

An Interpretable Machine Learning Framework for Rainfall-Related Slope Instability Susceptibility Assessment in Aba Prefecture, China: Sampling Strategy, Spatial Validation, and Susceptibility Score Mapping

Liping Zheng

zhengliping@scujj.edu.cn

Sichuan University Jinjiang College

Yanwen Yang

Sichuan University Jinjiang College

Shuanglin Deng

Sichuan University Jinjiang College

Kaiwen Xiao

Sichuan University Jinjiang College

Moran Qiu

Lanzhou Jiaotong University

Feng Ning

Sichuan University Jinjiang College

Liangjun Zhao

Sichuan University of Science and Engineering

Article

Keywords: landslide susceptibility, sampling bias, explainable machine learning, spatial cross-validation, SHAP, Aba Prefecture

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11134982/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

An Interpretable Machine Learning Framework for Rainfall-Related Slope Instability Susceptibility Assessment in Aba Prefecture, China: Sampling Strategy, Spatial Validation, and Susceptibility Score Mapping

Liping Zheng^{1*}, Yanwen Yang^{1†}, Shuanglin Deng^{1†},
Kaiwen Xiao¹, Moran Qiu², Feng Ning¹, Liangjun Zhao³

¹School of Computer Science, Sichuan University Jinjiang College,
Meishan, 620860, Sichuan, China.

²School of Traffic and Transportation, Lanzhou Jiaotong University,
Lanzhou, 730070, Gansu, China.

³Sichuan University of Science & Engineering, Yibin, 644000, Sichuan,
China.

*Corresponding author. E-mail: zhengliping@scujj.edu.cn;

[†]These authors contributed equally to this work.

Abstract

This study addresses sampling bias and exposure-induced shortcut variables in landslide susceptibility modelling for rainfall-related slope failures. A data-driven framework integrates dual-source mechanism-contrast sample construction with two-stage susceptibility-exposure separation, evaluated on 10,644 samples (5,322 instability records and 5,322 constructed contrasts) from Aba Prefecture, south-western China. The positive inventory comprises debris flows (41.4%), landslides (39.0%), and unstable slopes (19.6%), and 66.1% are compound earthquake-rainfall events. The dual-source contrast strategy gives a random-forest distinguishability area under the receiver operating characteristic curve (AUC-ROC) of 0.965, indicating separable but not trivially separable classes, and distance to the nearest active fault is included as a formal feature for both classes. Among seventeen models, the AUC-weighted voting ensemble attains the highest F1 (0.923), and the highest AUC-ROC of 0.973 is achieved by default LightGBM,

tuned LightGBM, and tuned XGBoost. A paired random-versus-spatial cross-validation comparison (5-fold 2D KMeans blocks) shows that spatial blocking reduces AUC by only 1.2–1.6% and leaves F1 essentially unchanged, while a coordinate ablation changes AUC by at most 0.004 when longitude and latitude are removed. Group-level Shapley additive explanations (SHAP) identify terrain (59.3%), rainfall (19.1%), and vegetation (6.4%) as the dominant feature families, with coordinates contributing 5.8% and fault distance 3.3%; these are model-learned associations rather than verified causal mechanisms. Classifying out-of-fold susceptibility probabilities into five relative classes highlights the southeastern study area as the highest-scoring region. Priority weighting is restricted to records carrying documented exposure data, which are unavailable for spatially sampled locations.

Keywords: landslide susceptibility, sampling bias, explainable machine learning, spatial cross-validation, SHAP, Aba Prefecture

1 Introduction

Landslides cause significant casualties and economic losses in mountainous areas. Rainfall is a leading trigger in southwestern China (Zhang et al. 2015; Zhou and Tang 2014). Aba Prefecture, on the eastern margin of the Qinghai–Tibet Plateau, is highly prone to slope instability because of its complex terrain and intense summer monsoon rainfall.

Traditional susceptibility maps rely on heuristic methods such as the analytical hierarchy process (AHP) (Pourghasemi et al. 2012) and logistic regression (Dai and Lee 2002). Machine learning approaches, including random forests (RF) (He et al. 2021), support vector machines (SVM) (Yao et al. 2008), gradient boosting (Sahin 2022), and comparative reviews (Pradhan 2013; Pourghasemi and Rossi 2017; Merghadi et al. 2020; Reichenbach et al. 2018), have shown superior performance. Deep learning and ensemble neural-network approaches, such as convolutional neural networks (Yi et al. 2020) and neural-network ensembles (Bragagnolo et al. 2020), have also been applied, and recent research has increasingly emphasised interpretable machine learning and spatial prediction (Sun et al. 2023; Chen 2025; Lundberg et al. 2020; Roscher et al. 2020). However, two challenges persist: (1) landslide inventories contain only positive samples, and the construction of contrast samples strongly affects performance (Rabby et al. 2023; Liu 2023; Zhu et al. 2019; Gu et al. 2024); (2) exposure features (population, property) can become shortcut variables that inflate accuracy while obscuring physical mechanisms.

This study investigates how contrast sample composition influences landslide susceptibility modelling under spatially heterogeneous conditions, and proposes an interpretable framework that separates susceptibility prediction from exposure assessment. The target is *rainfall-related* instability: 98.5% of the inventory lists rainfall among its documented triggering factors, encompassing both rainfall-only and compound earthquake–rainfall events. The main contributions are: (1) a dual-source

mechanism-contrast sample construction strategy with random-forest distinguishability analysis (AUC 0.965), used as an internal quality check on contrast-sample difficulty; (2) a two-stage susceptibility–exposure separation that prevents exposure features from acting as shortcut variables; (3) multi-scale rainfall features incorporating a rainfall anomaly index, together with a paired random-versus-spatial cross-validation protocol (2D KMeans blocking with per-fold dispersion (Yoo and Koo 2024; Roberts et al. 2017; Ploton et al. 2020)) for evaluating generalization under geographic heterogeneity; and (4) a group-level SHAP interpretation that summarises feature contributions by family while acknowledging that correlated predictors can redistribute attribution (Lundberg and Lee 2017; Lundberg et al. 2020).

2 Study Area and Data

Aba Prefecture in northwestern Sichuan spans 99.8°E–104.6°E and 30.5°N–34.4°N, with elevations from below 1,000 m to over 6,000 m. Intense summer monsoon rainfall creates favourable conditions for slope instability, and the region was strongly affected by the 2008 Wenchuan earthquake (Qi et al. 2011; Xu et al. 2014; Yin et al. 2016).

2.1 Landslide Inventory and Definition of Rainfall-Related Instability

The primary data source is the regional geological hazard census maintained by the Sichuan Geological Survey. Following filtering, outlier removal, and spatial deduplication, the inventory used in this study contains 5,322 records. Each record carries a *triggering factor* field that documents the causal process (for example, rainfall, earthquake, human activity, or combinations thereof).

The cleaning keeps the three rainfall-dominated instability types recorded in the census, discards records with missing triggering-factor or threat-target fields, removes outliers in elevation, aspect, rainfall, topographic position and river distance, and merges coincident coordinates, giving 5,322 records. Of these, 5,240 (98.5%) list rainfall as a contributing trigger (31.6% rainfall alone, 66.1% compound earthquake–rainfall events), and the remaining 82 records belong to the same rainfall-dominated instability types. The inventory includes three instability types: debris flows (41.4%), landslides (39.0%), and unstable slopes (19.6%). Landslides and debris flows are documented failure events; unstable slopes are pre-failure slope bodies identified during the census and have no discrete failure date. The inventory is therefore treated as a rainfall-dominated instability inventory rather than a strictly rainfall-triggered landslide inventory, and no claim is made that all records represent slope failure at a single point in time.

2.2 Temporal Coverage and Matching

Each record carries a *discovery date*, i.e. the date on which the feature was recorded during the census, used here as a proxy for the occurrence date. Dates span 1905–2020, with 92.2% of records documented between 2005 and 2020; 89 records predate 1984 and 187 fall in 1984–1999. Two consequences follow.

First, the rainfall predictors are not event-time observations. They are interpolated from the meteorological station network and represent conditions in a window surrounding each record; for records without a reliable day-level date (notably unstable slopes) the window is anchored to the record date. Rainfall values should therefore be read as an environmental covariate at the documented time rather than as the measured triggering rainfall.

Second, the vegetation indices are not strictly contemporaneous with the records. Vegetation was supplied with the census together with a source-sensor field, which for the 5,322 records comprises Landsat 7 ETM+ (3,736 records), Landsat 8 OLI (1,136), and Landsat 5 TM (423), with 27 records unassigned. The sensor distribution is not aligned with record decades, and 89 records predate the Landsat 5 launch (1984), so no contemporaneous Landsat observation exists for them. Vegetation and land-cover layers are therefore treated as quasi-static covariates rather than event-time measurements, and this mismatch is revisited in Section 5.5.

The provenance, epoch, and spatial resolution of each data source, together with the temporal-matching strategy adopted, are summarised in Online Resource 1 (Table S1).

2.3 Environmental Predictors

Terrain attributes are derived from the 30m Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) (Farr et al. 2007). Slope angle is computed with the Horn (1981) third-order finite difference on a 3×3 window and is expressed in degrees; plan and profile curvature are computed from the second derivatives of the same window and are reported in units of 0.01 m^{-1} (relative to the cell size); the topographic position index (TPI) is the difference between each cell and the mean of its surrounding 3×3 neighbourhood, in metres. All derivatives use the geographic cell size converted to metres at the study-area mid-latitude (approximately 32.5°N).

Rainfall features (3-day, 7-day, and 30-day cumulative rainfall and an antecedent precipitation index, API) are interpolated from meteorological station observations using inverse-distance weighting (IDW) with a power parameter of 2.0. The API is a 30-day antecedent precipitation index supplied with the census, formed as a lag-weighted accumulation of antecedent daily rainfall, $API_t = \sum_{k=1}^K k^{-1} r_{t-k}$, in which the weight of lag k is $1/k$; the derived *rain_anomaly* ratio is defined as 3-day cumulative rainfall divided by the API and is clipped at 10.0, so a value of 2 indicates that the recent 3-day total is twice the lag-weighted antecedent accumulation. Vegetation indices (normalized difference vegetation index, NDVI; normalized difference water index, NDWI) come from the Landsat composites described above. Land use was classified from ESA WorldCover into seven categories (grassland, trees, bare/sparse vegetation, built-up, cropland, permanent water bodies, moss and lichen). One-hot encoding is applied with *bare/sparse vegetation* as the reference level, so six indicator variables enter the model; the moss-and-lichen class is retained as one of those indicators but accounts for very few records. Fault distance is computed from the GEM Global Active Faults Database (CC-BY-4.0) as the planar distance to the nearest mapped active fault trace.

Fig. 1 shows the spatial distribution of key environmental features across the study area together with the locations of the instability records.

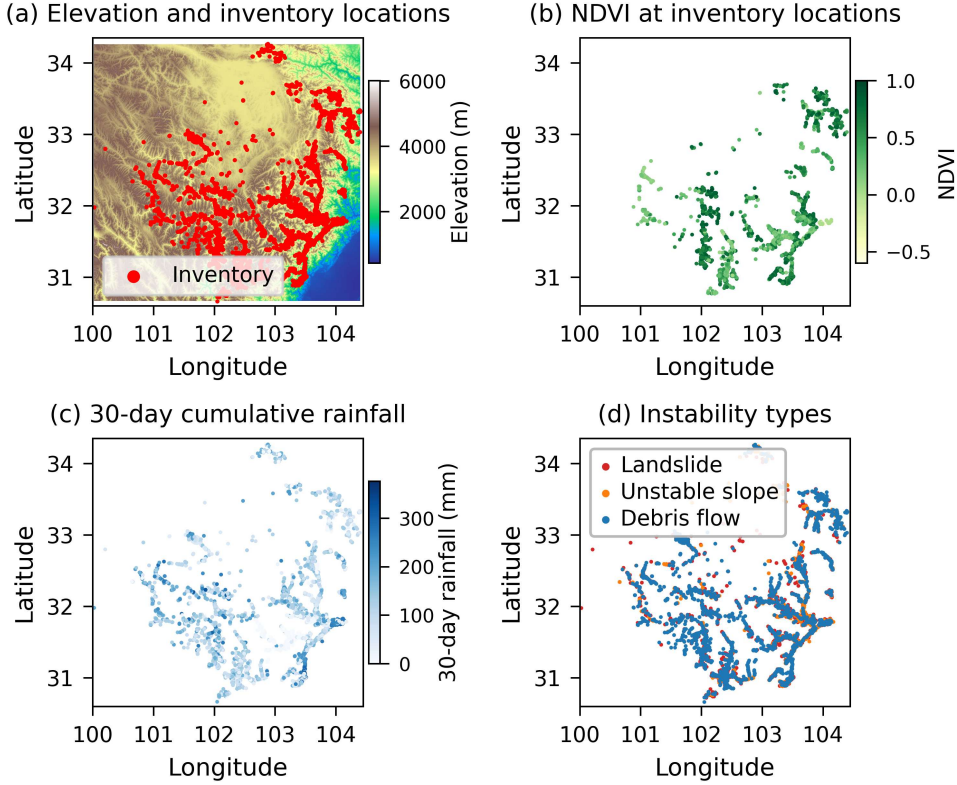


Fig. 1 Study area and inventory: (a) SRTM elevation with inventory locations (red); (b) recorded NDVI, plotted on the physical NDVI range (2,486 of the 5,322 inventory records have no NDVI measurement and are omitted from this panel; they are imputed before modelling); (c) 30-day cumulative rainfall; and (d) instability types (landslide $n = 2,075$, unstable slope $n = 1,043$, debris flow $n = 2,204$). All panels use geographic coordinates

Data preprocessing includes hazard filtering, outlier removal, spatial deduplication, and missing-value completion. Missing values are imputed using random-forest regression for NDVI (cross-validated $R^2 = 0.75$), linear regression on NDVI for NDWI, and Multiple Imputation by Chained Equations (MICE) for rainfall features.

3 Methodology

3.1 Overall Framework

The framework follows a two-stage paradigm (Fig. 2). Stage 1 trains classification models using 32 susceptibility features. Stage 2 computes a composite exposure index

from three aggregated exposure variables (potentially affected population, households, and threatened property) at each documented hazard location. The final exposure-adjusted priority score combines susceptibility probability and exposure. Without this separation, exposure features act as shortcut variables that inflate accuracy without capturing physical mechanisms.

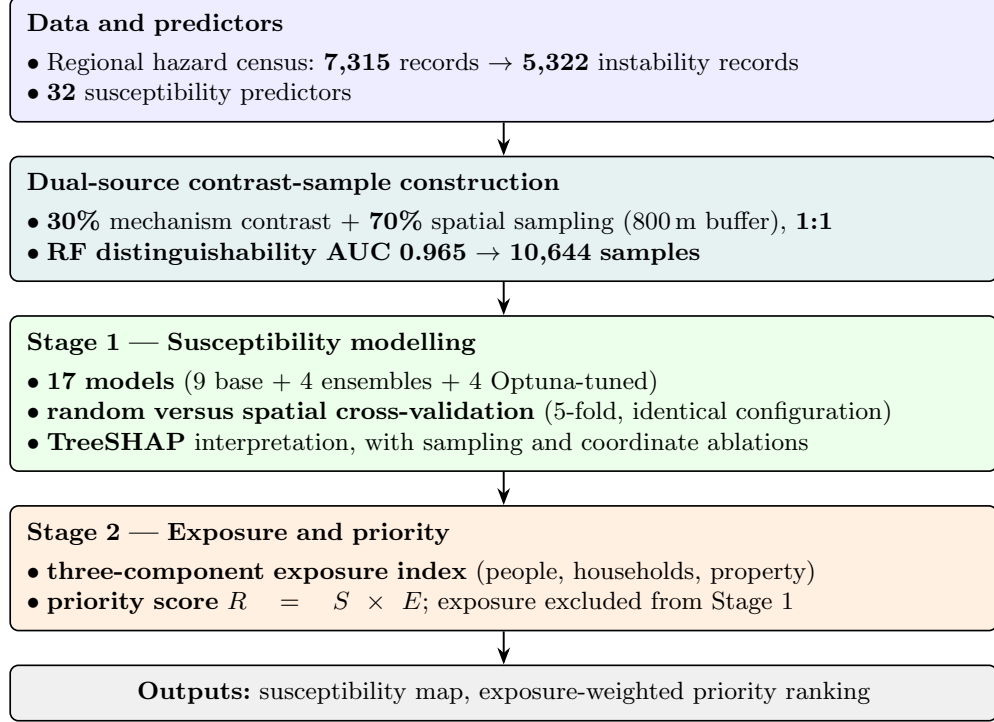


Fig. 2 Overall framework: inventory and predictor assembly, dual-source contrast-sample construction with a random-forest distinguishability check, Stage-1 susceptibility modelling with paired random and spatial cross-validation and SHAP interpretation, and Stage-2 exposure-based priority mapping

3.2 Feature Engineering

This section constructs the features that capture the physical drivers of rainfall-related slope instability. Online Resource 1 (Table S2) lists all 32 susceptibility features in seven categories: terrain, fault, rainfall, vegetation, coordinate, derived, and land-use variables. Two derived features are newly proposed: *rain_anomaly* (the ratio of 3-day cumulative rainfall to the antecedent precipitation index, clipped at 10.0) and *aspect_rain_abs* ($|\cos(\text{aspect})| \times \text{rain}_{3d}$, capturing the aspect-rainfall interaction). *slope_angle*, *plan_curvature*, and *profile_curvature* are derived from the SRTM 30 m DEM using the standard [Horn \(1981\)](#) formulation rather than being newly proposed. In addition, *fault_distance_km* (distance to the nearest active fault from the GEM

Global Active Faults Database, CC-BY-4.0) is included as a formal feature. Unlike the earlier scheme in which it was excluded and evaluated only post-hoc, fault distance is now computed for both positive samples and generated contrast samples via spatial queries to the GEM fault traces, so no information leakage arises. Ratio-type rainfall features (*rain_ratio_3d_30d*, *rain_ratio_7d_30d*, *rain_intensity*) are clipped at 1.0. Coordinate features (longitude, latitude) capture regional spatial heterogeneity that may reflect unmeasured environmental variations across the study area; they also risk absorbing spatial autocorrelation, which is why spatial cross-validation (CV) is examined separately in Section 3.5.

3.3 Dual-Source Mechanism-Contrast Sample Construction

The features described above characterise positive samples only. Landslide inventories lack contrast labels. A dual-source hybrid strategy is proposed, building on recent work on absence-sample design (Rabby et al. 2023; Liu 2023; Zhu et al. 2019; Gu et al. 2024). Source 1 (30%) uses mechanism-contrast samples drawn from other instability types recorded in the same census, namely rockfalls and ground collapses. These represent failure processes whose mechanics differ from rainfall-driven sliding and flow, and they are used to make the contrast set harder rather than to assert that the locations are stable; importantly, many of these records also list rainfall among their triggering factors, so this source should be understood as a *mechanism* contrast and not as a rainfall/no-rainfall contrast. Source 2 (70%) uses BallTree spatial random sampling outside an 800 m buffer around positives. Feature values for these samples are assigned as follows. Continuous environmental features (elevation, TPI, distance to rivers, 3/7/30-day rainfall, API, NDVI, and NDWI) are inherited from the three nearest inventory records by inverse-distance weighting. It is noted explicitly that land use is *not* interpolated: because it is categorical, one of the three nearest records' land-use classes is drawn at random, so only valid classes occur and no meaningless fractional class codes are created. Slope angle and curvature are not inherited but are computed directly from the SRTM DEM at the sampled location, and fault distance is likewise computed directly from the GEM fault traces, so that positive and contrast samples are derived from the same underlying geospatial layers. Aspect is angular and is inherited by *circular* weighted averaging, $\text{atan2}(\sum w_i \sin \theta_i, \sum w_i \cos \theta_i)$, because the arithmetic mean of bearings near the $0^\circ/360^\circ$ discontinuity would be badly biased. Aggregated exposure fields are set to zero for sampled locations; this prevents fabricated exposure values from entering the risk stage, but it also means that the risk score is undefined for those cells (Section 5.5). In the final configuration, random-forest distinguishability analysis yields $\text{AUC} = 0.965$, indicating that the two classes are separable but not trivially so. This diagnostic is treated as an internal quality check rather than as a universal acceptance threshold. The final dataset contains 5,322 positive and 5,322 contrast samples (1:1 ratio). The 800 m buffer is comparable to typical landslide runout distances in the region; a sensitivity test comparing 0.8 km and 3.0 km buffers changed the RF distinguishability AUC by at most 0.001 and the spatial-CV AUC by at most 0.002. Because the census contains only 1,693 rockfall and ground-collapse records (1,622 of which survive the same cleaning and imputation applied to the positives), a 1:1 contrast set can contain at most 1,622 such records: with 5,322

contrast samples the mechanism-contrast share is therefore bounded by 30.5%, and this ceiling defines the range explored in Section 4.5. No resampling is applied because the classes are balanced.

3.4 Model Training

Training proceeds in two phases. **Phase 1** constructs nine base models spanning commonly used linear, kernel-based, bagging, and boosting paradigms in landslide susceptibility studies: logistic regression (LR), decision tree (DT), SVM (radial basis function kernel), k-nearest neighbours (KNN), RF, gradient boosting (GBoost), CatBoost (Prokhorenkova et al. 2018), LightGBM (Ke et al. 2017), and XGBoost (Chen and Guestrin 2016), all with Platt scaling probability calibration. Four ensemble strategies are then developed: AUC-weighted soft voting (temperature 0.05), simple averaging, stacking, and blending.

For stacking, an XGBoost meta-learner (`max_depth=3`, `n_estimators=100`) is trained on out-of-fold (OOF) probabilities from 5-fold StratifiedKFold concatenated with 32 standardized original features, producing an $N \times 41$ meta-feature matrix (9 base-model probabilities + 32 features). To prevent data leakage, Platt scaling and hyperparameter search are performed strictly within each training fold: the training fold is further split for calibration and hyperparameter optimization, and the validation fold remains untouched until final prediction. Blending uses the same meta-learner but trained on an 80/20 split instead of OOF.

Phase 2 investigates whether fine-grained hyperparameter optimization further improves performance. Based on the Phase 1 observation that tree-based models dominate the top ranks, the four strongest tree learners (RF, XGBoost, LightGBM, CatBoost) are selected for Optuna optimization with the tree-structured Parzen estimator (TPE; 30 trials per model, 3-fold CV targeting F1). All other models retain default hyperparameters. This two-phase design separates the question of *which model family* is most suitable from the question of *whether hyperparameter tuning* adds value. Classification thresholds are not fixed a priori: for each model the decision threshold is selected from 100 candidates between 0.05 and 0.95 by maximising $0.6F1 + 0.4\text{specificity}$ on the validation partition, and is then applied unchanged to the test partition. The random-versus-spatial comparison of Section 3.5 instead uses a fixed threshold of 0.5, so that the comparison isolates the effect of spatial blocking rather than of threshold choice.

3.5 Spatial Cross-Validation

Random CV may overestimate performance due to spatial autocorrelation (Goetz et al. 2015; Yoo and Koo 2024; Roberts et al. 2017; Ploton et al. 2020): nearby samples share environmental conditions, allowing models to exploit spatial proximity, and spatial resampling is therefore recommended whenever predictions are transferred to unseen locations (Roberts et al. 2017; Ploton et al. 2020; Meyer and Pebesma 2021). To quantify rather than merely assert this effect, both a 5-fold random StratifiedKFold CV and a 5-fold spatial CV are computed on the identical training set using identical model configurations, so that any difference isolates the effect of spatial blocking. The spatial

CV uses 2D KMeans blocks, which partition the study area into spatially contiguous clusters based on both longitude and latitude and force prediction on unseen regions; this is more stringent than 1D longitude binning. Four tree-based models (XGBoost, LightGBM, CatBoost, and Random Forest) and the Stacking ensemble are evaluated. Per-fold sample counts, class composition, threshold, precision, recall, and F1 are reported alongside the aggregate scores. For every mean score, the standard deviation across the five folds is also reported. It is stressed that this dispersion is a *descriptive* summary of the spread among the five folds: it does not represent independent spatial partitions, and because the training sets of different folds overlap it cannot capture the full uncertainty associated with the choice of spatial blocking. It should be read as a stability indicator, not as a significance test.

3.6 SHAP Interpretability

After verifying spatial generalization, the next step is to interpret what the model has learned. SHAP (Lundberg and Lee 2017; Lundberg et al. 2020) provides model-agnostic interpretation grounded in cooperative game theory, and interpretable machine learning has been increasingly adopted in landslide susceptibility research (Sun et al. 2023; Roscher et al. 2020). The Shapley value ϕ_i of feature i measures its average marginal contribution across all possible feature coalitions:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

where N is the full feature set, S is a subset excluding feature i , and $f(S)$ is the model prediction with only features in S . A positive ϕ_i indicates that feature i increases predicted susceptibility.

TreeSHAP is applied to the XGBoost tree model. Global feature importance is computed by averaging absolute SHAP values across all samples: $I_i = \frac{1}{n} \sum_{j=1}^n |\phi_i^{(j)}|$. Because correlated features can redistribute attribution among themselves, group-level aggregation is additionally reported to summarise the contribution of feature families (terrain, rainfall, vegetation, coordinate, derived, land use, and fault distance). The resulting rankings should be read as associations learned by the model rather than as causal physical mechanisms.

3.7 Susceptibility Classification

Susceptibility probability alone does not fully characterize priority. Two areas with identical susceptibility but different population density require different management priorities. Susceptibility probabilities are classified into five levels using quantile breaks. Because the whole dataset (including generated contrast samples) enters the classification, the resulting map describes the distribution of model scores over the sampled locations rather than an independent validation over the full administrative area. The priority score combines susceptibility S and a composite exposure index $E = 0.4\hat{P} + 0.3\hat{H} + 0.3\hat{V}$, where $\hat{P}$, $\hat{H}$, $\hat{V}$ are max-normalised population, households, and property. This index is a single composite indicator built from three location-based

aggregated exposure components. The weights reflect local disaster-management priorities, with population receiving the highest weight (0.4) because life safety is the primary concern. A sensitivity analysis perturbing each weight by ± 0.1 is reported in Section 5.1. Aggregated exposure—persons, households, and threatened property at each hazard site—is recorded only for the inventory and mechanism-contrast records, not for spatially sampled locations; it is treated as location-based aggregated statistics rather than individual-level human data. The priority analysis is therefore restricted to records with documented exposure, and $R = S \cdot E$ is reported only for that subset.

4 Results

4.1 Model Performance

Table 1 summarizes test-set performance under an 80/20 stratified split, organized in three tiers: base models with default hyperparameters, ensemble strategies, and Optuna-tuned variants of the four strongest tree learners.

Base models. Tree-based ensembles occupy the top ranks: XGBoost (F1 0.920), Random Forest (0.919), LightGBM (0.910), GBoost (0.899), and CatBoost (0.891) all outperform the instance-based and linear alternatives. This indicates that nonlinear ensemble methods capture the multivariate interactions among terrain, rainfall, and vegetation features more effectively than the other model families in this dataset. The decision tree (0.873) and KNN (0.847) are intermediate, while the radial-basis-function SVM (F1 0.644) and logistic regression (0.704) show the weakest performance here. These results characterise the specific configurations evaluated and should not be generalised to the model families as a whole without further tuning.

Ensemble models. Stacking attains the highest AUC-PR among the ensemble strategies (0.967), consistent with out-of-fold aggregation smoothing individual model biases, while AUC-weighted voting attains the best F1 (0.923) and matches stacking on AUC-ROC (0.972 for both). Voting also edges simple averaging (F1 0.923 vs. 0.921; AUC-ROC 0.972 vs. 0.965), indicating that performance-weighted aggregation exploits complementary model strengths for ranking. Blending (0.901) is the weakest ensemble, which is attributed to its meta-learner being fitted on a smaller hold-out partition than the cross-validated stacking design.

Tuned models. Optuna tuning again yields model-dependent results. The tuned CatBoost improves from F1 0.891 to 0.921 and the tuned LightGBM from 0.910 to 0.920, whereas the tuned XGBoost (0.920 to 0.913) and Random Forest (0.919 to 0.873) degrade. The ranking-based metrics are essentially unaffected: AUC-ROC changes by at most 0.002 and AUC-PR by at most 0.003 for every model (Table 1), so the F1 shifts mainly reflect a re-balancing of the precision–recall trade-off rather than improved discrimination. Within the search budget considered here, tuning therefore provides no consistent or practically meaningful improvement over the default configurations, and none of the tuned variants displaces the default XGBoost, the default LightGBM, or the voting ensemble at the top of the table. The strongest single models remain the default XGBoost (F1 0.920) and the tuned LightGBM (0.920), which are tied at the reported precision; the best overall F1 is obtained by the AUC-weighted voting ensemble (0.923), and the best ranking performance by the default LightGBM

Table 1 Performance comparison of all 17 models on the test set (32 features). The table is organized in three tiers: nine base models with default hyperparameters, four ensemble strategies^a, and Optuna-tuned variants of the four strongest tree-based learners^b. Bold marks the best value in each metric (ties included). The AUC-weighted voting ensemble attains the highest F1 (0.923), the default LightGBM the highest AUC-ROC (0.973), and the tuned XGBoost the highest area under the precision–recall curve (AUC-PR, 0.969). These margins lie at the third–fourth decimal place, within the variability introduced by the contrast-sample construction and the single stratified split; the ordering should therefore not be read as evidence of a meaningful performance difference among the leading models. Thresholds are selected on the validation partition as described in Section 3.4

Rank	Model	F1	AUC-ROC	AUC-PR	Precision	Recall
<i>(a) Base Models (default hyperparameters)</i>						
1	XGBoost	0.920	0.971	0.966	0.892	0.950
2	Random Forest	0.919	0.969	0.961	0.887	0.954
3	LightGBM	0.910	0.973	0.969	0.905	0.915
4	GBoost	0.899	0.971	0.967	0.916	0.883
5	CatBoost	0.891	0.972	0.968	0.920	0.865
6	Decision Tree	0.873	0.955	0.934	0.904	0.843
7	KNN	0.847	0.909	0.886	0.804	0.895
8	Logistic Regression	0.704	0.828	0.815	0.795	0.633
9	SVM	0.644	0.787	0.779	0.776	0.551
<i>(b) Ensemble Models^a</i>						
1	Voting	0.923	0.972	0.967	0.897	0.951
2	Ensemble_Avg	0.921	0.965	0.954	0.900	0.944
3	Stacking	0.920	0.972	0.967	0.897	0.945
4	Blending	0.901	0.956	0.925	0.913	0.890
<i>(c) Optuna-Tuned Models^b</i>						
1	CatBoost	0.921	0.972	0.968	0.890	0.955
2	LightGBM	0.920	0.973	0.968	0.892	0.949
3	XGBoost	0.913	0.973	0.969	0.899	0.928
4	Random Forest	0.873	0.968	0.961	0.927	0.824

^a AUC-weighted soft voting, simple averaging, stacking, and blending.

^b Tuned with Optuna (tree-structured Parzen estimator, 30 trials, 3-fold cross-validation on F1).

(AUC-ROC 0.973). The absence of a consistent gain across the four learners indicates that, within the search budget and feature space examined here, tuning gives limited and uneven benefit, and that 3-fold cross-validated F1 does not always translate into held-out improvement—an observation consistent with the spatial cross-validation results in Section 4.2. Because the budget (30 trials per model) and search space are finite, this should not be read as evidence that tuning is unnecessary in general. Both default and tuned variants are therefore reported, rather than treating tuning as universally beneficial.

Overall, the performance gap among the leading tree-based models is narrow—differences appear only at the third and fourth decimal place (Table 1), and the tuned

variants are numerically very close to their default counterparts—indicating that the feature set and evaluation design, rather than the specific model choice, drive most of the predictive accuracy; the narrow margins should not be over-interpreted as a stable model ranking.

4.2 Random versus Spatial Cross-Validation

Table 2 reports a paired comparison in which random and spatial cross-validation are computed on the identical training set with identical model configurations, so that the difference isolates the effect of spatial blocking. Under 2D KMeans spatial CV, every model shows a modest AUC reduction relative to random CV (1.2–1.6%), with CatBoost attaining the highest spatial AUC (0.957; Δ AUC -1.2%) and LightGBM degrading most (-1.6%), while XGBoost, Random Forest, and Stacking show intermediate reductions (-1.5% , -1.2% , and -1.3% , respectively). Under this design, random CV therefore overestimates geographic transferability only slightly, and ranking performance is largely retained across space. This gap is a property of the current feature set and sample construction as a whole; because earlier and current designs differ simultaneously in several respects, it is not attributed to any single feature without a controlled ablation—the coordinate ablation in Section 4.6 provides such a control.

Table 2 Paired random vs. spatial cross-validation (5-fold, identical training set and model configuration, 32 features). Values are mean AUC-ROC with the standard deviation across folds

Model	Random CV AUC	Spatial CV AUC	Δ AUC
XGBoost	0.968 $\pm$ 0.005	0.953 $\pm$ 0.054	-1.5%
LightGBM	0.969 $\pm$ 0.004	0.954 $\pm$ 0.052	-1.6%
CatBoost	0.969 $\pm$ 0.004	0.957 $\pm$ 0.050	-1.2%
Random Forest	0.965 $\pm$ 0.004	0.953 $\pm$ 0.053	-1.2%
Stacking	0.969 $\pm$ 0.004	0.956 $\pm$ 0.050	-1.3%

Online Resource 1 (Table S3) reports the per-fold detail. The folds are far from uniform: they range from 848 to 3,313 samples, and the class balance varies from 1,998:1,315 (positive:contrast) in the largest block to 126:767 in the most contrast-dominated block, because KMeans blocks are defined geometrically rather than by class. The consequences are visible directly. The best-separated blocks reach AUC 0.993–0.998, whereas the block containing the densest part of the inventory (the largest, most balanced block) gives AUC 0.866–0.876, and the F1 score at a fixed threshold of 0.5 ranges from 0.82 to 0.98 across folds and models. In the low-AUC block the models maintain moderate precision (0.77–0.83) with recall of 0.82–0.97, whereas in the contrast-dominated small blocks recall saturates at 1.0 with precision near 0.95. This confirms that the spatial CV spread reflects both genuine geographic heterogeneity and the uneven class composition of the blocks, which is why per-fold composition is reported rather than only aggregate means.

The dispersion reveals an important nuance that point estimates alone conceal: although random CV scores are tightly concentrated (standard deviation ≈ 0.004 – 0.005), spatial CV scores vary widely across folds (standard deviation ≈ 0.050 – 0.054). Spatial performance is therefore highly fold-dependent, and a single spatial split can give a misleading impression. The reported dispersion summarises this fold-to-fold spread only; it is not a significance test and does not capture uncertainty over alternative spatial partitions, so overlapping spreads should not be read as evidence that two models are statistically indistinguishable. At the fixed threshold of 0.5, mean F1 under spatial blocking (0.912–0.933) is in fact comparable to random CV (0.912–0.915), so the aggregate penalty appears mainly in ranking (AUC) rather than in thresholded accuracy; per-fold F1 nevertheless ranges from 0.82 to 0.98, indicating that the aggregate similarity masks strong fold dependence. Per-fold sample counts, class composition, thresholds, precision, recall, and F1 are reported in Online Resource 1 (Table S3). The models are therefore described as retaining most of their geographic transferability, although the deviations are largest in the largest and most class-balanced block, which also contains the densest part of the inventory. Fig. 3 visualizes the paired comparison.

4.3 SHAP Feature Importance

Fig. 4 shows the global SHAP feature importance computed with TreeSHAP on the XGBoost tree model. The top five features are slope angle (2.373), plan curvature (0.769), profile curvature (0.763), 30-day cumulative rainfall (0.671), and elevation (0.463), followed by the antecedent precipitation index (0.311), NDWI (0.284), fault distance (0.271) and latitude (0.246). Slope angle and curvature jointly dominate the terrain signal: slope governs gravitational shear stress, while curvature distinguishes convergent (hollow) from divergent (nose) locations that control water concentration. The vegetation indices (NDWI, NDVI) contribute more modestly here (0.284 and 0.235), because they proxy surface moisture and land-cover conditions but overlap with the terrain and rainfall signals. The 30-day cumulative rainfall is the leading rainfall predictor and ranks fourth overall (0.671). The coordinate features rank ninth (latitude) and twelfth (longitude), and a controlled ablation that removes them leaves performance essentially unchanged (Section 4.6); they are therefore not interpreted as physical drivers but as weak spatial proxies whose attribution overlaps with elevation, rainfall and terrain. Fault distance ranks eighth (0.271)—a modest but physically meaningful contribution. These associations are consistent with the geological setting of the study area, but they should be read as model-learned relationships rather than proof of causal mechanism: correlated predictors can redistribute attribution among themselves even after group aggregation.

Fig. 5 shows the SHAP values aggregated by feature group, using the same group definitions as Online Resource 1 (Table S2). Terrain features contribute the most (59.3%), followed by rainfall (19.1%), vegetation (6.4%) and the spatial coordinates (5.8%); derived interactions account for 5.5%, fault distance for 3.3%, and land-use variables for 0.6%. Within the rainfall family, 30-day cumulative rainfall (0.671) and the API (0.311) rank highest; the smaller contributions of the ratio-type rainfall

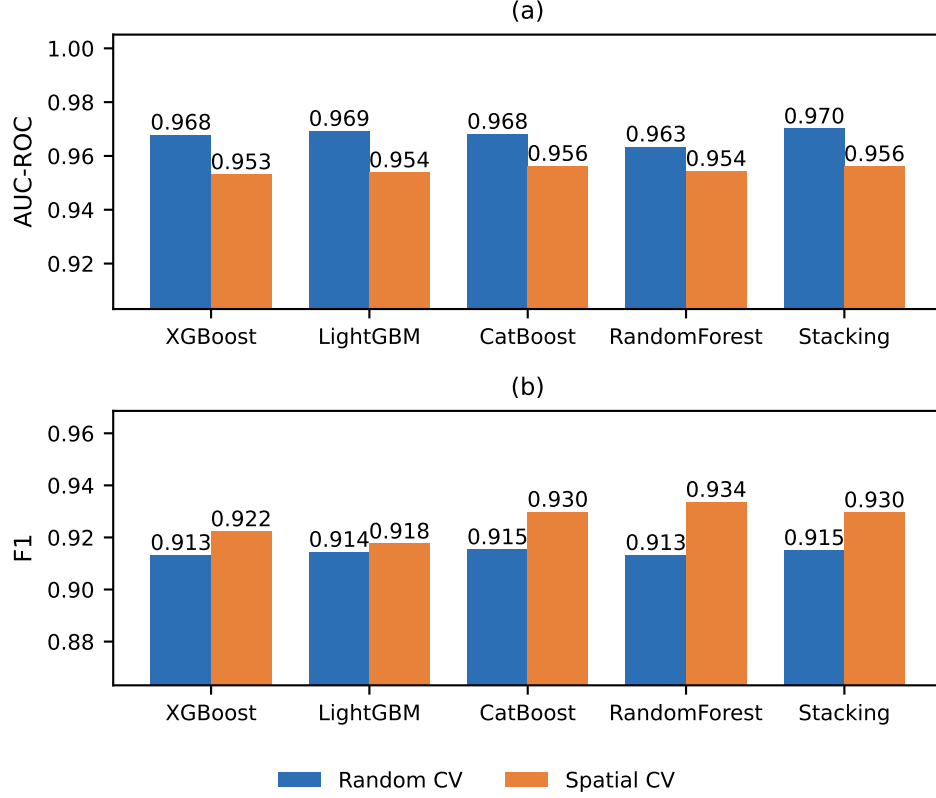


Fig. 3 Random CV vs. spatial CV comparison for five models: (a) AUC-ROC, (b) F1 score. Both evaluations use the identical training set and model configurations

features (0.040–0.158) reflect importance redistribution among correlated rainfall variables. This group-level view indicates that terrain and rainfall, rather than regional position, dominate susceptibility in the current feature set, with vegetation and fault proximity as secondary contributions.

Fig. 6 shows ROC and precision–recall curves for all 17 models, including four Optuna-tuned variants. The curves are computed from the same calibrated test-set probabilities that generate Table 1, so the legend values and the table agree numerically. The leading models are tightly clustered; the precision–recall curves separate them somewhat more clearly than the ROC curves.

4.4 Susceptibility Classification and Priority Mapping

All 10,644 samples are classified into five relative susceptibility classes using quantile breaks, a procedure that by construction assigns approximately 20% of samples to each class. The probabilities used here are out-of-fold predictions from 5-fold stratified cross-validation (OOV AUC-ROC = 0.969), so no location is scored by a model

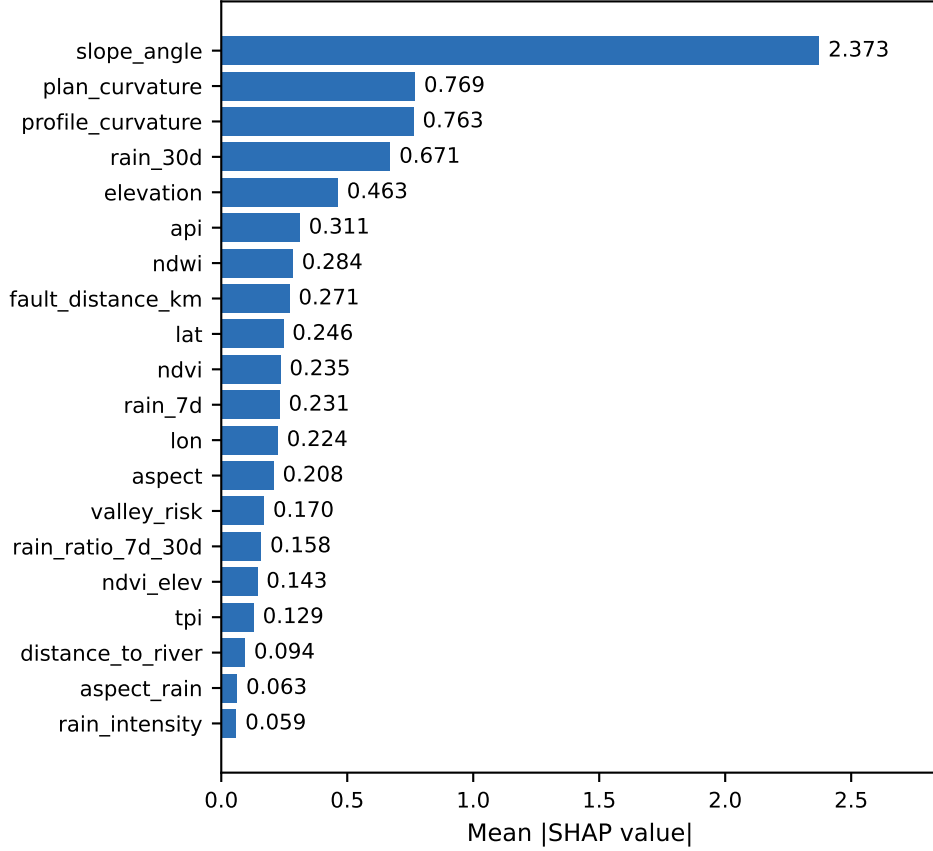


Fig. 4 SHAP global feature importance (top 20 features) for the XGBoost tree model

that was trained on it; applying a fitted model to its own training data produces almost entirely 0/1 probabilities, whereas the out-of-fold version places 40.1% of values between 0.01 and 0.99. Online Resource 1 (Table S4) shows that mean susceptibility increases monotonically from 0.000 (very low) to 0.997 (very high). The distribution remains strongly bimodal: the two lowest and the two highest classes are each tightly clustered (standard deviations 0.000–0.021), whereas the middle class carries most of the internal spread (mean 0.603, standard deviation 0.265). This bimodality is a property of the constructed contrast sample: because contrast locations inherit their environmental features from nearby inventory records, the model separates them confidently, and the intermediate probability range is sparsely populated. The five-class classification should therefore be read as a description of the relative distribution of model scores over the sampled locations, not as a calibrated probability of failure at any individual location. Fig. 7 shows the spatial distribution of the five classes and the corresponding distribution of out-of-fold probabilities.

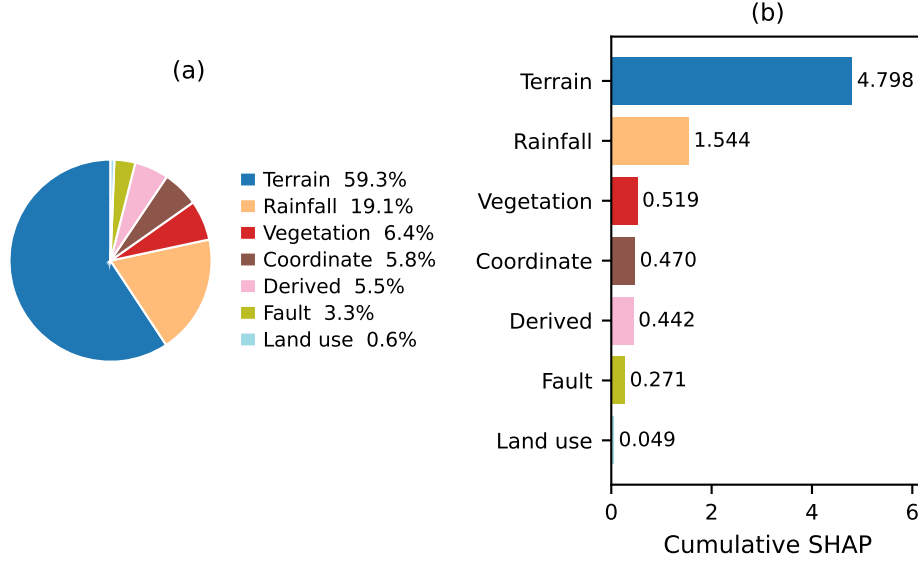


Fig. 5 SHAP feature group contribution: (a) pie chart of percentage contribution, (b) cumulative SHAP importance by group

4.5 Mechanism-Contrast Sample Composition Sensitivity Analysis

To evaluate the effect of contrast-sample composition, the mechanism-contrast share was varied over its full realisable range, from 0% (pure spatial sampling) to 30.5%, the maximum attainable share given that the census supplies only 1,622 usable mechanism-contrast records (Section 3.3). The total contrast-sample count remained fixed at 5,322, and the buffer distance was set to 0.8 km. Table 3 summarises the results.

Pure spatial sampling (share = 0%) produced trivially separable classes: the RF distinguishability AUC reached 1.000 and XGBoost attained a random-CV AUC of 1.000 and a spatial-CV AUC of 0.9993, indicating that the positive and constructed samples have almost entirely distinct feature distributions. Increasing the mechanism-contrast share degraded separability monotonically: the distinguishability AUC fell to 0.9872, 0.9759 and 0.9650, and the spatial-CV AUC from 0.9993 to 0.9817, 0.9646 and 0.9531, at shares of 10%, 20% and 30%. At the 30.5% ceiling (the maximum attainable share) the two metrics are numerically almost identical to the 30% configuration (0.9644 and 0.9525), so the response flattens only at the ceiling. Because the design space is bounded by the availability of mechanism-contrast records (Section 3.3), no interior optimum can be identified, and the share is treated as a design choice rather than an optimised parameter. That the distinguishability AUC falls from unity to about 0.96 shows the classes became harder to separate; it does not by itself demonstrate that the model learned rainfall-specific mechanisms.

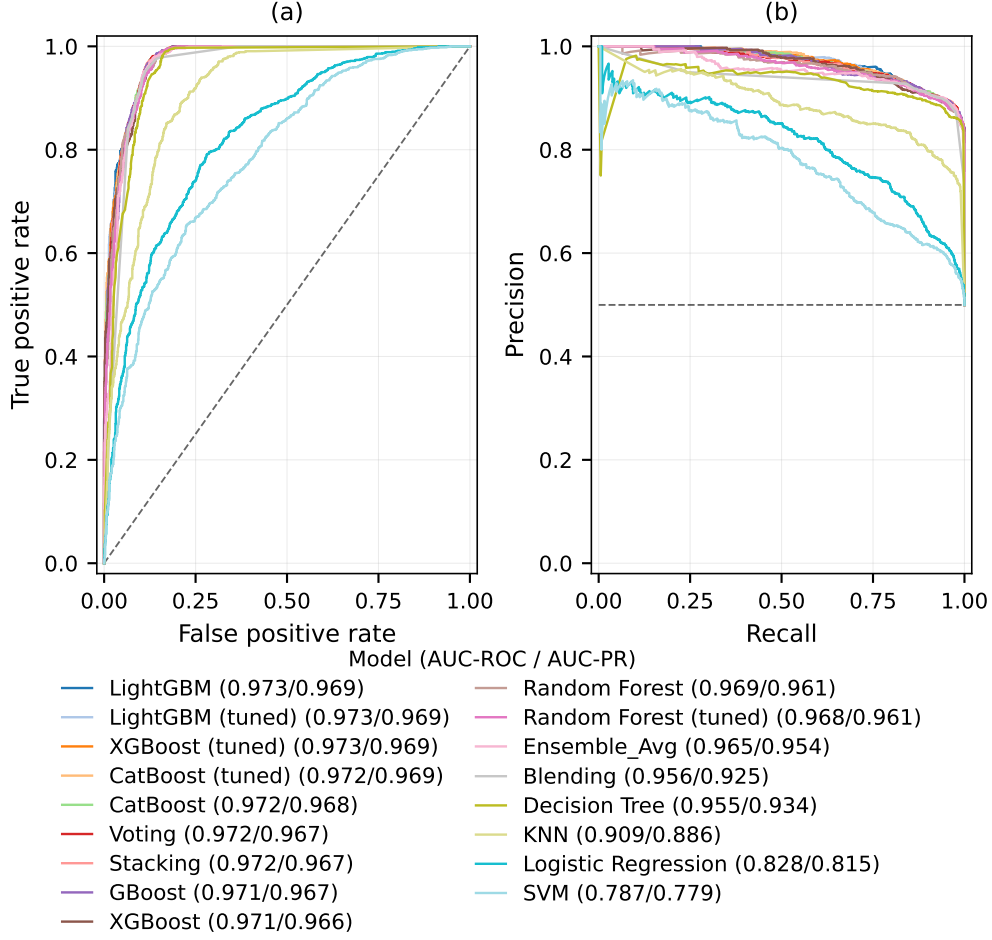


Fig. 6 ROC and precision–recall curves for all 17 models: (a) ROC curves, (b) precision–recall curves; the dashed line marks the chance level in (a) and the prevalence floor in (b)

The buffer distance was found to have a negligible effect: comparing 0.8 km and 3.0 km buffers changed the RF distinguishability AUC by at most 0.001 (0.9650 vs. 0.9651 at a 30% share, and 1.000 in both cases at 0%) and the spatial-CV AUC by at most 0.002 (0.9531 vs. 0.9511 at a 30% share, and 0.9993 vs. 1.000 at 0%). Accordingly, the 0.8 km buffer is retained as the working configuration.

4.6 Coordinate-Feature Ablation

Because the coordinate features (longitude and latitude) are the most likely carriers of spatial-autocorrelation signal, they were examined in a controlled ablation rather than by inspection of their SHAP values alone. The full feature set (32 features) was compared with a reduced set in which both coordinates were removed (30 features); random and spatial cross-validation were run on identical folds with identical

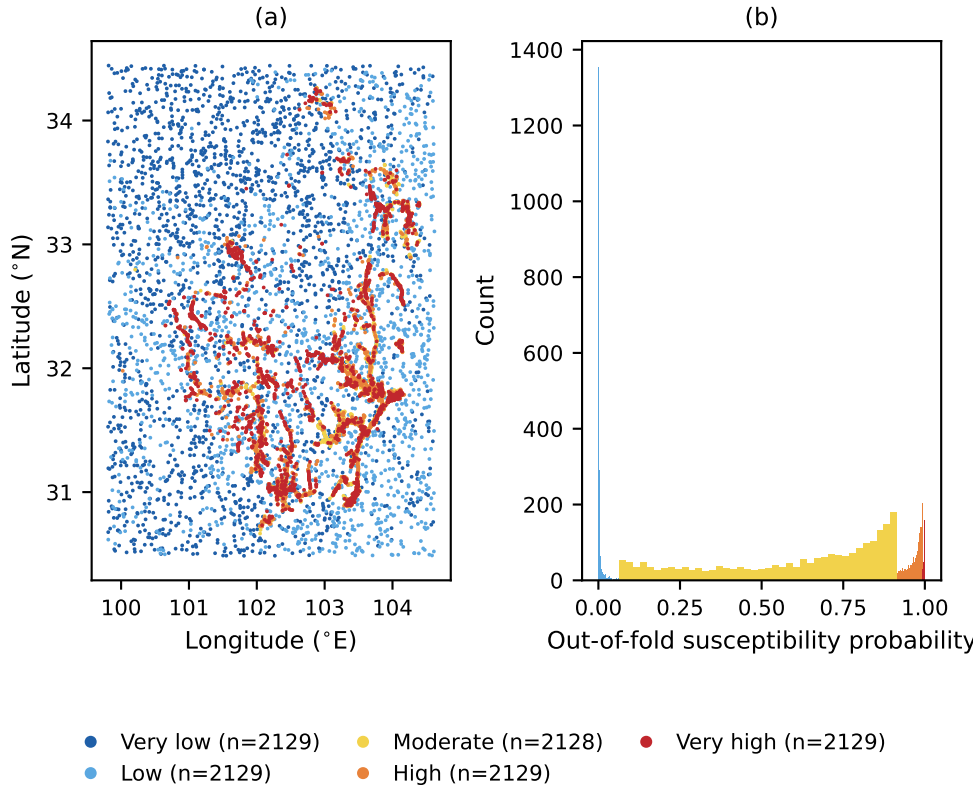


Fig. 7 Out-of-fold model predictions across the 10,644 sampled locations: (a) spatial distribution of the five relative susceptibility classes defined by quantile breaks, (b) distribution of the out-of-fold probabilities. Points are sample locations, not an independently validated susceptibility classification

Table 3 Sensitivity analysis of mechanism-contrast sample composition over its full realisable range. The share is the realised mechanism-contrast proportion of the 5,322 contrast samples; the census supplies only 1,622 usable mechanism-contrast records, so 30.5% is the maximum attainable share. The ceiling configuration (RF QC 0.9644; spatial CV 0.9525) is not tabulated separately, as it differs from the 30% configuration by less than 0.001 AUC. Buffer distance = 0.8 km. RF QC: random-forest distinguishability-check AUC (internal diagnostic; values approaching 1.0 indicate trivially separable classes). Spatial CV: 5-fold 2D KMeans spatial cross-validation AUC-ROC for XGBoost

Mech-Contrast (%)	RF QC AUC	Diagnosis	Spatial CV AUC
0	1.0000	Trivially separable	0.9993
10	0.9872	Highly separable	0.9817
20	0.9759	Highly separable	0.9646
30	0.9650	Acceptable separability	0.9531

model configurations for both variants. Online Resource 1 (Table S5) reports the full comparison, and Fig. 8 visualises it.

Removing the coordinates changed random-CV AUC by -0.0007 to -0.0023 , spatial-CV AUC by $+0.0001$ to $+0.0038$, and test-set AUC by -0.0029 to $+0.0015$ across the five models. The coordinate features therefore contribute little independent predictive signal: their group-level SHAP contribution (5.8% of the total) appears largely redundant with elevation, rainfall and terrain, and removing them leaves both cross-validation protocols essentially unchanged (all differences within ± 0.004 AUC).

The ablation also clarifies the random-versus-spatial gap of Section 4.2. If the coordinates were responsible for the modest random-CV advantage by memorising location, their removal should narrow the gap; in fact the gap is essentially unchanged when they are removed (for example, XGBoost: -1.5% with coordinates and -1.3% without). The spatial penalty is thus a property of the environmental features and the sample construction, not of the coordinate features. Coordinates are consequently retained only as a convenience for mapping, and are not interpreted as physical drivers.

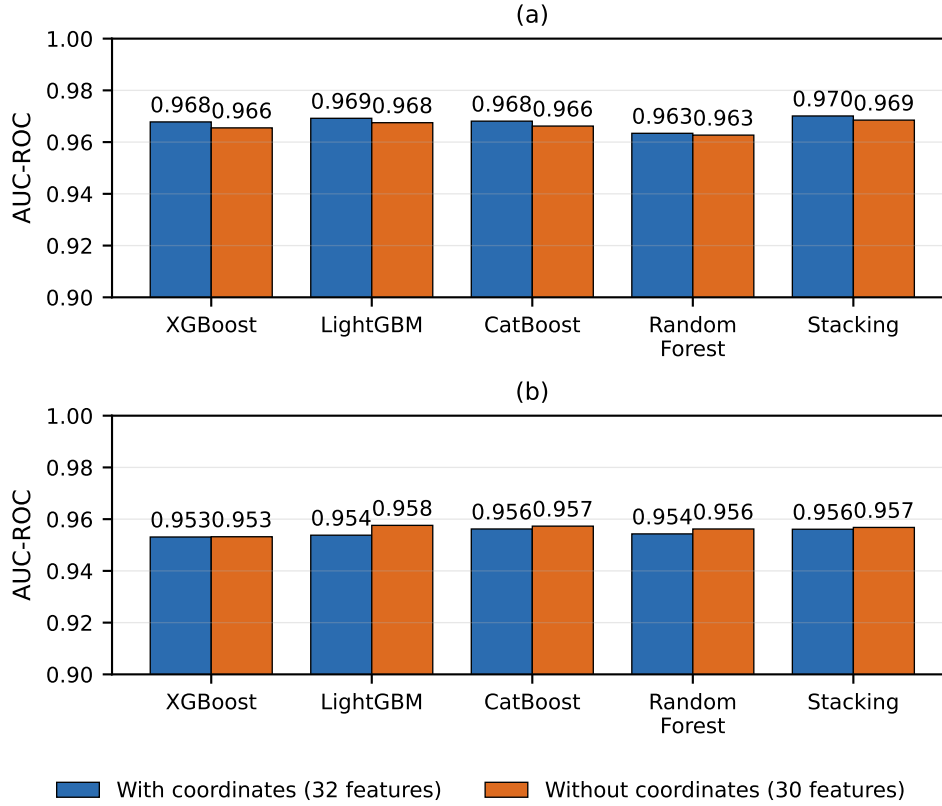


Fig. 8 Coordinate-feature ablation. AUC-ROC under (a) random and (b) spatial cross-validation, comparing the full 32-feature model (FULL; blue) with the 30-feature model that excludes longitude and latitude (NOCOORD; orange). Bars are means across the five models' folds

5 Discussion

5.1 Mechanism-Contrast Sample Construction and Susceptibility–Exposure Separation

The sensitivity analysis in Section 4.5 indicates that pure spatial sampling produces contrasts that are trivially separable (RF distinguishability AUC 1.000), whereas introducing mechanism-contrast samples reduces separability to a regime in which the classes remain distinguishable but are no longer separable by trivial distributional differences. The mechanism-contrast samples are not assumed to represent stable terrain; rather, they represent areas affected by different failure mechanisms (e.g., rockfall, ground collapse) that help the model distinguish rainfall-related slope instability from other instability processes. The objective is therefore not to identify safe areas, but to construct contrast samples that encourage the model to rely more on features associated with rainfall-dominated instability. It is emphasised that the distinguishability diagnostic is an internal quality check; the numeric range observed here should not be read as a universally applicable acceptance criterion.

The susceptibility–exposure separation further ensures that predictions reflect physical susceptibility rather than demographic proxies. To quantify the robustness of the priority ranking, each exposure weight was perturbed by ± 0.1 and the weights were re-normalised to unity. Across all perturbed schemes, the Kendall rank correlation with the default ranking remained above 0.96 when restricted to samples with non-zero exposure, and the top-10% priority set overlapped by 96–99%. The priority ranking is therefore robust to moderate misspecification of the exposure weights *among records that carry exposure data*. Aggregated exposure is recorded only for the inventory and mechanism-contrast records, not for spatially sampled locations, whose exposure index is set to zero; absent exposure does not imply absent population or property, so these cells are excluded from the priority analysis, and the risk map cannot be read as a complete picture of priority across the prefecture.

5.2 Spatial Autocorrelation

The paired validation shows that spatial blocking reduces AUC by 1.2–1.6% relative to random CV, so random CV overestimates geographic transferability only modestly and ranking performance largely survives spatial blocking. CatBoost attains the highest spatial AUC (0.957; Δ AUC -1.2%), and LightGBM degrades most (-1.6%), suggesting that gradient-boosted trees with leaf-wise growth are the most sensitive to spatial clustering in the training folds. The coordinate ablation (Section 4.6) shows that this residual gap is not caused by the coordinate features, because removing them changes it only marginally (for example, XGBoost: -1.5% with coordinates and -1.3% without). The per-fold results also show that the spatial folds differ strongly in size and class balance, and that the lowest-AUC fold is the one that contains the densest part of the inventory; part of the spatial penalty therefore reflects the block design rather than class imbalance alone. The modest magnitude of the spatial–random gap indicates that the models rely only partly on spatially clustered information, and that the current feature set supports reasonably robust spatial extrapolation within the study

area. F1 at the fixed threshold of 0.5 is approximately unchanged under spatial CV, so the penalty appears mainly in ranking rather than in thresholded accuracy, while per-fold F1 still varies widely.

The dispersion reported here summarises variability among the five spatial folds only. Because the folds share overlapping training data and because only one spatial partition was drawn, these figures are descriptive and should not be treated as significance tests or as a complete account of uncertainty over spatial partitioning; differences of a few tenths between models are therefore not interpreted as reliable model rankings. The 2D KMeans blocking provides a practical approximation of spatial independence; however, it may not capture all spatial structures (e.g., topographic corridors, fault-aligned valleys), and alternative strategies such as spatial leave-one-out or geographically weighted CV could be explored (Roberts et al. 2017). A complementary safeguard is the area-of-applicability framework, which flags locations whose predictor values fall outside the training domain (Meyer and Pebesma 2021).

5.3 Model Complexity vs. Performance Gain

Stacking does not improve AUC-ROC over the best base model: its AUC-ROC is 0.972 compared with 0.973 for default LightGBM, tuned LightGBM, and tuned XGBoost. The cost is nine base models plus a meta-learner. Training is offline, the meta-learner is lightweight (100 trees, depth 3), and inference requires nine forward passes plus one meta-learner prediction. The primary benefit is variance reduction: out-of-fold aggregation smooths individual model biases. In this study the tuned single model (LightGBM tuned with Optuna) matches Stacking on F1 (0.920 vs. 0.920) and exceeds it on AUC-ROC (0.973 vs. 0.972), so stacking offers little practical advantage here. Its spatial-AUC reduction (−1.3%) is in fact slightly larger than that of CatBoost (−1.2%) and comparable to Random Forest (−1.2%). The single models therefore remain preferable on both simplicity and spatial robustness. LightGBM, however, is the least spatially robust model (−1.6%), illustrating that strong in-distribution performance does not guarantee the best spatial transferability.

5.4 Comparison with Previous Studies

The present framework differs from most published landslide susceptibility studies in three respects: the dual-source mechanism-contrast strategy explicitly addresses the well-documented sensitivity of model performance to contrast-sample design (Goetz et al. 2015; Rabby et al. 2023; Liu 2023; Zhu et al. 2019; Gu et al. 2024); the susceptibility–exposure separation prevents demographic features from acting as shortcut variables, which single-stage designs that keep exposure covariates inside the classifier cannot separate from the physical signal (Sun et al. 2023); and the paired random-versus-spatial validation with per-fold detail follows the recommendation that spatially structured data be validated against spatially independent folds (Roberts et al. 2017; Ploton et al. 2020), adding fold-level dispersion whereas most susceptibility studies report only a single random-CV estimate. The modest random-to-spatial gap (1.2–1.6%) is smaller than the degradation frequently reported for large-scale environmental models (Ploton et al. 2020), consistent with evidence that optimistic bias

depends on the strength of spatial autocorrelation (Yoo and Koo 2024; Roberts et al. 2017).

The highest AUC-ROC (0.973) is comparable to gradient-boosting variants (Sahin 2022), neural-network ensembles (Bragagnolo et al. 2020), and convolutional networks (Yi et al. 2020), and falls within the range summarised in comparative reviews (Merghadi et al. 2020; Reichenbach et al. 2018). Studies in Sichuan report comparable or slightly lower discrimination (Chen 2025; Xiong et al. 2020). Direct comparison remains difficult because of differences in study areas, sample sizes, mapping units, and because many published figures use random CV only. Most published frameworks combine susceptibility and exposure within a single model and report no explicit check on contrast-sample separability, whereas the random-forest distinguishability diagnostic used here provides a quantitative and reproducible indication. The interpretation follows the model-agnostic view that explanations describe model behaviour rather than physical causation (Lundberg et al. 2020; Roscher et al. 2020), with group-level aggregation addressing attribution redistribution among correlated predictors (Lundberg and Lee 2017; Lundberg et al. 2020).

5.5 Limitations

This study has several limitations: (1) fault distance from the GEM Global Active Faults Database is included as a formal feature, but it captures only proximity to mapped active fault traces; fault type and slip rate, which modulate seismic shaking potential, are not incorporated, and the GEM database may underrepresent secondary or buried faults. Lithology is not included: a global 0.5° lithological map was evaluated but dismissed because it yielded only five lithological classes across the study area, providing insufficient discriminating variance while adding dimensionality; (2) the inventory aggregates debris flows, landslides, and unstable slopes into a single class, which is appropriate for reconnaissance-level susceptibility mapping but ignores differences in failure mechanics; no per-type analysis was conducted because individual subsets are imbalanced and because the inventory does not record failure depth or volume for most records; (3) the static predictor layers (terrain, land use, fault distance) are treated as time-invariant while the inventory spans several decades, so individual records may be matched to conditions that differ from those at the time of failure—particularly for land use and vegetation, which change over time; (4) KMeans spatial blocks may not capture all spatial structures, and their geometric definition produces folds of very unequal size and class balance, which contributes to the wide spread of spatial-CV scores; (5) the RF distinguishability check quantifies separability between positive and contrast samples but does not verify geological realism; (6) exposure weights are expert-defined, and although the ranking is robust to ± 0.1 perturbations, this does not establish their correctness; aggregated exposure is absent for spatially sampled locations, so the priority analysis covers only records with exposure data; (7) no temporal external validation using independent landslide events from different regions or later periods is included; and (8) the framework is evaluated in a single study area, and transferability to regions with different geological settings requires further investigation; no formal test of the area of applicability (Meyer and Pebesma 2021) was performed; and (9) the environmental features for the contrast samples

(both spatially sampled and mechanism-contrast records) are not field-measured but reconstructed via spatial interpolation—inverse-distance weighting for continuous variables (elevation, rainfall, NDVI, NDWI), nearest-neighbour assignment for categorical land use, and circular weighted averaging for aspect—so the interpolated values carry unquantified uncertainty that may introduce noise into the contrast set.

6 Conclusion

This study proposes an interpretable machine learning framework for rainfall-related slope instability susceptibility assessment that addresses sampling bias and spatial generalization through dual-source mechanism-contrast sample construction and two-stage susceptibility–exposure separation. The dual-source strategy is evaluated by random-forest distinguishability analysis (AUC 0.965), indicating that the contrast samples are separable but not trivially so. Among seventeen models, the AUC-weighted voting ensemble attains the highest F1 (0.923), and the highest AUC-ROC (0.973) is achieved by LightGBM; hyperparameter tuning with Optuna did not change these rankings and improved discrimination by at most 0.002 AUC-ROC. A paired random-versus-spatial cross-validation shows only a modest AUC reduction (1.2–1.6%) under spatial blocking, with F1 at a fixed threshold approximately unchanged, and a coordinate-feature ablation confirms that removing longitude and latitude leaves performance within ± 0.004 AUC. Group-level SHAP identifies terrain (59.3%), rainfall (19.1%), and vegetation (6.4%) as the dominant feature families, with coordinates contributing 5.8% and fault distance 3.3%; these are model-learned associations rather than verified causal mechanisms. The highest susceptibility class clusters in the south-eastern study area. Priority weighting is restricted to records carrying documented exposure data, which are unavailable for spatially sampled locations.

Future work will incorporate fine-grained fault type and slip-rate data, conduct temporal external validation using recent landslide events, extend the exposure-weight sensitivity analysis across multiple study areas, and examine failure-type-specific susceptibility models where inventory size permits.

Supplementary information. The article contains one online-only supplementary file (Online Resource 1) that provides the data-source table (Table S1), the complete list of the 32 susceptibility features (Table S2), the per-fold spatial cross-validation detail (Table S3), the five-class classification statistics (Table S4), and the coordinate-feature ablation (Table S5).

References

- Bragagnolo L, da Silva RV, Grzybowski JMV (2020) Artificial neural network ensembles applied to the mapping of landslide susceptibility. *Catena* 184:104240. <https://doi.org/10.1016/j.catena.2019.104240>
- Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 785–794, <https://doi.org/10.1145/2939672.2939785>

- Chen Y (2025) Spatial prediction and mapping of landslide susceptibility using machine learning models. *Natural Hazards* 121(7):8367–8385. <https://doi.org/10.1007/s11069-025-07132-3>
- Dai F, Lee CF (2002) Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. *Geomorphology* 42(3–4):213–228. [https://doi.org/10.1016/S0169-555X\(01\)00087-3](https://doi.org/10.1016/S0169-555X(01)00087-3)
- Farr TG, Rosen PA, Caro E, et al (2007) The Shuttle Radar Topography Mission. *Reviews of Geophysics* 45(2):RG2004. <https://doi.org/10.1029/2005RG000183>
- Goetz JN, Brenning A, Petschko H, et al (2015) Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences* 81:1–11. <https://doi.org/10.1016/j.cageo.2015.04.007>
- Gu T, Duan P, Wang M, et al (2024) Effects of non-landslide sampling strategies on machine learning models in landslide susceptibility mapping. *Scientific Reports* 14(1):7201. <https://doi.org/10.1038/s41598-024-57964-5>
- He Q, Wang M, Liu K (2021) Rapidly assessing earthquake-induced landslide susceptibility on a global scale using random forest. *Geomorphology* 391:107889. <https://doi.org/10.1016/j.geomorph.2021.107889>
- Horn BKP (1981) Hill shading and the reflectance map. *Proceedings of the IEEE* 69(1):14–47. <https://doi.org/10.1109/PROC.1981.11918>
- Ke G, Meng Q, Finley T, et al (2017) LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp 3146–3154, URL <https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
- Liu C (2023) Optimization of negative sample selection for landslide susceptibility mapping based on machine learning using K-means-KNN algorithm. *Earth Science Informatics* 16(4):4131–4152. <https://doi.org/10.1007/s12145-023-01151-z>
- Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp 4765–4774, URL <https://arxiv.org/abs/1705.07874>
- Lundberg SM, Erion G, Chen H, et al (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1):56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- Merghadi A, Yunus AP, Dou J, et al (2020) Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews* 207:103225. <https://doi.org/10.1016/j.earscirev.2020.103225>

- Meyer H, Pebesma E (2021) Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution* 12(9):1620–1633. <https://doi.org/10.1111/2041-210X.13650>
- Ploton P, Mortier F, Réjou-Méchain M, et al (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* 11(1):4540. <https://doi.org/10.1038/s41467-020-18321-y>
- Pourghasemi HR, Rossi M (2017) Landslide susceptibility modeling in a landslide prone area in Mazandarn province, north of Iran: a comparison between GLM, GAM, MARS, and M-AHP methods. *Theoretical and Applied Climatology* 130:609–633. <https://doi.org/10.1007/s00704-016-1919-2>
- Pourghasemi HR, Pradhan B, Gokceoglu C (2012) Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. *Natural Hazards* 63(2):965–996. <https://doi.org/10.1007/s11069-012-0217-2>
- Pradhan B (2013) A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers & Geosciences* 51:350–365. <https://doi.org/10.1016/j.cageo.2012.08.023>
- Prokhorenkova L, Gusev G, Vorobev A, et al (2018) CatBoost: Unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems* (NeurIPS), pp 6638–6648, URL <https://arxiv.org/abs/1706.09516>
- Qi S, Xu Q, Zhang B, et al (2011) Source characteristics of long runout rock avalanches triggered by the 2008 Wenchuan earthquake, China. *Journal of Asian Earth Sciences* 40(4):896–906. <https://doi.org/10.1016/j.jseas.2010.05.010>
- Rabby YW, Li Y, Hilafu H (2023) An objective absence data sampling method for landslide susceptibility mapping. *Scientific Reports* 13(1):1740. <https://doi.org/10.1038/s41598-023-28991-5>
- Reichenbach P, Rossi M, Malamud BD, et al (2018) A review of statistically-based landslide susceptibility models. *Earth-Science Reviews* 180:60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Roberts DR, Bahn V, Ciuti S, et al (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>
- Roscher R, Bohn B, Duarte MF, et al (2020) Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8:42200–42216. <https://doi.org/10.1109/ACCESS.2020.2976199>

- Sahin EK (2022) Comparative analysis of gradient boosting algorithms for landslide susceptibility mapping. *Geocarto International* 37(9):2441–2465. <https://doi.org/10.1080/10106049.2020.1831623>
- Sun D, Chen D, Zhang J, et al (2023) Landslide susceptibility mapping based on interpretable machine learning from the perspective of geomorphological differentiation. *Land* 12(5):1018. <https://doi.org/10.3390/land12051018>
- Xiong K, Adhikari BR, Stamatopoulos CA, et al (2020) Comparison of different machine learning methods for debris flow susceptibility mapping: A case study in the Sichuan province, China. *Remote Sensing* 12(2):295. <https://doi.org/10.3390/rs12020295>
- Xu C, Xu X, Yao X, et al (2014) Three (nearly) complete inventories of landslides triggered by the May 12, 2008 Wenchuan Mw 7.9 earthquake of China and their spatial distribution statistical analysis. *Landslides* 11(3):441–461. <https://doi.org/10.1007/s10346-013-0404-6>
- Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on support vector machine: A case study on natural slopes of Hong Kong, China. *Geomorphology* 101(4):572–582. <https://doi.org/10.1016/j.geomorph.2008.02.011>
- Yi Y, Zhang Z, Zhang W, et al (2020) Landslide susceptibility mapping using multiscale sampling strategy and convolutional neural network: A case study in Jiuzhaigou region. *Catena* 195:104851. <https://doi.org/10.1016/j.catena.2020.104851>
- Yin Y, Cheng Y, Liang J, et al (2016) Heavy-rainfall-induced catastrophic rockslide-debris flow at Sanxicun, Dujiangyan, after the Wenchuan Ms 8.0 earthquake. *Landslides* 13(1):9–23. <https://doi.org/10.1007/s10346-015-0554-9>
- Yoo M, Koo H (2024) Exploring the impact of spatial autocorrelation on optimistic bias in cross-validation and assessing the effectiveness of spatial cross-validation. *Cartography and Geographic Information Science* 52(5):596–609. <https://doi.org/10.1080/15230406.2024.2422593>
- Zhang M, Yin Y, Huang B (2015) Mechanisms of rainfall-induced landslides in gently inclined red beds in the eastern Sichuan basin, SW China. *Landslides* 12(5):973–983. <https://doi.org/10.1007/s10346-015-0611-4>
- Zhou W, Tang C (2014) Rainfall thresholds for debris flow initiation in the Wenchuan earthquake-stricken area, southwestern China. *Landslides* 11(5):877–887. <https://doi.org/10.1007/s10346-013-0421-5>
- Zhu AX, Miao Y, Liu J, et al (2019) A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods. *CATENA* 183:104188. <https://doi.org/10.1016/j.catena.2019.104188>

Statements and Declarations

Funding. This work was supported by the Sichuan University Jinjiang College Youth Fund (QNJJ-2025-A07).

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics statement. This study used aggregated, non-identifiable exposure statistics from documented geological hazard sites. No human participants, individual-level data, or personally identifiable information were involved; therefore, ethics approval and informed consent were not applicable. All methods complied with relevant guidelines and the data-use conditions of the Sichuan Geological Survey.

Data availability. The geological hazard inventory and associated aggregated exposure data are not publicly available because access is restricted by the data provider (Sichuan Geological Survey). These data are location-based aggregated statistics and contain no individual-level personally identifiable information. Access may be granted by the data owner upon reasonable request; data requests should be directed to siry39747@gmail.com, subject to the data owner's permission. SRTM DEM data are publicly available from NASA (<https://www.earthdata.nasa.gov>); the GEM Global Active Faults Database is available under a CC-BY-4.0 licence (<https://github.com/GEMScienceTools/gem-global-active-faults>). The processed feature dataset is available from the corresponding author on reasonable request, subject to the restrictions of the underlying data.

Code availability. The processing and modelling source code is available from the corresponding author on reasonable request.

Author Contributions. Liping Zheng contributed to the study conception and design, research methodology, project supervision, funding acquisition, and manuscript review and revision. Yanwen Yang was responsible for model development and training, spatial validation, and writing the first draft of the manuscript. Shuanglin Deng contributed to data organization, feature construction, and results visualization. Kaiwen Xiao contributed to data preprocessing and investigation. Moran Qiu contributed to geological analysis and data validation. Feng Ning contributed to manuscript review and critical revision. Liangjun Zhao contributed to the interpretation of results and research guidance. All authors commented on previous versions of the manuscript and read and approved the final manuscript.