applied sciences

MDPI

Article

Evaluating Conditioning Factor Selection and Reduction for Landslide Susceptibility Mapping in the Hualong–Xunhua Region, Upper Yellow River

Heming Yang [1], Wenhui Liu [1,*], Yabin Liu [1,*], Sha Yang [1], Qifa Li [2], Caiqi Leng [1] and Jingjing Wang [1]

[1] School of Geological Engineering, Qinghai University, Xining 810016, China; ys240859010316@mailbox.qhu.edu.cn (H.Y.)
[2] Qinghai Provincial Natural Resources Remote Sensing Center, Xining 810001, China
* Correspondence: 2019990102@qhu.edu.cn (W.L.); 2019990011@qhu.edu.cn (Y.L.)

Featured Application

The proposed framework supports reliable conditioning factor selection and reduction in regional landslide susceptibility mapping, providing a practical basis for hazard screening and land use planning.

Abstract

Conditioning factor reduction can decrease redundancy and modelling complexity in landslide susceptibility mapping. These factors describe the environmental conditions associated with slope instability. Statistical screening, redundancy diagnostics, importance rankings and internal validation do not establish that the remaining conditioning factors can be safely removed. Whether aggressive reduction preserves the environmental information required for reliable regional landslide susceptibility mapping therefore remains uncertain. This study developed an integrated framework for evaluating the reliability of conditioning-factor reduction in the Hualong–Xunhua region of the Upper Yellow River, China, combining statistical diagnostics, model interpretation, GeoDetector analysis, and random and spatial validation. The compact system selected during development was further compared with the complete system using a reserved evaluation subset, followed by regional susceptibility mapping and spatial comparison. The results showed that, first, precipitation showed weak univariate but strong model and interaction evidence, illustrating their complementarity. Second, Top5, comprising normalised difference vegetation index (NDVI), distance to rivers, precipitation, lithology and terrain ruggedness index (TRI), met development criteria: mean area under the receiver operating characteristic curve (AUC) values were 0.866 under random validation and 0.846 under spatial validation, compared with 0.875 and 0.853 for the complete 17-factor system (Full17), respectively. Third, in the post-selection reserved comparison, mean AUC was 0.865 for Top5 and 0.895 for Full17, with paired uncertainty supporting this decline; favourable development performance did not ensure reliable substitution. Finally, Full17 was retained for final mapping; absolute score differences greater than 0.10 covered 32.02% of mapped area, indicating potential spatial information loss. Similar predictive performance did not imply spatial equivalence. These findings show that factor reduction requires evaluation of predictive performance, generalisation, model dependence, and spatial information preservation.

Keywords: landslide susceptibility; conditioning factor selection; factor reduction; spatial validation; machine learning

Check for updates

Academic Editor: Salvador García-Ayllón Veintimilla

Received: 11 August 2026
Revised: 20 September 2026
Accepted: 25 September 2026
Published: 26 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Landslides are among the most widespread geological hazards in mountainous regions, posing persistent threats to settlements, transportation infrastructure, and land-use activities. Landslide susceptibility mapping estimates where slope failures are more likely to occur under a given set of terrain and environmental conditions, without explicitly considering their timing or magnitude, and thus provides an important basis for regional screening, spatial planning, and disaster reduction [1–3]. Conditioning factors are environmental variables describing the background conditions associated with slope instability; triggering events, such as intense rainfall or earthquakes, initiate failure under those conditions. With the growing availability of remote-sensing products, geographic information systems, and multi-source environmental datasets, data-driven susceptibility assessment has progressively shifted from heuristic and conventional statistical approaches towards machine learning methods capable of capturing complex and nonlinear associations between mapped landslides and their surrounding environments [4–6]. Nevertheless, increasingly sophisticated algorithms do not necessarily produce more reliable susceptibility estimates. Model outcomes remain sensitive to the quality and spatial representation of landslide inventories, the construction of pseudo-absence samples, the composition of conditioning factor systems, and the validation strategies used to assess predictive performance [7,8].

Regional landslide susceptibility models commonly integrate topographic, geological, hydrological, climatic, vegetation, and anthropogenic variables to represent the heterogeneous environmental conditions associated with slope instability in complex mountainous terrain [9–11]. Machine learning studies have compared tree-based and other data-driven algorithms across contrasting landslide settings [5,12]. As the number of input variables increases, however, multicollinearity, information redundancy, weakly informative factors, and modelling complexity may also increase. Consequently, statistical screening, redundancy diagnostics, feature-selection methods, and model-based importance measures have been widely employed to identify more compact conditioning-factor subsets [1,7,13–15]. Feature-selection studies have compared alternative screening rules within XGBoost [13], while factor-removal experiments have shown different performance responses among tree and ensemble models [15]. Inventory completeness and the construction of landslide-reference and comparison samples can also materially influence susceptibility estimates [16–18]. Factor reduction may retain or improve predictive performance when genuinely redundant or weakly informative variables are removed; however, excessive reduction can also yield only marginal gains or reduce performance when complementary information is discarded, and factor rankings may vary across study areas, algorithms, and selection procedures [4,5,7]. This inconsistency highlights an important distinction: a factor may appear weak or redundant when evaluated individually, yet still provide complementary information that is spatially localised or arises from interactions within a multivariate model. Therefore, identifying a small set of highly ranked factors is not equivalent to demonstrating that the remaining factors can be safely removed.

Reliable evaluation of conditioning-factor reduction therefore requires more than demonstrating comparable predictive performance under a single random split. Random partitioning may place spatially adjacent or environmentally similar samples in both the training and validation subsets, potentially producing optimistic estimates of model generalisation; spatially structured validation can provide a more conservative assessment of regional transferability [19–22]. Moreover, similar aggregate predictive metrics do not necessarily indicate equivalent susceptibility maps, because different models or factor systems may produce systematic local overestimation and underestimation despite achieving comparable overall discrimination [6,18]. Nevertheless, the spatial consequences of

reduced factor sets and the preservation of susceptibility score patterns have received comparatively limited attention [1,4].

This study evaluates whether a compact conditioning factor system can adequately represent the environmental information contained in a complete factor system in the Hualong–Xunhua region of the Upper Yellow River, China. An integrated framework combining statistical diagnostics, model interpretation, GeoDetector analysis, repeated random cross-validation, and spatial block cross-validation was developed to construct and evaluate candidate factor systems. The compact system selected during development was further compared with the complete system using a reserved evaluation subset. Regional landslide susceptibility mapping was then conducted using an ensemble of machine learning models, and the susceptibility maps produced by the two factor systems were compared to evaluate the spatial information associated with factor reduction. The results of this study can provide a methodological reference for conditioning factor selection, validation design, and factor reduction strategies in regional landslide susceptibility mapping.

2. Data and Methods

2.1. Study Area

The study area covers Hualong Hui Autonomous County and Xunhua Salar Autonomous County in eastern Qinghai Province, China, and is located in the upper reaches of the Yellow River (Figure 1). The vector boundary encompasses approximately 4526.38 km^2. The region is characterised by strongly dissected mountain–valley terrain, pronounced topographic gradients, heterogeneous lithological conditions, and intensive interactions among river incision, surface cover, and human activities. These environmental contrasts generate marked spatial heterogeneity in the terrain and environmental conditions associated with slope instability. Such heterogeneity also provides an appropriate regional setting for examining whether a compact conditioning factor system can adequately represent the information contained in a broader multi-source factor system. A total of 281 landslide polygons were mapped within the study area and used as the basis for subsequent sample construction and susceptibility modelling. The location, topography, administrative setting, and spatial distribution of the mapped landslides are shown in Figure 1. A recent study of the Xunhua–Hualong basin reported a mean annual precipitation of approximately 366 mm, with more than 80% of the annual total concentrated from May to September, indicating a pronounced seasonal concentration of rainfall in the study region [23]. The official intensity assessment of the 18 December 2023 Jishishan magnitude-6.2 earthquake placed parts of both counties in intensity-VII zones [24]. Mean annual precipitation and distance to faults describe climatic and structural conditioning. Because event dates were unavailable for the inventory, event-specific attribution was not assessed.

2.2. Data and Sample Construction

Sample construction provides consistent landslide-reference and comparison locations for evaluating conditioning factors and susceptibility predictions. One reference point per mapped polygon avoids disproportionate contributions from larger landslides.

A landslide inventory was compiled by manually identifying and delineating landslide polygons from Sentinel-2 optical imagery acquired during the summer of 2025. The principal visible and near-infrared bands have a nominal resolution of 10 m. High-resolution Google Earth imagery was used as an auxiliary reference for cross-checking during inventory interpretation. Sentinel-2 data were accessed and processed through the Google Earth Engine platform [25,26]. The geometric centroid of each polygon was extracted, yielding 281 landslide-reference samples, with one sample per polygon. Because locations without

mapped landslides cannot be assumed to represent confirmed stable terrain, comparison samples were defined as pseudo-absence rather than non-landslide samples [27]. A 500 m exclusion buffer was constructed around the landslide reference points, and a minimum inter-point distance of 90 m, equivalent to three 30 m analysis cells, was imposed to reduce immediate neighbourhood effects and local clustering. A total of 281 pseudo-absence points were randomly generated outside the exclusion buffers using a fixed random seed of 42. The resulting balanced dataset contained 562 records, comprising 281 landslide-reference samples and 281 pseudo-absence samples.

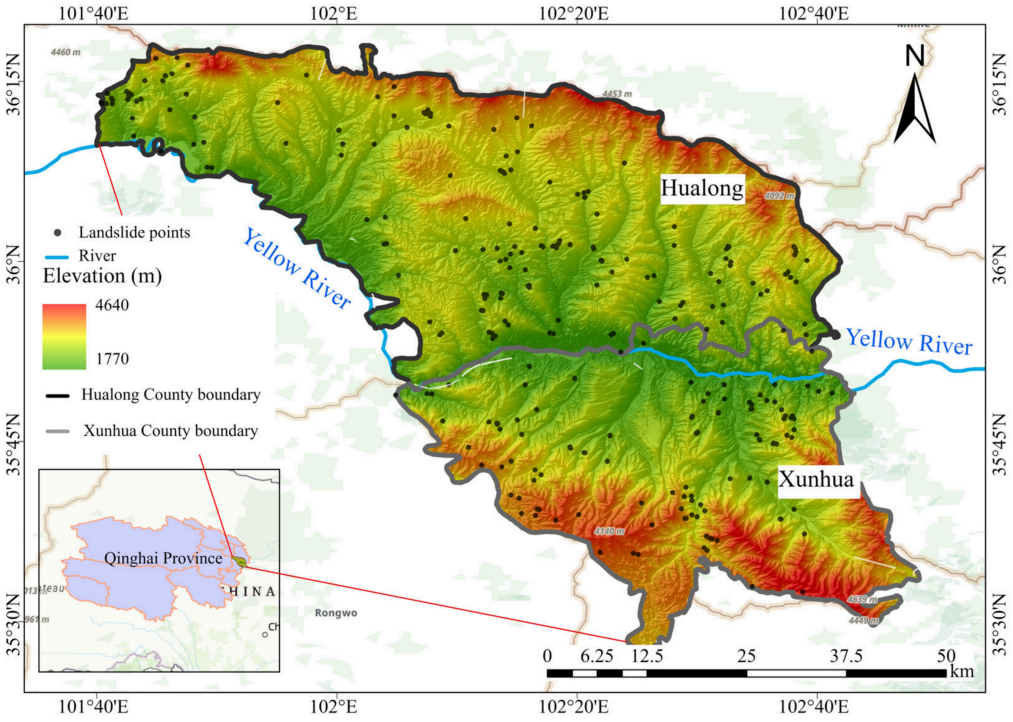


Figure 1. Location, topography, and spatial distribution of mapped landslides in the Hualong–Xunhua region, Qinghai Province, China.

Seventeen conditioning factors were assembled to represent the principal environmental dimensions commonly considered in data-driven landslide susceptibility assessment [1,4]. Their data sources, acquisition periods, and original spatial resolutions, processing methods, and final grid are summarised in Table 1. Topographic, vegetation, and climatic factors were derived from SRTM, Sentinel-2, and CHIRPS data, respectively [25,28,29]. Land-cover data were obtained from ESA WorldCover v200 [30]. River and road data were obtained from HydroRIVERS and OpenStreetMap, respectively [31,32], whereas lithology and fault data were obtained from the National Geological Archives of China. Terrain factors include elevation, slope, sine- and cosine-transformed aspect, plan and profile curvature, relief, terrain ruggedness index (TRI), topographic position index (TPI), and topographic wetness index (TWI). All factor layers were projected to WGS 84/UTM Zone 48N (EPSG:32648) and aligned at 30 m to the elevation reference grid. This resolution maintains the DEM-based terrain framework and aligns heterogeneous predictors. The aligned CHIRPS layer retained the information scale of the original 0.05° product. Continuous DEM, Sentinel-2 red/near-infrared, and annual CHIRPS rasters used bilinear interpolation. Terrain derivatives were computed from the aligned DEM; NDVI was calculated from aligned bands and composited by the July 2025 pixel-wise median, and precipitation was averaged over 2015–2024. WorldCover and image quality masks used nearest-neighbour resampling. Lithology used categorical rasterisation; river, fault, and road vectors were rasterised before

distance transformation. These operations distinguish continuous interpolation from class-preserving and vector processing [33]. Lithology and land cover were treated as categorical variables, whereas the remaining factors were treated as continuous variables.

Table 1. Data sources and spatial characteristics of the conditioning factors.

Factor	Category	Data Source	Time	Original Resolution	Processing/Final Grid
Elevation	Topographic	USGS SRTM 30 m DEM	-	30 m	Bilinear/30 m
Slope, sine- and cosine-transformed aspect, plan curvature, profile curvature, relief, TRI, TPI, and TWI	Topographic	Derived from USGS SRTM 30 m DEM	-	30 m	DEM derivatives/30 m
Normalized difference vegetation index (NDVI)	Vegetation	Sentinel-2 Level-2A surface-reflectance imagery	July 2025	10 m	Bilinear bands; median NDVI/30 m
Lithology	Geological	National Geological Archives of China	–	1:500,000	Categorical rasterisation/30 m
Mean annual precipitation	Climatic	CHIRPS v2.0 annual precipitation	2015–2024	0.05°	Bilinear; 10-year mean/30 m
Land cover	Land-surface	ESA WorldCover v200	2021	10 m	Nearest neighbour/ 30 m
Distance to rivers	Hydrological	HydroRIVERS Version 1.0, Central and South-East Asia tile	October 2019	Vector	Rasterise; distance/30 m
Distance to faults	Geological	National Geological Archives of China	–	Vector	Rasterise; distance/30 m
Distance to roads	Anthropogenic	Geofabrik Qinghai OpenStreetMap road extract	OSM snapshot: 3 June 2026	Vector	Rasterise; distance/30 m

2.3. Multi-Source Evidence Framework for Factor Evaluation

Factor analysis was organised as a staged evidence framework rather than as a single automatic screening procedure. An initial descriptive analysis was conducted before the 70:30 split using the complete 562-sample factor table to characterise sample group contrasts. For the 15 continuous factors, differences between landslide-reference and pseudo-absence samples were evaluated using two-sided Mann–Whitney U tests [34], with the resulting 15 p values adjusted using the Benjamini–Hochberg false-discovery-rate procedure [35]. Because statistical significance alone does not indicate the magnitude of group separation, Cliff's delta was also calculated, with positive values indicating higher factor values among landslide-reference samples; absolute delta values of <0.147, 0.147–0.329, 0.330–0.473, and $\geq$0.474 were interpreted as negligible, small, medium, and large effects, respectively [36,37]. Lithology and land cover were treated as categorical variables, and their associations with the binary sample label were evaluated using Pearson's chi-square test and Cramér's V. Because both contingency tables contained sparse categories and violated conventional expected-frequency assumptions, empirical p values were estimated from 10,000 random permutations of class labels using a fixed random seed of 42, with class balance and category frequencies preserved in every permutation. These analyses based on the full sample were treated as descriptive evidence within the factor diagnostic framework rather than as model performance estimates or automatic criteria for factor removal. For factor selection, the contrast, categorical-association, correlation, and variance-inflation analyses were repeated using only the 393 development samples. The corresponding development-only diagnostics are reported separately in Section 3.1.

Potential redundancy among the 15 continuous factors was assessed using Pearson and Spearman correlation coefficients. Factor pairs with an absolute Pearson coefficient or Spearman rank correlation coefficient of at least 0.70 were identified as strongly associated.

The variance inflation factor (VIF) was subsequently calculated for continuous factors only. For the ith continuous conditioning factor, VIF was defined as follows:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (1)$$

where R_i^2 is the coefficient of determination obtained by regressing the ith conditioning factor against all remaining continuous factors. Larger VIF values indicate stronger multicollinearity among the conditioning factors. In this study, values greater than 10 were considered indicative of potentially problematic multicollinearity [38]. Correlation and VIF analyses were used to identify candidate redundancy structures and inform alternative factor combinations rather than to perform automatic variable deletion. This strategy was adopted because correlated conditioning factors may still provide complementary information for tree-based models through nonlinear responses or factor interactions. Therefore, redundancy diagnostics were incorporated as one component of the staged evidence framework rather than being used as an independent criterion for factor exclusion.

After completion of the initial full sample diagnostics, the 562 samples were divided through stratified random sampling into a development dataset containing 393 samples and a reserved evaluation subset containing 169 samples using a fixed random seed of 42. The development dataset contained 393/562 samples (69.93%), including 197 landslide-reference samples (50.13%) and 196 pseudo-absence samples (49.87%); the reserved subset contained 169/562 samples (30.07%), including 84 landslide-reference samples (49.70%) and 85 pseudo-absence samples (50.30%). The reserved subset was excluded from subsequent model fitting, original-factor permutation importance, SHapley Additive exPlanations (SHAP) analysis, GeoDetector analysis, threshold determination, and candidate evaluation during development. However, because its factor values and labels had already contributed to the initial full-sample contrast and redundancy diagnostics, it was not considered a completely independent dataset for construction of the factor systems. Within the development dataset, Random Forest, XGBoost, LightGBM, and CatBoost were trained using repeated stratified five-fold cross-validation with five repeats, resulting in 25 validation folds for each model [39–42]. These algorithms were compared to test model dependence of factor reduction: Random Forest provides a bagged-tree baseline, XGBoost uses regularised gradient boosting, and LightGBM and CatBoost provide distinct boosting implementations. All data-dependent preprocessing was fitted separately within each training fold. Lithology and land cover were one-hot encoded, and previously unseen categorical levels during validation were handled by ignoring the corresponding encoded categories rather than removing samples.

Two complementary model interpretation approaches were applied using the validation folds generated during development. First, grouped original-factor permutation importance was estimated by jointly permuting all encoded columns belonging to each original factor with the same row permutation within each validation fold and measuring the resulting decrease in validation AUC. Thirty permutation repeats were conducted for each original factor using a reproducible seed of 42 plus the internal split index. Negative permutation values were retained, whereas non-negative values were normalised within each fold before summarisation across validation folds and Base4 models. Second, SHAP values were calculated using TreeExplainer for each fitted fold model on its corresponding validation subset, and mean absolute SHAP values of encoded columns were aggregated back to their original conditioning factors [43]. Within-fold normalised SHAP values and rankings were summarised across validation folds and Base4 models. Both permutation importance and SHAP were interpreted using consistency across folds and models rather than individual model rankings. The integrated ranking averaged the cross-model mean

permutation and SHAP ranks. Permutation-importance and SHAP analyses were based exclusively on the development samples. GeoDetector was additionally applied only to the 393-sample development dataset to evaluate differentiation at the factor level and pairwise interactions [44,45]. Continuous factors were discretised into five strata based on quantiles under the primary scheme, whereas lithology and land cover retained their original categorical classes. For each factor, the q statistic used to quantify differentiation of the binary sample labels among factor strata was calculated as follows:

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2} \quad (2)$$

where h = 1, L denotes the strata of a conditioning factor, N_h and ${\sigma_h}^2$ are the sample size and variance of the binary sample labels within stratum h, respectively, and N and σ^2 are the corresponding values for the entire development sample. The q statistic ranges from 0 to 1, with larger values indicating stronger explanatory power for the spatial differentiation of landslide occurrence. Statistical significance was assessed using 5000 label permutations followed by Benjamini–Hochberg correction across the 17 factor tests. Four- and six-quantile discretisations were additionally used as sensitivity analyses for continuous-factor stratification. Pairwise q statistics were calculated for all 136 combinations of the 17 conditioning factors. To correspond with the interaction gain reported in the results, the gain for factors X_i and X_j was defined as follows:

$$\Delta q_{ij} = q(X_i \cap X_j) - \max[q(X_i), q(X_j)] \quad (3)$$

where $q(X_i \cap X_j)$ represents the explanatory power of the interaction between factors X_i and X_j. A positive Δq_{ij} indicates that their joint explanatory power exceeds that of the stronger individual factor. Pairwise interactions were classified as independent, weakened, bivariate enhancement, or nonlinear enhancement according to the implemented GeoDetector interaction rules. Interactions were flagged as sparse when at least 25% of the joint strata contained fewer than five samples and were excluded from primary interpretation. GeoDetector results were interpreted as statistical differentiation within the development dataset rather than as causal effects or landslide occurrence probabilities. These analyses were integrated with the development-only diagnostics and model-interpretation results to construct the candidate factor systems. Candidate definitions and thresholds were fixed before model evaluation. Fixed model settings were used throughout the analysis: Random Forest used 500 trees with square-root feature sampling; XGBoost used 500 estimators with a maximum depth of 3, learning rate of 0.03, and subsample and column-sampling ratios of 0.8; LightGBM used 500 estimators with 31 leaves, a learning rate of 0.03, and subsample and column-sampling ratios of 0.8; and CatBoost used 500 iterations with a depth of 4 and learning rate of 0.03. Class balancing was applied to Random Forest and LightGBM, and a fixed random seed of 42 was used throughout. The overall analytical workflow is summarised in Figure 2.

2.4. Factor Combination Design and Development Validation

Twelve candidate conditioning factor combinations were predefined for evaluation during development. Full17 retained all 17 conditioning factors and served as the reference system. Slope and TRI were both retained in Full17. Two 16-factor combinations (VIF A16 and VIF B16) represented alternative treatments of the strong slope–TRI collinearity pattern by excluding slope or TRI, respectively. Four Core combinations represented TRI- and slope-based alternatives, with or without land cover, and were constructed from the combined evidence provided by the preceding sample contrasts, redundancy diagnostics,

permutation importance, SHAP analysis, and GeoDetector results. Top-N systems followed the combined permutation and SHAP ranking; the development-only analysis yielded the same order. Five progressively reduced Top-N systems contained 5, 8, 10, 12, and 15 factors, respectively. Top5 consisted of NDVI, distance to rivers, precipitation, lithology, and TRI; Top8 additionally included distance to roads, slope, and cosine-transformed aspect; Top10 further added distance to faults and TWI; Top12 added TPI and relief; and Top15 further incorporated land cover, elevation, and profile curvature. The membership of each candidate system was defined before model evaluation and is summarised in Table 2.

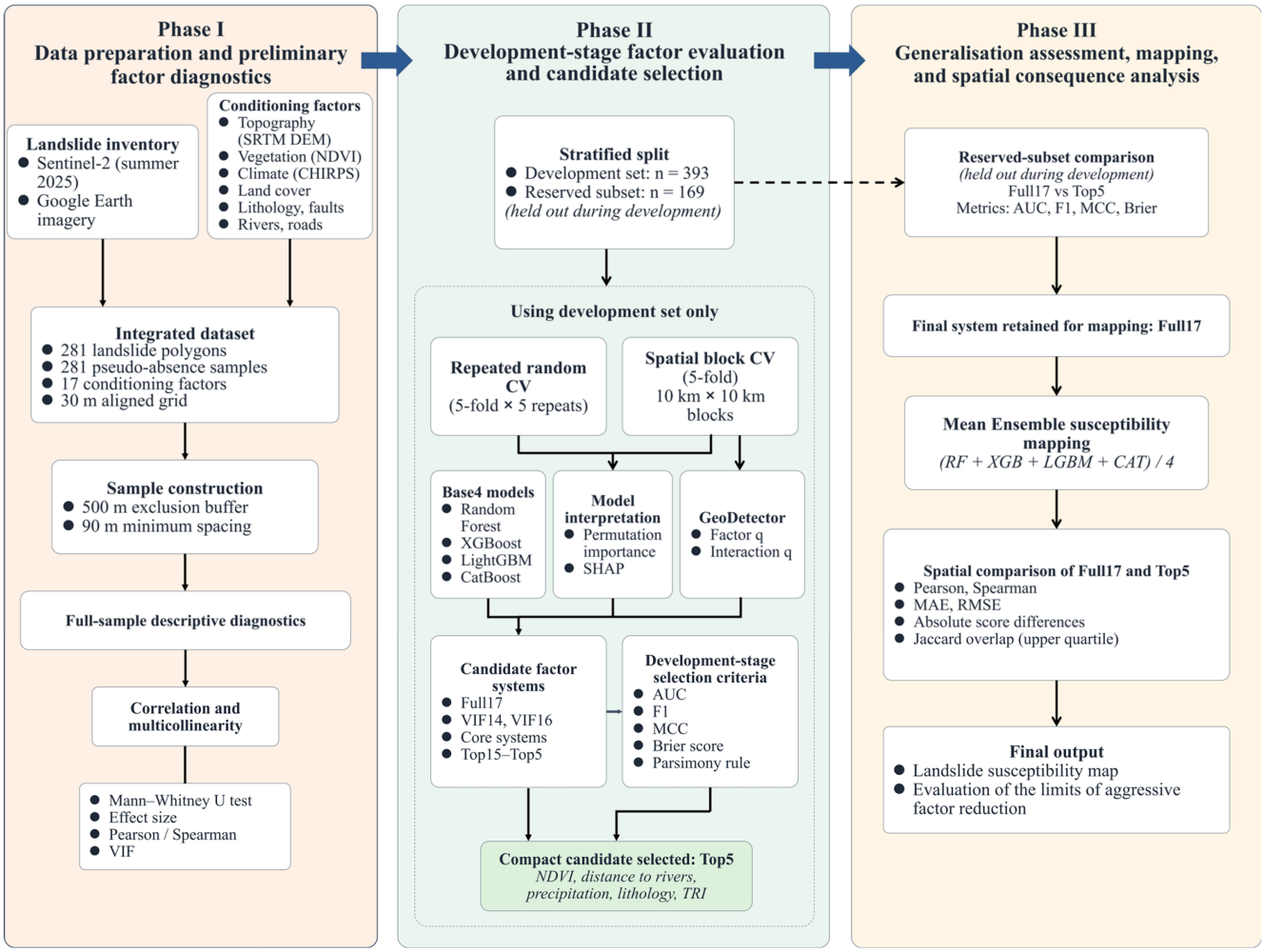


Figure 2. Technical roadmap.

All candidate systems were evaluated using only the 393-sample development dataset to assess predictive performance across models and validation designs. Conventional predictive performance was assessed using repeated stratified five-fold cross-validation with five repeats, yielding 25 validation folds for each combination of model and factor system. Spatial generalisation was additionally evaluated using five-fold spatial block cross-validation [19,20]. The 393 development samples were grouped into 10 km × 10 km spatial blocks and allocated to folds using StratifiedGroupKFold, with all samples from the same block retained within a single fold. Training and validation sets therefore contained non-overlapping blocks within each fold. The 10 km value denotes block side length rather than a minimum separation distance, so samples in adjacent blocks could remain geographically close. The grouping strategy preserved block-level separation between training and validation data while maintaining adequate sample availability and class representation across the five folds

under the spatial validation design. This block size represented a study-specific practical compromise between reducing local spatial dependence and maintaining sufficient spatial groups and class representation for five-fold validation. Identical random and spatial folds were used consistently across candidate systems to ensure direct comparability, and all data-dependent preprocessing was fitted exclusively to the corresponding training subsets before application to validation data.

Table 2. Definitions of the twelve candidate conditioning factor combinations.

Combination	*n*	Definition
Full17	17	All 17 conditioning factors
VIF A16	16	Full17 excluding Slope
VIF B16	16	Full17 excluding TRI
Core TRI9	9	Top5 plus cosine-transformed aspect, TWI, distance to roads, and distance to faults
Core Slope9	9	Core TRI9 with Slope replacing TRI
Core TRI LC10	10	Core TRI9 plus Land cover
Core Slope LC10	10	Core Slope9 plus Land cover
Top5	5	NDVI, distance to rivers, precipitation, lithology, and TRI
Top8	8	Top5 plus distance to roads, Slope, and cosine-transformed aspect
Top10	10	Top8 plus distance to faults and TWI
Top12	12	Top10 plus TPI and Relief
Top15	15	Top12 plus Land cover, Elevation, and Profile curvature

Random Forest, XGBoost, LightGBM, and CatBoost were collectively termed Base4 and formed the basis of formal candidate evaluation, whereas the arithmetic Mean Ensemble was retained only for descriptive score aggregation and regional susceptibility mapping. Performance was assessed using AUC, F1 score, Matthews correlation coefficient (MCC), and Brier score. The Brier score measured the mean squared difference between predicted probabilities and observed binary outcomes [46] and was calculated as follows:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \quad (4)$$

where p_i is the predicted landslide probability for sample i, y_i is the observed binary outcome, and N is the number of samples. Lower Brier scores indicate better probabilistic prediction performance. MCC was used to evaluate classification performance while accounting for all four elements of the confusion matrix and was defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. MCC ranges from −1 to 1, with values closer to 1 indicating better classification performance. A probability threshold of 0.50 was applied to F1 and MCC calculations. Relative to Full17, a reduced factor system was considered acceptable when the decrease in Base4 mean AUC did not exceed 0.010 under either random or spatial validation, the decrease in spatial F1 or MCC did not exceed 0.020, no individual model spatial AUC decreased by more than 0.030, and the spatial Brier rejection criterion was not triggered. That criterion required both an increase for every base model and a Base4 mean increase greater than 0.010. When multiple systems satisfied these criteria, the system

containing the fewest factors was retained according to the predefined parsimony rule. To further characterise the effect of progressive factor reduction, model-specific responses were examined across Full17, Top15, Top12, Top10, Top8, and Top5 using spatial validation results for individual folds. For each Base4 model, paired differences in AUC, F1, MCC, and Brier score were calculated relative to Full17 using the same five spatial folds, and mean differences with descriptive 95% t-based intervals were summarised across folds. Training sets overlap across these five folds, limiting independent-replicate interpretation. Positive differences indicate improvement relative to Full17 for AUC, F1, and MCC, whereas negative differences indicate improvement for the Brier score. This gradient analysis was used to characterise model-specific sensitivity to the degree of factor reduction and did not constitute an additional criterion for candidate selection.

2.5. Generalisation Assessment and Spatial Consequence Analysis

After development stage candidate evaluation, the selected compact factor system was compared with Full17 using the 169-sample reserved evaluation subset to assess whether the reduction selected during development remained robust beyond the resampling procedure used for candidate selection. The reserved subset had not been used for model fitting, model interpretation, GeoDetector analysis, threshold determination, or candidate evaluation after the 70:30 split. However, because these samples had contributed to the initial full sample descriptive diagnostics, the comparison was treated as a post-selection generalisation assessment rather than as a strictly independent validation. Both factor systems were fitted using the 393-sample development dataset under identical Base4 configurations and preprocessing procedures, and their performance was compared using AUC, F1 score, MCC, and Brier score. The Mean Ensemble was retained only as a descriptive summary and did not participate in the formal comparison.

Paired differences were defined as Top5 minus Full17. For each of 10,000 bootstrap iterations (seed 42), 84 landslide-reference and 85 pseudo-absence observations were resampled with replacement from the saved reserved predictions, using identical indices across factor systems and models. The Base4 difference was calculated by averaging the four model differences within each iteration. The 2.5th and 97.5th percentiles defined the paired 95% intervals. These intervals quantify uncertainty conditional on the fitted predictions and observed class counts, while the reserved comparison remains post-selection. Individual-model and Base4 mean intervals are presented with the reserved-subset results.

Spatial map comparison was further used to determine whether factor reduction preserved the regional susceptibility pattern represented by the complete factor system. Following the reserved subset assessment, the factor system retained for formal regional mapping was refitted using all 562 samples. For each 30 m prediction cell, the Mean Ensemble probability was calculated as the arithmetic mean of the susceptibility probabilities produced by the four Base4 models:

$$P_{ensemble} = \frac{P_{RF} + P_{XGB} + P_{LGBM} + P_{CAT}}{4} \quad (6)$$

where the four terms in the numerator are the predicted landslide probabilities produced by Random Forest, XGBoost, LightGBM, and CatBoost, respectively. This refitting with all samples was used exclusively for regional mapping and was not regarded as an additional performance evaluation. A corresponding surface from the compact factor system was generated for diagnostic comparison with Full17. The surfaces were compared on the same 5,035,125 valid 30 m cells. Pearson and Spearman coefficients measured linear and rank association. With D = Top5 − Full17, the mean of D measured signed difference, mean absolute error (MAE) was the mean of $|D|$, and root-mean-square error (RMSE)

was the square root of the mean of D^2; the proportions of the mapped area with absolute differences greater than 0.05, 0.10, and 0.20 were also calculated. Each surface defined its upper-quartile set using its own 75th-percentile score; the Jaccard index was the number of cells in the intersection divided by that in the union of these sets. Together, these metrics were used to assess whether apparent predictive similarity between factor systems was accompanied by comparable spatial susceptibility patterns.

3. Results

3.1. Environmental Contrasts, Redundancy and Cross-Model Factor Evidence

The initial full-sample contrasts characterised landslide-reference locations by lower NDVI, higher relief, terrain ruggedness and slope, and shorter river distances. Nine of the 15 continuous conditioning factors differed significantly between landslide-reference and pseudo-absence samples after false-discovery-rate correction (Figure 3a). NDVI showed the strongest separation, with a markedly lower median for landslide-reference samples than for pseudo-absence samples (0.194 versus 0.530; Cliff's delta = −0.581), corresponding to a large effect. Relief, TRI, and slope showed medium positive effects, with Cliff's delta values of 0.398, 0.388, and 0.384, respectively. TWI, cosine-transformed aspect, and distance to rivers showed small negative effects (−0.216, −0.213, and −0.204), whereas elevation and sine-transformed aspect were statistically significant but had negligible effect sizes (−0.143 and 0.107). Plan curvature, profile curvature, TPI, distance to faults, precipitation, and distance to roads showed no significant univariate differences after correction. Lithology and land cover also differed significantly between the two sample groups, with Cramér's V values of 0.440 and 0.347, respectively, and permutation p values of 0.0001 for both variables. Because their contingency tables contained sparse categories, these statistics were interpreted at the overall factor level rather than for individual rare classes.

Six strongly associated continuous-factor pairs were identified (Figure 3b), with the most pronounced correlations occurring among slope, TRI, and relief. Slope and TRI yielded Pearson and Spearman coefficients of 0.975 and 0.992, respectively; the corresponding coefficients were 0.939 and 0.944 for slope and relief, and 0.937 and 0.957 for relief and TRI. Strong associations were also observed for profile curvature–TPI, elevation–precipitation, and plan curvature–TPI. Consistent with these correlations, the initial VIF values reached 27.45 for slope and 25.57 for TRI, whereas relief had a VIF of 9.97 (Figure 3c). The redundancy diagnostics therefore identified slope, TRI, and relief as the most strongly overlapping components of the initial factor system, but no factor was removed solely on the basis of correlation or VIF.

The development-only analysis identified eight FDR-significant continuous factors and confirmed the principal environmental contrasts (Table 3). Initial development VIF reached 67.02, while full-sample VIF values are reported descriptively in Figure 3. The development-only analysis reproduced the same Top5 subset and ranking order: NDVI, distance to rivers, precipitation, lithology, and TRI.

Interpretation across the four models produced a broadly consistent factor evidence structure for Full17 across Random Forest, XGBoost, LightGBM, and CatBoost despite differences in their detailed rankings (Figure 4). NDVI, distance to rivers, and precipitation formed the high support group in the integrated permutation importance and SHAP evidence. Slope, cosine-transformed aspect, TRI, TPI, TWI, distance to faults, distance to roads, and lithology received moderate support, whereas elevation, sine-transformed aspect, plan curvature, profile curvature, relief, and land cover received comparatively low support. The evidence patterns were not identical to the univariate or redundancy results: precipitation showed no significant univariate difference but received high support

across models, whereas relief showed a medium univariate effect but comparatively low model-based support after the remaining factors were considered.

Table 3. Development-only factor diagnostics and integrated ranking (n = 393; 197 landslide-reference samples and 196 pseudo-absence samples).

Factor	Combined Rank	Effect Statistic	p Value	Initial VIF
NDVI	1.000	$\delta = -0.5914$	5.488×10^{-23}	1.8171
Distance to rivers	3.135	$\delta = -0.2153$	0.0005562	1.6737
Precipitation	3.370	$\delta = 0.0835$	0.2285	2.4316
Lithology	6.050	V = 0.4395	0.0001	—
TRI	7.400	$\delta = 0.3843$	2.219×10^{-10}	67.0157
Distance to roads	8.275	$\delta = -0.0348$	0.636	1.2683
Slope	8.600	$\delta = 0.3874$	2.219×10^{-10}	60.5999
Aspect cosine	8.900	$\delta = -0.2027$	0.001092	1.1990
Distance to faults	8.910	$\delta = 0.0911$	0.1972	1.2999
TWI	9.910	$\delta = -0.2308$	0.0002284	2.5832
TPI	10.455	$\delta = -0.0584$	0.3958	3.3426
Relief	11.245	$\delta = 0.3669$	1.189×10^{-9}	11.4131
Land cover	11.775	V = 0.3484	0.0001	—
Elevation	12.790	$\delta = -0.1601$	0.01138	4.4802
Profile curvature	13.010	$\delta = 0.0132$	0.8798	2.5294
Plan curvature	13.475	$\delta = -0.0027$	0.9628	1.5795
Aspect sine	14.700	$\delta = 0.0790$	0.2393	1.0824

For continuous factors, δ is Cliff's delta and p values are Benjamini–Hochberg adjusted. For lithology and land cover, V is Cramér's V and p values are from 10,000 label permutations. VIF applies to continuous factors. Combined rank averages cross-model mean permutation and SHAP ranks. The development-only ranking identified NDVI, distance to rivers, precipitation, lithology, and TRI as the Top5 factors.

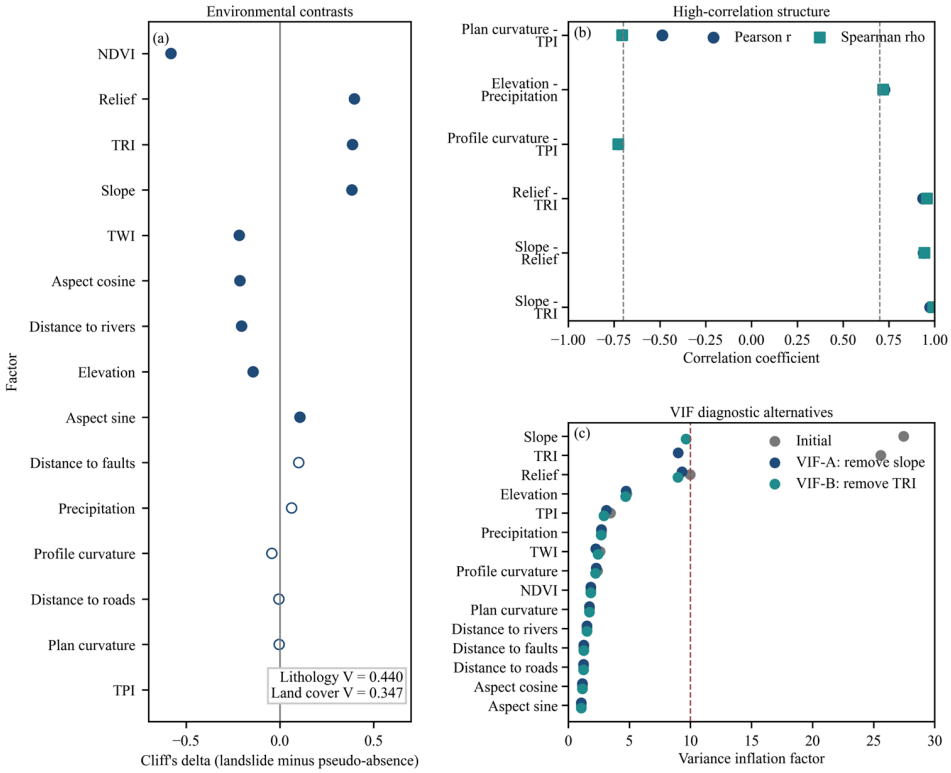


Figure 3. Initial conditioning-factor diagnostics based on the full 562-sample table: (**a**) Cliff's delta for continuous-factor contrasts, with filled circles indicating FDR-adjusted $p < 0.05$ and open circles indicating non-significant contrasts; (**b**) Pearson and Spearman coefficients for strongly correlated continuous-factor pairs; and (**c**) VIF-based diagnostic alternatives.

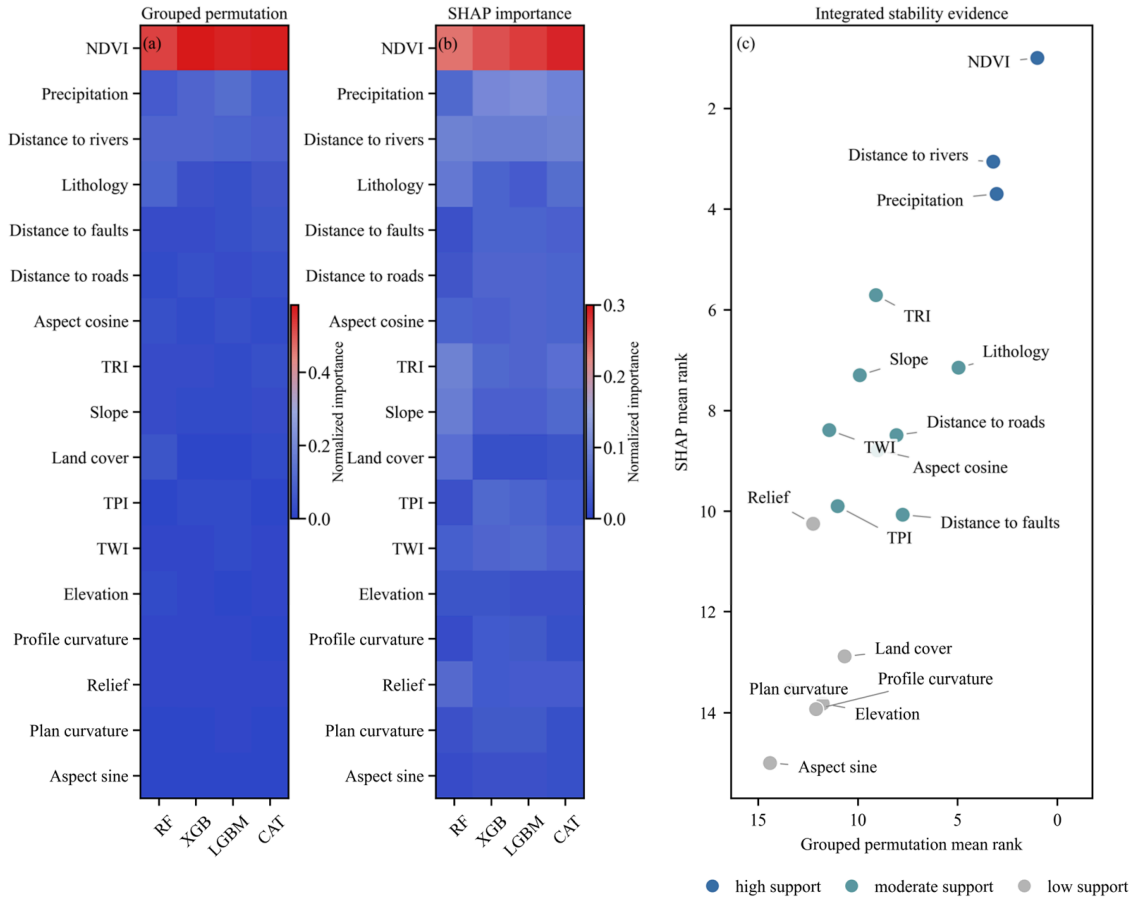


Figure 4. Factor evidence across models during development: (**a**) grouped original-factor permutation importance, (**b**) aggregated SHAP importance, and (**c**) integrated rankings from permutation importance and SHAP. Lower rank values in (**c**) indicate stronger support across models.

3.2. GeoDetector Factor Differentiation and Interaction Patterns

Under the primary five-quantile discretisation, NDVI had the highest GeoDetector q value (q = 0.278), followed by lithology (q = 0.193), land cover (q = 0.121), TRI (q = 0.113), slope (q = 0.111), and relief (q = 0.090) (Figure 5a). After 5000 label permutations and Benjamini–Hochberg correction across the 17 conditioning factors, significant differentiation was identified for slope, cosine-transformed aspect, plan curvature, relief, TRI, TWI, distance to rivers, NDVI, lithology, and land cover. The relative factor ranking remained highly stable under alternative discretisation schemes, with Spearman ρ values of 0.996 between the four- and five-quantile schemes, 0.964 between the four- and six-quantile schemes, and 0.968 between the five- and six-quantile schemes.

All 136 pairwise factor combinations were classified as enhancement interactions, comprising 116 nonlinear enhancement and 20 bivariate enhancement relationships. Among interactions without sparse joint stratum warnings, NDVI–precipitation produced the strongest nonlinear enhancement, with a joint q value of 0.386 and an interaction gain of 0.109 relative to the larger individual-factor q value (Figure 5b). Elevation–NDVI also showed nonlinear enhancement (q = 0.354, gain = 0.076). The leading bivariate enhancement relationships were TRI–NDVI (q = 0.374, gain = 0.097), slope–NDVI (q = 0.369, gain = 0.091), relief–NDVI (q = 0.353, gain = 0.075), distance to rivers–NDVI (q = 0.336, gain = 0.058), and TWI–NDVI (q = 0.323, gain = 0.045). Thus, several factors with modest individual q values exhibited substantially stronger differentiation when considered jointly with NDVI.

The highest overall interaction q value occurred for NDVI–lithology (q = 0.473, gain = 0.196), which was classified as nonlinear enhancement. However, this interaction was flagged as sparse because at least 25% of its joint strata contained fewer than five samples. Several other lithology-related combinations, including relief–lithology, TRI–lithology, slope–lithology, TWI–lithology, and cosine-transformed aspect–lithology, also had at least 25% of joint strata containing fewer than five samples. These sparse interactions were therefore retained as descriptive results but were not treated as primary evidence in constructing the candidate factor systems.

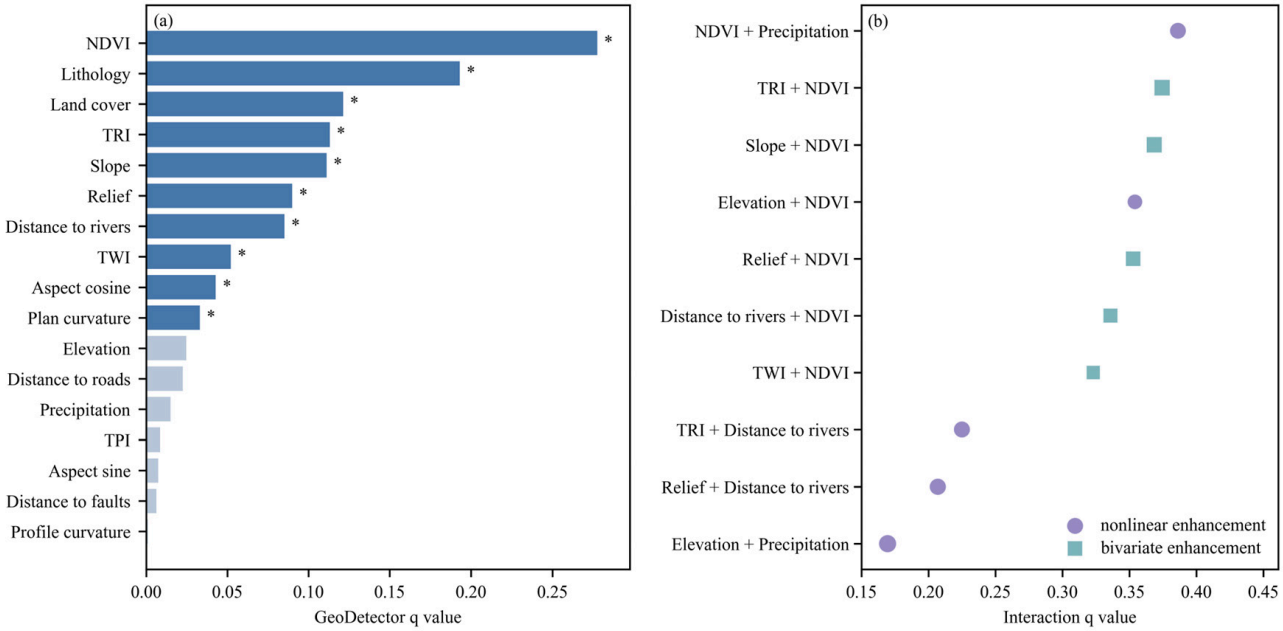


Figure 5. GeoDetector results for the development dataset: (**a**) factor level q values under the primary five-quantile discretisation and (**b**) selected pairwise interactions without sparse strata. Dark bars in (**a**) indicate FDR-significant factors and light bars indicate non-significant factors. Asterisks indicate FDR-adjusted significance, and marker size in (**b**) represents interaction gain relative to the larger individual-factor q value.

3.3. Development Stage Performance and Candidate Selection

The 12 conditioning factor systems showed relatively small differences in aggregate performance within the development dataset (Figure 6; Table 4). Under the predefined Base4 retention criteria, all 11 reduced systems remained within the acceptable range relative to Full17 under the combined random and spatial validation framework. Among these systems, Top5 represented the most aggressive reduction, retaining only NDVI, distance to rivers, precipitation, lithology, and TRI.

Top5 achieved Base4 mean AUC values of 0.866 under repeated random cross-validation and 0.846 under spatial block cross-validation, compared with 0.875 and 0.853 for Full17, respectively. Top5 therefore satisfied all predefined retention criteria. Because all reduced systems met these requirements, candidate selection was determined by the predefined parsimony rule rather than by the highest observed validation score. Top5 contained the fewest conditioning factors and was consequently retained as the compact candidate for subsequent reserved subset assessment. This selection indicates only that Top5 remained within the predefined tolerance limits during development validation; its performance beyond the development dataset was assessed separately using the reserved subset.

Table 4. Base4 mean development-stage performance of the twelve conditioning factor systems under random and spatial cross-validation.

Combination	*n*	Random AUC	Random F1	Random MCC	Random Brier	Spatial AUC	Spatial F1	Spatial MCC	Spatial Brier
Full17	17	0.875	0.792	0.580	0.152	0.853	0.772	0.541	0.164
VIF A16	16	0.875	0.793	0.582	0.153	0.852	0.779	0.552	0.165
VIF B16	16	0.876	0.796	0.586	0.151	0.855	0.779	0.555	0.163
Core TRI9	9	0.883	0.806	0.607	0.144	0.855	0.787	0.566	0.162
Core Slope9	9	0.885	0.805	0.604	0.143	0.859	0.782	0.559	0.160
Core TRI LC10	10	0.886	0.809	0.613	0.142	0.857	0.788	0.571	0.161
Core Slope LC10	10	0.888	0.809	0.611	0.141	0.861	0.791	0.574	0.159
Top5	5	0.866	0.785	0.568	0.155	0.846	0.769	0.532	0.169
Top8	8	0.881	0.797	0.591	0.144	0.858	0.781	0.557	0.162
Top10	10	0.883	0.803	0.601	0.144	0.858	0.786	0.570	0.161
Top12	12	0.878	0.797	0.589	0.147	0.848	0.777	0.544	0.165
Top15	15	0.880	0.801	0.598	0.147	0.854	0.779	0.554	0.164

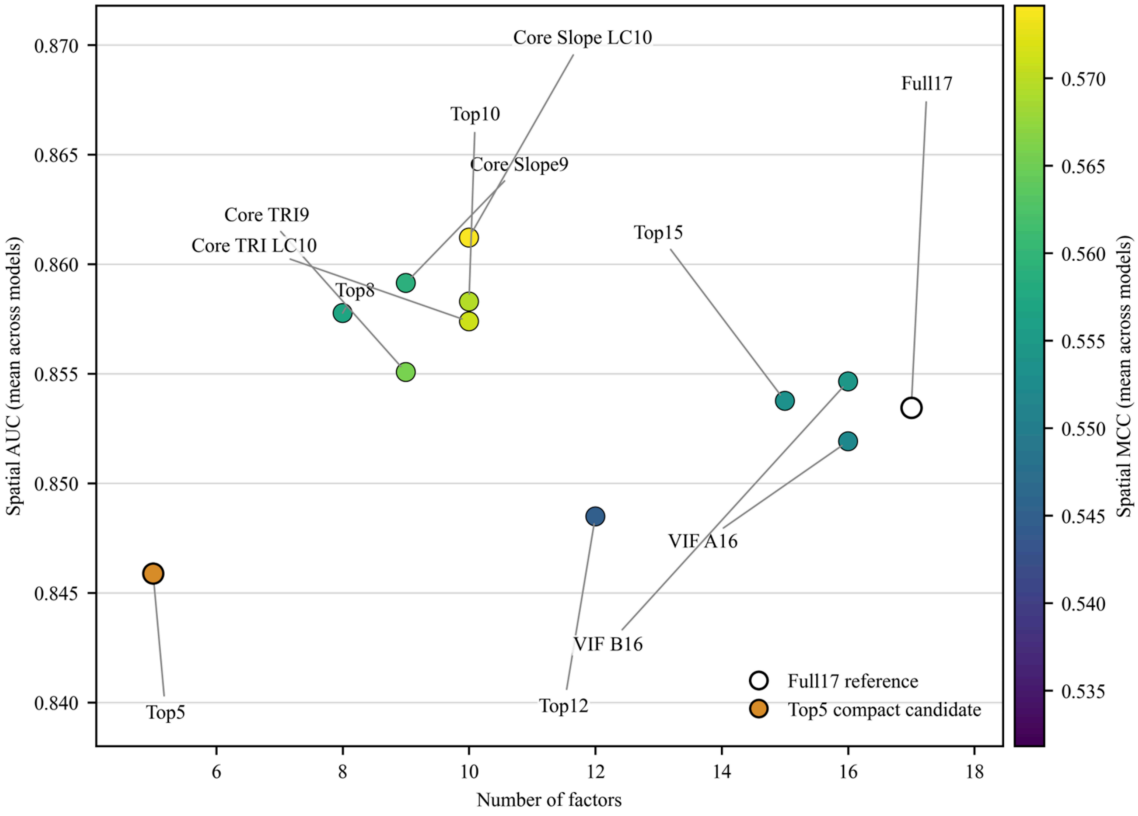


Figure 6. Development-stage performance-complexity comparison of the twelve candidate conditioning factor combinations. Point position represents the Base4 mean spatial AUC, and colour represents the Base4 mean spatial MCC.

3.4. Reserved Subset Comparison and Observed Performance Loss

After the selection during development had been fixed, Full17 and Top5 were compared using the 169-sample reserved evaluation subset (Figure 7; Table 5). Because these samples had contributed to the initial full sample descriptive diagnostics, the comparison represents a post-selection generalisation assessment rather than a strictly independent validation. On the reserved subset, Full17 achieved Base4 mean AUC, F1, MCC, and Brier values of 0.895, 0.819, 0.629, and 0.133, respectively, whereas Top5 achieved corresponding values of 0.865, 0.777, 0.548, and 0.153. Relative to Full17, Top5 therefore showed decreases

of 0.0301 in AUC, 0.0416 in F1, and 0.0812 in MCC, together with a 0.0195 increase in Brier score. The paired Base4 AUC difference was −0.03009 (95% interval [−0.06157, −0.00107]); all four Base4 mean-difference intervals favoured Full17 (Table 6).

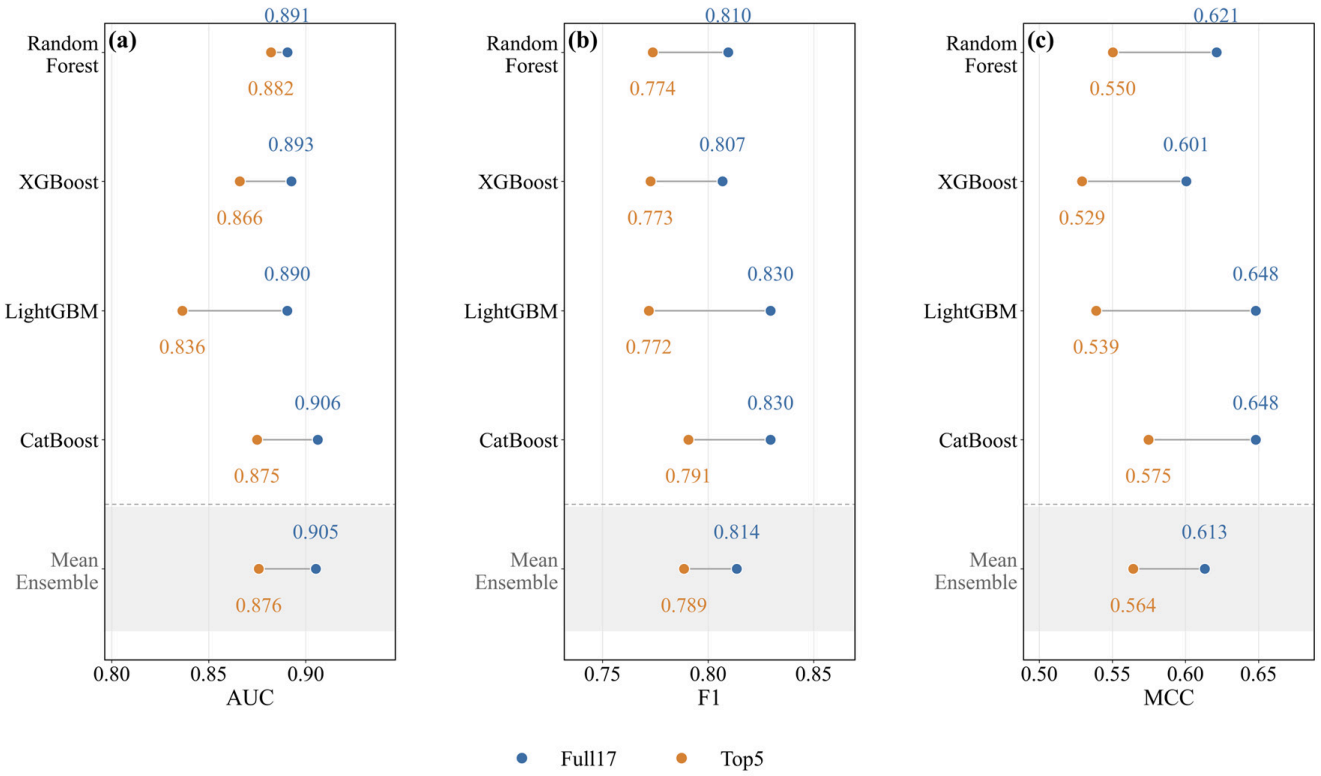


Figure 7. Reserved subset comparison of Full17 and Top5 across individual models: (**a**) AUC, (**b**) F1 score, and (**c**) MCC. Paired points represent results for the same model under the two factor systems and are connected to highlight their differences. The grey shaded area indicates the Mean Ensemble.

Table 5. Base4 performance of Full17 and Top5 in the reserved subset comparison.

Model	Factor System	AUC	F1	MCC	Brier
Random Forest	Full17	0.891	0.810	0.621	0.144
	Top5	0.882	0.774	0.550	0.139
XGBoost	Full17	0.893	0.807	0.601	0.130
	Top5	0.866	0.773	0.529	0.148
LightGBM	Full17	0.890	0.830	0.648	0.139
	Top5	0.836	0.772	0.539	0.177
CatBoost	Full17	0.906	0.830	0.648	0.121
	Top5	0.875	0.791	0.575	0.147
Base4 mean	Full17	0.895	0.819	0.629	0.133
	Top5	0.865	0.777	0.548	0.153
	Difference (Top5 − Full17)	−0.0301	−0.0416	−0.0812	+0.0195

The magnitude of the performance difference varied among the four base models (Figure 7). LightGBM showed the largest AUC decrease, from 0.890 with Full17 to 0.836 with Top5, corresponding to a reduction of 0.0541. The other base models also produced lower AUC values on the reserved subset under Top5 than under Full17, although the magnitude of the decrease differed among algorithms. Overall, the reserved subset comparison

showed consistently lower Base4 aggregate performance for Top5 than for Full17 across discrimination, classification, and probabilistic evaluation metrics. Only the LightGBM AUC interval excluded zero; the other three AUC intervals and all individual-model F1 and MCC intervals crossed zero (Table 6). LightGBM and CatBoost Brier intervals favoured Full17, whereas RF and XGBoost Brier intervals crossed zero. These intervals distinguish the average Base4 loss from individual-model uncertainty.

Table 6. Paired differences (Top5 − Full17) and 95% bootstrap intervals for the reserved subset (n = 169).

Model	Metric	Difference	95% Lower	95% Upper
Base4 mean	AUC	−0.03009	−0.06157	−0.00107
Base4 mean	F1	−0.04157	−0.08388	−0.00203
Base4 mean	MCC	−0.08121	−0.16392	−0.00293
Base4 mean	Brier	+0.01946	+0.00050	+0.03965
RF	AUC	−0.00847	−0.04195	+0.02577
RF	F1	−0.03571	−0.09736	+0.02443
RF	MCC	−0.07101	−0.18975	+0.04625
RF	Brier	−0.00495	−0.02121	+0.01183
XGBoost	AUC	−0.02661	−0.06163	+0.00658
XGBoost	F1	−0.03409	−0.09404	+0.02418
XGBoost	MCC	−0.07129	−0.18884	+0.04258
XGBoost	Brier	+0.01861	−0.00323	+0.04072
LightGBM	AUC	−0.05406	−0.09874	−0.01331
LightGBM	F1	−0.05762	−0.12243	+0.00155
LightGBM	MCC	−0.10918	−0.23601	+0.00990
LightGBM	Brier	+0.03786	+0.00492	+0.07187
CatBoost	AUC	−0.03123	−0.06681	+0.00238
CatBoost	F1	−0.03885	−0.09265	+0.01190
CatBoost	MCC	−0.07335	−0.17827	+0.03064
CatBoost	Brier	+0.02633	+0.00482	+0.04939

3.5. Final Susceptibility Mapping and Spatial Comparison

Following the reserved subset comparison, Full17 was retained as the factor system for formal regional susceptibility mapping (Figure 8). The four Base4 models were refitted using all 562 samples, and their continuous susceptibility scores were averaged to generate the final Mean Ensemble susceptibility surface. This refitting with all samples was used exclusively for regional mapping and was not treated as an additional performance evaluation. The valid rasterised mapping area was 4531.61 km^2. The slight difference from the vector boundary area of 4526.38 km^2 resulted from rasterisation to the 30 m analysis grid and the inclusion of boundary-intersecting valid cells. Very low susceptibility occupied 1837.14 km^2 (40.54%), followed by low susceptibility at 959.41 km^2 (21.17%) and moderate susceptibility at 630.85 km^2 (13.92%). High and very high susceptibility covered 524.83 km^2 (11.58%) and 579.39 km^2 (12.79%), respectively. Together, the high and very high classes accounted for 1104.22 km^2, or 24.37% of the valid mapped area. Overall, the mapped area was therefore dominated by the very low and low susceptibility classes, which together accounted for 61.71%, while approximately one quarter of the region was assigned to the two highest susceptibility classes.

A diagnostic comparison between the Full17 and Top5 Mean Ensemble surfaces was further conducted to quantify the spatial consequences of factor reduction (Figure 9). The two continuous susceptibility surfaces remained strongly associated at the regional scale, with Pearson and Spearman correlation coefficients of 0.901 and 0.922, respectively. Nevertheless, the mean absolute error and root-mean-square error between the two surfaces

were 0.090 and 0.133. The mean signed difference, calculated as Top5 minus Full17, was 0.023, indicating a slightly higher average susceptibility score for Top5 across the valid mapping area. Absolute susceptibility score differences greater than 0.05 occurred across 50.83% of the mapped area, while differences greater than 0.10 and 0.20 covered 32.02% and 13.43%, respectively. In addition, the upper quartile susceptibility zones derived from the two systems had a Jaccard overlap of 0.652, indicating only partial correspondence between their highest-ranked spatial areas. Broad map correlation therefore did not imply spatial equivalence; these differences indicate potential loss of spatial susceptibility information after reduction.

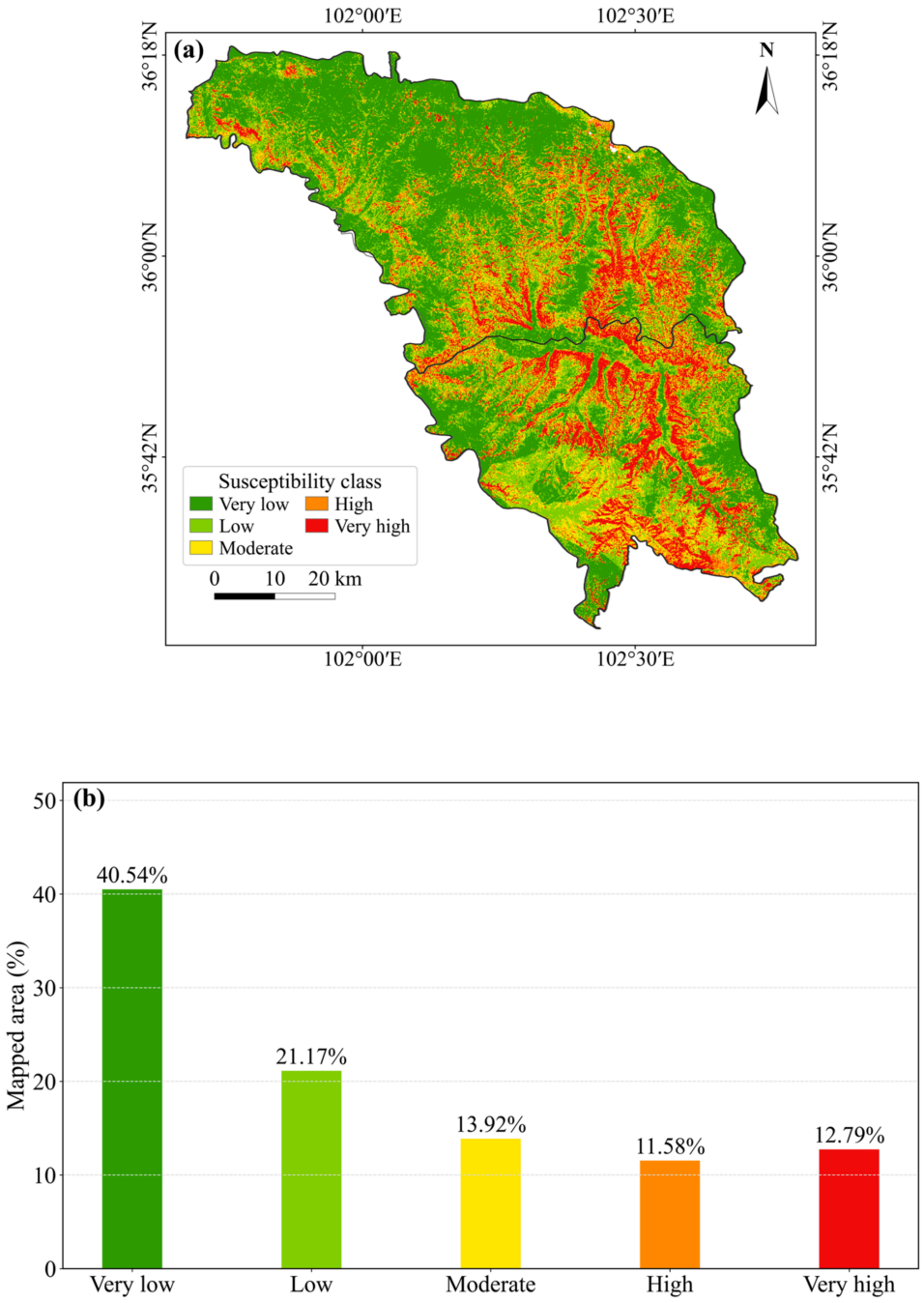


Figure 8. Regional landslide susceptibility mapping based on the Full17 Mean Ensemble: (**a**) five-class susceptibility map and (**b**) mapped-area shares of the five susceptibility classes.

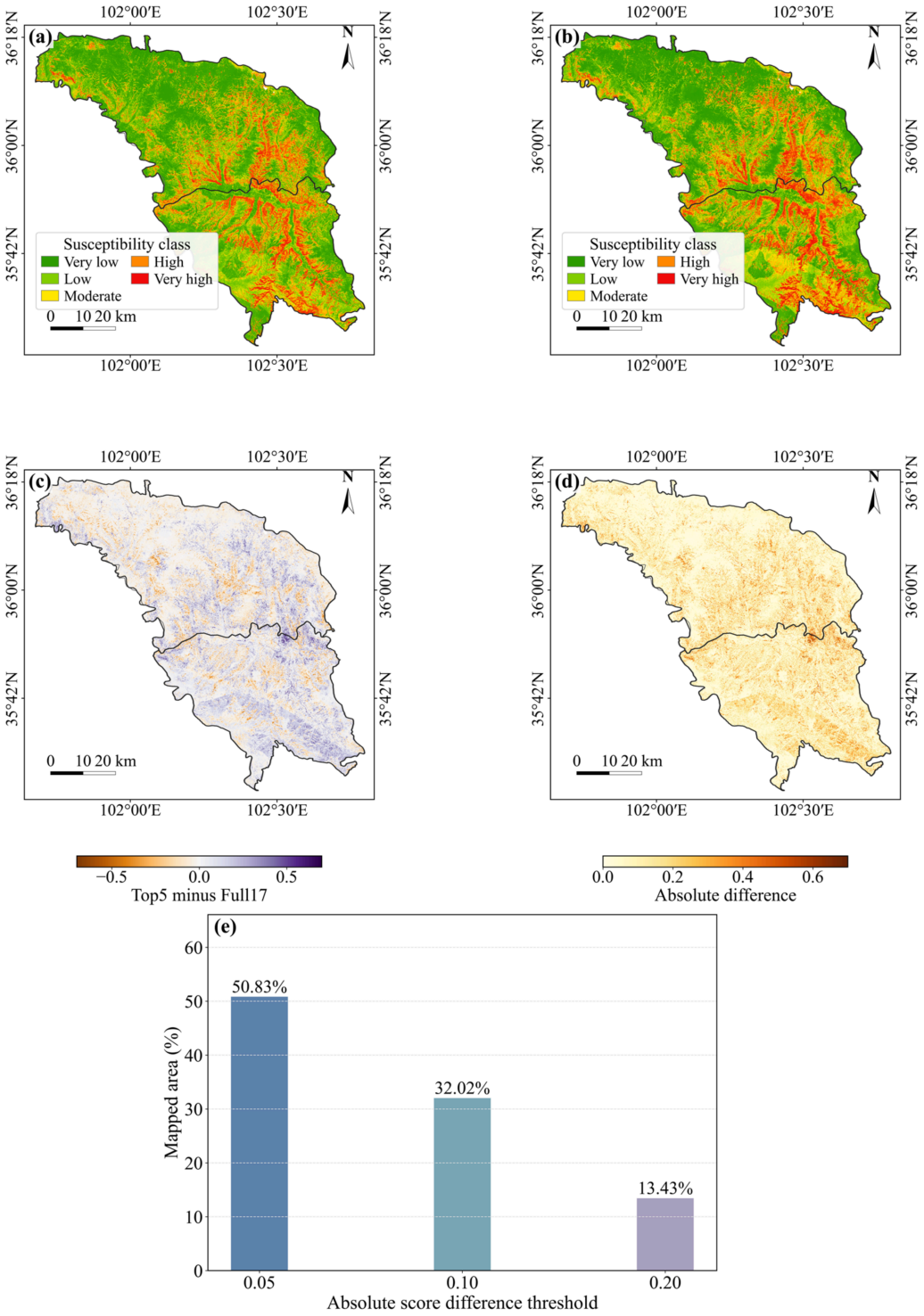


Figure 9. Diagnostic spatial comparison of the Full17 and Top5 Mean Ensemble susceptibility score surfaces: (**a**) Full17 five-class susceptibility pattern, (**b**) Top5 diagnostic five-class susceptibility pattern, (**c**) Top5 minus Full17 signed score difference, (**d**) absolute score difference, and (**e**) mapped-area proportions exceeding selected absolute-difference thresholds. Together, the panels illustrate the spatial consequences of aggressive factor reduction.

4. Discussion

4.1. Model-Specific Responses and the Limits of Aggressive Factor Reduction

The progressive reduction experiment revealed clear model-specific and nonmonotonic responses to removal of conditioning factors under spatial validation (Figure 10). Random Forest improved at Top5, while the three boosting models had lower AUC

than Full17. More importantly, the intermediate Top-N systems did not produce monotonic trajectories for any of the four algorithms. These results indicate that the effect of factor reduction depended on both the learning algorithm and the degree of reduction, and that a small change in the Base4 average can conceal substantially different model-specific responses.

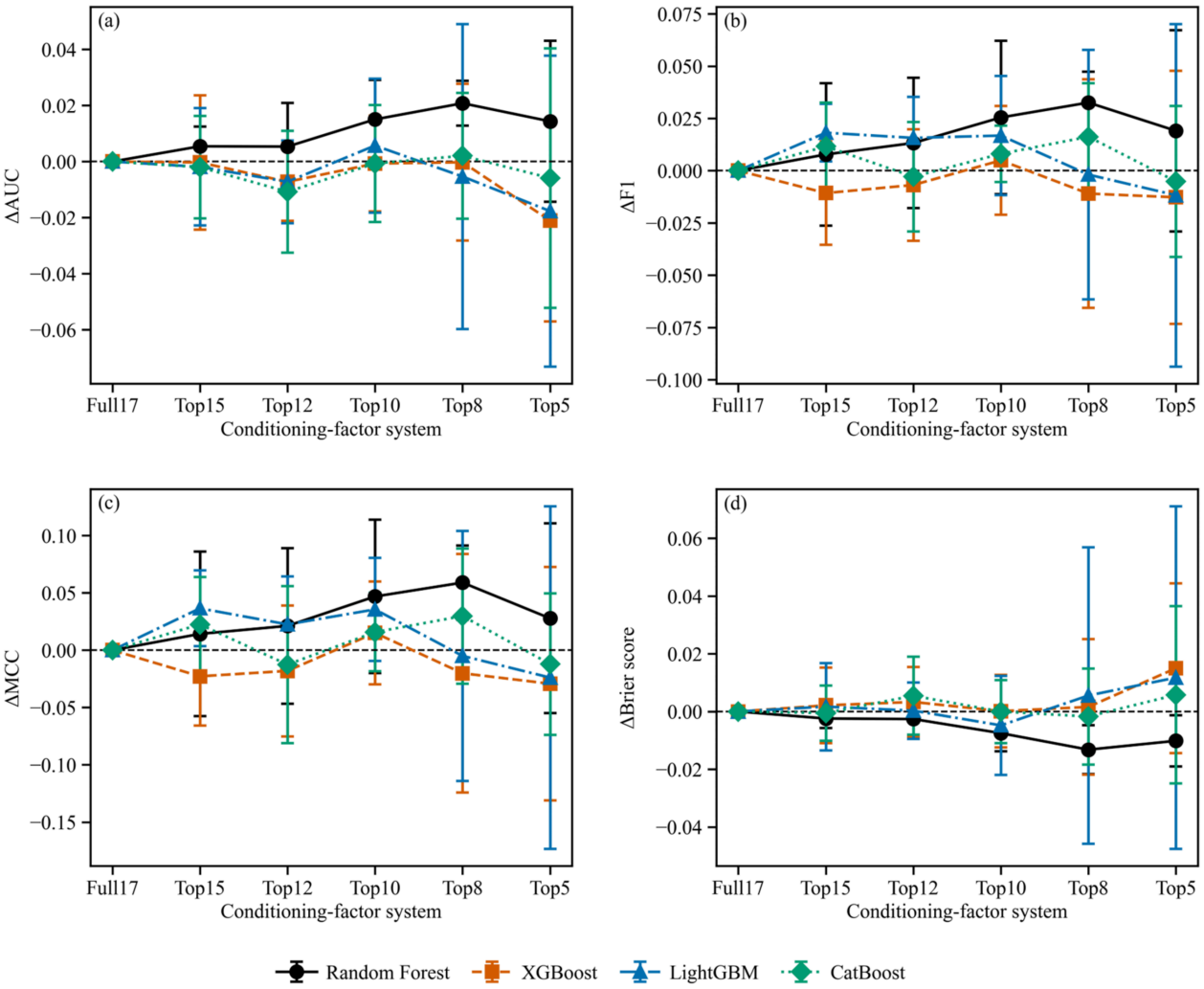


Figure 10. Model-specific responses to progressive conditioning-factor reduction under spatial block cross-validation: changes relative to Full17 in (**a**) AUC, (**b**) F1 score, (**c**) MCC, and (**d**) Brier score. Error bars indicate descriptive 95% t-based intervals of paired differences across the five spatial folds; training sets overlap across folds. For the plotted differences, positive values indicate improvement for AUC, F1, and MCC, whereas negative values indicate improvement for Brier score.

The reserved subset comparison further showed that apparent stability during development validation did not establish that the compact factor system could substitute for the complete representation. All four Base4 mean-difference intervals favoured Full17, while most individual-model intervals crossed zero (Tables 5 and 6). The development-only analysis reproduced Top5 (Table 3). The reserved-subset evidence remains post-selection because the initial full-sample diagnostics contributed to candidate construction.

The spatial results provide complementary evidence for evaluating factor reduction alongside aggregate predictive metrics. Despite strong map correlation, absolute score differences exceeded 0.10 across 32.02% of mapped area, with only partial overlap of upper-quartile areas (Figure 9). Broad map correlation therefore did not imply local spatial equivalence. These differences identify locations where regional screening priorities depend

on the factor system. Reliability of a reduced factor system therefore requires consistent evidence across algorithms, validation schemes, and spatial comparisons.

4.2. Implications for Conditioning Factor Selection in Landslide Susceptibility Modelling

Conditioning-factor selection therefore benefits from integrating complementary statistical and model-based evidence. Different diagnostics characterised different components of factor information and did not produce a simple one-to-one correspondence. Precipitation, for example, did not show a significant univariate difference between landslide-reference and pseudo-absence samples (Figure 3; Table 3), yet it was classified within the high support group by the integrated cross-model permutation importance and SHAP evidence (Figure 4). Relief showed the opposite pattern: it exhibited a medium univariate effect but comparatively low model-based support after the remaining factors were considered. GeoDetector provided additional evidence of complementarity: NDVI–precipitation produced the strongest non-sparse nonlinear enhancement (Figure 5). Factor importance describes contribution within a fitted model, not the exclusivity of that information.

These contrasts show that weak marginal differentiation, strong correlation, or a low ranking under one interpretation method does not necessarily demonstrate that a factor lacks complementary information. Recent studies have likewise evaluated feature selection, core factor-set identification, and factor-engineering strategies for reducing redundant conditioning information in landslide susceptibility modelling [13–15]. Redundancy does not by itself establish dispensability. These diagnostics should be used to construct candidate subsets rather than to delete variables automatically. Each candidate should be evaluated separately across multiple algorithms and under random and spatial validation, because factor removal produced different responses among models (Table 4; Figure 10).

Regional process knowledge should remain an explicit constraint throughout this procedure. Full17 was retained for final mapping in the present study because it provided broader environmental representation and stronger reserved-subset performance (Table 5; Figures 8 and 9). The maps can guide initial image review and field investigation priorities; individual slope stability requires site-level assessment.

4.3. Limitations and Future Perspectives

The proposed factor reduction framework was developed and tested within a single mountainous region, and its applicability to areas with different geological settings, climatic conditions, and geomorphological processes requires further verification. The OpenStreetMap road dataset was obtained from a June 2026 snapshot, whereas the landslide inventory was interpreted from imagery acquired in 2025. Potential road-network changes during this interval may have introduced limited temporal inconsistency into the distance-to-roads factor. The 30 m grid aligns inputs but does not recover detail absent from coarser source data.

Pseudo-absence samples represented comparison locations rather than field-confirmed stable slopes. They were located outside 500 m exclusion buffers around the landslide reference points, although unmapped or inactive landslides may remain within the pseudo-absence domain. The inventory was constructed from remote sensing interpretation and may still contain uncertainties associated with landslide boundary delineation, historical activity, and incomplete detection of small or vegetated landslides. One centroid per polygon does not retain internal geomorphic characteristics or activity state. Four lithological categories absent from the development dataset occurred in five reserved evaluation samples. These samples were retained by the unknown-category handling procedure during one-hot encoding, but category-specific effects for the unseen lithological classes could not be learned during model fitting.

The reserved evaluation subset came from the same inventory through a stratified random split and did not provide temporal validation or validation across regions. Initial full sample contrasts and redundancy diagnostics also informed the candidate evidence chain, creating a selection-independence limitation. The development-only diagnostics excluded reserved observations, although their use in the initial full-sample diagnostics keeps the reserved comparison post-selection. The paired intervals also condition on the fitted predictions. Future work should implement candidate construction within training folds and independent temporal or geographically separated validation.

5. Conclusions

This study evaluated whether a compact conditioning factor system could adequately replace a complete factor system for landslide susceptibility mapping in the Hualong–Xunhua region of the Upper Yellow River. First, different factor diagnostic methods provided complementary but not identical evidence. The precipitation example combined weak marginal differentiation with strong model and interaction support, indicating that single importance rankings and redundancy diagnostics alone cannot determine factor removability. Second, the compact Top5 system satisfied all predefined retention criteria during development and approached Full17 performance, demonstrating factor-reduction potential within the development setting, although responses to progressive reduction differed among models. Third, in the post-selection reserved comparison, Top5 showed weaker Base4 mean performance, supported by paired uncertainty estimates, while individual-model uncertainty varied. This result indicates that favourable development performance did not ensure reliable substitution of the complete system. Finally, Full17 was retained for final regional susceptibility mapping. Although the Full17 and Top5 maps remained strongly correlated, local score differences and incomplete upper-quartile overlap indicated potential loss of spatial information after reduction. Overall, aggressive factor reduction should not be justified solely by factor importance, redundancy diagnostics, or favourable internal validation performance.

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

Funding: This research was funded by the Qinghai University Research Ability Enhancement Project (Grant No. 2025KTSA04) and the Qinghai University Graduate Student Research and Practice Innovation Project (Grant No. qdkc-2553).

Institutional Review Board Statement: Not applicable. This study did not involve humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The publicly available data sources used in this study are described in Section 2.2 and Table 1.

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI ChatGPT (GPT-5.6) for English-language editing and improvement of clarity. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Reichenbach, P.; Rossi, M.; Malamud, B.D.; Mihir, M.; Guzzetti, F. A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **2018**, *180*, 60–91. [CrossRef]
2. van Westen, C.J.; Castellanos, E.; Kuriakose, S.L. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. *Eng. Geol.* **2008**, *102*, 112–131. [CrossRef]
3. Fell, R.; Corominas, J.; Bonnard, C.; Cascini, L.; Leroi, E.; Savage, W.Z. Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Eng. Geol.* **2008**, *102*, 85–98. [CrossRef]
4. Lima, P.; Steger, S.; Glade, T.; Murillo-García, F.G. Literature review and bibliometric analysis on data-driven assessment of landslide susceptibility. *J. Mt. Sci.* **2022**, *19*, 1670–1698. [CrossRef]
5. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; ThaiPham, B.; Bui, D.T.; Avtar, R.; Abderrahmane, B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci. Rev.* **2020**, *207*, 103225. [CrossRef]
6. Zêzere, J.L.; Pereira, S.; Melo, R.; Oliveira, S.C.; Garcia, R.A.C. Mapping landslide susceptibility using data-driven methods. *Sci. Total Environ.* **2017**, *589*, 250–267. [CrossRef] [PubMed]
7. Goetz, J.N.; Brenning, A.; Petschko, H.; Leopold, P. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci.* **2015**, *81*, 1–11. [CrossRef]
8. Chen, W.; Peng, J.; Hong, H.; Shahabi, H.; Pradhan, B.; Liu, J.; Zhu, A.X.; Pei, X.; Duan, Z. Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Sci. Total Environ.* **2018**, *626*, 1121–1135. [CrossRef] [PubMed]
9. Murillo-García, F.G.; Steger, S.; Alcántara-Ayala, I. Landslide susceptibility: A statistically-based assessment on a depositional pyroclastic ramp. *J. Mt. Sci.* **2019**, *16*, 561–580. [CrossRef]
10. Chen, T.; Zhu, L.; Niu, R.Q.; Trinder, C.J.; Peng, L.; Lei, T. Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models. *J. Mt. Sci.* **2020**, *17*, 670–685. [CrossRef]
11. Hussain, M.L.; Shafique, M.; Bacha, A.S.; Chen, X.Q.; Chen, H.Y. Landslide inventory and susceptibility assessment using multiple statistical approaches along the Karakoram highway, northern Pakistan. *J. Mt. Sci.* **2021**, *18*, 583–598. [CrossRef]
12. Dou, J.; Yunus, A.P.; Tien Bui, D.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.W.; Khosravi, K.; Yang, Y.; Pham, B.T. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. Total Environ.* **2019**, *662*, 332–346. [CrossRef] [PubMed]
13. Pawłuszek-Filipiak, K.; Lewandowski, T. The Impact of Feature Selection on XGBoost Performance in Landslide Susceptibility Mapping Using an Extended Set of Features: A Case Study from Southern Poland. *Appl. Sci.* **2025**, *15*, 8955. [CrossRef]
14. Yu, X.; Wang, H. Research on Core Factor Sets for Landslide Susceptibility Mapping Based on Interpretable Machine Learning Methods. *Appl. Sci.* **2026**, *16*, 4219. [CrossRef]
15. Zhang, L.; Ye, S.; He, D.; Wang, L.; Li, W.; Jin, B.; Zeng, T. Examination of Landslide Susceptibility Modeling Using Ensemble Learning and Factor Engineering. *Appl. Sci.* **2025**, *15*, 6192. [CrossRef]
16. Steger, S.; Brenning, A.; Bell, R.; Glade, T. The influence of systematically incomplete shallow landslide inventories on statistical susceptibility models and suggestions for improvements. *Landslides* **2017**, *14*, 1767–1781. [CrossRef]
17. Hussin, H.Y.; Zumpano, V.; Reichenbach, P.; Sterlacchini, S.; Micu, M.; van Westen, C.; Bălteanu, D. Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. *Geomorphology* **2016**, *253*, 508–523. [CrossRef]
18. Jacobs, L.; Kervyn, M.; Reichenbach, P.; Rossi, M.; Marchesini, I.; Alvioli, M.; Dewitte, O. Regional susceptibility assessments with heterogeneous landslide information: Slope unit- vs. pixel-based approach. *Geomorphology* **2020**, *356*, 107084. [CrossRef]
19. 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.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
20. Schratz, P.; Muenchow, J.; Iturritxa, E.; Richter, J.; Brenning, A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecol. Model.* **2019**, *406*, 109–120. [CrossRef]
21. Kumar, C.; Walton, G.; Santi, P.; Luza, C. Random Cross-Validation Produces Biased Assessment of Machine Learning Performance in Regional Landslide Susceptibility Prediction. *Remote Sens.* **2025**, *17*, 213. [CrossRef]
22. Barella, C.F.; Zêzere, J.L.; Fernandes, N.F. Validation and prediction challenges in data-driven landslide susceptibility mapping: Insights from random sampling of training and test data in machine learning models. *Bull. Eng. Geol. Environ.* **2026**, *85*, 253. [CrossRef]
23. Xing, S.; Guo, H.; Hu, X. Sources and enrichment processes of groundwater arsenite and arsenate in fissured bedrock aquifers in the Xunhua-Hualong basin, China. *Appl. Geochem.* **2023**, *155*, 105708. [CrossRef]
24. Ministry of Emergency Management of the People's Republic of China. Intensity Map of the Gansu Jishishan Magnitude-6.2 Earthquake. 22 December 2023. Available online: https://www.mem.gov.cn/xw/yjglbgzdt/202312/t20231222_472849.shtml (accessed on 13 September 2026). (In Chinese)

25. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
26. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
27. Dornik, A.; Drăguţ, L.; Oguchi, T.; Hayakawa, Y.; Micu, M. Influence of sampling design on landslide susceptibility modeling in lithologically heterogeneous areas. *Sci. Rep.* **2022**, *12*, 2106. [CrossRef] [PubMed]
28. Farr, T.G.; Rosen, P.A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. The Shuttle Radar Topography Mission. *Rev. Geophys.* **2007**, *45*, 2005RG000183. [CrossRef]
29. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A.; et al. The climate hazards infrared precipitation with stations—A new environmental record for monitoring extremes. *Sci. Data* **2015**, *2*, 150066. [CrossRef]
30. Zanaga, D.; Van De Kerchove, R.; Daems, D.; De Keersmaecker, W.; Brockmann, C.; Kirches, G.; Wevers, J.; Cartus, O.; Santoro, M.; Fritz, S.; et al. *ESA WorldCover 10 m 2021 v200*; Zenodo: Geneva, Switzerland, 2022. [CrossRef]
31. Lehner, B.; Grill, G. Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. *Hydrol. Processes* **2013**, *27*, 2171–2186. [CrossRef]
32. Haklay, M.; Weber, P. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
33. GDAL/OGR Contributors. gdalwarp: Image Reprojection and Warping. GDAL Documentation. Available online: https://gdal.org/en/stable/programs/gdalwarp.html (accessed on 13 September 2026).
34. Mann, H.B.; Whitney, D.R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]
35. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B* **1995**, *57*, 289–300. [CrossRef]
36. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* **1993**, *114*, 494–509. [CrossRef]
37. Romano, J.; Kromrey, J.D.; Coraggio, J.; Skowronek, J. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? In Proceedings of the Annual Meeting of the Florida Association of Institutional Research, Cocoa Beach, FL, USA, 1–3 February 2006; pp. 1–33.
38. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J.R.G.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. [CrossRef]
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Chen, T.; Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
41. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 3146–3154. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (accessed on 13 September 2026).
42. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 6639–6649. Available online: https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html (accessed on 13 September 2026).
43. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 4765–4774. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed on 13 September 2026).
44. Wang, J.F.; Li, X.H.; Christakos, G.; Liao, Y.L.; Zhang, T.; Gu, X.; Zheng, X.Y. Geographical Detectors-Based Health Risk Assessment and its Application in the Neural Tube Defects Study of the Heshun Region, China. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 107–127. [CrossRef]
45. Wang, J.F.; Zhang, T.L.; Fu, B.J. A measure of spatial stratified heterogeneity. *Ecol. Indic.* **2016**, *67*, 250–256. [CrossRef]
46. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.