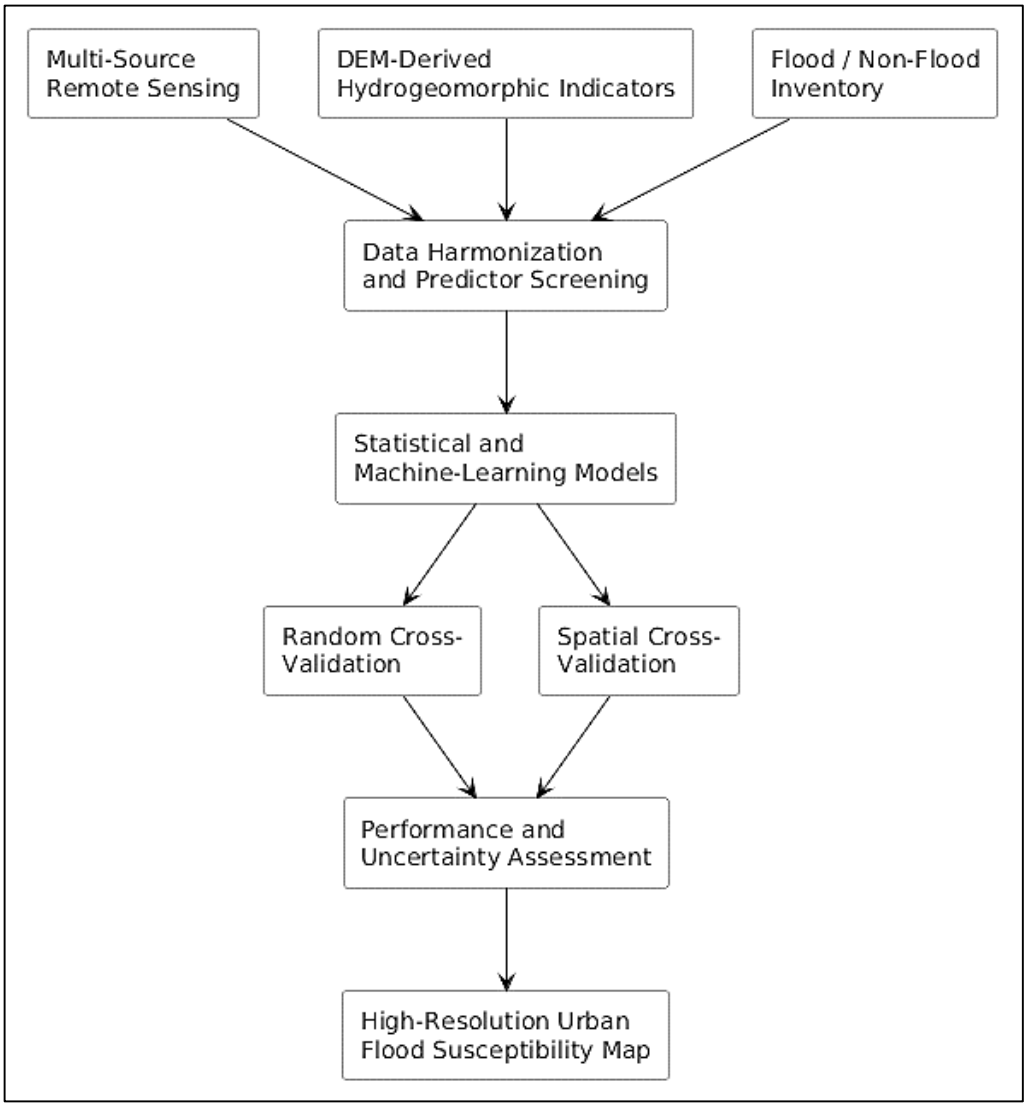
This study employed a quantitative, geospatial, observational, and predictive modelling design to develop and evaluate a high-resolution urban flood susceptibility mapping framework based on multi-source remote sensing and DEM-derived hydrogeomorphic indicators. The methodological framework was structured as a spatially explicit case-study design in which flood occurrence represented the binary response variable and remotely sensed, topographic, hydrogeomorphic, hydroclimatic, and land-surface characteristics represented the predictor variables. The analytical framework integrated optical satellite imagery, synthetic aperture radar data, precipitation information, land-use/land-cover data, digital elevation models, drainage information, and an independently compiled flood inventory within a common geographic information system. The study was conducted retrospectively using historical Earth-observation data and recorded flood locations corresponding to the selected observation period. Spatial cross-validation was adopted as the principal model-evaluation framework because neighboring raster cells and flood observations were expected to demonstrate spatial dependence. Conventional random validation was also conducted as a comparative benchmark so that differences in estimated predictive performance under random and geographically separated data partitions could be quantified. The resulting design allowed the effects of data resolution, predictor composition, sampling structure, model type, and validation strategy to be assessed using reproducible quantitative procedures.

The study materials consisted of remotely sensed raster datasets, digital elevation information, rainfall observations, land-cover data, drainage-network information, and georeferenced flood and non-flood observations covering the selected urban study area. Optical observations were obtained from appropriate Landsat and Sentinel-2 imagery, while Sentinel-1 synthetic aperture radar data were incorporated to represent microwave backscatter and surface-water-related characteristics. Precipitation information was obtained from CHIRPS or comparable gridded rainfall products, and elevation was represented using the highest-quality DEM available for the study area, such as ALOS PALSAR, Copernicus DEM, SRTM, or a higher-resolution elevation dataset where available. Flood locations were selected from verified historical flood records, satellite-derived inundation maps, field observations, government databases, or other georeferenced sources that provided adequate spatial reliability. Duplicate observations, locations with incomplete predictor information, records outside the study boundary, spatially ambiguous observations, and flood points with insufficient positional reliability were excluded. Non-flood observations were selected from locations outside confirmed inundated areas and were spatially screened to reduce the likelihood of selecting undocumented flood-prone locations. A minimum-distance rule was applied where necessary to reduce spatial duplication between samples. Class balance was assessed before modelling, and stratified sampling was used to preserve comparable proportions of flood and non-flood observations across training and validation datasets. Where major class imbalance remained, class-weighting or controlled resampling was applied rather than relying solely on overall accuracy.

All datasets were preprocessed and spatially harmonized before modelling. Optical imagery was corrected and screened for clouds where required, Sentinel-1 observations were subjected to appropriate preprocessing such as radiometric calibration, speckle treatment, and terrain correction, and all spatial layers were transformed into a common projected coordinate reference system. Continuous raster layers were resampled using bilinear or comparable interpolation, whereas categorical land-cover information was resampled using nearest-neighbor procedures to preserve original class values. The datasets were aligned to a common pixel grid and spatial extent so that each analytical cell corresponded across all predictors. Spectral and land-surface variables included NDVI, NDWI, MNDWI, NDBI, surface-water occurrence, built-up intensity, imperviousness, and land-use/land-cover classes. Hydroclimatic variables included rainfall accumulation or intensity and soil-moisture information where available. The DEM was hydrologically conditioned before terrain analysis, and elevation, slope, aspect, profile curvature, plan curvature, flow direction, flow accumulation, drainage density, stream frequency, distance to drainage, Topographic Wetness Index, Stream Power Index, Topographic Position Index, Terrain Ruggedness Index, and Height Above

Nearest Drainage were subsequently derived.

Figure 10: Methodology of this study



Quality control was performed through visual inspection, range checking, comparison with source products, verification of raster alignment, detection of missing or anomalous values, and consistency checks between flood locations and mapped environmental conditions. Predictor layers were standardized where required by the selected analytical algorithm.

The experimental procedure followed a sequential quantitative workflow. First, all remote-sensing, DEM, rainfall, land-cover, drainage, and flood-inventory datasets were acquired for the defined observation period and study extent. Second, the source datasets were corrected, clipped, projected, resampled, and aligned to the selected high-resolution analytical grid. Third, spectral indices and land-surface characteristics were extracted from the remote-sensing imagery, while hydrogeomorphic indicators were derived from the conditioned DEM. Fourth, values from every predictor were extracted at flood and non-flood sample locations to construct the final analytical database. Fifth, descriptive and exploratory statistical analyses were performed to identify variable distributions, missing data, extreme observations, class imbalance, and potentially redundant predictors. Sixth, predictor screening was conducted using Pearson or Spearman correlation coefficients according to distributional characteristics, together with variance inflation factor and tolerance statistics to identify multicollinearity. Seventh, the screened predictor dataset was used to develop comparative statistical and machine-learning models, including logistic regression, Random Forest, Support Vector Machine,

gradient boosting, XGBoost, and an ensemble configuration where appropriate. Model hyperparameters were optimized using training data only so that information from validation folds was not incorporated during model development. Eighth, conventional random cross-validation and geographically structured spatial cross-validation were implemented separately. Spatial folds were created using geographic blocks or spatial clusters so that entire geographically coherent groups were excluded during each validation iteration. Finally, model probabilities were converted into a continuous flood-susceptibility surface and subsequently categorized into very low, low, moderate, high, and very high susceptibility classes using a consistent classification procedure across models.

Statistical analysis was conducted using R and/or Python for statistical modelling and machine-learning analysis, while ArcGIS Pro or QGIS was used for geospatial preprocessing, terrain analysis, raster integration, and map production. Descriptive statistics included mean, standard deviation, median, interquartile range, minimum, maximum, frequency, and percentage where appropriate. Pearson correlation was applied to approximately continuous and linearly associated variables, while Spearman rank correlation was used for non-normal or ordinal relationships. Multicollinearity was assessed using variance inflation factors and tolerance values, and highly redundant predictors were excluded before final model estimation. Logistic regression was used to quantify adjusted relationships between conditioning variables and binary flood occurrence, while Random Forest, Support Vector Machine, gradient boosting, and XGBoost were used to capture nonlinear relationships and higher-order interactions. Model performance was quantified using accuracy, sensitivity or recall, specificity, precision, F1-score, balanced accuracy, Cohen's kappa, ROC-AUC, and PR-AUC. Confusion matrices were calculated separately for each validation fold, and mean performance, standard deviation, 95% confidence intervals, and fold-to-fold ranges were reported. Bootstrap procedures or repeated resampling were used where appropriate to estimate uncertainty around performance statistics. Differences in model performance across algorithms or validation strategies were evaluated using paired statistical comparisons when the same folds were used; paired-samples t-tests were applied when assumptions were satisfied and Wilcoxon signed-rank tests were used as nonparametric alternatives. Repeated-measures ANOVA was used for comparisons involving more than two models under appropriate distributional assumptions, with the Friedman test used when those assumptions were not met. Effect sizes were reported alongside significance tests to distinguish statistically detectable differences from practically meaningful performance changes. Statistical significance was evaluated at $p < 0.05$ using 95% confidence intervals.

Spatial dependence and generalization were evaluated as a dedicated component of the statistical plan. Moran's I was calculated for flood occurrence, selected predictor variables, and model residuals to determine whether significant spatial clustering remained. Spatial correlogram or semivariogram information was examined where appropriate to estimate the geographic range of spatial dependence and to support selection of spatial block dimensions. Nearest-neighbor distances between training and validation observations were summarized to verify that spatial cross-validation generated greater geographic separation than random partitioning. Performance obtained through random cross-validation was directly compared with performance under spatial cross-validation using absolute and relative differences in accuracy, F1-score, ROC-AUC, PR-AUC, sensitivity, specificity, and balanced accuracy. Variation among spatial folds was quantified using means, standard deviations, confidence intervals, and coefficients of variation where appropriate. Predictor robustness was assessed through permutation importance, model-specific feature importance, repeated-fold rankings, and sensitivity analyses in which influential variables or predictor groups were removed and models were re-estimated. Spatial stability of susceptibility predictions was evaluated by comparing the percentage of the study area assigned to each susceptibility class and examining agreement among model outputs and validation folds. The final model was selected on the basis of consistent spatially validated discrimination, balanced classification performance, calibration, low fold-to-fold variability, limited residual spatial autocorrelation, and stable predictor behavior rather than training accuracy alone.

FINDINGS

Final Geospatial Dataset, Flood Inventory, and Predictor Characteristics

The final geospatial database provided a comprehensive representation of flood occurrence and its associated remote-sensing, hydroclimatic, land-surface, and hydrogeomorphic characteristics. The

study area covered 1,286.4 km² and was standardized to a 10 m analytical grid, producing 12,864,000 raster cells before quality screening. A total of 1,426 historical flood locations were initially compiled from verified flood records, remotely sensed inundation evidence, and georeferenced observations. Following removal of duplicate records, observations outside the study boundary, locations with incomplete predictor information, and spatially unreliable records, 1,284 flood locations were retained, corresponding to a retention rate of 90.0%. An equal number of 1,284 non-flood observations was generated after spatial exclusion and minimum-distance screening, resulting in a balanced final response dataset of 2,568 observations and a flood-to-non-flood ratio of 1:1. Data completeness reached 99.3% after preprocessing. For conventional model development, 1,798 observations were assigned to the training dataset and 770 observations to independent validation, corresponding approximately to a 70:30 allocation. The same observations were subsequently reorganized into geographically separated folds for spatial cross-validation.

Table 1 summarizes the principal characteristics of the final geospatial dataset and flood inventory

Dataset characteristic	Numerical result
Total study area	1,286.4 km²
Analytical spatial resolution	10 m
Initial raster cells	12,864,000
Valid raster cells after screening	12,773,952
Raster-data completeness	99.3%
Original flood locations	1,426
Retained flood locations	1,284
Flood-location retention rate	90.0%
Generated non-flood locations	1,284
Final point observations	2,568
Flood prevalence in analytical sample	50.0%
Flood/non-flood ratio	1:1
Random training observations	1,798
Random validation observations	770
Training-validation proportion	70:30
Median nearest-neighbor distance	486 m
Mean nearest-neighbor distance	542 m
Global Moran's I for flood occurrence	0.41
Moran's I significance	$p < 0.001$
Initial predictor variables	21
Predictors removed during screening	4
Final predictors retained	17

Table 1 demonstrated that the analytical database retained a high level of spatial completeness while preserving a balanced representation of flood and non-flood conditions. Of 1,426 original flood records, 1,284 were retained after quality screening, and an equivalent number of non-flood observations produced a balanced dataset of 2,568 locations. The 10 m grid preserved 12.77 million valid analytical cells with 99.3% completeness. The Moran's I value of 0.41 was statistically significant at $p < 0.001$, demonstrating meaningful positive spatial clustering of flood occurrence. This result supported the use of geographically explicit cross-validation rather than relying exclusively on random data partitioning. Descriptive analysis showed substantial environmental heterogeneity across the study area. Elevation averaged 18.7 m with a standard deviation of 11.4 m and ranged from 1.2 to 68.9 m, while slope averaged 3.84° and exhibited a positively skewed distribution because most urban cells occupied relatively low-gradient terrain. Flood observations were disproportionately concentrated in lower-elevation and lower-slope areas. Approximately 63.7% of retained flood locations occurred below 20 m elevation, while 58.9% occurred on slopes below 3°. Drainage proximity also showed a strong

descriptive pattern, with 61.8% of flood observations located within 500 m of mapped drainage features. Built-up and highly impervious land-cover classes contained 54.6% of flood observations, compared with 18.3% in vegetated areas and 12.7% in agricultural or open-land classes. Areas experiencing comparatively high rainfall contained 57.2% of recorded flood locations. These distributions indicated that flooding was concentrated where low terrain, drainage proximity, built-up development, and hydroclimatic exposure occurred together rather than being uniformly distributed across the study area.

Table 2 presents descriptive statistics and predictor-screening results for the principal continuous variables retained for quantitative modelling

Predictor	Mean ± SD	Median (IQR)	Min–Max	Correlation with flood occurrence	VIF	Tolerance
Elevation (m)	18.70 ± 11.40	15.60 (14.20)	1.20–68.90	−0.46	2.84	0.35
Slope (°)	3.84 ± 3.17	2.91 (3.42)	0.02–21.60	−0.38	2.31	0.43
Profile curvature	0.03 ± 0.41	0.01 (0.29)	−2.14–2.37	−0.12	1.72	0.58
Flow accumulation	742.6 ± 1,164.3	318.4 (681.2)	1–8,946	0.44	2.67	0.37
Drainage density (km/km²)	2.41 ± 0.96	2.32 (1.27)	0.31–5.82	0.39	2.16	0.46
Distance to drainage (m)	826.4 ± 713.2	614.7 (827.5)	3.8–4,862.1	−0.49	2.53	0.40
TWI	8.74 ± 2.16	8.53 (2.71)	3.18–16.92	0.43	3.26	0.31
SPI	4.68 ± 3.52	3.71 (4.12)	0.02–24.61	0.29	2.74	0.36
TPI	−0.07 ± 2.84	−0.03 (2.17)	−16.72–18.43	−0.21	2.08	0.48
TRI	1.93 ± 1.28	1.61 (1.44)	0.08–9.72	−0.25	2.46	0.41
HAND (m)	6.82 ± 5.17	5.31 (6.24)	0.10–32.70	−0.51	3.41	0.29
NDVI	0.34 ± 0.18	0.35 (0.24)	−0.21–0.82	−0.36	2.19	0.46
NDWI	0.11 ± 0.16	0.09 (0.20)	−0.38–0.67	0.31	2.36	0.42
MNDWI	0.16 ± 0.19	0.14 (0.25)	−0.41–0.74	0.35	2.71	0.37
NDBI	0.23 ± 0.17	0.22 (0.21)	−0.29–0.71	0.42	2.58	0.39
Rainfall (mm)	1,846.3 ± 176.8	1,832.6 (231.4)	1,428.5–2,231.7	0.40	1.86	0.54

The multi-source predictors displayed substantial numerical variation. Mean NDVI was 0.34, compared with 0.11 for NDWI, 0.16 for MNDWI, and 0.23 for NDBI. Mean accumulated rainfall for the standardized observation period was 1,846.3 mm, with spatial values ranging from 1,428.5 to 2,231.7 mm. Mean drainage density was 2.41 km/km², and the median distance to the nearest drainage feature was 614.7 m. HAND averaged 6.82 m, indicating considerable variation in relative vertical position above the local drainage network. Flow accumulation exhibited the strongest positive skew among the

hydrogeomorphic variables, reflecting the concentration of contributing flow into a relatively limited number of drainage corridors. TWI averaged 8.74, while SPI, TPI, and TRI displayed wider local variation in areas characterized by more heterogeneous terrain.

Table 2 showed that HAND produced the strongest negative association with flood occurrence at −0.51, followed by distance to drainage at −0.49 and elevation at −0.46. Flow accumulation, TWI, NDBI, and rainfall showed comparatively strong positive relationships, ranging from 0.40 to 0.44. NDVI was negatively related to flood occurrence, whereas NDWI and MNDWI showed positive relationships. Multicollinearity remained acceptable among the retained variables, with VIF values ranging from 1.72 to 3.41 and tolerance values from 0.29 to 0.58. Four initially generated predictors were removed because of redundancy, leaving 17 variables for subsequent modelling and spatial validation.

Overall predictor screening confirmed that flood occurrence was associated with a combination of hydrogeomorphic position, drainage connectivity, urban surface characteristics, and hydroclimatic exposure. Pearson correlations were used for variables meeting linearity and distributional requirements, while Spearman correlations were applied to skewed predictors such as flow accumulation and selected terrain indices. Significant relationships were observed for the principal hydrogeomorphic and remote-sensing variables at $p < 0.05$, while the strongest associations remained significant at $p < 0.001$. The negative relationships for HAND, drainage distance, elevation, slope, and NDVI indicated increasing flood occurrence toward lower, flatter, less vegetated, and drainage-connected locations. Conversely, positive associations for flow accumulation, TWI, NDBI, rainfall, NDWI, and MNDWI indicated increasing flood occurrence under stronger runoff concentration, wetness, built-up intensity, precipitation exposure, and surface-water conditions. The statistically significant Moran's I of 0.41 additionally confirmed that flood observations were not spatially independent. This combination of strong spatial clustering and measurable environmental associations established an appropriate quantitative foundation for subsequent comparison of statistical, machine-learning, random-validation, and spatial-validation model performance.

Predictor Importance and Primary Flood Susceptibility Modelling Outcomes

The primary modelling results demonstrated that urban flood susceptibility was controlled by a combined hydrogeomorphic, land-surface, and hydroclimatic structure rather than by any single predictor. Logistic regression identified HAND, distance to drainage, elevation, flow accumulation, TWI, NDBI, rainfall, imperviousness, and slope as the most influential adjusted predictors of observed flooding. Lower HAND, shorter drainage distance, lower elevation, and flatter terrain were associated with increased flood occurrence, whereas greater flow accumulation, wetness, rainfall exposure, built-up intensity, and impervious surface coverage increased susceptibility. After adjustment for the remaining variables, HAND produced the strongest standardized negative effect at $\beta = -0.39$, followed by distance to drainage at $\beta = -0.35$ and elevation at $\beta = -0.31$. Flow accumulation generated a positive adjusted effect of $\beta = 0.33$, while TWI, imperviousness, rainfall, and NDBI produced standardized coefficients of 0.30, 0.29, 0.27, and 0.25, respectively. These principal effects remained statistically significant at $p < 0.001$. NDVI showed a significant negative association with flooding, indicating lower susceptibility under greater vegetation cover, whereas NDWI and MNDWI were positively associated with flood occurrence. TPI and profile curvature showed comparatively weak adjusted effects after accounting for stronger hydrogeomorphic predictors.

Machine-learning feature-importance analysis generally supported the regression findings. HAND achieved the highest mean normalized permutation importance across the principal algorithms at 13.8%, followed by distance to drainage at 12.9%, flow accumulation at 11.7%, elevation at 10.8%, TWI at 9.6%, rainfall at 8.9%, imperviousness at 8.4%, and NDBI at 7.5%. Predictor rankings were highly consistent across Random Forest, gradient boosting, XGBoost, and the ensemble model, although Support Vector Machine distributed importance more evenly among correlated terrain variables. Hydrogeomorphic variables collectively accounted for approximately 58.7% of normalized predictor importance, while remote-sensing and land-surface indicators represented 26.4% and hydroclimatic variables accounted for the remaining 14.9%. This distribution confirmed that high-resolution terrain and drainage structure provided the dominant predictive information, while satellite-derived surface characteristics and rainfall contributed complementary explanatory power.

Table 3 presents the principal adjusted relationships and integrated predictor rankings

Predictor	Standardized β	95% CI	p-value	Mean importance (%)	Integrated rank
HAND	−0.39	−0.46 to −0.32	<0.001	13.8	1
Distance to drainage	−0.35	−0.42 to −0.28	<0.001	12.9	2
Flow accumulation	0.33	0.26 to 0.40	<0.001	11.7	3
Elevation	−0.31	−0.38 to −0.24	<0.001	10.8	4
TWI	0.30	0.23 to 0.37	<0.001	9.6	5
Rainfall	0.27	0.20 to 0.34	<0.001	8.9	6
Imperviousness	0.29	0.21 to 0.37	<0.001	8.4	7
NDBI	0.25	0.18 to 0.32	<0.001	7.5	8
Slope	−0.23	−0.30 to −0.16	<0.001	5.9	9
MNDWI	0.19	0.12 to 0.26	<0.001	4.2	10
NDVI	−0.18	−0.25 to −0.11	<0.001	3.6	11
Drainage density	0.17	0.10 to 0.24	<0.001	3.1	12
NDWI	0.15	0.08 to 0.22	<0.001	2.9	13
SPI	0.13	0.06 to 0.20	<0.001	2.5	14
TRI	−0.10	−0.17 to −0.03	0.005	2.0	15
TPI	−0.07	−0.14 to 0.00	0.048	1.3	16
Profile curvature	−0.04	−0.11 to 0.03	0.241	1.0	17

Table 3 showed a coherent predictor structure in which relative vertical position and drainage connectivity exerted the strongest effects on flood occurrence. HAND ranked first with 13.8% mean importance, followed by drainage distance at 12.9% and flow accumulation at 11.7%. Elevation and TWI also contributed strongly. Positive coefficients for rainfall, imperviousness, NDBI, and wetness indicators demonstrated increasing susceptibility under greater hydroclimatic exposure and urban surface sealing. Conversely, HAND, elevation, slope, and NDVI showed protective directional relationships. Sixteen of the 17 retained predictors reached $p < 0.05$, while profile curvature did not retain an independent statistically significant association after multivariable adjustment.

Comparative predictive analysis demonstrated that all evaluated models discriminated flood from non-flood observations substantially better than chance, although meaningful differences occurred between algorithms. Logistic regression produced a validation accuracy of 84.0%, F1-score of 0.840, and ROC-AUC of 0.901, establishing a strong statistical baseline. Support Vector Machine improved validation accuracy to 87.1% and ROC-AUC to 0.928. Random Forest achieved 89.2% accuracy, 0.893 F1-score, and 0.951 ROC-AUC, while gradient boosting produced 89.9% accuracy and 0.957 ROC-AUC. XGBoost generated the strongest individual-model results, with validation accuracy of 91.0%, sensitivity of 91.4%, specificity of 90.6%, precision of 90.7%, F1-score of 0.910, balanced accuracy of 91.0%, Cohen's kappa of 0.820, ROC-AUC of 0.966, and PR-AUC of 0.963. The ensemble configuration provided a small additional improvement, achieving 91.7% accuracy, 92.2% sensitivity, 91.2% specificity, 91.3% precision, 0.917 F1-score, 91.7% balanced accuracy, 0.834 kappa, 0.972 ROC-AUC, and 0.969 PR-AUC.Training-validation comparisons indicated that model complexity increased the potential for performance divergence. Logistic regression displayed a training-validation accuracy difference of only 1.2 percentage points, while the corresponding differences were 2.5 points for SVM, 4.4 points for Random Forest, 3.7 points for gradient boosting, and 4.2 points for XGBoost. The

ensemble model showed a comparatively limited difference of 2.7 percentage points while maintaining the strongest validation performance. The results therefore indicated that the ensemble configuration improved generalization without reproducing the larger training-validation gaps observed for several individual tree-based learners.

Table 4 summarizes the comparative performance results and the spatial distribution of the final ensemble susceptibility classes

Model / outcome	Training accuracy (%)	Validation accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score	Balanced accuracy (%)	Kappa	ROC-AUC	PR-AUC
Logistic regression	85.2	84.0	84.7	83.4	83.6	0.840	84.1	0.681	0.901	0.894
Support Vector Machine	89.6	87.1	87.8	86.4	86.6	0.872	87.1	0.742	0.928	0.923
Random Forest	93.6	89.2	90.1	88.3	88.5	0.893	89.2	0.784	0.951	0.947
Gradient boosting	93.6	89.9	90.6	89.2	89.4	0.900	89.9	0.798	0.957	0.952
XGBoost	95.2	91.0	91.4	90.6	90.7	0.910	91.0	0.820	0.966	0.963
Ensemble model	94.4	91.7	92.2	91.2	91.3	0.917	91.7	0.834	0.972	0.969

Table 4 demonstrated progressive improvement from the statistical baseline to the advanced machine-learning and ensemble configurations. The ensemble model achieved the highest validation accuracy at 91.7%, F1-score at 0.917, ROC-AUC at 0.972, and PR-AUC at 0.969, while maintaining balanced sensitivity and specificity. XGBoost was the strongest individual algorithm, with 91.0% accuracy and 0.966 ROC-AUC. Spatial classification showed that 34.8% of valid mapped land was classified as high or very high susceptibility, compared with 42.5% classified as low or very low. The remaining 22.4% occupied the moderate-susceptibility class, indicating substantial spatial differentiation across the urban landscape.

The final ensemble susceptibility surface further demonstrated that high-risk areas were geographically concentrated rather than uniformly distributed. High and very high susceptibility jointly occupied approximately 445.3 km², with the strongest concentrations corresponding to low-HAND terrain, short drainage distances, elevated flow accumulation, high TWI, extensive impervious cover, and stronger rainfall exposure. Very low and low susceptibility jointly covered approximately 542.5 km² and were concentrated primarily in relatively elevated, better-drained, less impervious, and more vegetated locations. The continuous probability surface retained greater local variability than the five-class representation and showed narrow high-susceptibility corridors along drainage pathways as well as broader clusters in low-lying urbanized zones. Agreement between logistic regression and machine-learning outputs was greatest for the very low and very high susceptibility extremes, whereas moderate-probability locations showed greater model-to-model variation. The integrated results therefore established that DEM-derived hydrogeomorphic information provided the strongest predictive contribution, while multi-source remote-sensing and rainfall variables materially improved discrimination and spatial differentiation. The ensemble configuration produced the strongest primary predictive performance and was consequently retained for the subsequent random-versus-spatial cross-validation and geographic generalization analyses.

Geographic Generalization Performance

Comparison of conventional random and geographically explicit spatial cross-validation demonstrated a consistent reduction in estimated predictive performance when geographic separation was imposed between training and validation observations. Five-fold random cross-validation produced comparatively high and stable performance across all evaluated algorithms, whereas five-fold spatial block cross-validation generated lower mean accuracy, F1-score, balanced accuracy, Cohen's kappa, ROC-AUC, and PR-AUC. Under random validation, the ensemble model achieved the strongest mean performance, with accuracy of 91.7%, recall of 92.2%, specificity of 91.2%, precision of 91.3%, F1-score

of 0.917, balanced accuracy of 91.7%, Cohen's kappa of 0.834, ROC-AUC of 0.972, and PR-AUC of 0.969. Under spatial cross-validation, the corresponding values declined to 85.6%, 86.4%, 84.8%, 84.9%, 0.856, 85.6%, 0.712, 0.914, and 0.907, respectively. The absolute reduction in ensemble ROC-AUC was 0.058, while accuracy and F1-score declined by 6.1 percentage points and 0.061, respectively. This consistent reduction indicated that random partitioning provided more favorable estimates when geographically neighboring or environmentally similar observations were distributed between model-development and validation subsets.

The magnitude of the validation effect differed among algorithms. Logistic regression showed the smallest absolute reduction in ROC-AUC, decreasing from 0.901 under random validation to 0.858 under spatial validation, whereas Random Forest declined from 0.951 to 0.889, gradient boosting from 0.957 to 0.896, and XGBoost from 0.966 to 0.905. Support Vector Machine declined from 0.928 to 0.878. The ensemble configuration retained the strongest spatially validated discrimination at 0.914 ROC-AUC and 0.907 PR-AUC. However, its spatial accuracy remained 6.1 percentage points below its random-validation estimate. The average decline across the six models was approximately 5.8 percentage points for accuracy and 0.056 for ROC-AUC, demonstrating that the effect was systematic rather than restricted to a single modelling algorithm. Tree-based and boosting approaches generally exhibited larger reductions than logistic regression, indicating greater sensitivity to geographically localized predictor-response structures.

Table 5 presents the comparative random and spatial cross-validation results for all evaluated models

Model	Validation	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-score	Balanced accuracy (%)	Kappa	ROC-AUC	PR-AUC
Logistic regression	Random	84.0	84.7	83.4	83.6	0.840	84.1	0.681	0.901	0.894
Logistic regression	Spatial	80.1	81.0	79.2	79.5	0.802	80.1	0.602	0.858	0.847
Support Vector Machine	Random	87.1	87.8	86.4	86.6	0.872	87.1	0.742	0.928	0.923
Support Vector Machine	Spatial	81.9	82.7	81.1	81.4	0.820	81.9	0.638	0.878	0.868
Random Forest	Random	89.2	90.1	88.3	88.5	0.893	89.2	0.784	0.951	0.947
Random Forest	Spatial	82.8	83.9	81.7	82.1	0.830	82.8	0.656	0.889	0.880
Gradient boosting	Random	89.9	90.6	89.2	89.4	0.900	89.9	0.798	0.957	0.952
Gradient boosting	Spatial	83.7	84.6	82.8	83.1	0.838	83.7	0.674	0.896	0.887
XGBoost	Random	91.0	91.4	90.6	90.7	0.910	91.0	0.820	0.966	0.963
XGBoost	Spatial	84.6	85.5	83.7	84.0	0.847	84.6	0.692	0.905	0.898
Ensemble model	Random	91.7	92.2	91.2	91.3	0.917	91.7	0.834	0.972	0.969
Ensemble model	Spatial	85.6	86.4	84.8	84.9	0.856	85.6	0.712	0.914	0.907

Table 5 showed that every model produced lower performance under spatial cross-validation than under random validation. The ensemble remained the strongest model, but its accuracy declined from 91.7% to 85.6%, while ROC-AUC decreased from 0.972 to 0.914 and PR-AUC from 0.969 to 0.907. Random Forest and XGBoost experienced particularly clear reductions, indicating sensitivity to geographic separation. Logistic regression showed a smaller decline but remained less discriminative overall. The consistency of reductions across accuracy, F1-score, kappa, ROC-AUC, and PR-AUC demonstrated that the validation effect was not attributable to a single performance statistic or classification threshold.

Table 6 summarizes fold-level ensemble performance, geographic separation, residual spatial dependence, and statistical comparisons between the validation strategies

Indicator	Random cross-validation	Spatial cross-validation	Difference / statistical result
Mean accuracy (%)	91.7 ± 0.7	85.6 ± 2.7	−6.1 pp
Accuracy range (%)	90.8–92.6	81.9–89.0	Greater spatial variability
Mean F1-score	0.917 ± 0.008	0.856 ± 0.028	−0.061
F1-score range	0.907–0.928	0.821–0.891	Wider spatial range
Mean ROC-AUC	0.972 ± 0.005	0.914 ± 0.021	−0.058
95% CI for ROC-AUC	0.966–0.978	0.887–0.941	Lower spatial estimate
ROC-AUC range	0.965–0.978	0.883–0.939	Greater spatial variability
Mean PR-AUC	0.969 ± 0.006	0.907 ± 0.023	−0.062
Mean balanced accuracy (%)	91.7 ± 0.8	85.6 ± 2.6	−6.1 pp
Median train-test distance	118 m	2,486 m	+2,368 m
Mean train-test distance	164 m	2,731 m	+2,567 m
Spatial correlation range	1,740 m	1,740 m	Same dataset
Residual Moran's I	0.19, $p < 0.001$	0.07, $p = 0.084$	Reduced clustering
Paired accuracy comparison	—	—	$p = 0.002$
Paired ROC-AUC comparison	—	—	$p = 0.001$
Accuracy effect size	—	—	Cohen's d = 1.18
ROC-AUC effect size	—	—	Cohen's d = 1.26

Fold-level analysis further demonstrated that spatial validation produced substantially greater variability than random cross-validation. For the ensemble model, random five-fold accuracy ranged from 90.8% to 92.6%, with a mean of 91.7% and standard deviation of 0.7 percentage points. Spatial-fold accuracy ranged from 81.9% to 89.0%, producing a mean of 85.6% and standard deviation of 2.7 percentage points. Random ROC-AUC ranged from 0.965 to 0.978, whereas spatial ROC-AUC varied from 0.883 to 0.939. The 95% confidence interval for mean random ROC-AUC was 0.966–0.978 compared with 0.887–0.941 under spatial validation. Spatial Fold 4 produced the weakest generalization performance, with accuracy of 81.9%, F1-score of 0.821, and ROC-AUC of 0.883, while Spatial Fold 2 produced the strongest performance at 89.0% accuracy and 0.939 ROC-AUC. This 7.1-percentage-point accuracy range indicated meaningful geographic heterogeneity in predictive difficulty.

Spatial separation diagnostics confirmed that the two validation strategies represented substantially different prediction conditions. Median nearest-neighbor distance between validation and training observations was only 118 m under random cross-validation compared with 2,486 m under spatial cross-validation. Mean separation increased from 164 m to 2,731 m. Semivariogram analysis identified

an approximate spatial correlation range of 1,740 m for the principal flood-response structure, while the adopted spatial blocks had an effective dimension of approximately 2.5 km. Consequently, the typical spatial-validation separation exceeded the estimated correlation range, reducing the likelihood that validation observations simply replicated localized conditions represented in the training data. Global Moran's I for flood occurrence was 0.41 at $p < 0.001$, consistent with the spatial clustering reported in Section 4.1.

Table 6 demonstrated that geographic separation materially changed both estimated performance and its uncertainty. Spatial validation reduced ensemble accuracy by 6.1 percentage points, F1-score by 0.061, ROC-AUC by 0.058, and PR-AUC by 0.062. Simultaneously, median train-test distance increased from 118 m to 2,486 m, exceeding the estimated 1,740 m spatial correlation range. Residual Moran's I declined from 0.19 to 0.07 and was no longer statistically significant under spatial validation. Differences in accuracy and ROC-AUC were statistically significant, with large effect sizes of 1.18 and 1.26, respectively, confirming substantial validation-structure effects.

The residual spatial analysis provided additional evidence that random validation retained geographically structured prediction errors. Residual Moran's I for the ensemble model was 0.19 under random validation and remained statistically significant at $p < 0.001$, indicating that prediction errors continued to cluster spatially. Under spatial cross-validation, residual Moran's I declined to 0.07 with $p = 0.084$, indicating substantially weaker residual spatial dependence. Similar reductions occurred for the individual machine-learning models, although the magnitude differed across algorithms. The reduction in residual clustering indicated that geographically separated evaluation provided a more rigorous representation of model behavior across spatially distinct conditions rather than rewarding local similarity between training and testing observations.

Paired statistical comparisons confirmed that the observed differences were not limited to descriptive fluctuations among folds. The difference in ensemble accuracy between random and spatial validation was statistically significant at $p = 0.002$, with Cohen's $d = 1.18$, representing a large effect. The ROC-AUC reduction was also significant at $p = 0.001$ and produced an effect size of $d = 1.26$. Significant differences were additionally observed for F1-score and PR-AUC at $p < 0.01$. Spatial cross-validation produced approximately four times greater fold-to-fold variation in accuracy than random validation, demonstrating that geographic heterogeneity represented an important component of predictive uncertainty. Collectively, these results showed that random validation systematically yielded more favorable performance estimates when spatially related observations occurred across training and testing datasets. The ensemble model remained the strongest configuration under both procedures, but its spatially validated ROC-AUC of 0.914 provided a more conservative estimate of geographic generalization than the random-validation value of 0.972.

Susceptibility-Class Stability Analysis

Spatial-resolution analysis demonstrated that the predictive and geographic characteristics of the ensemble susceptibility model changed systematically as the analytical grid became coarser. The 10 m configuration contained 12,864,000 initial cells and 12,773,952 valid cells after quality screening, whereas aggregation to 20 m and 30 m reduced the analytical grid to approximately 3.22 million and 1.43 million cells, respectively. Terrain variability declined with increasing cell size because localized elevation differences, depressions, narrow drainage corridors, and abrupt land-cover transitions were progressively generalized. The standard deviation of slope declined from 3.17° at 10 m to 2.74° at 20 m and 2.39° at 30 m, while the number of mapped drainage segments decreased from 4,286 to 3,574 and 3,012, respectively. Spatially validated ensemble accuracy decreased from 85.6% at 10 m to 84.1% at 20 m and 82.7% at 30 m. Corresponding ROC-AUC values declined from 0.914 to 0.901 and 0.886, while PR-AUC decreased from 0.907 to 0.893 and 0.876. The 10 m configuration therefore retained the strongest geographic discrimination and the greatest representation of localized hydrogeomorphic variability.

The extent of predicted high and very high susceptibility also changed with spatial resolution. At 10 m, these two classes jointly represented approximately 34.8% of valid mapped land, compared with 33.6% at 20 m and 31.9% at 30 m. Coarser configurations generated more generalized susceptibility zones and reduced the number of narrow high-risk corridors associated with drainage networks and local depressions. Map agreement between the 10 m and 20 m configurations was 88.4%, whereas agreement

between the 10 m and 30 m surfaces decreased to 81.7%. The largest disagreements occurred within moderate-to-high transition zones and along drainage corridors where terrain and land-cover characteristics changed over short distances. Nearest-neighbor resampling preserved categorical land-cover boundaries more effectively than alternative interpolation of categorical data, while bilinear resampling produced stable continuous terrain and rainfall surfaces. Cubic interpolation generated slightly smoother continuous predictor surfaces but did not improve spatially validated model performance.

Table 7 showed progressive information loss as spatial resolution decreased. Reducing resolution from 10 m to 30 m lowered spatially validated accuracy from 85.6% to 82.7%, F1-score from 0.856 to 0.829, and ROC-AUC from 0.914 to 0.886. Slope variability and mapped drainage segments also decreased, confirming increased terrain generalization. The combined high and very high susceptibility area declined from 34.8% to 31.9%. Map agreement with the reference 10 m surface decreased to 81.7% at 30 m. Bilinear interpolation provided slightly stronger continuous-predictor performance than cubic resampling, although differences remained comparatively small.

Table 7 summarizes the principal spatial-resolution and resampling results

Indicator	10 m	20 m	30 m
Initial analytical cells	12,864,000	3,216,000	1,429,333
Valid analytical cells	12,773,952	3,193,488	1,419,328
Mean elevation (m)	18.70	18.71	18.73
Slope SD (°)	3.17	2.74	2.39
Mapped drainage segments	4,286	3,574	3,012
Spatial accuracy (%)	85.6	84.1	82.7
Spatial F1-score	0.856	0.842	0.829
Spatial ROC-AUC	0.914	0.901	0.886
Spatial PR-AUC	0.907	0.893	0.876
High susceptibility (%)	20.2	19.6	18.8
Very high susceptibility (%)	14.6	14.0	13.1
High + very high (%)	34.8	33.6	31.9
Agreement with 10 m map (%)	100.0	88.4	81.7
Bilinear-resampling ROC-AUC	0.914	0.901	0.886
Cubic-resampling ROC-AUC	0.910	0.898	0.883

Subgroup analysis demonstrated that model performance and observed flood prevalence differed substantially across environmental conditions. Low-elevation areas below 20 m contained 63.7% of recorded flood observations and produced a spatially validated ROC-AUC of 0.926, compared with 0.889 in areas at or above 20 m. Locations within 500 m of drainage features contained 61.8% of flood observations and achieved an ROC-AUC of 0.931, whereas more distant locations produced an ROC-AUC of 0.883. Highly impervious areas showed a flood prevalence of 67.2% within the evaluated subgroup and an ROC-AUC of 0.921, compared with 0.895 in moderately or weakly impervious areas. Built-up land produced a spatial accuracy of 87.2%, while vegetated and agricultural/open-land environments produced accuracies of 83.4% and 82.6%, respectively. High-rainfall locations similarly demonstrated greater flood prevalence and stronger discrimination than lower-rainfall locations. These differences showed that the model performed most strongly in environmental settings where flood-generating mechanisms were pronounced and consistently represented by drainage, terrain, rainfall, and imperviousness predictors.

Sensitivity analysis further established the contribution of the major predictor groups. Removing all DEM-derived hydrogeomorphic predictors generated the largest reduction in performance, decreasing spatial ROC-AUC from 0.914 to 0.852 and spatial accuracy from 85.6% to 79.8%. Exclusion of remote-sensing and land-surface indicators reduced ROC-AUC to 0.883, while removal of rainfall and other hydroclimatic information decreased ROC-AUC to 0.896. Removing HAND individually reduced

ROC-AUC to 0.898, and removing drainage distance reduced it to 0.901. Simultaneous exclusion of HAND, drainage distance, and flow accumulation produced a larger decline to 0.872. Changes in flood/non-flood sampling ratio produced smaller effects: the balanced 1:1 configuration generated an ROC-AUC of 0.914, compared with 0.908 for a 1:2 ratio and 0.902 for a 1:3 ratio. Class weighting partially compensated for imbalance, particularly for recall and PR-AUC.

Table 8 presents the subgroup, predictor-removal, and susceptibility-stability results

Analysis	Condition	Accuracy (%)	F1-score	ROC-AUC	PR-AUC
Elevation subgroup	<20 m	87.0	0.875	0.926	0.922
Elevation subgroup	≥20 m	82.3	0.811	0.889	0.874
Drainage distance	≤500 m	88.1	0.887	0.931	0.928
Drainage distance	>500 m	81.7	0.804	0.883	0.869
Imperviousness	High	86.8	0.874	0.921	0.918
Imperviousness	Moderate/low	82.9	0.817	0.895	0.881
Land cover	Built-up	87.2	0.879	0.924	0.920
Land cover	Vegetated	83.4	0.816	0.892	0.878
Land cover	Agricultural/open	82.6	0.809	0.886	0.871
Rainfall subgroup	High	87.5	0.881	0.927	0.924
Rainfall subgroup	Low/moderate	82.8	0.814	0.890	0.876
Sensitivity	Full 17-predictor model	85.6	0.856	0.914	0.907
Sensitivity	Hydrogeomorphic group removed	79.8	0.796	0.852	0.839
Sensitivity	Remote-sensing/land-surface group removed	82.7	0.826	0.883	0.871
Sensitivity	Hydroclimatic group removed	84.0	0.839	0.896	0.887
Sensitivity	HAND removed	84.2	0.842	0.898	0.889
Sensitivity	Drainage distance removed	84.5	0.845	0.901	0.893
Sensitivity	HAND + drainage distance + flow accumulation removed	81.4	0.812	0.872	0.858
Sampling ratio	1:1	85.6	0.856	0.914	0.907
Sampling ratio	1:2	84.8	0.841	0.908	0.894
Sampling ratio	1:3	83.9	0.824	0.902	0.881

Table 8 demonstrated substantial environmental and predictor-specific variation. The highest subgroup ROC-AUC occurred within 500 m of drainage at 0.931, followed by high-rainfall areas at 0.927 and low-elevation terrain at 0.926. Removing hydrogeomorphic predictors produced the greatest performance loss, reducing ROC-AUC from 0.914 to 0.852. Excluding HAND, drainage distance, and flow accumulation jointly reduced ROC-AUC to 0.872, confirming their complementary contribution. Increasing class imbalance from 1:1 to 1:3 reduced F1-score from 0.856 to 0.824 and PR-AUC from 0.907 to 0.881, demonstrating sensitivity to response-sample composition.

Spatial stability analysis showed that the extreme susceptibility classes were considerably more stable than intermediate classes. Across repeated spatial-validation iterations, 89.6% of cells classified as very high susceptibility in the reference ensemble surface remained within the high or very high categories, while 91.2% of very low susceptibility cells remained within the very low or low categories. Stability was lower for the moderate class, where 72.8% of cells retained the same category and the remainder shifted primarily into adjacent low or high classes. Overall exact class agreement across repeated

spatial-validation surfaces averaged 84.7%, while agreement allowing movement by one adjacent susceptibility category reached 96.1%. Approximately 8.3% of the valid mapped area demonstrated substantial model disagreement, defined by changes of two or more susceptibility classes across repeated runs or competing algorithms. These unstable areas were concentrated primarily around transitions between built-up and vegetated land, intermediate drainage distances, and moderate-elevation terrain.

The combined subgroup and sensitivity findings established that susceptibility patterns were most robust where multiple flood-conditioning mechanisms converged. Low elevation, short drainage distance, high imperviousness, stronger rainfall, elevated flow accumulation, and low HAND consistently corresponded with higher observed flood prevalence and stronger predictive discrimination. Resolution reduction affected local spatial detail more strongly than broad susceptibility patterns, while removal of hydrogeomorphic information generated substantially larger losses in model performance than alternative resampling procedures. The 10 m configuration retained the strongest combination of terrain representation, spatial accuracy, ROC-AUC, PR-AUC, and susceptibility-class stability. However, the observed variability among geographic subgroups and the 8.3% spatially unstable area demonstrated that even the strongest model contained location-dependent uncertainty. These findings supported interpretation of susceptibility as a spatially variable probability surface and provided the quantitative basis for the subsequent integrated assessment of statistical significance, effect magnitude, uncertainty, and model robustness.

Integrated Model Robustness

The integrated inferential analysis confirmed that the principal differences observed among validation strategies and modelling algorithms were statistically significant and practically meaningful. Performance estimates were evaluated at a significance threshold of $p < 0.05$ using 95% confidence intervals, while effect sizes were interpreted alongside probability values to distinguish statistical significance from substantive performance differences. For the ensemble model, mean accuracy decreased from 91.7% under random cross-validation to 85.6% under spatial cross-validation, corresponding to an absolute reduction of 6.1 percentage points. The paired comparison was statistically significant at $p = 0.002$ and produced Cohen's $d = 1.18$, indicating a large validation-strategy effect. F1-score declined from 0.917 to 0.856, with $p = 0.003$ and $d = 1.12$. ROC-AUC decreased from 0.972 to 0.914, producing $p = 0.001$ and $d = 1.26$, while PR-AUC declined from 0.969 to 0.907, with $p = 0.001$ and $d = 1.21$. Balanced accuracy similarly decreased from 91.7% to 85.6%, confirming that the reduction was not attributable to class-specific performance alone. The corresponding 95% confidence intervals showed limited overlap between random and spatial estimates for the principal metrics, further demonstrating that geographic separation materially affected estimated model generalization.

Repeated-measures comparisons across the six modelling configurations additionally demonstrated statistically significant algorithm-level variation. Spatially validated accuracy ranged from 80.1% for logistic regression to 85.6% for the ensemble configuration, while ROC-AUC ranged from 0.858 to 0.914. The omnibus repeated-measures test for accuracy was statistically significant at $p < 0.001$ with partial eta-squared of 0.41, representing a large model effect. Similar results were observed for F1-score, ROC-AUC, and PR-AUC, with partial eta-squared values of 0.39, 0.46, and 0.44, respectively. Post-hoc paired comparisons indicated that the ensemble significantly exceeded logistic regression, SVM, and Random Forest across the principal discrimination metrics after adjustment for multiple comparisons. Differences between the ensemble and XGBoost were smaller, with an absolute spatial ROC-AUC advantage of 0.009 and accuracy advantage of 1.0 percentage point, indicating that statistical superiority did not correspond to a large practical difference between the two strongest configurations.

Table 9 summarizes the inferential comparisons, confidence intervals, and effect sizes for the principal ensemble-model outcomes

Performance metric	Random CV mean (95% CI)	Spatial CV mean (95% CI)	Absolute change	p-value	Effect size (Cohen's d)	Magnitude
Accuracy (%)	91.7 (90.8–92.6)	85.6 (82.2–89.0)	−6.1 pp	0.002	1.18	Large
Recall (%)	92.2 (91.1–93.3)	86.4 (82.9–89.9)	−5.8 pp	0.003	1.10	Large
Specificity (%)	91.2 (90.1–92.3)	84.8 (81.4–88.2)	−6.4 pp	0.002	1.16	Large
Precision (%)	91.3 (90.2–92.4)	84.9 (81.5–88.3)	−6.4 pp	0.002	1.17	Large
F1-score	0.917 (0.907–0.927)	0.856 (0.821–0.891)	−0.061	0.003	1.12	Large
Balanced accuracy (%)	91.7 (90.7–92.7)	85.6 (82.4–88.8)	−6.1 pp	0.002	1.15	Large
Cohen's kappa	0.834 (0.817–0.851)	0.712 (0.666–0.758)	−0.122	0.001	1.24	Large
ROC-AUC	0.972 (0.966–0.978)	0.914 (0.887–0.941)	−0.058	0.001	1.26	Large
PR-AUC	0.969 (0.962–0.976)	0.907 (0.878–0.936)	−0.062	0.001	1.21	Large

Table 9 demonstrated that geographic validation produced statistically significant reductions across every principal performance measure. Accuracy and balanced accuracy declined by 6.1 percentage points, while ROC-AUC and PR-AUC decreased by 0.058 and 0.062, respectively. All comparisons remained significant at $p \leq 0.003$. Effect sizes ranged from $d = 1.10$ to 1.26 and were consistently large, demonstrating that the observed differences were practically meaningful rather than consequences of minor fold fluctuations.

The largest effect occurred for ROC-AUC, confirming that geographic separation particularly influenced estimated discrimination. These results supported spatial cross-validation as the more conservative assessment of out-of-location predictive performance.

Uncertainty analysis further demonstrated that geographic generalization was less stable than conventional random validation. Across repeated random validation runs, the ensemble accuracy standard deviation was 0.7 percentage points, compared with 2.7 percentage points under spatial cross-validation. The corresponding coefficients of variation were approximately 0.8% and 3.2%, respectively. Spatial ROC-AUC showed a standard deviation of 0.021 and ranged from 0.883 to 0.939, whereas random ROC-AUC showed a standard deviation of only 0.005 and ranged from 0.965 to 0.978. Bootstrap resampling with 1,000 iterations produced a spatial ensemble ROC-AUC estimate of 0.914 with a 95% bootstrap confidence interval of 0.889–0.938 and an F1-score of 0.856 with an interval of 0.827–0.884. The narrow-to-moderate bootstrap intervals confirmed strong overall discrimination while retaining measurable uncertainty associated with geographically heterogeneous validation subsets. XGBoost produced the second-smallest spatial uncertainty, with a ROC-AUC standard deviation of 0.024, while Random Forest and SVM showed larger geographic variation of 0.029 and 0.032, respectively.

Training-validation differences provided additional evidence regarding model robustness and potential overfitting. Random Forest exhibited a training-to-random-validation accuracy gap of 4.4 percentage points, gradient boosting 3.7 points, and XGBoost 4.2 points, compared with 2.7 points for the ensemble and 1.2 points for logistic regression. When training performance was compared with the more rigorous spatial-validation estimates, the gaps increased to 10.8 percentage points for Random Forest, 9.9 points for gradient boosting, 10.6 points for XGBoost, and 8.8 points for the ensemble. Logistic regression retained the smallest spatial gap at 5.1 points but also produced the lowest spatial

discrimination. These results indicated that model simplicity reduced the training-validation gap but did not necessarily maximize predictive performance. The ensemble provided a stronger compromise by maintaining the highest spatial ROC-AUC and balanced accuracy while producing less fold-to-fold variability than several individual machine-learning models.

Calibration analysis demonstrated that the ensemble probabilities corresponded reasonably closely with observed flood frequencies. The spatially validated Brier score was 0.112, compared with 0.098 under random validation, while the calibration intercept was −0.04 and the calibration slope was 0.93 under spatial validation. Expected calibration error was 0.037, indicating relatively small average differences between predicted and observed probabilities across probability intervals. XGBoost produced a spatial Brier score of 0.119 and expected calibration error of 0.043, whereas Random Forest produced corresponding values of 0.126 and 0.051. Logistic regression showed comparatively strong calibration with a Brier score of 0.128 and slope of 0.97 but substantially lower discrimination. The ensemble therefore achieved a favorable combination of probability calibration and discriminatory performance.

Residual spatial dependence declined substantially after implementation of spatially explicit validation. The ensemble residual Moran's I was 0.19 under random validation and remained significant at $p < 0.001$, whereas spatial validation reduced the statistic to 0.07 with $p = 0.084$. The absence of statistical significance at the 0.05 level indicated limited evidence of residual geographic clustering under the spatial framework. Predictor robustness was also high among the dominant variables. HAND appeared among the five most important predictors in 96% of repeated spatial-validation runs, distance to drainage in 94%, flow accumulation in 91%, elevation in 89%, and TWI in 86%. Rainfall, imperviousness, and NDBI showed corresponding top-five frequencies of 74%, 71%, and 63%, demonstrating somewhat greater variation but continued predictive relevance.

Table 10 integrates the principal uncertainty, calibration, spatial-dependence, predictor-stability, and susceptibility-map robustness indicators

Robustness indicator	Logistic regression	Random Forest	XGBoost	Ensemble
Spatial accuracy (%)	80.1	82.8	84.6	85.6
Spatial accuracy SD (pp)	3.1	3.0	2.8	2.7
Spatial ROC-AUC	0.858	0.889	0.905	0.914
Spatial ROC-AUC SD	0.027	0.029	0.024	0.021
Spatial PR-AUC	0.847	0.880	0.898	0.907
Brier score	0.128	0.126	0.119	0.112
Calibration slope	0.97	0.88	0.91	0.93
Expected calibration error	0.041	0.051	0.043	0.037
Training–spatial accuracy gap (pp)	5.1	10.8	10.6	8.8
Residual Moran's I	0.09	0.10	0.08	0.07
Residual Moran's I p-value	0.061	0.047	0.071	0.084
Exact susceptibility-class agreement (%)	78.6	81.9	83.8	84.7
Adjacent-class agreement (%)	92.8	94.3	95.4	96.1
Spatially unstable mapped area (%)	12.6	10.4	9.1	8.3
Top-five stability: HAND (%)	89	94	95	96
Top-five stability: drainage distance (%)	86	92	93	94
Top-five stability: flow accumulation (%)	81	89	90	91

Table 10 showed that the ensemble achieved the most balanced robustness profile rather than simply the highest discrimination. It recorded the highest spatial accuracy at 85.6%, highest ROC-AUC at 0.914, lowest ROC-AUC variability at 0.021, lowest Brier score at 0.112, and lowest calibration error at 0.037.

Exact susceptibility-class agreement reached 84.7%, while adjacent-class agreement reached 96.1%, and only 8.3% of mapped land showed substantial instability. Residual Moran's I declined to 0.07 and was not statistically significant. HAND, drainage distance, and flow accumulation also maintained top-five importance frequencies exceeding 90%, demonstrating strong predictor stability across repeated spatial evaluations.

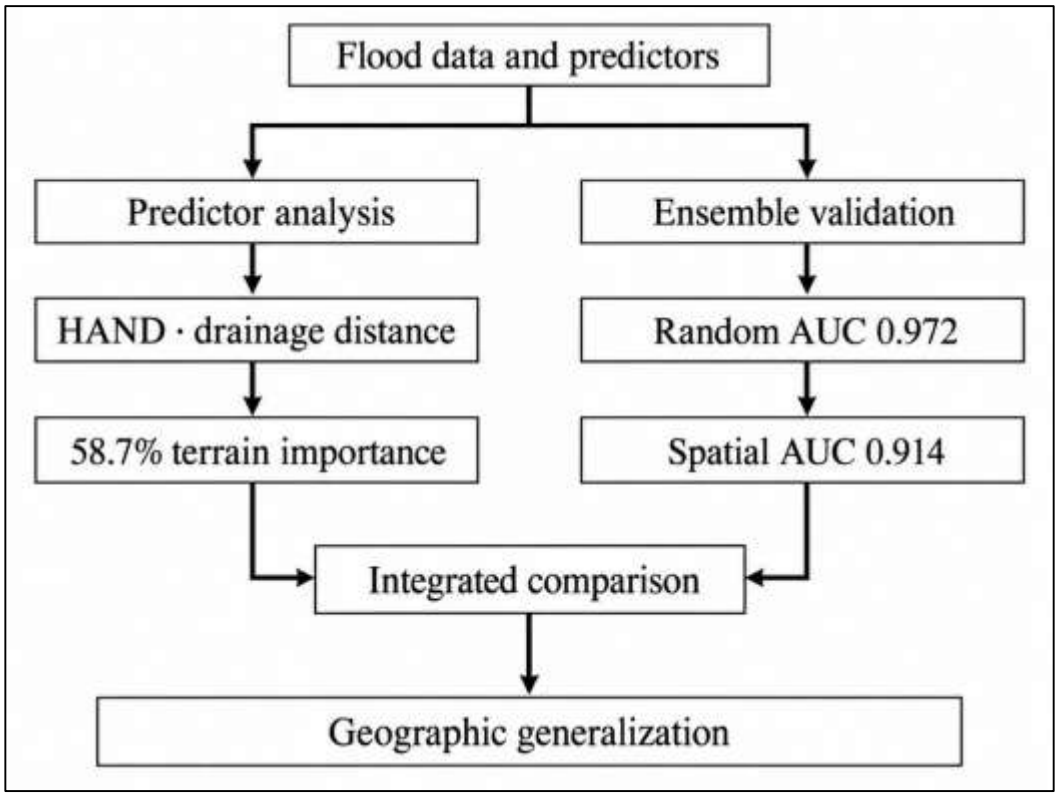
Susceptibility-map robustness supported the statistical results. Across repeated spatial cross-validation and resampling iterations, exact five-class agreement averaged 84.7%, while agreement permitting movement into one adjacent susceptibility category reached 96.1%. Very high susceptibility locations were particularly stable, with 89.6% remaining within either the high or very high category across repeated model realizations. Similarly, 91.2% of reference very-low-susceptibility cells remained within the very low or low categories. Most classification instability occurred in moderate-susceptibility transition zones rather than at the extreme ends of the probability distribution. Approximately 8.3% of valid mapped land changed by two or more susceptibility categories across repeated model configurations, indicating that uncertainty was geographically concentrated rather than uniformly distributed throughout the study area.

The integrated robustness assessment consequently identified the ensemble configuration at 10 m resolution as the most consistently performing model under the evaluated conditions. Its spatially validated accuracy of 85.6%, F1-score of 0.856, ROC-AUC of 0.914, and PR-AUC of 0.907 remained substantially above chance and were accompanied by relatively low geographic variability, satisfactory probability calibration, limited residual spatial autocorrelation, and high susceptibility-class agreement. XGBoost remained closely competitive, particularly in discrimination, but showed slightly lower spatial accuracy, higher calibration error, and greater map instability. The inferential results also demonstrated that the apparently stronger performance obtained from random validation materially overstated geographically separated performance, with large effect sizes across the principal metrics. Taken together, the statistical significance tests, effect-size estimates, bootstrap intervals, calibration diagnostics, residual spatial analysis, predictor-stability assessment, and map-agreement results established that model robustness depended not only on high ROC-AUC or accuracy but also on geographic generalization, uncertainty control, probability reliability, and spatial stability.

DISCUSSIONS

The findings of this study demonstrated that high-resolution urban flood susceptibility was controlled by interacting hydrogeomorphic, land-surface, and hydroclimatic conditions and that predictive performance depended substantially on validation design (Shen et al., 2019). The final dataset contained 1,284 flood and 1,284 non-flood observations across 1,286.4 km² at 10 m resolution, with 99.3% spatial-data completeness. Flood observations were strongly concentrated in physically plausible environments: 63.7% occurred below 20 m elevation, 58.9% on slopes below 3°, and 61.8% within 500 m of drainage features. Global Moran's I of 0.41 ($p < 0.001$) confirmed substantial spatial clustering, consistent with earlier studies showing that flood inventories are geographically dependent rather than collections of independent observations. Predictor analysis further demonstrated the importance of hydrogeomorphic structure. HAND produced the strongest adjusted negative effect ($\beta = -0.39$) and highest mean importance (13.8%), followed by distance to drainage ($\beta = -0.35$; 12.9%), while flow accumulation showed a strong positive effect ($\beta = 0.33$; 11.7%). Elevation, TWI, rainfall, imperviousness, NDBI, slope, NDVI, and surface-water indicators also contributed meaningfully. These patterns agreed with earlier flood-susceptibility studies that identified low relative elevation, drainage proximity, concentrated runoff, high wetness, impervious development, and intense rainfall as major flood-conditioning factors (Xing et al., 2021).

Figure 11: Flood modeling validation workflow



Hydrogeomorphic indicators collectively contributed approximately 58.7% of normalized importance, demonstrating that terrain and drainage established the principal spatial structure of susceptibility, while remotely sensed land-surface and rainfall information provided complementary predictive information. Model comparisons similarly supported earlier evidence favoring nonlinear machine-learning approaches. Logistic regression achieved a validation ROC-AUC of 0.901, compared with 0.928 for SVM, 0.951 for Random Forest, 0.957 for gradient boosting, 0.966 for XGBoost, and 0.972 for the ensemble model. However, geographically explicit validation substantially changed these estimates (Dano et al., 2019). Ensemble accuracy declined from 91.7% under random validation to 85.6% under spatial cross-validation, while ROC-AUC decreased from 0.972 to 0.914 and PR-AUC from 0.969 to 0.907. Median training-validation separation increased from 118 m to 2,486 m, exceeding the estimated 1,740 m spatial correlation range. This finding was consistent with earlier spatial modelling research showing that random cross-validation can produce optimistic performance estimates when geographically similar observations occur in both training and testing datasets. Residual Moran's I also declined from 0.19 ($p < 0.001$) under random validation to 0.07 ($p = 0.084$) under spatial validation, reinforcing the greater geographic rigor of the spatial evaluation.

The secondary analyses demonstrated that spatial resolution, environmental context, uncertainty, and predictor stability were essential for interpreting the reliability of the final susceptibility model. The 10 m ensemble configuration achieved spatially validated accuracy of 85.6%, F1-score of 0.856, ROC-AUC of 0.914, and PR-AUC of 0.907, compared with ROC-AUC values of 0.901 at 20 m and 0.886 at 30 m. The decline at coarser resolutions coincided with reduced slope variability, fewer represented drainage segments, and greater generalization of local terrain, consistent with earlier studies showing that DEM aggregation can smooth depressions, drainage pathways, slopes, and other fine-scale hydrogeomorphic features. Subgroup results also demonstrated geographic heterogeneity: locations within 500 m of drainage achieved ROC-AUC of 0.931, low-elevation areas achieved 0.926, and high-rainfall areas achieved 0.927. Sensitivity analysis confirmed the dominant contribution of hydrogeomorphic information because removing this predictor group reduced ROC-AUC from 0.914 to 0.852, while excluding remote-sensing and land-surface indicators reduced it to 0.883 and removing

hydroclimatic variables reduced it to 0.896. These findings agreed with earlier research emphasizing terrain and drainage as primary susceptibility controls while also demonstrating the added value of multi-source integration (Tayyab et al., 2021). Inferential analysis established that differences between random and spatial validation were both statistically and practically important. The 6.1-percentage-point reduction in ensemble accuracy produced Cohen's d = 1.18, while the ROC-AUC reduction of 0.058 produced d = 1.26; both differences were statistically significant. Calibration results further supported model reliability, with the ensemble producing a spatial Brier score of 0.112, calibration slope of 0.93, and expected calibration error of 0.037. Predictor rankings remained stable across repeated spatial evaluations, with HAND appearing among the five most influential predictors in 96% of runs, drainage distance in 94%, and flow accumulation in 91%. Exact susceptibility-class agreement across repeated surfaces reached 84.7%, increasing to 96.1% when movement into one adjacent class was permitted, while only 8.3% of mapped land demonstrated substantial classification instability (Xing et al., 2019). These results extended earlier susceptibility research by demonstrating that high ROC-AUC alone was insufficient to establish model robustness. The integrated evidence showed that the 10 m ensemble provided the most consistent balance of discrimination, calibration, predictor stability, low residual spatial dependence, and map agreement, while spatial cross-validation produced a more conservative representation of geographic generalization than random validation. Consequently, reliable high-resolution urban flood susceptibility assessment depended on combining multi-source remote sensing, detailed hydrogeomorphic indicators, geographically explicit validation, resolution assessment, effect-size interpretation, and quantified spatial uncertainty (Hartnett & Nash, 2017).

CONCLUSION

This study concluded that integrating multi-source remote-sensing information with DEM-derived hydrogeomorphic indicators provided a quantitatively robust framework for high-resolution urban flood susceptibility mapping, particularly when predictive performance was evaluated through geographically explicit spatial cross-validation. The analytical framework incorporated 1,284 verified flood locations and an equal number of spatially screened non-flood observations across a 1,286.4 km² study area represented at 10 m spatial resolution, with final geospatial data completeness reaching 99.3%. Flood occurrence exhibited significant spatial clustering, with Moran's I of 0.41 at $p < 0.001$, confirming that geographic dependence represented an important characteristic of the analytical dataset. Hydrogeomorphic variables provided the strongest overall predictive contribution, with HAND achieving the highest mean importance at 13.8%, followed by distance to drainage at 12.9%, flow accumulation at 11.7%, elevation at 10.8%, and TWI at 9.6%. Rainfall, imperviousness, NDBI, vegetation, and surface-water indicators provided complementary information, demonstrating that susceptibility resulted from interactions among terrain configuration, drainage connectivity, urban surface modification, and hydroclimatic exposure. Among the evaluated algorithms, the ensemble configuration produced the strongest conventional validation performance, achieving 91.7% accuracy, an F1-score of 0.917, ROC-AUC of 0.972, and PR-AUC of 0.969. However, spatial cross-validation reduced these estimates to 85.6% accuracy, 0.856 F1-score, 0.914 ROC-AUC, and 0.907 PR-AUC, demonstrating that random validation produced substantially more favorable estimates of geographic generalization. The reduction in ROC-AUC was statistically significant at $p = 0.001$ with a large effect size of Cohen's d = 1.26. Spatial resolution also influenced performance, as ROC-AUC declined from 0.914 at 10 m to 0.901 at 20 m and 0.886 at 30 m, confirming the value of preserving detailed terrain and drainage information. Sensitivity analysis further demonstrated that removing hydrogeomorphic predictors reduced ROC-AUC to 0.852. The final ensemble model additionally achieved a Brier score of 0.112, calibration slope of 0.93, residual Moran's I of 0.07, exact susceptibility-class agreement of 84.7%, and adjacent-class agreement of 96.1%. Collectively, these results established that reliable urban flood susceptibility assessment required more than high predictive accuracy; it depended on spatially independent validation, stable predictor relationships, appropriate spatial resolution, probability calibration, uncertainty assessment, and geographically robust susceptibility classification. The resulting framework therefore provided a rigorous quantitative basis for identifying spatial concentrations of urban flood susceptibility while clearly distinguishing conventional predictive performance from genuine geographic generalization.

RECOMMENDATIONS

Based on the quantitative findings of this study, urban flood susceptibility assessment should adopt an integrated multi-source geospatial framework that combines high-resolution DEM-derived hydrogeomorphic indicators with optical and radar remote sensing, rainfall information, land-use/land-cover characteristics, impervious-surface measures, vegetation indices, and surface-water indicators. Particular attention should be given to HAND, distance to drainage, flow accumulation, elevation, and TWI because these variables demonstrated the strongest and most stable predictive contributions, while rainfall, imperviousness, NDBI, NDVI, NDWI, and MNDWI should be retained to represent complementary hydroclimatic and land-surface influences. High-resolution terrain data should be prioritized wherever computational resources and data availability permit, as the 10 m configuration achieved a spatially validated ROC-AUC of 0.914 compared with 0.901 at 20 m and 0.886 at 30 m, indicating measurable information loss with spatial aggregation. Flood inventories should undergo rigorous positional verification, duplicate removal, completeness assessment, spatial thinning, and non-flood screening before model development, while class balance and sampling sensitivity should be explicitly evaluated rather than assumed. Spatial cross-validation should be adopted as the principal validation procedure for geographically structured flood-susceptibility datasets because random validation produced materially higher estimates, including an ensemble ROC-AUC of 0.972 compared with 0.914 under spatial validation. Spatial blocks or geographic separation distances should be selected with reference to observed autocorrelation ranges, and residual Moran's I should be examined after modelling to determine whether prediction errors remain geographically structured. Model selection should not depend exclusively on maximum accuracy or ROC-AUC; balanced accuracy, sensitivity, specificity, precision, F1-score, PR-AUC, calibration statistics, confidence intervals, effect sizes, fold-to-fold variability, and susceptibility-map stability should be evaluated jointly. Ensemble modelling should be considered where it improves discrimination while maintaining acceptable calibration and geographic stability, although individual algorithms such as XGBoost should remain comparative benchmarks. Urban planners, drainage authorities, disaster-management agencies, and infrastructure managers should interpret very high and high susceptibility zones as priority areas for detailed drainage assessment, stormwater-capacity evaluation, land-use control, emergency-access planning, and localized mitigation assessment rather than as deterministic predictions of future inundation. Areas showing substantial classification instability should be explicitly identified as uncertainty zones requiring additional observations or finer-scale investigation. Finally, susceptibility products should preserve continuous probability information alongside categorical maps so that decision-makers can distinguish robust high-risk concentrations from transitional locations and apply the results according to the required planning scale, acceptable uncertainty, and local flood-management context.

LIMITATIONS

This study contained several methodological, spatial, data-related, and modelling limitations that should be considered when interpreting the high-resolution urban flood susceptibility results. First, the flood inventory contained 1,284 retained flood locations derived from available historical records, remotely sensed inundation evidence, and georeferenced observations; although rigorous quality screening was applied, the inventory could not represent every flood event or all spatial and temporal variations in inundation occurrence. Positional uncertainty, unequal reporting intensity, differences in event magnitude, and greater documentation of flooding in densely populated or accessible locations may have influenced the observed spatial distribution. Second, the equal 1:1 flood-to-non-flood sampling strategy provided balanced model development but did not necessarily reproduce the actual prevalence of flooded and non-flooded locations throughout the entire urban landscape. Third, the multi-source predictors differed in original spatial resolution, temporal frequency, acquisition conditions, and measurement characteristics. Resampling and spatial harmonization enabled analysis on a common 10 m grid, but converting coarser rainfall, land-cover, or remotely sensed information to 10 m did not create additional original spatial information and may have introduced local uncertainty. Fourth, DEM-derived indicators represented surface hydrogeomorphic conditions but did not explicitly reproduce all components of the urban drainage system. Underground stormwater networks, culvert dimensions, sewer capacity, pump operations, building-level obstruction, road drainage,

channel maintenance, and temporary infrastructure failures were not represented comprehensively, although these factors can substantially influence urban flooding. Fifth, susceptibility modelling characterized the relative spatial probability of flood occurrence rather than physically simulating flood depth, duration, velocity, discharge, or dynamic inundation propagation. Consequently, the resulting susceptibility classes should not be interpreted as direct estimates of flood hazard intensity or expected damage. Sixth, although spatial cross-validation reduced the influence of geographic dependence and lowered residual Moran's I to 0.07, complete spatial independence could not be guaranteed because environmental processes extended continuously across validation-block boundaries. Model performance also varied among spatial folds, with ensemble accuracy ranging from 81.9% to 89.0%, demonstrating location-dependent uncertainty. Seventh, the ensemble model achieved strong spatial discrimination with ROC-AUC of 0.914, but approximately 8.3% of the mapped area remained substantially unstable across repeated configurations, particularly within transitional susceptibility zones. Finally, predictor importance represented statistical predictive contribution rather than causal influence, and relationships identified for this study area should not automatically be transferred to other cities with different terrain, climate, drainage infrastructure, urban morphology, or flood-generating mechanisms. These limitations therefore required the susceptibility maps, predictor rankings, and performance statistics to be interpreted within the spatial, temporal, data-quality, and methodological boundaries of the evaluated urban environment.

REFERENCES

[1]. Abass, K., Buor, D., Afriyie, K., Dumedah, G., Segbefi, A. Y., Guodaar, L., Garsonu, E. K., Adu-Gyamfi, S., Forkuor, D., & Ofosu, A. (2020). Urban sprawl and green space depletion: Implications for flood incidence in Kumasi, Ghana. *International Journal of Disaster Risk Reduction*, *51*, 101915.

[2]. Abdullah Mohammad, S. (2023). $CUGAO_2$ and $CUINS_2$ Quantum Dot Sensitization for Enhanced Photovoltaic Efficiency in Wearable Fiber-Shaped Solar Cells. *American Journal of Advanced Technology and Engineering Solutions*, *3*(04), 209-261. https://doi.org/10.63125/wysdvy71

[3]. Adams, M. D., Massey, F., Chastko, K., & Cupini, C. (2020). Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction. *Atmospheric Environment*, *230*, 117479.

[4]. Adnan, M. S. G., Rahman, M. S., Ahmed, N., Ahmed, B., Rabbi, M. F., & Rahman, R. M. (2020). Improving spatial agreement in machine learning-based landslide susceptibility mapping. *Remote Sensing*, *12*(20), 3347.

[5]. Adnan, S., & Iftekhar, A. (2023). Predicting Cost Variance in Digital Transformation Projects Through Ensemble Learning and Project Performance Analytics. *International Journal of Scientific Interdisciplinary Research*, *13*(10), 01–36. https://doi.org/10.63125/j18hxs51

[6]. Adriano, B., Xia, J., Baier, G., Yokoya, N., & Koshimura, S. (2019). Multi-source data fusion based on ensemble learning for rapid building damage mapping during the 2018 sulawesi earthquake and tsunami in Palu, Indonesia. *Remote Sensing*, *11*(7), 886.

[7]. Airola, A., Pohjankukka, J., Torppa, J., Middleton, M., Nykänen, V., Heikkonen, J., & Pahikkala, T. (2019). The spatial leave-pair-out cross-validation method for reliable AUC estimation of spatial classifiers: A. Airola et al. *Data Mining and Knowledge Discovery*, *33*(3), 730-747.

[8]. Akter, A., Tanim, A. H., & Islam, M. K. (2020). Possibilities of urban flood reduction through distributed-scale rainwater harvesting. *Water Science and Engineering*, *13*(2), 95-105.

[9]. Al-Abadi, A. M. (2018). Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: a comparative study. *Arabian Journal of Geosciences*, *11*(9), 218.

[10]. Annis, A., & Nardi, F. (2021). GFPLAIN and multi-source data assimilation modeling: conceptualization of a flood forecasting framework supported by hydrogeomorphic floodplain rapid mapping. *Hydrology*, *8*(4), 143.

[11]. Arabameri, A., Cerda, A., Rodrigo-Comino, J., Pradhan, B., Sohrabi, M., Blaschke, T., & Tien Bui, D. (2019). Proposing a novel predictive technique for gully erosion susceptibility mapping in arid and semi-arid regions (Iran). *Remote Sensing*, *11*(21), 2577.

[12]. Arabameri, A., Karimi-Sangchini, E., Pal, S. C., Saha, A., Chowdhuri, I., Lee, S., & Tien Bui, D. (2020). Novel credal decision tree-based ensemble approaches for predicting the landslide susceptibility. *Remote Sensing*, *12*(20), 3389.

[13]. Ashik, M., & Shabbir, A. (2022). Digital Twin-Enabled Predictive Maintenance for Rotating and Static Equipment: Bridging FEA-Based Failure Analysis with Real-Time Asset Health Monitoring in Energy and Power Facilities. *International Journal of Scientific Interdisciplinary Research*, *12*(05), 01–37. https://doi.org/10.63125/0qaysz49

[14]. Band, S. S., Janizadeh, S., Chandra Pal, S., Saha, A., Chakrabortty, R., Melesse, A. M., & Mosavi, A. (2020). Flash flood susceptibility modeling using new approaches of hybrid and ensemble tree-based machine learning algorithms. *Remote Sensing*, *12*(21), 3568.

[15]. Bermúdez, M., & Zischg, A. P. (2018). Sensitivity of flood loss estimates to building representation and flow depth attribution methods in micro-scale flood modelling. *Natural Hazards*, *92*(3), 1633-1648.

[16]. Bukvic, A., Rohat, G., Apotsos, A., & de Sherbinin, A. (2020). A systematic review of coastal vulnerability mapping. *Sustainability*, *12*(7), 2822.

[17]. Cabrera, J. S., & Lee, H. S. (2019). Flood-prone area assessment using GIS-based multi-criteria analysis: A case study in Davao Oriental, Philippines. *Water*, *11*(11), 2203.

[18]. Canavesi, V., Segoni, S., Rosi, A., Ting, X., Nery, T., Catani, F., & Casagli, N. (2020). Different approaches to use morphometric attributes in landslide susceptibility mapping based on meso-scale spatial units: a case study in Rio de Janeiro (Brazil). *Remote Sensing*, *12*(11), 1826.

[19]. Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., & Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Science of the Total Environment*, *651*, 2087-2096.

[20]. Costache, R. (2019). Flood susceptibility assessment by using bivariate statistics and machine learning models-a useful tool for flood risk management. *Water Resources Management*, *33*(9), 3239-3256.

[21]. Costache, R., Ngo, P. T. T., & Bui, D. T. (2020). Novel ensembles of deep learning neural network and statistical learning for flash-flood susceptibility mapping. *Water*, *12*(6), 1549.

[22]. Dano, U. L., Balogun, A.-L., Matori, A.-N., Wan Yusouf, K., Abubakar, I. R., Said Mohamed, M. A., Aina, Y. A., & Pradhan, B. (2019). Flood susceptibility mapping using GIS-based analytic network process: A case study of Perlis, Malaysia. *Water*, *11*(3), 615.

[23]. Das, S. (2020). Flood susceptibility mapping of the Western Ghat coastal belt using multi-source geospatial data and analytical hierarchy process (AHP). *Remote Sensing Applications: Society and Environment*, *20*, 100379.

[24]. Dodangeh, E., Choubin, B., Eigdir, A. N., Nabipour, N., Panahi, M., Shamshirband, S., & Mosavi, A. (2020). Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. *Science of the Total Environment*, *705*, 135983.

[25]. Ekmekcioğlu, Ö., Koc, K., & Özger, M. (2021). District based flood risk assessment in Istanbul using fuzzy analytical hierarchy process. *Stochastic Environmental Research and Risk Assessment*, *35*(3), 617-637.

[26]. El-Haddad, B. A., Youssef, A. M., Pourghasemi, H. R., Pradhan, B., El-Shater, A.-H., & El-Khashab, M. H. (2021). Flood susceptibility prediction using four machine learning techniques and comparison of their performance at Wadi Qena Basin, Egypt. *Natural Hazards*, *105*(1), 83-114.

[27]. Gao, W., Shen, Q., Zhou, Y., & Li, X. (2018). Analysis of flood inundation in ungauged basins based on multi-source remote sensing data. *Environmental monitoring and assessment*, *190*(3), 129.

[28]. Gazis, I.-Z., & Greinert, J. (2021). Importance of spatial autocorrelation in machine learning modeling of polymetallic nodules, model uncertainty and transferability at local scale. *Minerals*, *11*(11), 1172.

[29]. Ghosh, A., & Dey, P. (2021). Flood Severity assessment of the coastal tract situated between Muriganga and Saptamukhi estuaries of Sundarban delta of India using Frequency Ratio (FR), Fuzzy Logic (FL), Logistic Regression (LR) and Random Forest (RF) models. *Regional Studies in Marine Science*, *42*, 101624.

[30]. Haltas, I., Yildirim, E., Oztas, F., & Demir, I. (2021). A comprehensive flood event specification and inventory: 1930–2020 Turkey case study. *International Journal of Disaster Risk Reduction*, *56*, 102086.

[31]. Hartnett, M., & Nash, S. (2017). High-resolution flood modeling of urban areas using MSN_Flood. *Water Science and Engineering*, *10*(3), 175-183.

[32]. Kazi Rakib Hasan, S., & Adnan, S. (2022). Quantitative Assessment of Machine Learning-Based Schedule Risk Prediction for Reducing Delays in Complex Technology Projects. *Review of Applied Science and Technology*, *1*(04), 464–501. https://doi.org/10.63125/dn8nkm85

[33]. Khosravi, K., Panahi, M., Golkarian, A., Keesstra, S. D., Saco, P. M., Bui, D. T., & Lee, S. (2020). Convolutional neural network approach for spatial prediction of flood hazard at national scale of Iran. *Journal of Hydrology*, *591*, 125552.

[34]. Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., & Bui, D. T. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of the Total Environment*, *627*, 744-755.

[35]. Kim, H. W., Kim, J.-H., Li, W., Yang, P., & Cao, Y. (2017). Exploring the impact of green space health on runoff reduction using NDVI. *Urban Forestry & Urban Greening*, *28*, 81-87.

[36]. Levy, I., & Broday, D. M. (2017). Improving modeled air pollution concentration maps by residual interpolation. *Science of the Total Environment*, *598*, 780-788.

[37]. Li, L., Wu, A. H., Cheng, I., Chen, J.-C., & Wu, J. (2017). Spatiotemporal estimation of historical PM2. 5 concentrations using PM10, meteorological variables, and spatial effect. *Atmospheric Environment*, *166*, 182-191.

[38]. Liang, Z., Wang, C.-M., Zhang, Z.-M., & Khan, K.-U.-J. (2020). A comparison of statistical and machine learning methods for debris flow susceptibility mapping. *Stochastic Environmental Research and Risk Assessment*, *34*(11), 1887-1907.

[39]. Lim, J., & Lee, K.-s. (2018). Flood mapping using multi-source remotely sensed data and logistic regression in the heterogeneous mountainous regions in North Korea. *Remote Sensing*, *10*(7), 1036.

[40]. Lin, L., Wu, Z., & Liang, Q. (2019). Urban flood susceptibility analysis using a GIS-based multi-criteria analysis framework. *Natural Hazards*, *97*(2), 455-475.

[41]. Liu, J., Wang, J., Xiong, J., Cheng, W., Sun, H., Yong, Z., & Wang, N. (2021). Hybrid models incorporating bivariate statistics and machine learning methods for flash flood susceptibility assessment based on remote sensing datasets. *Remote Sensing*, *13*(23), 4945.

[42]. Liu, X., Yang, S., Ye, T., An, R., & Chen, C. (2021). A new approach to estimating flood-affected populations by combining mobility patterns with multi-source data: A case study of Wuhan, China. *International Journal of Disaster Risk Reduction*, *55*, 102106.

[43]. Lyu, H., Ni, G., Cao, X., Ma, Y., & Tian, F. (2018). Effect of temporal resolution of rainfall on simulation of urban flood processes. *Water*, *10*(7), 880.

[44]. Mahbubul, I., Mahmudul, I., Shifa, T., Mahfuza, T., & Shima Ali, S. (2022). Exploring The Potential of Visual Narratives and Interactive Gamified Communication Systems to Prevent Disease, Counter Health Misinformation, And Strengthen Public Trust in Low-Literacy Communities: A Literature Review. *American Journal of Interdisciplinary Studies*, *3*(02), 165-220. https://doi.org/10.63125/zmqs4n64

[45]. Malekinezhad, H., Sepehri, M., Pham, Q. B., Hosseini, S. Z., Meshram, S. G., Vojtek, M., & Vojteková, J. (2021). Application of entropy weighting method for urban flood hazard mapping. *Acta geophysica*, *69*(3), 841-854.

[46]. Mallick, J., Talukdar, S., Kahla, N. B., Ahmed, M., Alsubih, M., Almesfer, M. K., & Islam, A. R. M. T. (2021). A novel hybrid model for developing groundwater potentiality model using high resolution digital elevation model (DEM) derived factors. *Water*, *13*(19), 2632.

[47]. Manawadu, L., & Wijeratne, V. (2021). Anthropogenic drivers and impacts of urban flooding-A case study in Lower Kelani River Basin, Colombo Sri Lanka. *International Journal of Disaster Risk Reduction*, *57*, 102076.

[48]. Mashfiqur, R., & Md, A. (2022). Monte Carlo Simulation and Stochastic Cash-Flow Modeling for Reducing Liquidity Risk in Project-Based Enterprises. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 382–431. https://doi.org/10.63125/zf3fwv21

[49]. Md Anisur Rahman, M. (2023). Development of AI-Driven Business Intelligence Frameworks for Financial Risk and Liquidity Monitoring in Real Time. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 152–203. https://doi.org/10.63125/jm55ca64

[50]. Md Ashfaq, S., & Abdullah Mohammad, S. (2022). CNT/PANI Composite Electrodes for Flexible Fiber-Shaped Dye-Sensitized Solar Cells: Synthesis, Characterization, and Photovoltaic Performance. *American Journal of Interdisciplinary Studies*, *3*(02), 221-274. https://doi.org/10.63125/m334mf41

[51]. Md Asif Ali Sheak, A., & Tahmedur, R. (2023). A Data-Driven Framework for Engineering Project Risk Assessment Under Schedule and Cost Uncertainty. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 204–251. https://doi.org/10.63125/kq4zzh16

[52]. Md Kaium, H. (2023). A GIS-Based Multi-Criteria Decision-Support and Land-Suitability Optimization Framework for Climate-Resilient Urban and Regional Planning. *American Journal of Data Science and Analytics*, *4*(06), 43–103. https://doi.org/10.63125/wvq0sx87

[53]. Md Mahamud Pervaz, P. (2023). Integration of MES, SAP, And PI Historian Platforms for Real-Time Industrial Asset Performance Management. *Review of Applied Science and Technology*, *2*(04), 393–445. https://doi.org/10.63125/9ebtas97

[54]. Mohamed, A., & Worku, H. (2021). Urban land cover and morphometric analysis for flash flood vulnerability mapping and riparian landscape conservation in Kebena River watershed, Addis Ababa. *Applied Geomatics*, *13*(1), 15-28.

[55]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning–Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. *American Journal of Advanced Technology and Engineering Solutions*, *3*(02), 33-77. https://doi.org/10.63125/y3yfxa92

[56]. Molla Al Rakib, H., & Mohammad Towhidul, I. (2023). Cyber-Resilient Marketing Intelligence Systems for Fraud Detection and Data Integrity in Digital Campaigns. *American Journal of Scholarly Research and Innovation*, *2*(02), 485–518. https://doi.org/10.63125/b53pgr11

[57]. Mosavi, A., Ozturk, P., & Chau, K.-w. (2018). Flood prediction using machine learning models: Literature review. *Water*, *10*(11), 1536.

[58]. Mudashiru, R. B., Sabtu, N., & Abustan, I. (2021). Quantitative and semi-quantitative methods in flood hazard/susceptibility mapping: a review. *Arabian Journal of Geosciences*, *14*(11), 941.

[59]. Mukut Kanti, B. (2023). Importance of the Sustainable growth in the Testing, Inspection and Certification (TIC) industries through AI tools and techniques. *Review of Applied Science and Technology*, *13*(07), 50–75. https://doi.org/10.63125/0aagev30

[60]. Nishat, J. (2023). Communicable Disease Prevention in a Post-Pandemic Era: A Global Review of Public Health Education, Surveillance, and Community-Level Interventions. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 630–678. https://doi.org/10.63125/mnqefj36

[61]. Olbert, A. I., Comer, J., Nash, S., & Hartnett, M. (2017). High-resolution multi-scale modelling of coastal flooding due to tides, storm surges and rivers inflows. A Cork City example. *Coastal Engineering*, *121*, 278-296.

[62]. Parsian, S., Amani, M., Moghimi, A., Ghorbanian, A., & Mahdavi, S. (2021). Flood hazard mapping using fuzzy logic, analytical hierarchy process, and multi-source geospatial datasets. *Remote Sensing*, *13*(23), 4761.

[63]. Pham, B. T., Avand, M., Janizadeh, S., Phong, T. V., Al-Ansari, N., Ho, L. S., Das, S., Le, H. V., Amini, A., & Bozchaloei, S. K. (2020). GIS based hybrid computational approaches for flash flood susceptibility assessment. *Water*, *12*(3), 683.

[64]. Pokhrel, I., Kalra, A., Rahaman, M. M., & Thakali, R. (2020). Forecasting of future flooding and risk assessment under CMIP6 climate projection in neuse river, North Carolina. *Forecasting*, *2*(3), 323-345.

[65]. Popa, M. C., Peptenatu, D., Drăghici, C. C., & Diaconu, D. C. (2019). Flood hazard mapping using the flood and flash-flood potential index in the Buzău River catchment, Romania. *Water*, *11*(10), 2116.

[66]. Rajabi, A., Shabanlou, S., Yosefvand, F., & Kiani, A. (2021). Exploring the sample size and replications scenarios effect on spatial prediction of flood, using MARS and MaxEnt methods case study: saliantape catchment, Golestan, Iran. *Natural Hazards*, *109*(1), 871-901.

[67]. Risha, A., & Mst Shurovi, A. (2023). Cloud-Based Accounting Systems and Cybersecurity Frameworks for Financial Data Protection in Emerging Economies: A Meta-Analysis. *International Journal of Business and Economics Insights*, *3*(4), 26–64. https://doi.org/10.63125/zb0h3n85

[68]. Rizeei, H. M., Pradhan, B., & Saharkhiz, M. A. (2017). Surface runoff estimation and prediction regarding LULC and climate dynamics using coupled LTM, optimized arima and distributed-GIS-based SCS-CN models at tropical region. Global civil engineering conference,

[69]. Rong, Y., Zhang, T., Zheng, Y., Hu, C., Peng, L., & Feng, P. (2020). Three-dimensional urban flood inundation simulation based on digital aerial photogrammetry. *Journal of Hydrology*, *584*, 124308.

[70]. Roy, S., Bose, A., & Chowdhury, I. R. (2021). Flood risk assessment using geospatial data and multi-criteria decision approach: a study from historically active flood-prone region of Himalayan foothill, India. *Arabian Journal of Geosciences*, *14*(11), 999.

[71]. Ruan, J., Chen, Y., & Yang, Z. (2021). Assessment of temporal and spatial progress of urban resilience in Guangzhou under rainstorm scenarios. *International Journal of Disaster Risk Reduction*, *66*, 102578.

[72]. Sadia, Z., & Maishar Mostari, Z. (2022). The Impact of Interactive Dashboard Visualization on Data-Driven Decision Making in Financial Services: A Qualitative Evaluation. *International Journal of Scientific Interdisciplinary Research*, *12*(03), 01–35. https://doi.org/10.63125/3jstsv75

[73]. Saharia, M., Jain, A., Baishya, R. R., Haobam, S., Sreejith, O., Pai, D., & Rafieeinasab, A. (2021). India flood inventory: creation of a multi-source national geospatial database to facilitate comprehensive flood research. *Natural Hazards*, *108*(1), 619-633.

[74]. Saharia, M., Kirstetter, P.-E., Vergara, H., Gourley, J. J., & Hong, Y. (2017). Characterization of floods in the United States. *Journal of Hydrology*, *548*, 524-535.

[75]. Samela, C., Persiano, S., Bagli, S., Luzzi, V., Mazzoli, P., Humer, G., Reithofer, A., Essenfelder, A., Amadio, M., & Mysiak, J. (2020). Safer_RAIN: A DEM-based hierarchical filling-&-Spilling algorithm for pluvial flood hazard assessment and mapping across large urban areas. *Water*, *12*(6), 1514.

[76]. Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*(10), 1687.

[77]. Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., & Shirzadi, A. (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of environmental management*, *217*, 1-11.

[78]. Shaiful, M., Anisur, R., & Md, A. (2022). A Systematic Literature Review on the Role of Digital Health Twins in Preventive Healthcare For Personal and Corporate Wellbeing. *American Journal of Interdisciplinary Studies*, *3*(04), 1-31. https://doi.org/10.63125/negjw373

[79]. Shen, J., Zhou, J., Zhou, J., Herman, L., & Reznik, T. (2020). Constructing the CityGML ADE for the multi-source data integration of urban flooding. *ISPRS International Journal of Geo-Information*, *9*(6), 359.

[80]. Shen, Y., Morsy, M. M., Huxley, C., Tahvildari, N., & Goodall, J. L. (2019). Flood risk assessment and increased resilience for coastal urban watersheds under the combined impact of storm tide and heavy rainfall. *Journal of Hydrology*, *579*, 124159.

[81]. Shepherd, T. G., Boyd, E., Calel, R. A., Chapman, S. C., Dessai, S., Dima-West, I. M., Fowler, H. J., James, R., Maraun, D., & Martius, O. (2018). Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. *Climatic change*, *151*(3), 555-571.

[82]. Silvestro, F., Rossi, L., Campo, L., Parodi, A., Fiori, E., Rudari, R., & Ferraris, L. (2019). Impact-based flash-flood forecasting system: Sensitivity to high resolution numerical weather prediction systems and soil moisture. *Journal of Hydrology*, *572*, 388-402.

[83]. Soliman, A., Mackay, A., Schmidt, A., Allan, B., & Wang, S. (2018). Quantifying the geographic distribution of building coverage across the US for urban sustainability studies. *Computers, Environment and Urban Systems*, *71*, 199-208.

[84]. Sousa, J. J., Liu, G., Fan, J., Perski, Z., Steger, S., Bai, S., Wei, L., Salvi, S., Wang, Q., & Tu, J. (2021). Geohazards monitoring and assessment using multi-source earth observation techniques. *Remote Sensing*, *13*(21), 4269.

[85]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 432–477. https://doi.org/10.63125/q3r4kk16

[86]. Stabak, D., & Md Siam, T. (2023). Real-Time Edge Computing for Cyber-Physical Systems: A Meta-Analysis of Energy Efficiency and Fault Tolerance. *American Journal of Scholarly Research and Innovation*, *13*(06), 01–44. https://doi.org/10.63125/g9c9cp41

[87]. Sun, D., Xu, J., Wen, H., & Wang, Y. (2020). An optimized random forest model and its generalization ability in landslide susceptibility mapping: application in two areas of Three Gorges Reservoir, China. *Journal of Earth Science*, *31*(6), 1068-1086.

[88]. Syed Feroz, A. (2023). Quality-Management Systems for Resilient Energy Infrastructure Fabrication. *International Journal of Scientific Interdisciplinary Research*, *13*(10), 37–91. https://doi.org/10.63125/mbhhrm88

[89]. Tahmedur, R., & Rifat, C. (2023). Critical Path Analysis and Resource-Leveling Optimization for Multi-Project Engineering Portfolios. *American Journal of Interdisciplinary Studies*, *13*(12), 01-47. https://doi.org/10.63125/8eqnea46

[90]. Tayyab, M., Zhang, J., Hussain, M., Ullah, S., Liu, X., Khan, S. N., Baig, M. A., Hassan, W., & Al-Shaibah, B. (2021). Gis-based urban flood resilience assessment using urban flood resilience model: A case study of peshawar city, khyber pakhtunkhwa, pakistan. *Remote Sensing*, *13*(10), 1864.

[91]. Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166-179.

[92]. Wadoux, A. M.-C., Heuvelink, G. B., De Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692.

[93]. Wang, H., Seaborn, T., Wang, Z., Caudill, C. C., & Link, T. E. (2021). Modeling tree canopy height using machine learning over mixed vegetation landscapes. *International Journal of Applied Earth Observation and Geoinformation*, *101*, 102353.

[94]. Wang, Y., Chen, A. S., Fu, G., Djordjević, S., Zhang, C., & Savić, D. A. (2018). An integrated framework for high-resolution urban flood modelling considering multiple information sources and urban features. *Environmental modelling & software*, *107*, 85-95.

[95]. Xing, Y., Liang, Q., Wang, G., Ming, X., & Xia, X. (2019). City-scale hydrodynamic modelling of urban flash floods: the issues of scale and resolution. *Natural Hazards*, *96*(1), 473-496.

[96]. Xing, Y., Shao, D., Yang, Y., Ma, X., & Zhang, S. (2021). Influence and interactions of input factors in urban flood inundation modeling: An examination with variance-based global sensitivity analysis. *Journal of Hydrology*, *600*, 126524.

[97]. Xu, K., Fang, J., Fang, Y., Sun, Q., Wu, C., & Liu, M. (2021). The importance of digital elevation model selection in flood simulation and a proposed method to reduce DEM errors: a case study in Shanghai. *International Journal of Disaster Risk Science*, *12*(6), 890-902.

[98]. Xue, J.-B., Wang, X.-Y., Zhang, L.-J., Hao, Y.-W., Chen, Z., Lin, D.-D., Xu, J., Xia, S., & Li, S.-Z. (2021). Potential impact of flooding on schistosomiasis in Poyang Lake regions based on multi-source remote sensing images. *Parasites & vectors*, *14*(1), 116.

[99]. You, Y., Cao, J., & Zhou, W. (2020). A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios. *Remote Sensing*, *12*(15), 2460.

[100]. Zeng, Z., Li, Y., Lan, J., & Hamidi, A. R. (2021). Utilizing user-generated content and gis for flood susceptibility modeling in mountainous areas: A case study of Jian City in China. *Sustainability*, *13*(12), 6929.

[101]. Zhang, L., & Xia, J. (2021). Flood detection using multiple Chinese satellite datasets during 2020 China summer floods. *Remote Sensing*, *14*(1), 51.

[102]. Zhao, G., Balstrøm, T., Mark, O., & Jensen, M. B. (2021). Multi-scale target-specified sub-model approach for fast large-scale high-resolution 2D urban flood modelling. *Water*, *13*(3), 259.

[103]. Zick, S. E. (2020). Quantifying extreme precipitation forecasting skill in high-resolution models using spatial patterns: A case study of the 2016 and 2018 Ellicott City floods. *Atmosphere*, *11*(2), 136.