Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Flood Susceptibility Mapping in Yola Metropolis, Nigeria, Using Random Forest, Terrain-Derived Predictors, and Nested Spatial Cross-Validation

Temitope Emmanuel Paul
u23944359@tuks.co.za
University of Pretoria
Adedayo Kelvin ADELEKE
University of Pretoria
Peace Olubukunmi AWOLEYE
University of Pretoria
Akintunde Ayo AKINSE
Adekunle Ajasin University

Research Article

Keywords: Flood susceptibility, Random Forest, Digital Elevation Model (DEM), Spatial cross-validation, Nested cross-validation, Yola Metropolis

Posted Date: October 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11363913/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Flood susceptibility mapping is essential for disaster risk reduction, particularly in rapidly urbanizing, data-constrained regions of sub-Saharan Africa. This study developed a Random Forest-based flood susceptibility model for Yola Metropolis, northeastern Nigeria, a low-lying urban area within the Upper Benue River basin. Predictor variables included elevation, conventional slope, 2022 land-use/land-cover (LULC), engineered terrain metrics, and distance-to-stream variables. An archived Sentinel-1 event inventory provided 852 labelled locations, of which 832 complete samples (409 Flood and 423 NonFlood) were retained for modelling. Model performance was evaluated using 5-km spatial blocks and nested cross-validation to separate hyperparameter tuning from outer-fold assessment. Five predictor configurations were compared. The model with the highest balanced accuracy used elevation, conventional slope, and LULC, achieving 92.43% accuracy, 91.69% sensitivity, 93.14% specificity, 92.42% balanced accuracy, and a ROC-AUC of 0.943. Permutation analysis identified elevation as the dominant predictor, while slope and LULC contributed limited independent information. Susceptibility classes were defined from nested out-of-fold scores as low (< 0.1869), moderate (0.1869–<0.7331), and high (≥ 0.7331). Of the 981.06 km² model-supported area, 82.83% was classified as low susceptibility, 9.41% as moderate, and 7.76% as high, while 5.26 km² remained unsupported because one mapped LULC class was absent from training. Although the framework is reproducible, the external LULC product lacked local validation, and no independent 2025 flood polygon was available. The resulting map should therefore be interpreted as a susceptibility screening product rather than a calibrated hazard or risk estimate. The map supports flood-susceptibility screening and planning prioritization in Yola and provides a workflow that can be evaluated in other data-limited urban settings subject to local validation.

1. Introduction

Flooding remains a persistent challenge in rapidly growing cities, where expansion of impervious surfaces, occupation of low-lying terrain, inadequate drainage infrastructure, and intense rainfall or river overflow can substantially increase flood impacts. These pressures are especially pronounced in sub-Saharan Africa, where rapid and often unplanned urbanization, limited drainage capacity, and constrained disaster-preparedness systems increase the vulnerability of urban settlements to hydrological extremes (Islam et al., 2025). In response, flood-susceptibility studies increasingly combine remote sensing, terrain information, and machine-learning techniques because these approaches can identify spatial patterns even where dense hydrometric observations are unavailable (Islam et al., 2025; Kaya & Derin, 2023).

Nigeria has experienced recurrent and damaging floods with consequences for livelihoods, infrastructure, agriculture, and economic activity. Yola Metropolis in Adamawa State, northeastern Nigeria, is particularly relevant because expanding urban development occurs alongside low-lying terrain associated with the Upper Benue system. In such a data-constrained setting, there is a strong need for spatially explicit tools that can identify areas with environmental characteristics associated with flooding and support prioritization of locations requiring more detailed investigation.

Flood susceptibility should, however, be distinguished from flood hazard and flood risk. Susceptibility reflects a location's relative propensity to experience flooding based on its environmental characteristics; it does not directly quantify flood frequency, water depth, exposure, or vulnerability. Event-specific hydraulic modelling remains necessary where the objective is to estimate flood extent, depth, or probability under defined hydrological conditions (Teng et al., 2017). This distinction is important for urban decision-making because a susceptibility map is best used as a screening and prioritization layer rather than as a substitute for engineering flood-hazard analysis.

Machine-learning models are particularly useful when the relationship between flood occurrence and environmental conditioning factors is nonlinear, although their apparent performance can be overstated if model evaluation ignores spatial dependence (Mosavi et al., 2018; Roberts et al., 2017). Machine-learning approaches have also demonstrated considerable potential for flood-susceptibility assessment in data-constrained environments, as they can capture complex, nonlinear relationships among environmental conditioning factors (Mosavi et al., 2018; Islam et al., 2025).

Random Forest (RF) is widely used for flood susceptibility because it accommodates nonlinear interactions, mixed predictor types and complex decision boundaries without requiring a prespecified functional form (Breiman, 2001). Previous flood-mapping studies have demonstrated the value of topography, land cover, and proximity to drainage as conditioning factors, while also showing that predictor importance can vary substantially between landscapes (Tehrany et al., 2014; Tehrany et al., 2015). More recent work has emphasized interpretability and has shown that a small set of terrain and land-cover variables can sometimes perform strongly when the dominant flood controls are topographic (Alcântara et al., 2025).

An equally important issue is validation design. Conventional random or leave-one-out validation can place geographically neighbouring observations in both training and testing sets, which may produce optimistic accuracy when spatial autocorrelation is present. Spatially structured cross-validation reduces this leakage by keeping groups of nearby observations together during resampling (Roberts et al., 2017). When hyperparameters are tuned, a second source of optimism can arise if the same resampling results are used both to select the model and to report its final accuracy. Nested validation addresses this problem by separating inner model selection from outer performance estimation (Varma & Simon, 2006; Cawley & Talbot, 2010).

The aim of this study is to produce and evaluate a spatially explicit RF-based flood susceptibility model for the Yola Metropolis study area. Specifically, the study (i) harmonizes a 30 m digital elevation model, conventional slope and 2022 Sentinel-2 LULC; (ii) evaluates alternative predictor sets, including an archived engineered slope layer and distance to streams; (iii) tunes and assesses RF models using nested five-kilometre spatial cross-validation; (iv) quantifies predictor contribution using held-out permutation importance; and (v) converts the continuous RF flood-class score into low, moderate and high susceptibility classes using thresholds derived from out-of-fold validation scores.

The study, therefore, emphasizes spatially explicit validation, predictor ablation, and reproducible class thresholds rather than assuming that additional engineered variables necessarily improve performance.

2. Materials and Methods

2.1 Study area

The analysis covers the Yola Metropolis study boundary in Adamawa State, northeastern Nigeria, as shown in Fig. 1. The final projected boundary is referenced to WGS 1984 UTM Zone 33N (EPSG:32633) and spans approximately 12.08°–12.61° E and 9.04°–9.44° N.

Rasterization on the 30 m modelling grid produced an analytical footprint of 986.32 km². The study area includes low-lying surfaces adjacent to the Benue River as well as relatively higher interior terrain, providing a suitable setting for examining topographic controls on flood susceptibility.

Because this work evaluates susceptibility rather than event-specific hydraulic hazard, the study boundary is treated as the spatial domain for comparing environmental conditions. The final susceptibility statistics are therefore reported separately for model-supported cells and for cells where prediction was withheld because the mapped LULC category was absent from model training.

2.2 Spatial data and predictor preparation

The workflow used terrain, land-cover, and event-reference data that could be traced to the archived project and independently checked. Table 1 summarizes the datasets and their roles. All spatial processing and modelling were carried out in WGS 1984 UTM Zone 33N so that distances and raster cell sizes were expressed in meters.

Table 1
Spatial datasets and predictor preparation used in the flood-susceptibility workflow.

Dataset/variable	Source	Year/date	Native resolution	Processing	Role
Elevation	SRTM GL1 via OpenTopography / NASA	SRTM mission: 2000	~ 30 m (1 arc-second)	Projected to EPSG:32633; clipped to study domain	Retained in final model
Conventional slope	Derived from SRTM DEM	Derived for this study	30 m	Calculated in degrees on the DEM grid	Retained in final model
2022 LULC	ArcGIS Living Atlas; Impact Observatory, Microsoft & Esri	2022	10 m	Tiles exported/mosaicked; nearest-neighbour alignment to 30 m grid	Retained as a categorical predictor
Flood/NonFlood reference inventory	Archived Sentinel-1 event inventory	23 Aug 2022 pre-event; 4 Sep 2022 event	Point data	852 labels; NoData screening retained 832 samples	Training and spatial validation labels
Distance to streams	HydroRIVERS	Version 1.0	Vector -> 30 m raster	Projected to UTM 33N; planar distance generated and sampled	Ablation/sensitivity predictor
Engineered slope	Archived terrain derivative	Derived from DEM	30 m-equivalent raster	Retained only as a comparator because the conventional slope was more reproducible	Ablation/sensitivity predictor

2.2.1 Digital elevation model and conventional slope

Elevation was represented by the Shuttle Radar Topography Mission Global 1-arc-second product (SRTM GL1), obtained from OpenTopography. SRTM was acquired during the February 2000 Shuttle Radar Topography Mission and provides near-global elevation data at approximately 1-arc-second (about 30 m) spacing (Farr et al., 2007). The DEM was projected to WGS 1984 UTM Zone 33N (EPSG:32633), clipped to the study domain, and used as the reference grid for subsequent raster harmonization. Conventional slope was derived from the DEM in degrees and retained as the reproducible terrain-gradient variable in the final modelling workflow.

The DEM corresponds to the NASA SRTM Global 1 arc-second Version 3 collection (SRTMGL1.003; NASA JPL, 2013; DOI: 10.5067/MEASURES/SRTM/SRTMGL1.003). The standard OpenTopography SRTM GL1 product is horizontally referenced to WGS 84 and elevationally referenced to the EGM1996 geoid (EPSG:5773). The archived project did not retain the original OpenTopography job identifier or the source granule/tile filename; therefore, tile-level provenance cannot be reconstructed retrospectively.

The analysis treats SRTM as a regional terrain representation rather than survey-grade topography. No independent local vertical-accuracy campaign or tile-specific error assessment was available for the Yola subset, and the study therefore avoids claims that depend on centimeter- or meter-level vertical precision. This limitation is particularly relevant in very low-relief floodplain terrain, where small vertical errors can affect the interpretation of local drainage.

2.2.2 Land use/land cover

Land-use/land-cover information was obtained from the 2022 Sentinel-2 10 m Land Use/Land Cover Time Series distributed through ArcGIS Living Atlas of the World and produced through the Esri-Impact Observatory-Microsoft partnership. The product is derived from ESA Sentinel-2 imagery using a deep-learning semantic-segmentation workflow trained on billions of human-labeled pixels (Karra et al., 2021; Esri, Impact Observatory, & Microsoft, 2023). The 2022 annual layer was selected because it corresponds to the year represented by the event-based flood reference inventory.

Tiles covering the Yola Metropolis were exported from ArcGIS Living Atlas, mosaicked where necessary, and clipped to the study boundary. The classes present within the study area, as shown in Fig. 2, were Water (1), Trees (2), Flooded Vegetation (4), Crops (5), Built Area (7), Bare Ground (8), and Rangeland (11). LULC was treated as a categorical model input rather than as an ordinal numeric variable. For wall-to-wall prediction, the native 10 m LULC raster was aligned with the 30 m DEM grid using nearest-neighbor resampling to preserve class codes and avoid introducing artificial intermediate categories (Congalton & Green, 2019).

Because this study used an externally produced global LULC product, no new local classifier was trained, and a study-specific LULC confusion matrix was not generated. Independent global assessment has shown that the Esri product performs comparatively well but that accuracy varies by class and geographic context; Venter et al. (2022), for example, reported 75% overall accuracy for the 2020 Esri product under their global validation design. That global figure is not treated here as a Yola-specific accuracy estimate. Instead, local processing uncertainty was checked directly: 822 of the 832 modelling locations (98.80%) retained the same LULC class after 10 m-to-30 m harmonization, while 10 locations (1.20%) changed class, mainly at class boundaries. This agreement test evaluates grid harmonization rather than thematic classification accuracy, and the absence of an independent local LULC validation set remains a limitation.

Because the annual 2022 LULC product may incorporate Sentinel-2 observations acquired at different times during the year, it was treated as a broad categorical surface-condition predictor rather than as a strictly pre-event land-cover snapshot.

2.2.3 Sentinel-1 reference inventory

The analysis used an archived event-based flood/non-flood reference inventory associated with Sentinel-1 observations acquired in 2022. Metadata associated with the inventory identified a pre-event acquisition on 23 August 2022 and an event acquisition on 4 September 2022, both from a descending pass with orbit identifier 153.

Sentinel-1 SAR is well suited to flood mapping because open water commonly exhibits lower radar backscatter than surrounding land surfaces, enabling flood detection under cloud cover (Twele et al., 2016). The archive contained 852 labelled points together with class labels, sample identifiers, projected coordinates, acquisition dates, and relevant Sentinel-1 metadata. After predictor extraction, 20 records containing − 9999/NoData in one or more required fields were excluded, leaving 832 samples comprising 409 Flood and 423 NonFlood locations, as shown in Fig. 3. Sample identifiers and projected coordinates were retained for reproducible spatial blocking. The original pixel-level thresholding procedure used to derive the event labels was not documented in sufficient detail to permit independent reconstruction; therefore, the archived labels were used as provided, without inferring or recreating undocumented processing steps. The surviving project archive also did not retain sufficient documentation to reconstruct the original point-selection protocol. Accordingly, no undocumented sampling strategy is inferred or claimed in the present study.

2.2.4 Distance to streams

To test whether a hydrological proximity variable improved discrimination, HydroRIVERS v1.0 river segments were projected to UTM Zone 33N and clipped to a 10 km buffer around the study domain. A planar 30 m distance-to-stream raster was generated and then clipped to the Yola boundary. The final sampled distance values ranged from 0 to approximately 4.40 km. HydroRIVERS is derived from the HydroSHEDS framework and provides a consistent global river network suitable for large-area spatial analysis (Lehner & Grill, 2013). Distance to streams was evaluated in ablation experiments but was not retained in the final classifier.

2.3 Raster harmonization and quality control

The DEM defined the common 30 m modelling grid. Conventional slope was checked against that grid, while the native 10 m categorical LULC raster was aligned using nearest-neighbor resampling. Quality control at the 832 modelling points showed exact agreement between stored and raster-sampled DEM values and agreement within 0.001° for all conventional-slope values. After LULC alignment, 822 points retained their original 10 m class, and 10 points changed class, giving 98.80% agreement. Missing or invalid predictor values were not imputed as zero; records containing the − 9999 sentinel value were excluded before modelling. This avoided creating artificial flat terrain or false categorical values.

2.4 Predictor configurations and Random Forest modelling

Random Forest classification was implemented in Python using scikit-learn (Breiman, 2001; Pedregosa et al., 2011). LULC was treated as a categorical variable and one-hot encoded within the modelling pipeline, while the continuous predictors were passed directly to the classifier. Five feature configurations were evaluated under the same spatial-validation framework: M0 (DEM + LULC), M1 (DEM + conventional slope + LULC), M2 (DEM + engineered slope + LULC), M3 (DEM + LULC + distance to streams), and M4 (DEM + engineered slope + LULC + distance to streams). This ablation design was used to test, rather than assume, whether slope engineering or river proximity provided measurable incremental value.

2.5 Nested spatial cross-validation and hyperparameter tuning

Model assessment used spatial rather than random resampling because neighbouring observations can share environmental structure and thereby inflate conventional cross-validation scores (Roberts et al., 2017). Projected coordinates were grouped into 5-km square spatial blocks. Five outer folds were created so that all samples from the same block remained together, and each outer fold contained both Flood and NonFlood observations.

Hyperparameter tuning was performed only within the outer training data using four predefined inner spatial folds. In this nested design, the outer fold was untouched until final evaluation, reducing selection bias that can arise when model tuning and performance reporting use the same resampling results (Varma & Simon, 2006; Cawley & Talbot, 2010).

Randomized hyperparameter search used balanced accuracy as the selection criterion because the objective was to maintain comparable discrimination of Flood and NonFlood classes rather than optimize overall accuracy alone. Eight parameter combinations were evaluated within each outer-fold/model search. The search varied the number of trees, maximum tree depth, minimum samples required to split an internal node, minimum samples per leaf, and the number of features considered at each split. The Gini criterion, balanced-subsample class weighting, and random state 42 were held fixed. After nested evaluation identified M1 as the leading threshold-based configuration, a separate full-data spatial search was conducted solely to select production parameters for the final wall-to-wall model; its cross-validation score was not reported as an independent performance estimate.

Table 2
Random Forest hyperparameter search space and final production settings.

Parameter	Nested spatial search	Final production search	Selected production value
Number of trees (n_estimators)	200, 300, 500	200, 300, 500, 700	500
Maximum depth	None, 8, 12, 20	None, 8, 12, 16, 20	None
Minimum samples split	2, 5, 10	2, 5, 10	5
Minimum samples leaf	1, 2, 3, 5, 8	1, 2, 3, 5, 8	3
Maximum features	sqrt, 0.7, 1.0	sqrt, 0.7, 1.0	sqrt
Class weighting	balanced_subsample (fixed)	balanced_subsample (fixed)	balanced_subsample
Criterion	Gini (fixed)	Gini (fixed)	Gini
Random state	42	42	42

2.6 Final model, permutation importance, and production fitting

The final candidate was selected primarily on outer-fold balanced accuracy, with ROC-AUC used as a complementary ranking metric. For the selected feature set, permutation importance was calculated separately on each held-out outer fold using 30 permutations and balanced accuracy as the scoring function. Importance was then summarized across folds. This approach evaluates the change in predictive performance when a predictor is disrupted on unseen data, rather than relying only on impurity-based importance from the fitted trees. The production RF was subsequently refitted to all 832 samples using 500 trees, max_depth=None, min_samples_split = 5, min_samples_leaf = 3, max_features=sqrt, and class_weight=balanced_subsample.

2.7 Susceptibility score, model-support mask, and class thresholds

The fitted RF generated the probability assigned to the Flood class for each supported 30 m cell. Because these probabilities were not calibrated independently as event frequencies, they are interpreted here as continuous

susceptibility scores rather than as absolute flood probabilities. The final wall-to-wall predictor stack contained elevation, conventional slope, and the aligned 2022 LULC. LULC class 4 (Flooded Vegetation) occurred in the map but was absent from model training; those cells were therefore withheld from prediction and mapped as an unsupported model domain rather than silently extrapolated.

Low, moderate, and high susceptibility thresholds were derived from the nested out-of-fold predictions rather than from visual map breaks. The lower threshold was defined as the largest score for which no more than 5% of observed Flood samples fell below the cutoff. The upper threshold was defined as the smallest score for which no more than 5% of observed NonFlood samples occurred at or above the cutoff. This produced thresholds of 0.18694 and 0.73312, respectively.

Threshold sensitivity was evaluated by repeating the out-of-fold threshold derivation using tail tolerances of 2.5%, 5%, 7.5%, and 10%. For each tolerance, the lower threshold was defined as the largest score for which the specified proportion of observed Flood samples could fall below the cutoff, while the upper threshold was defined as the smallest score for which the specified proportion of observed NonFlood samples could occur at or above the cutoff. A threshold pair was considered admissible only when the lower cutoff remained below the upper cutoff, thereby preserving mutually exclusive Low, Moderate, and High susceptibility classes. The corresponding class areas were recalculated from the final model-supported probability surface.

The middle interval was retained as Moderate susceptibility, representing observations for which the two classes were less clearly separated. Thresholds were therefore linked directly to held-out classification behavior, not to Natural Breaks or arbitrary quantiles.

2.8 Reproducibility, support domain, and uncertainty checks

Reproducibility checks were built into the workflow. All spatial folds were predefined from projected coordinates, the RF random state was fixed at 42, categorical encoding was performed inside the modelling pipeline, and out-of-fold predictions were retained for every observation. The 3-km, 5-km, and 7.5-km block-size sensitivity analysis was used to examine whether the main conclusions depended strongly on a single spatial blocking scale. For the fixed-parameter DEM-LULC baseline, balanced accuracy remained at 93.61%, 92.88%, and 93.61% at 3, 5, and 7.5 km, respectively, indicating that the main discrimination pattern was not an artifact of any single block size.

For wall-to-wall prediction, a model-support mask was also retained. LULC class 4 (Flooded Vegetation) occurred in the mapped 2022 LULC layer but was absent from the training samples. Rather than allowing the one-hot encoder to silently extrapolate to an unseen category, those cells were withheld from prediction and mapped as an unsupported model domain. This affected 5,842 cells (5.26 km²; 0.53% of the rasterized analysis area).

Geospatial preprocessing and cartographic preparation were performed in ArcGIS Pro 3.4.0. Statistical modelling, spatial cross-validation, and raster prediction were implemented in Python 3.9.13 using pandas 2.3.3, NumPy 1.24.4, scikit-learn 1.6.1, rasterio 1.4.3, and matplotlib 3.5.2. Bootstrap sampling was not included among the tuned hyperparameters and remained enabled at the scikit-learn default (bootstrap=True) throughout Random Forest model fitting. Random-state control was applied consistently to support reproducibility.

3. Results

3.1 Nested spatial validation and model comparison

Nested spatial validation showed that all five feature configurations discriminated Flood and NonFlood locations well, but the ranking differed depending on the metric considered. M1 (DEM + conventional slope + LULC) produced the highest threshold-based performance, with 92.43% accuracy and 92.42% balanced accuracy. The simpler M0 model (DEM + LULC) was only 0.37 percentage points lower in balanced accuracy, showing that conventional slope added a modest incremental gain rather than a large improvement. M2, which substituted the archived engineered slope for a conventional slope, performed slightly worse than M0. M3 and M4 produced the highest or near-highest ROC-AUC values, but both reduced sensitivity and balanced accuracy at the 0.5 classification threshold. These comparisons provide direct ablation evidence that neither engineered slope nor distance to streams materially improved the final binary discrimination.

Table 3. Nested 5-km spatial cross-validation performance of the five Random Forest predictor configurations.

Model	Predictors	Accuracy (%)	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	ROC-AUC
M1	DEM + conventional slope + LULC	92.43	91.69	93.14	92.42	0.943
M0	DEM + LULC	92.07	91.20	92.91	92.05	0.940
M2	DEM + engineered slope + LULC	91.83	91.20	92.44	91.82	0.940
M3	DEM + LULC + distance to streams	90.14	86.55	93.62	90.08	0.944
M4	DEM + engineered slope + LULC + distance to streams	89.54	86.55	92.44	89.49	0.944

A supplementary block-size sensitivity analysis using 3 km, 5 km, and 7.5 km blocks showed that the principal DEM-LULC configuration remained stable, with balanced accuracy of approximately 93.61%, 92.88%, and 93.61%, respectively, under the fixed-parameter comparison. This supported the use of 5 km blocks as a reasonable spatial scale of separation for the nested analysis.

3.2 Performance of the selected M1 model

The selected M1 model achieved 92.43% overall accuracy, 91.69% sensitivity, 93.14% specificity, 92.82% precision, 92.25% F1-score, and 92.42% balanced accuracy. The pooled nested out-of-fold ROC-AUC was 0.943. The confusion matrix comprised 394 true negatives, 29 false positives, 34 false negatives, and 375 true positives (Table 4; Figure 4). Across the five outer spatial folds, balanced accuracy ranged from 87.80% to 96.38% (mean 92.18%, SD 3.05 percentage points), while ROC-AUC ranged from 0.919 to 0.973 (mean 0.944, SD 0.024). The geographic variation between folds is informative rather than incidental: it shows that model performance was not spatially uniform and reinforces the need for spatially explicit validation.

Table 4. Diagnostic performance of the selected M1 Random Forest model under nested 5-km spatial cross-validation.

Metric	Definition	Value
Accuracy	(TP + TN) / N	92.43%
Sensitivity / Recall	TP / (TP + FN)	91.69%
Specificity	TN / (TN + FP)	93.14%
Precision	TP / (TP + FP)	92.82%
F1-score	2 × Precision × Recall / (Precision + Recall)	92.25%
Balanced accuracy	(Sensitivity + Specificity) / 2	92.42%
ROC-AUC	Area under ROC curve	0.943
Confusion matrix	TN=394; FP=29; FN=34; TP=375	N=832

3.3 Predictor contribution and ablation

Held-out permutation importance showed that elevation carried most of the independent predictive information in M1, as shown in Figure 5. Permuting elevation reduced balanced accuracy by an average of 0.4102 (SD = 0.0294 across outer folds). LULC produced a much smaller mean decrease of 0.0017 (SD = 0.0050), while the conventional slope had a near-zero mean importance of -0.0009 (SD = 0.0031). The small negative value does not imply that slope physically reduces flooding; rather, it indicates that slope added little independent information once elevation and LULC were already available. This interpretation is consistent with the ablation results, in which M1 exceeded the simpler M0 by only 0.37 percentage points in balanced accuracy, as shown in Table 3.

3.4 Spatial distribution of flood susceptibility

The final continuous susceptibility score ranged from approximately 0.027 to 0.988. The two out-of-fold thresholds separated the reference observations into high-confidence tails and an intermediate zone. Only 20 of 409 observed Flood samples (4.89%) fell in the Low class, while only 21 of 423 observed NonFlood samples (4.97%) fell in the High class. The Moderate class contained the remaining ambiguous cases (Table 5). This validation-linked approach makes the class boundaries reproducible and avoids interpreting an arbitrary cartographic break as a scientific threshold.

Table 5. Observed class allocation under the susceptibility thresholds derived from nested out-of-fold predictions.

Susceptibility class	Score rule	Observed NonFlood	Observed Flood	Total
Low	< 0.18694	358	20	378
Moderate	0.18694 to < 0.73312	44	18	62
High	>= 0.73312	21	371	392
Total	-	423	409	832

Threshold-sensitivity analysis showed that susceptibility-class proportions varied with the allowable tail tolerance. At 2.5%, the lower and upper cutoffs were 0.10922 and 0.90853, respectively, producing 72.23% Low, 22.81%

Moderate, and 4.97% High susceptibility within the model-supported area. At the adopted 5% tolerance, the corresponding proportions were 82.83%, 9.41%, and 7.76%. Increasing the tolerance to 7.5% narrowed the intermediate interval substantially (0.43901–0.49419), reducing the Moderate class to 0.07% of the supported area, while Low and High accounted for 91.02% and 8.91%, respectively. At 10%, the derived lower threshold exceeded the upper threshold, so a mutually exclusive three-class solution was no longer admissible. The sensitivity analysis, therefore, confirms that mapped class proportions depend on the tolerance criterion and supports the 5% rule as a transparent compromise between the broader intermediate zone produced at 2.5% and the near-collapse of the Moderate class at 7.5%, as shown in Table 6 and Figure 6.

Table 6. Sensitivity of validation-derived susceptibility thresholds and mapped class proportions to alternative tail-tolerance criteria

Tail tolerance (%)	Lower cutoff	Upper cutoff	Low (%)	Moderate (%)	High (%)	Admissible
2.5	0.10922	0.90853	72.23	22.81	4.97	Yes
5.0	**0.18694**	**0.73312**	**82.83**	**9.41**	**7.76**	**Yes**
7.5	0.43901	0.49419	91.02	0.07	8.91	Yes
10.0	0.73675	0.33286	—	—	—	No

Note: Percentages refer to the model-supported area. The 10% rule was inadmissible because the lower threshold exceeded the upper threshold, resulting in overlapping class definitions.

Table 7. Area and proportion of the final flood-susceptibility classes within the study boundary.

Class	30 m cells	Area (km²)	% of analysis area	% of supported area
Unsupported model domain	5,842	5.2578	0.53	—
Low susceptibility	902,907	812.6163	82.39	82.83
Moderate susceptibility	102,618	92.3562	9.36	9.41
High susceptibility	84,546	76.0914	7.71	7.76
Total rasterized analysis area	1,095,913	986.3217	100.00	—
Model-supported area	1,090,071	981.0639	—	100.00

Table 7 shows that within the 981.06 km² model-supported domain, low susceptibility covered 812.62 km² (82.83%), moderate susceptibility covered 92.36 km² (9.41%), and high susceptibility covered 76.09 km² (7.76%). An additional 5.26 km², equivalent to 0.53% of the rasterized analysis area, was mapped as unsupported because LULC class 4 (Flooded Vegetation) was not represented among the training samples. Percentages for Low, Moderate, and High are therefore reported relative to the supported domain, while the unsupported area is reported separately rather than being misclassified as low susceptibility.

3.5 Spatial stability and uncertainty

The spatial validation results show that predictive skill varied across the study area. Outer-fold balanced accuracy ranged from 87.80% to 96.38%, and the lowest-performing fold was materially weaker than the best fold. This

spread is an important uncertainty signal: the pooled 92.42% balanced accuracy summarizes all held-out observations, but it should not be interpreted as evidence that every part of the metropolis is predicted equally well.

The block-size sensitivity experiment provided a second check on spatial dependence. For the DEM-LULC baseline, balanced accuracy changed by less than one percentage point between 3-km and 7.5-km blocking and was 92.88% at the 5-km scale used for nested tuning. Together with the fold-level variation and the explicit unsupported-domain mask, these checks provide a more transparent representation of model uncertainty than a single random-split accuracy value.

4. Discussion

4.1 What the validation changes

The central methodological result is not simply the level of model accuracy; it is the way that performance was estimated. Spatial blocking prevents neighbouring samples from being split arbitrarily between training and testing data, while nested resampling keeps hyperparameter selection inside the training data. Both choices make the reported accuracy more defensible for spatial prediction (Roberts et al., 2017; Cawley & Talbot, 2010). The strong and balanced sensitivity and specificity of M1 therefore provide a more informative summary of within-study-area discrimination than would be obtained from ordinary random resampling.

At the same time, the modest difference between M1 and M0 argues against presenting slope as a major optimization breakthrough. Conventional slope improved balanced accuracy only slightly, and the engineered slope did not improve the model at all. The interpretation is therefore deliberately conservative: elevation + LULC already captured most of the classification structure, while conventional slope provided a small incremental gain.

4.2 Why elevation dominates

The permutation result indicates that elevation is the principal independent predictor in the selected model. This is physically plausible in a low-relief riverine setting because small differences in relative elevation can influence whether water accumulates or drains away. Flood-susceptibility studies commonly identify elevation and related terrain variables among the strongest conditioning factors (Tehrany et al., 2014; Tehrany et al., 2015). Recent interpretable machine-learning work has likewise shown that topographic descriptors can dominate urban flood-susceptibility models when they reflect the principal controls on water accumulation (Alcântara et al., 2025).

The near-zero independent importance of slope should not be interpreted as evidence that slope is hydrologically irrelevant. Slope was derived from the same DEM and therefore shares information with elevation. Permutation importance measures incremental predictive information conditional on the other variables in the fitted model; correlated or derived predictors can appear weak even when they remain physically meaningful. The small gain from M1 over M0 is consistent with this redundancy.

4.3 Role of LULC and distance to streams

LULC contributed comparatively little independent permutation importance in M1, but it remains useful for representing broad differences in surface condition. Because categorical LULC was harmonized to a 30 m modelling grid, the observed 1.2% class disagreement at sample locations is best understood as an edge effect of resolution harmonization rather than a continuous-value interpolation error. Nearest-neighbour resampling was therefore retained because it preserves class identities (Congalton & Green, 2019).

Distance to streams slightly increased ROC-AUC but reduced sensitivity and balanced accuracy at the 0.5 decision threshold. This illustrates that a predictor can improve the ranking of susceptibility scores without improving the final binary classification. The variable was therefore treated as informative for sensitivity analysis but was not retained in the production model. This result also cautions against assuming that adding more hydrological predictors will automatically improve performance; their value depends on spatial scale, data quality, and the structure of the reference inventory.

4.4 Comparison with recent machine-learning flood-susceptibility studies

The present findings are broadly consistent with recent machine-learning flood-susceptibility studies showing that Random Forest can provide useful urban-scale discrimination while the dominant predictors vary with environmental setting and predictor availability. In the São Paulo Metropolitan Region, Alcântara et al. (2025) found that Random Forest outperformed several alternative machine-learning algorithms, while an explainability analysis identified valley depth, land cover, and geology as among the most influential variables. Similarly, Danumah et al. (2026) applied Random Forest to an urban watershed in Abidjan, Côte d'Ivoire, using 12 topographic, hydrological, land-cover, and climatic factors and reported an AUC of approximately 0.94 and an overall accuracy of 0.83, with drainage density, rainfall, and altitude among the principal explanatory variables. The Yola results likewise emphasize the importance of terrain, particularly elevation, but were obtained using a more parsimonious set of predictors and a nested spatial validation framework. Direct numerical ranking among these studies would be inappropriate because their flood inventories, predictor sets, study environments, class definitions, and validation procedures differ. The comparison instead indicates that strong model discrimination can arise from different combinations of conditioning variables and reinforces the importance of evaluating predictor contribution and validation design within each study context.

4.5 LULC uncertainty and transferability

The LULC predictor warrants a separate discussion of uncertainty because it is an externally produced global product rather than a locally trained classifier. Independent assessment of 10 m global land-cover products found that the Esri product achieved 75% overall accuracy under a global 2020 validation design, with substantial variation among classes (Venter et al., 2022). That evidence supports the product as a useful broad-scale predictor but does not establish the accuracy of individual Yola pixels. The 98.80% agreement measured in this study refers only to preservation of class labels during 10 m-to-30 m harmonization; it is not a substitute for local thematic accuracy assessment.

This distinction matters because a small amount of LULC error can propagate into susceptibility predictions, particularly near class boundaries or where rare classes are poorly represented in the reference inventory. The decision to mask Flooded Vegetation rather than extrapolate to an unseen training category is one response to that uncertainty. A future independent Yola LULC validation using contemporaneous reference imagery or field observations would provide a stronger local accuracy estimate and would help quantify how land-cover uncertainty propagates through the RF model.

4.6 Planning, interpretation, and limitations

The final map is most appropriate as a screening product for identifying locations whose terrain and land-cover characteristics resemble the flooded reference class. It should not be interpreted as a calibrated probability of inundation, a design-flood depth map, or a complete flood-risk map. Those applications require information on

event frequency, rainfall or discharge forcing, hydraulic routing, exposure, and vulnerability. The current map can nevertheless support prioritization of detailed surveys, drainage assessment, and land-use review, particularly where high susceptibility coincides with planned development.

Several limitations remain. First, the event reference inventory was retrieved from an archived Sentinel-1 project, and the original pixel-level thresholding and point-selection procedures were not completely preserved in the project archive. The observable acquisition metadata are therefore reported explicitly, and the inventory is interpreted as an archived event-labelled reference dataset rather than an independently reproduced long-term flood catalogue. Second, the model is dominated by elevation and does not include event rainfall intensity, drainage capacity, or river discharge. Third, the external LULC product was not independently validated within Yola, and 10 of the 832 sample locations changed class during 10 m-to-30 m grid harmonization. Fourth, mapped Flooded Vegetation was absent from training and was deliberately excluded from prediction, leaving a small unsupported domain. Fifth, the July 2025 flood could not be used as a quantitative independent validation because a reliable georeferenced inundation polygon was not available to the study. Finally, nested spatial cross-validation estimates transfer within the present study region; it does not demonstrate transferability to other cities, hydroclimatic regimes, or future flood events.

5. Conclusions

This study presents a spatially validated Random Forest approach to flood susceptibility mapping in the Yola Metropolis based on an 832-point event-labeled inventory, nested 5-km spatial cross-validation, explicit feature ablation, and validation-derived class thresholds. The selected model combined elevation, conventional slope, and 2022 Sentinel-2 LULC and achieved 92.43% accuracy, 91.69% sensitivity, 93.14% specificity, 92.42% balanced accuracy, and a ROC-AUC of 0.943.

Ablation and permutation analyses showed that elevation accounted for most of the model's independent predictive information. The conventional slope produced only a marginal improvement over the DEM-LULC baseline, whereas the archived engineered slope did not improve classification. Distance to streams slightly improved ROC-AUC but reduced threshold-based performance. These results do not support describing the model as "topographically optimized"; instead, they support a more defensible interpretation in which a parsimonious terrain-land-cover model provides strong within-study-area discrimination.

Using thresholds derived from nested out-of-fold scores, 82.83% of the model-supported area was classified as low susceptibility, 9.41% as moderate, and 7.76% as high. A further 5.26 km² was left unsupported because its LULC category was not included in the training. The resulting map is suitable for spatial screening and planning prioritization, provided it is not treated as a substitute for event-specific hydraulic hazard modelling or a complete risk assessment. The main methodological contribution is therefore not a claim of universal predictive transferability, but a transparent workflow that combines spatially blocked nested validation, predictor ablation, held-out importance analysis, reproducible thresholds, and an explicit model-support domain. Future work should test the framework against independent flood-event polygons and incorporate rainfall, discharge, drainage capacity, and locally validated land-cover information where reliable observations become available.

6. Practical implications

Planning agencies can use the final susceptibility map as an initial screening layer to flag locations that warrant more detailed flood investigation before development approval. High-susceptibility zones should trigger site-

specific drainage, topographic, and hydraulic assessments rather than automatic prohibition, because susceptibility does not account for event magnitude, structural exposure, or vulnerability. The unsupported model domain should remain clearly distinguished in all planning products so that users do not mistake the lack of model support for low susceptibility.

Declarations

Funding: This work was not supported by external funding.

Ethics approval and consent to participate: The study used geospatial and remotely sensed data and did not involve human or animal participants. Ethical approval and participant consent were therefore not required.

Consent for publication: Not applicable.

Competing interests: The authors declare no known competing financial or personal interests that could have influenced the work reported in this paper.

Author contributions: Temitope Emmanuel Paul: Conceptualization, methodology, geospatial analysis, machine-learning analysis, spatial validation, visualization, and original draft preparation.

Adedayo Kelvin Adeleke: Supervision, investigation, review and editing, and project administration.

Peace Olubukunmi Awoleye: Methodology, review, and editing.

Akintunde Ayo Akinse: Conceptualization, validation, review, and editing.

Data availability: The processed data tables and analysis outputs supporting the reported results can be made available by the corresponding author upon reasonable request, subject to the terms of the original data providers.

References

1. Alcântara, E., Guimarães, Y. C., Baião, C. F. P., & Mantovani, J. R. (2025). Interpretable machine learning for flood susceptibility mapping in the metropolitan region of São Paulo, Southeast Brazil. Discover Geoscience, 3, 243. https://doi.org/10.1007/s44288-025-00362-9
2. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324
3. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079-2107.
4. Congalton, R. G., & Green, K. (2019). Assessing the accuracy of remotely sensed data: Principles and practices (3rd ed.). CRC Press.
5. Danumah, J. H., Ataba, W. A., Jofack Sokeng, V. C., Akpa, Y. L., Saley, M. B., & Ogilvie, A. (2026). Assessing urban flood susceptibility using Random Forest machine learning and geospatial technologies: Application to the Bonoumin-Palmeraie Watershed, Abidjan (Côte d'Ivoire). Water, 18(3), 402. https://doi.org/10.3390/w18030402
6. Esri, Impact Observatory, & Microsoft. (2023). Sentinel-2 10m Land Use/Land Cover Time Series [Data set]. ArcGIS Living Atlas of the World. https://livingatlas.arcgis.com/landcover/

7. Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., & Alsdorf, D. E. (2007). The Shuttle Radar Topography Mission. Reviews of Geophysics, 45(2), RG2004. https://doi.org/10.1029/2005RG000183
8. 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
9. Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., & Brumby, S. P. (2021). Global land use/land cover with Sentinel-2 and deep learning. In 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 4704-4707). IEEE. https://doi.org/10.1109/IGARSS47720.2021.9553499
10. Kaya, C. M., & Derin, L. (2023). Parameters and methods used in flood susceptibility mapping: A review. Journal of Water and Climate Change, 14(6), 1935-1960. https://doi.org/10.2166/wcc.2023.035
11. Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15), 2171-2186. https://doi.org/10.1002/hyp.9740
12. Mosavi, A., Ozturk, P., & Chau, K.-W. (2018). Flood prediction using machine learning models: Literature review. Water, 10(11), 1536. https://doi.org/10.3390/w10111536
13. NASA JPL. (2013). *NASA Shuttle Radar Topography Mission Global 1 arc second* [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MEASURES/SRTM/SRTMGL1.003
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
15. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929. https://doi.org/10.1111/ecog.02881
16. Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2014). Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. Journal of Hydrology, 512, 332-343. https://doi.org/10.1016/j.jhydrol.2014.03.008
17. Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. Catena, 125, 91-101. https://doi.org/10.1016/j.catena.2014.10.017
18. Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F. W., Dutta, D., & Kim, S. (2017). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. Environmental Modelling & Software, 90, 201-216. https://doi.org/10.1016/j.envsoft.2017.01.006
19. Twele, A., Cao, W., Plank, S., & Martinis, S. (2016). Sentinel-1-based flood mapping: A fully automated processing chain. International Journal of Remote Sensing, 37(13), 2990-3004. https://doi.org/10.1080/01431161.2016.1192304
20. Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T., & Singh, G. (2022). Global 10 m land use land cover datasets: A comparison of Dynamic World, WorldCover and Esri Land Cover. Remote Sensing, 14(16), 4101. https://doi.org/10.3390/rs14164101

21. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

Figures

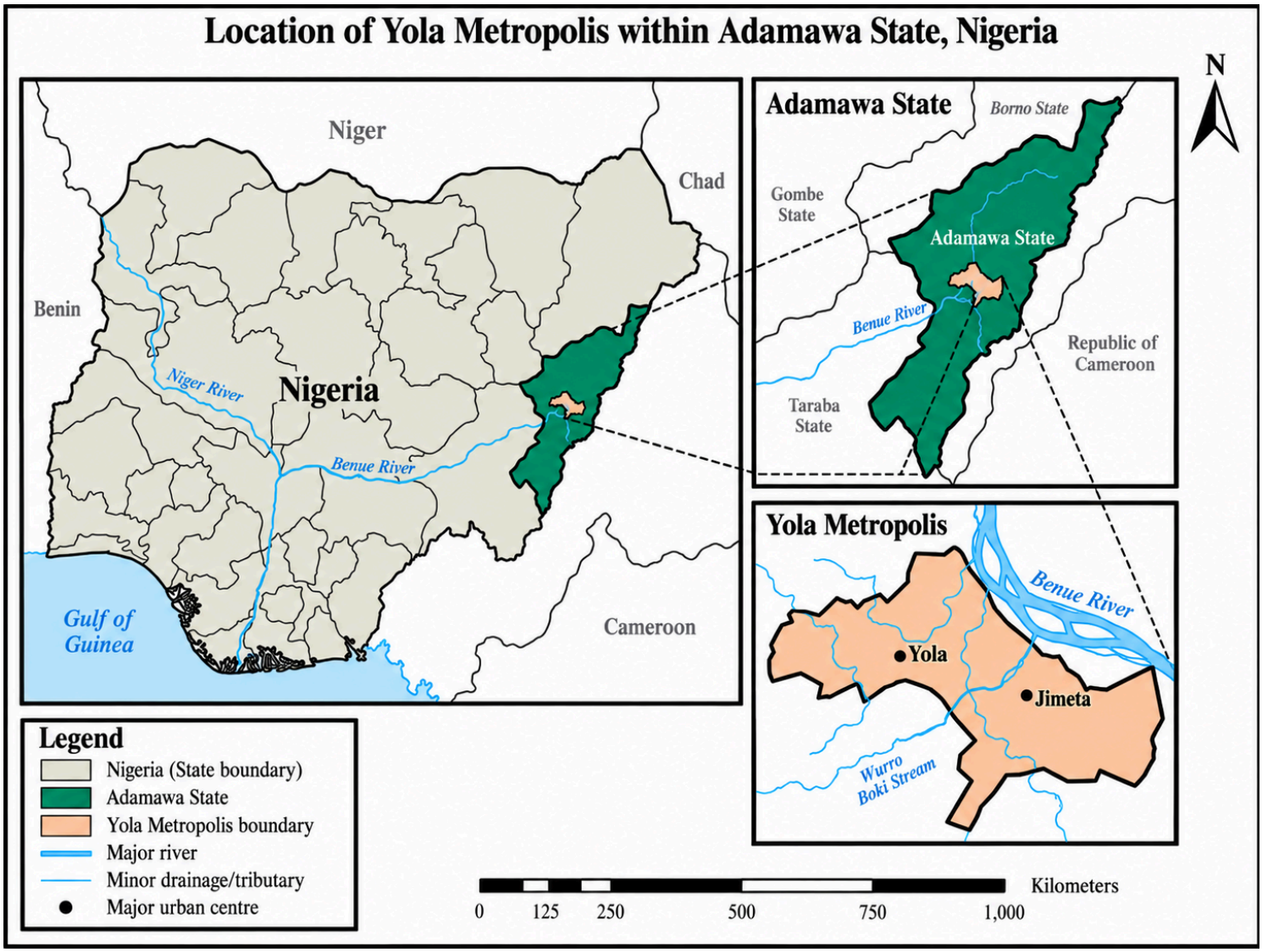


Figure 1

Location of the Yola Metropolis study area in Adamawa State, northeastern Nigeria.

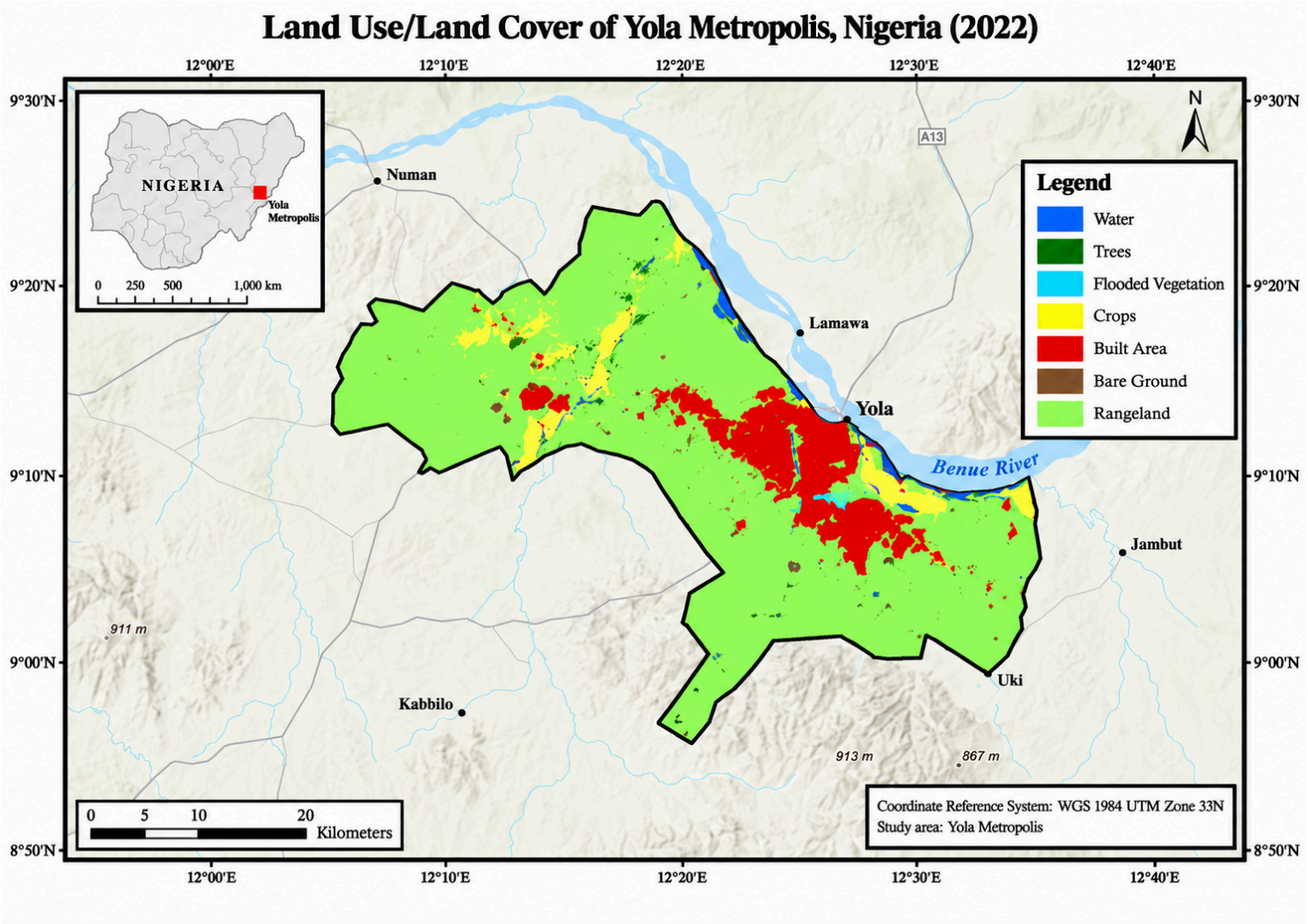


Figure 2

Spatial distribution of land-use/land-cover classes across Yola Metropolis in 2022, derived from the Sentinel-2 10 m Land Use/Land Cover Time Series available through ArcGIS Living Atlas.

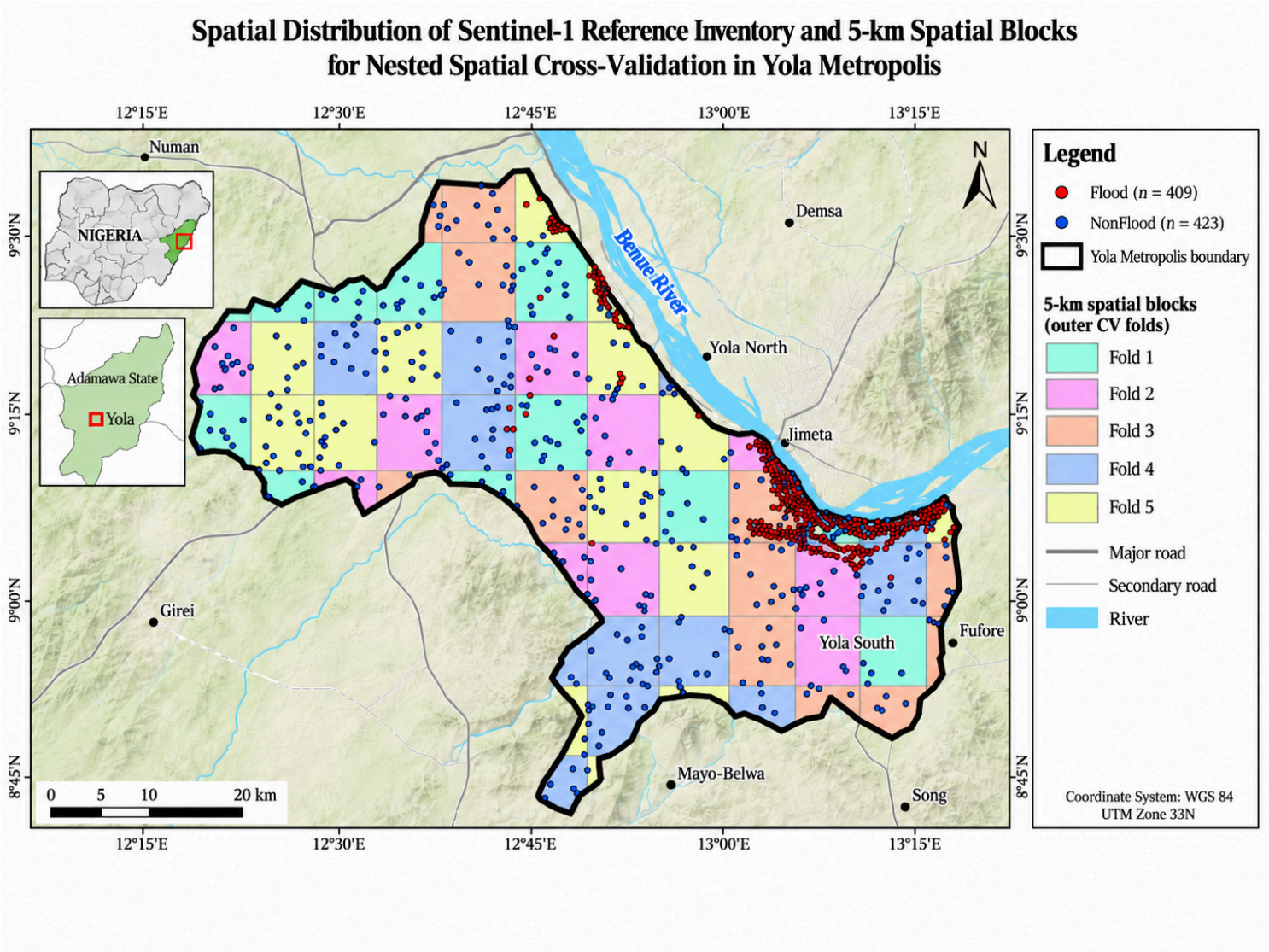


Figure 3

Spatial distribution of the 832-point Sentinel-1 reference inventory and predefined 5-km spatial blocks used for nested spatial cross-validation in Yola Metropolis. The inventory comprised 409 Flood and 423 NonFlood samples. Coloured polygons represent the five outer validation folds; block polygons were clipped to the study boundary for cartographic presentation.

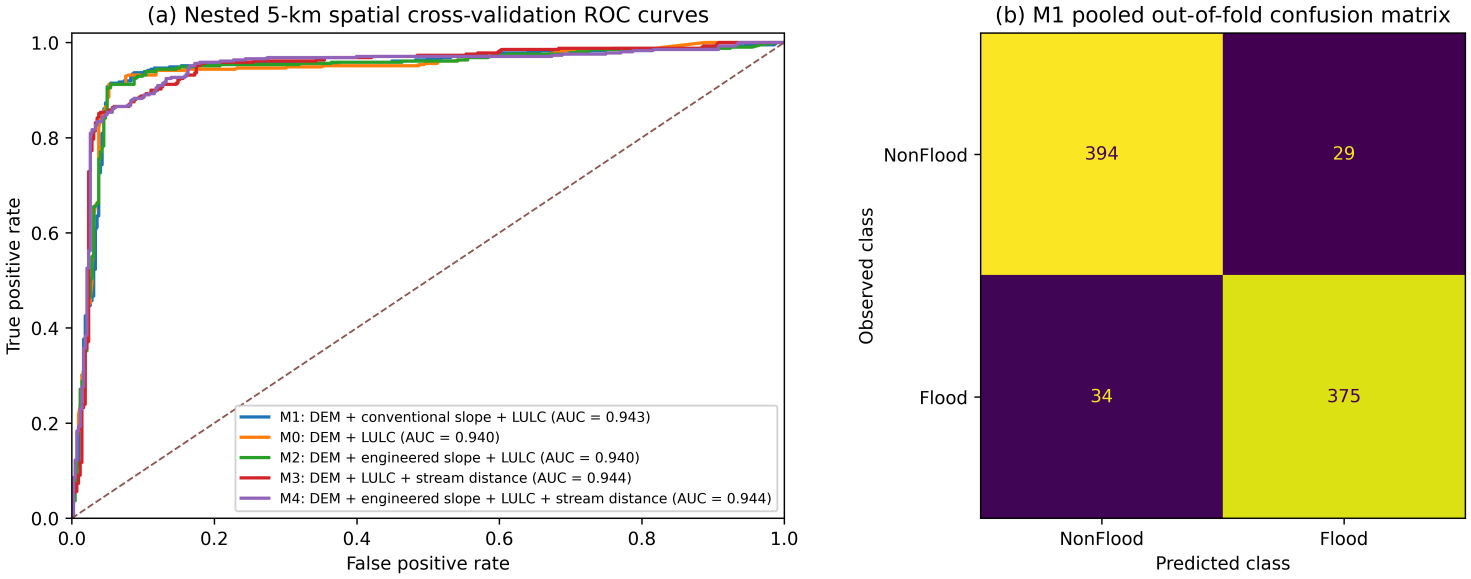


Figure 4

Nested 5-km spatial cross-validation results for the Random Forest models: (a) ROC curves for the five predictor configurations and (b) pooled out-of-fold confusion matrix for the selected M1 model (TN = 394, FP = 29, FN = 34, TP = 375)

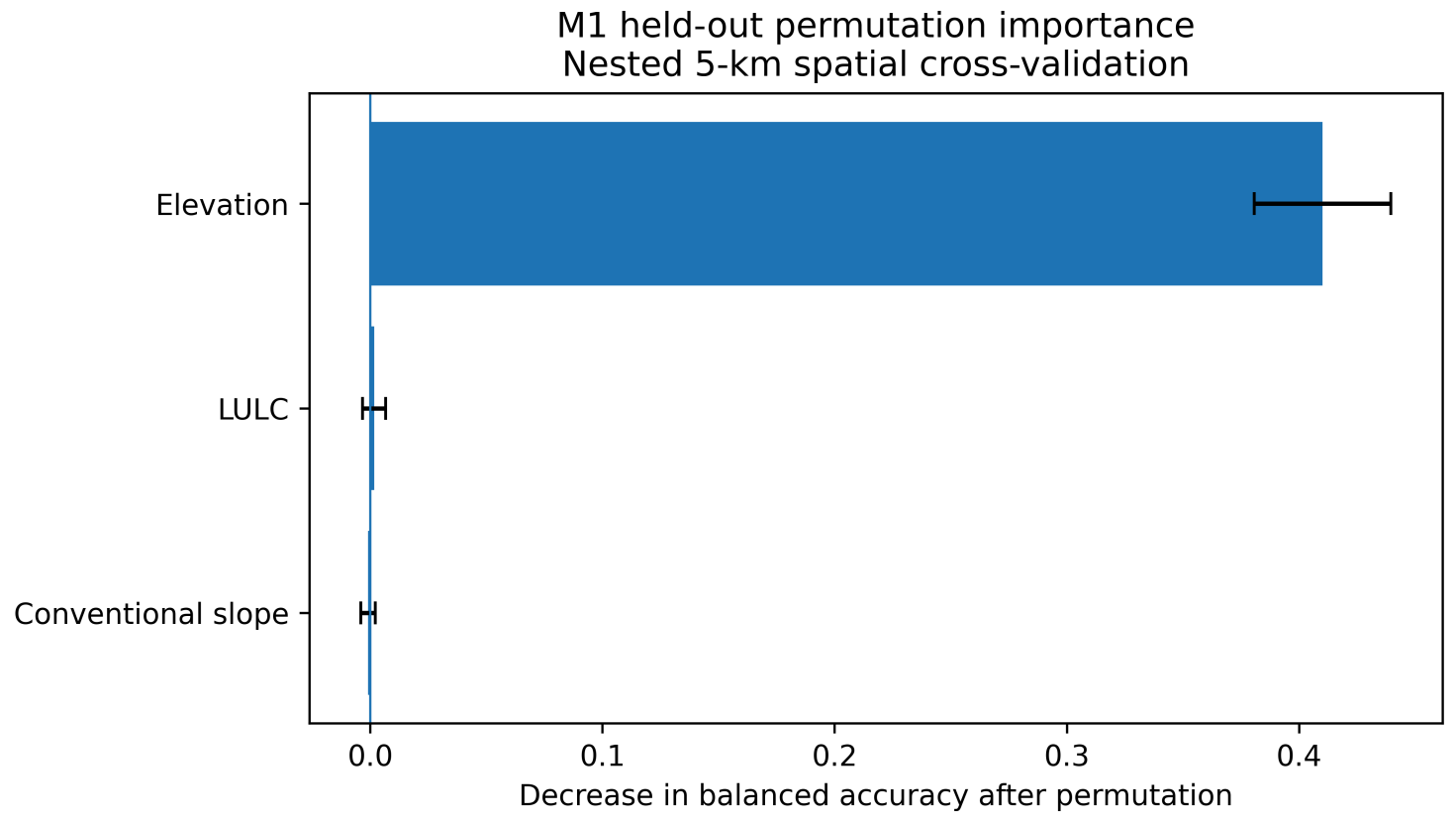


Figure 5

Held-out permutation importance of the selected M1 predictors across the five outer spatial folds. Bars represent the mean decrease in balanced accuracy after permutation; error bars show the standard deviation between folds.

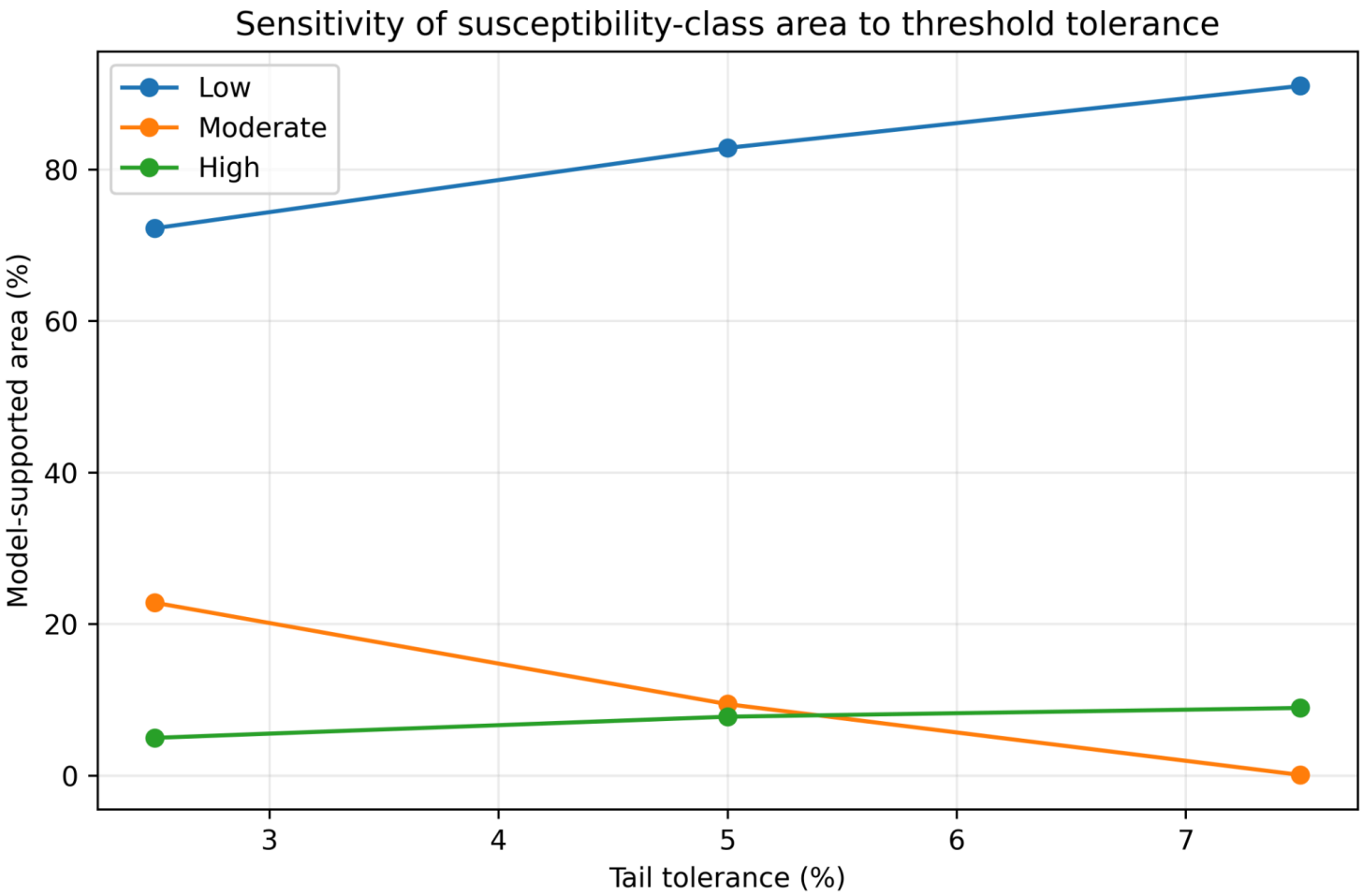


Figure 6

Sensitivity of the proportion of model-supported area assigned to Low, Moderate, and High flood-susceptibility classes under alternative validation-derived tail-tolerance criteria. The 10% tolerance is omitted because its lower threshold exceeded the upper threshold and therefore did not produce a mutually exclusive three-class solution.

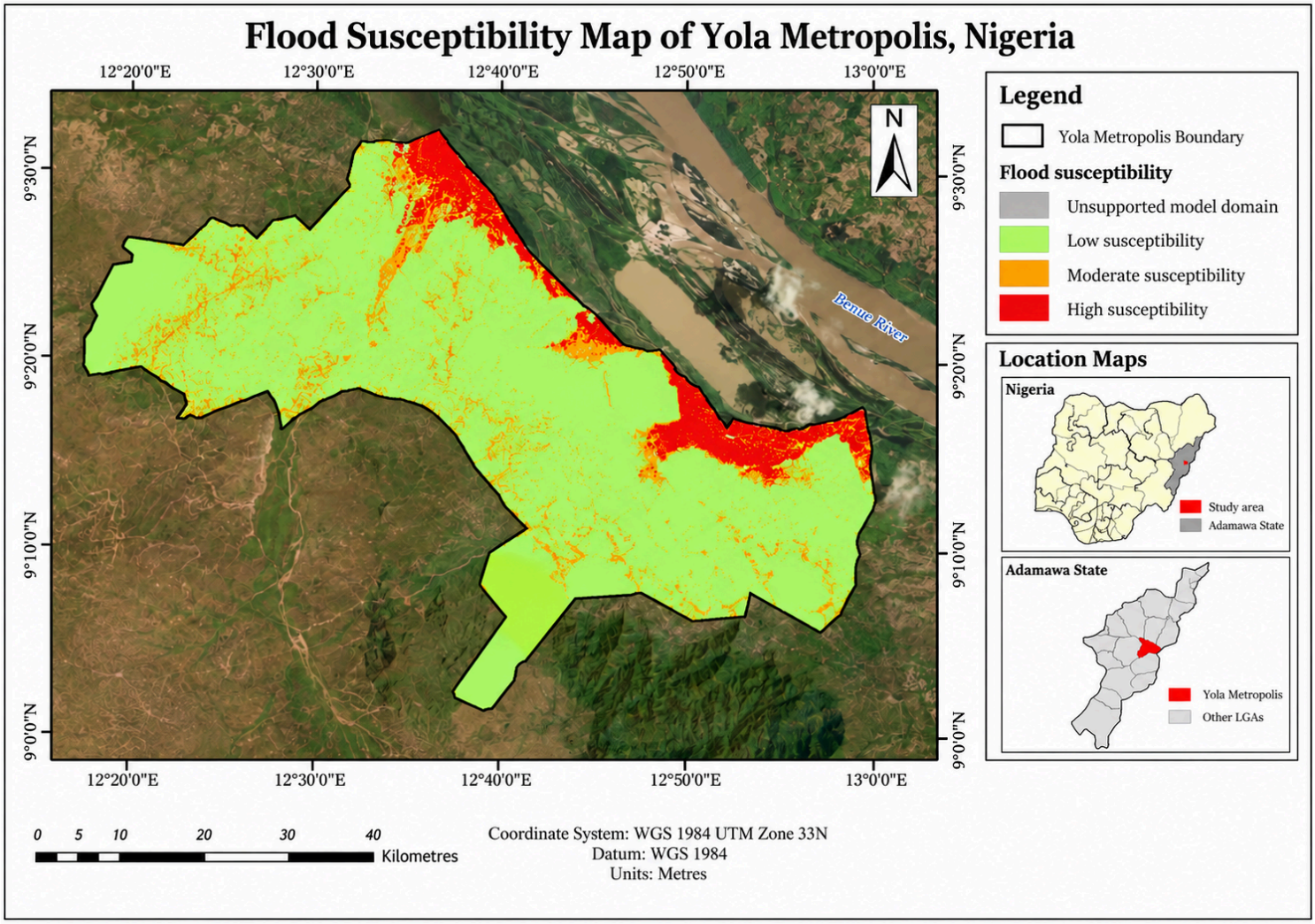


Figure 7

Random Forest flood susceptibility classes for the Yola Metropolis study area. Thresholds were derived from nested spatial out-of-fold scores: Low < 0.1869, Moderate = 0.1869-<0.7331, and High >= 0.7331.