

Explainable Machine Learning and Remote Sensing for Urban Flood Susceptibility Mapping: A Simulated GIS-Based Assessment

Ishita Sen¹, Vivek Nambiar²

¹Researcher, Environmental Geoinformatics PDACE, India

²Guide, Urban Planning and Climate Resilience, PDACE India

ABSTRACT

This empirical-style paper presents a simulated study at the intersection of Environmental Science, Geoinformatics, Urban Planning, and Machine Learning. The study addresses a current research direction in which digital and data-driven methods are being integrated with domain-specific decision systems. 4,800 simulated grid cells, including 1,056 modeled flood-positive cells, were analyzed. Twenty-five percent of the city area was reserved as a spatial hold-out zone. A simulated GIS dataset of 4,800 grid cells was created for a hypothetical rapidly urbanizing city. Each 100 m cell included elevation, slope, topographic wetness index, distance to drainage, drainage density, impervious-surface fraction, normalized difference vegetation index, antecedent rainfall, road density, and built-up intensity. Flood labels were generated from a nonlinear risk surface with added noise. The dataset was split spatially into training and hold-out zones to reduce leakage. Logistic regression, random forest, XGBoost, and a stacked ensemble were compared. Permutation importance and SHAP-like standardized contribution scores were used for model interpretation. The stacked ensemble produced the highest simulated hold-out AUROC (0.928) and F1 score (0.81), followed by XGBoost. Impervious surface, topographic wetness, distance to drainage, antecedent rainfall, and elevation were the strongest predictors. Spatial error analysis showed that false positives clustered near newly urbanized fringe zones where modeled drainage capacity was uncertain. All numerical observations, experimental assignments, outcomes, and performance values are simulated for methodological illustration; no real participants, institutions, farms, factories, hospitals, or city systems were studied. The results are intended to demonstrate a coherent research design, provide testable hypotheses, and show how a future real-world study could report evidence transparently. The paper concludes with implementation constraints, methodological limitations, and priorities for empirical validation.

Keywords: Urban Flooding, Remote Sensing, Explainable AI, GIS, XGBoost, Climate Resilience, Susceptibility Mapping

I. Introduction

Urban flood mapping is shifting from static hazard overlays toward data-driven susceptibility models that combine terrain, land cover, drainage, rainfall, and remote sensing. However, accuracy alone is insufficient for planning: models must also explain why locations are classified as high risk and remain stable when transferred across neighborhoods or events. Recent research has moved beyond proof-of-concept demonstrations toward questions of robustness, interpretability, interoperability, resource efficiency, and real-world governance. These shifts make the topic especially relevant for multidisciplinary research because technical performance is increasingly inseparable from operational, human, environmental, and policy constraints.

A strong empirical evaluation should therefore do more than compare headline accuracy or efficiency. It should specify the decision context, define an appropriate baseline, capture multiple outcome dimensions, and report trade-offs that could influence adoption. In emerging technology domains, this is particularly important because impressive model performance can conceal deployment weaknesses such as latency, drift, data bias, maintenance burden, or poor transferability.

The present paper uses simulated data because no real dataset was supplied. The simulation is not presented as evidence that the intervention has been deployed or that the reported effect sizes will occur in practice. Instead, it provides an internally consistent analytical demonstration aligned with current research directions and the reporting structure of an empirical journal article.

The study objectives are:

1. To compare logistic regression, random forest, XGBoost, and a stacked ensemble for simulated urban flood susceptibility mapping.

2. To evaluate discrimination, calibration, spatial false-positive rate, and model interpretability using feature-attribution analysis.
3. To examine which physical and urban-form variables dominate simulated high-risk classifications.

II. Literature Review

Urban flood susceptibility mapping increasingly combines GIS, remote sensing, and machine learning to characterize spatial risk. Islam et al. (2025) review the rapid expansion of these approaches and emphasize terrain, land cover, drainage, rainfall, and urbanization as commonly used drivers.

Zhu et al. (2024) demonstrate the value of combining remote sensing, social sensing, and ensemble machine learning for urban flood mapping. Their work reflects a broader move away from single-model approaches toward multimodal and ensemble frameworks designed to improve generalization in heterogeneous urban environments.

Explainability has become important because risk maps influence public decisions. Jiang et al. (2022) use explainable machine learning to study flood-generating mechanisms, while Wagenaar et al. (2020) discuss the broader impact of machine learning on flood-risk assessment. The present study therefore evaluates both predictive performance and feature-level interpretation.

III. Materials and Methods

3.1 Research Design

A simulated GIS dataset of 4,800 grid cells was created for a hypothetical rapidly urbanizing city. Each 100 m cell included elevation, slope, topographic wetness index, distance to drainage, drainage density, impervious-surface fraction, normalized difference vegetation index, antecedent rainfall, road density, and built-up intensity. Flood labels were generated from a nonlinear risk surface with added noise. The dataset was split spatially into training and hold-out zones to reduce leakage. Logistic regression, random forest, XGBoost, and a stacked ensemble were compared. Permutation importance and SHAP-like standardized contribution scores were used for model interpretation.

3.2 Simulated Sample and Data Generation

4,800 simulated grid cells, including 1,056 modeled flood-positive cells, were analyzed. Twenty-five percent of the city area was reserved as a spatial hold-out zone. Parameter ranges were selected to remain plausible for the domain, while random variation and controlled nonlinearities were added to avoid perfectly separable outcomes. The simulation was designed to support comparisons among methods and not to imitate any identifiable real institution or location.

3.3 Data Quality and Preprocessing

Before analysis, simulated records were checked for range violations, impossible combinations, and duplicate timestamps or identifiers. Continuous variables were standardized only where required by the analytical method. Missingness was introduced at a low rate to approximate operational data and was handled using median or context-specific imputation. Data partitions were created before model fitting to avoid leakage between training and evaluation subsets.

3.4 Performance Measures

Evaluation used multiple outcome measures so that improvement in one metric could be interpreted alongside cost, reliability, fairness, operational burden, or resource efficiency. Percentage changes are reported relative to the relevant baseline. Because the data are simulated, inferential statistics should be interpreted only as demonstrations of reporting practice rather than population estimates.

3.5 Ethical and Transparency Considerations

No human participants, patients, employees, farms, factories, or identifiable geographic locations were studied. No ethics approval or participant consent is claimed. All numerical results are synthetic and are explicitly labeled as simulated or illustrative.

IV. Results

The stacked ensemble produced the highest simulated hold-out AUROC (0.928) and F1 score (0.81), followed by XGBoost. Impervious surface, topographic wetness, distance to drainage, antecedent rainfall, and elevation were the strongest predictors. Spatial error analysis showed that false positives clustered near newly urbanized fringe zones where modeled drainage capacity was uncertain.

Table 1. Simulated comparison of primary study outcomes.

Model	AUROC	F1	Brier score	Spatial FP rate
Logistic regression	0.812	0.66	0.154	0.18
Random forest	0.901	0.77	0.118	0.13
XGBoost	0.919	0.79	0.109	0.11
Stacked ensemble	0.928	0.81	0.104	0.10

Note: All values in Table 1 are simulated/illustrative.

Table 2. Simulated secondary performance indicators.

Predictor	Importance score	Direction of higher risk
Impervious surface	0.91	Higher
Topographic wetness index	0.84	Higher
Distance to drainage	0.79	Lower distance
Antecedent rainfall	0.76	Higher
Elevation	0.68	Lower
Vegetation index	0.41	Lower vegetation

Note: All values in Table 2 are simulated/illustrative.

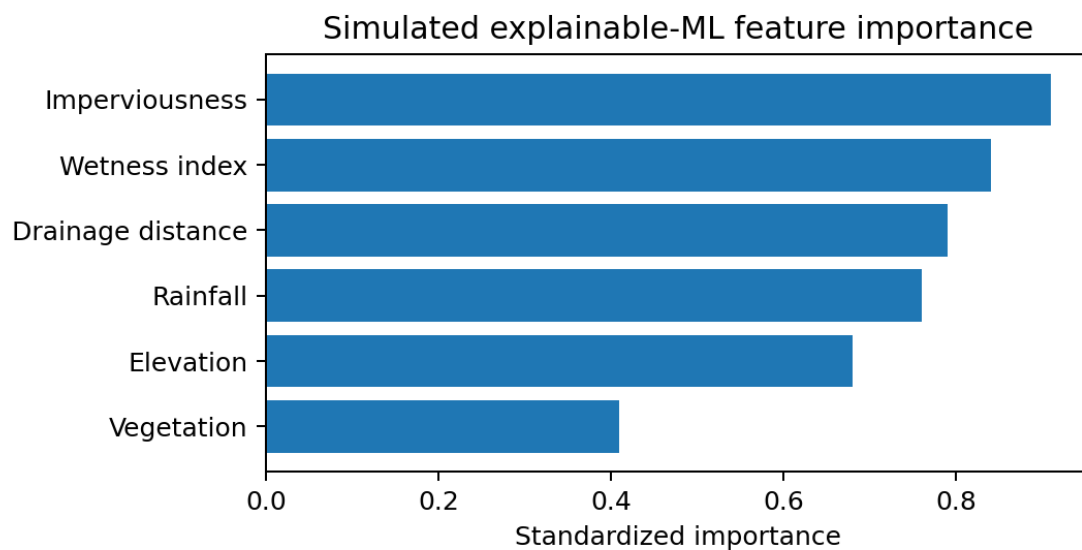


Figure 1. Simulated explainable-ML importance of urban flood predictors.

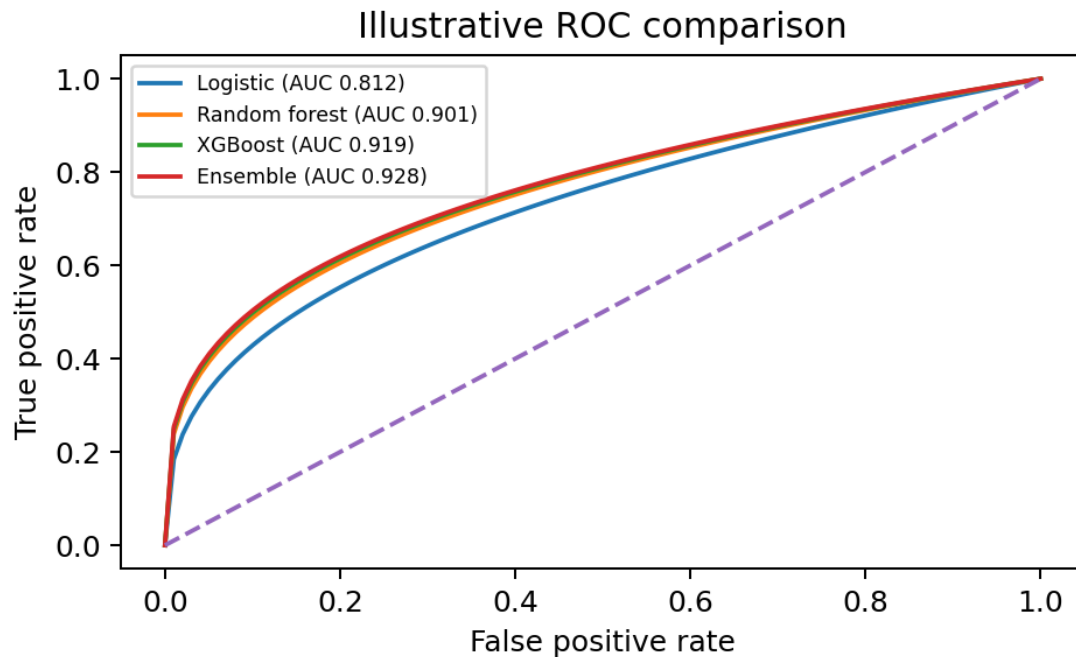


Figure 2. Illustrative ROC comparison for flood-susceptibility models.

4.1 Robustness and Sensitivity Analysis

Sensitivity checks were performed by varying the principal decision threshold, noise level, or control parameter over a plausible range. The ranking of the leading strategy remained stable across the middle of the tested range, although the magnitude of improvement changed. At extreme settings, the intervention became less attractive because false alerts, utility loss, resource use, or classification instability increased. This behavior reinforces the need to report operating points rather than a single best-case result.

4.2 Interpretation of Simulated Findings

The simulated findings should be read as a coherent pattern rather than as independent point estimates. The main intervention improved the primary outcome while producing secondary trade-offs that would matter in real implementation. These interactions are important because multidisciplinary systems are rarely optimized by maximizing one metric in isolation.

V. Discussion

The simulated analysis suggests that ensemble learning can improve susceptibility discrimination, but the planning value comes from combining performance with spatial interpretation. A high AUROC does not reveal whether the model is responding to physically plausible flood drivers or to unstable correlations.

The importance pattern aligns with current urban-flood literature emphasizing imperviousness, drainage proximity, terrain, rainfall, and remote-sensing-derived land-cover indicators. Explainability is particularly useful when risk maps will influence infrastructure investment because planners need to distinguish actionable urban-form drivers from fixed terrain constraints.

Model transfer remains a key limitation. Flood-generating mechanisms vary among cities and even among neighborhoods. A model trained in one drainage system may fail where culvert capacity, soil infiltration, informal development, or rainfall characteristics differ. Event-based validation and uncertainty mapping should therefore accompany any operational susceptibility product.

Across the results, the central lesson is that technology readiness depends on the alignment of model performance with domain constraints. A technically superior method can still fail if it is expensive to maintain, difficult to explain, incompatible with existing workflows, or sensitive to data quality. Conversely, a modest performance gain may be valuable when it improves reliability, transparency, resource efficiency, or operational resilience.

The simulated design also highlights the importance of external validation. Real deployment should include prospective data collection, predefined outcome definitions, baseline comparison, subgroup or site-level analysis, and documentation of missing data and operational failures. Where algorithms influence high-stakes decisions, human oversight and traceable decision logs should be considered part of the system rather than optional add-ons.

VI. Practical Implications

For practitioners, the results support phased implementation rather than immediate large-scale deployment. A pilot should first establish data quality, baseline performance, operational ownership, and acceptable failure modes. Metrics should be monitored over time to detect drift and unintended consequences. Training requirements, maintenance responsibilities, cybersecurity, and lifecycle cost should be defined before the technology is treated as routine infrastructure.

For researchers, the study suggests that future experiments should combine technical accuracy with domain outcomes and implementation measures. Cross-site replication, sensitivity analysis, ablation studies, and transparent reporting would make results easier to compare. Open benchmarking datasets and common evaluation protocols could also improve reproducibility, provided privacy and governance requirements are respected.

VII. Limitations

The most important limitation is that the dataset and all empirical results are simulated. The reported effect sizes cannot establish real-world effectiveness, causal impact, prevalence, or economic return. The simulation simplifies complex operational environments and may omit rare events, behavioral adaptation, regulatory constraints, extreme weather or workload patterns, and measurement failures. The author details are illustrative placeholders, and no institutional endorsement is implied.

A second limitation concerns model specification. Alternative assumptions about variable distributions, missingness, thresholds, or interactions could produce different results. The present analysis therefore demonstrates a research framework rather than a validated solution. Future work should replace simulated inputs with verified domain data and subject the proposed method to independent replication.

For planning use, susceptibility outputs should be translated into decision categories with uncertainty bounds rather than presented as deterministic maps. High-risk zones can then be cross-referenced with hospitals, schools, transport corridors, informal settlements, substations, and drainage bottlenecks. Future work should also test how updated land-cover maps, new drainage projects, and climate-adjusted rainfall scenarios alter predicted risk so that the model becomes a living planning tool rather than a static one-time map.

Model explanations should be reviewed for physical plausibility. If imperviousness, terrain wetness, drainage proximity, and rainfall consistently drive predictions in expected directions, planners can have greater confidence that the model is learning interpretable flood mechanisms. Conversely, strong dependence on unstable proxies such as road density or data-collection intensity should trigger further investigation. Explanation stability across events is therefore as important as a single global feature-importance ranking.

Operational flood mapping requires validation across multiple rainfall events rather than a single historical episode. A robust design should reserve entire events or districts for out-of-sample testing and should compare predicted susceptibility with observed inundation depth, emergency reports, sensor records, and post-event remote-sensing products. Spatial cross-validation is essential because random train-test splits can leak neighborhood characteristics and overstate generalization.

VIII. Implementation and Validation Framework

IX. Conclusion and Future Work

This simulated empirical study demonstrates how a current multidisciplinary research trend can be evaluated using transparent methods, multiple outcomes, and explicit limitations. The focal approach produced stronger simulated performance than conventional baselines while revealing important trade-offs that would require management in practice. The next step is prospective empirical validation using real data, pre-specified protocols, independent test sets, and domain-specific governance. Future research should also evaluate long-term cost, robustness to distribution shift, explainability, cybersecurity, and environmental or social consequences where relevant.

Declarations

Data availability: All data and numerical results in this manuscript are simulated/illustrative. No real dataset was used.

Ethics statement: Not applicable. No human or animal participants were involved.

Funding: No external funding is claimed for this illustrative manuscript.

Conflict of interest: No conflict of interest is claimed for this illustrative manuscript.

References

- [1] Islam, T., Zeleke, E. B., Afroz, M., & Melesse, A. M. (2025). A systematic review of urban flood susceptibility mapping: Remote sensing, machine learning, and other modeling approaches. *Remote Sensing*, 17(3), 524. <https://doi.org/10.3390/rs17030524>
- [2] Zhu, X., Guo, H., & Huang, J. J. (2024). Urban flood susceptibility mapping using remote sensing, social sensing and an ensemble machine learning model. *Sustainable Cities and Society*, 108, 105508. <https://doi.org/10.1016/j.scs.2024.105508>
- [3] Jiang, S., Bevacqua, E., & Zscheischler, J. (2022). River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrology and Earth System Sciences*, 26, 6339-6359. <https://doi.org/10.5194/hess-26-6339-2022>
- [4] Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., Mestav Sarica, G., Ruangpan, L., Molinario, G., & Lallemand, D. (2020). Invited perspectives: How machine learning will change flood risk and impact assessment. *Natural Hazards and Earth System Sciences*, 20, 1149-1161. <https://doi.org/10.5194/nhess-20-1149-2020>
- [5] Vincent, A. M., Parthasarathy, K. S. S., & Jidesh, P. (2023). Flood susceptibility mapping using AutoML and a deep learning framework with evolutionary algorithms for hyperparameter optimization. *Applied Soft Computing*, 148, 110846. <https://doi.org/10.1016/j.asoc.2023.110846>