

# Monte Carlo Positive-Unlabeled Two-Step Learning for Landslide Susceptibility Mapping under Inventory Uncertainty: Comparison with Hard-Threshold PU Two-Step and EIV in Northern Morocco

El harit Abdessamad ADRI

[adrielharitabdessamad@gmail.com](mailto:adrielharitabdessamad@gmail.com)

Hassania School of Public Works

SINAN Mohamed

Hassania School of Public Works

FALLAHI Azzeddine

Mohammed V University

---

## Research Article

**Keywords:** landslide susceptibility, PU learning, Monte Carlo, uncertainty quantification, spatial cross-validation, northern Morocco

**Posted Date:** September 9th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10894069/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Monte Carlo Positive-Unlabeled Two-Step Learning for Landslide Susceptibility Mapping under Inventory Uncertainty: Comparison with Hard-Threshold PU Two-Step and EIV in Northern Morocco

**Authors.** ADRI El harit Abdessamad<sup>1</sup>, SINAN Mohamed<sup>2</sup>, FALLAHI Azzeddine<sup>3</sup>

<sup>1</sup>Hydrogeology, Water Treatment and Purification, and Climate Research Team, Laboratory of Civil Engineering, Hydraulics, Environment and Climate, Hassania School of Public Works (EHTP), Casablanca, Morocco.

<sup>2</sup>Head of Department, Hassania School of Public Works (EHTP), Casablanca, Morocco.

<sup>3</sup>Energy Research Center (CRE), Faculty of Sciences, Mohammed V University in Rabat, Morocco.

*Corresponding author:* [adrielharitabdessamad@gmail.com](mailto:adrielharitabdessamad@gmail.com)

## Abstract

Landslide susceptibility mapping depends critically on the definition of the non-landslide class. In many study areas, including the Rif of northern Morocco, the completeness of landslide inventories cannot be guaranteed, making conventional random absence sampling unreliable. This study proposes **MC-PU-2S** (Monte Carlo Positive-Unlabeled Two-Step), a probabilistic extension of classical PU Two-Step learning, and compares it with Enhanced Information Value (**EIV**) and deterministic **PU Two-Step** on a database of 5,877 observations (2,320 mapped landslides, 3,557 unlabeled) described by 21 conditioning factors. All three methods share the same Random Forest classifier and are evaluated under a five-fold spatial block cross-validation to prevent spatial leakage. EIV selects reliable negatives deterministically from the lowest-susceptibility quantile of a preliminary model. PU Two-Step applies a hard threshold (30th percentile) to an out-of-bag PU Bagging score. MC-PU-2S converts those scores into Bernoulli membership probabilities  $\pi_i = 1 - g(u_i)$ , resamples  $M$  plausible reliable-negative sets, and aggregates  $M$  Random Forest predictions into a mean susceptibility surface  $\hat{u}(x)$  and an uncertainty surface  $\hat{\sigma}(x)$ . The five-fold cross-validation yielded mean ROC-AUC values of 0.641 (EIV), 0.637 (PU Two-Step), and 0.673 (MC-PU-2S); mean Brier scores of 0.434, 0.389, and 0.236, respectively. Calibration analysis confirms that EIV and PU Two-Step produce poorly calibrated probabilities (slope beta approx. 0.45), whereas MC-PU-2S approaches the ideal. Beyond predictive metrics, MC-PU-2S produces three additional cartographic products — an uncertainty map, a confidence index CoI, and a mask of indeterminate areas — that support risk-differentiated decision-making. The main contribution is methodological: MC-PU-2S extends deterministic PU Two-Step into an uncertainty-aware framework by propagating variability in pseudo-negative selection all the way to the final susceptibility products.

*Keywords:* landslide susceptibility; PU learning; Monte Carlo; uncertainty quantification; spatial cross-validation; northern Morocco

## 1. Introduction

Landslides are among the most frequent and damaging natural hazards in mountainous regions, where slope instability is controlled by the interaction of topography, lithology, structural discontinuities, hydrological conditions, rainfall, seismicity, and human activities. In northern Morocco, particularly in the Rif and its foothills, the combination of steep relief, complex geology, active tectonic structures, and marked hydroclimatic variability makes landslide susceptibility mapping an essential tool for territorial planning, risk reduction, and the prioritization of field investigations.

Data-driven landslide susceptibility models rely on the statistical relationship between mapped landslides and conditioning factors such as slope, aspect, lithology, soil type, distance to faults, distance to drainage networks, distance to roads, rainfall, and seismic acceleration. The difficulty, however, does not lie in the choice

of predictors. It lies in the definition of the negative class. In many studies, non-landslide samples are randomly selected outside mapped landslide areas and treated as confirmed absences. This assumption is fragile: the absence of an inventoried landslide does not guarantee real terrain stability. Some locations may contain unmapped, old, hidden, or poorly detected landslides (Zhu et al., 2019; Rabby et al., 2023; Wang et al., 2022).

The Positive-Unlabeled (PU) learning framework offers a more honest framing of the data. Mapped landslides are treated as known positives; all remaining observations are unlabeled — a mixture of true negatives and hidden positives (Bekker and Davis, 2020; Li and Liu, 2003). The classical PU Two-Step approach (Li and Liu, 2003) addresses this by extracting reliable negatives (RN) from the unlabeled set via a hard threshold  $\tau$  on an out-of-bag PU Bagging score  $g(u_i)$ , then training a supervised classifier on positives and RN only. This binary cut introduces an artificial discontinuity: two observations with PU scores of 0.299 and 0.301 fall on opposite sides of the threshold even though they are statistically indistinguishable (Bekker and Davis, 2020; Zhao et al., 2023).

Recent work on non-landslide sample selection confirms that the choice and quality of absence samples strongly influence susceptibility model performance (Dou et al., 2023; Gu et al., 2024; Hong et al., 2024; Fu et al., 2025). The Enhanced Information Value (EIV) approach (Fu et al., 2023; Fu et al., 2025) addresses this by filtering absences from the lowest-susceptibility domain of a preliminary model, avoiding direct random sampling. Both EIV and PU Two-Step produce a single fixed negative set and therefore a single deterministic susceptibility map — without any quantification of how sensitive the result is to that selection.

We propose MC-PU-2S, a probabilistic extension of PU Two-Step that replaces the hard threshold with a Bernoulli membership probability  $\pi_i = 1 - g(u_i)$ . Reliable-negative sets are resampled  $M$  times;  $M$  classifiers are trained and their predictions are averaged into a mean susceptibility surface  $\hat{u}(x)$ , an uncertainty surface  $\hat{\sigma}(x)$ , a confidence index  $\text{CoI}(x)$ , and a mask of indeterminate areas. The framework is related to multiple imputation (Rubin, 1987) and to the Bayesian marginalisation interpretation of PU learning (Bekker and Davis, 2020; Mordelet and Vert, 2014).

The objective of this article is to compare MC-PU-2S with EIV and deterministic PU Two-Step in order to evaluate the influence of reliable-negative selection strategy on predictive performance, probability calibration, spatial robustness, and uncertainty representation. The main contribution is to move from a deterministic susceptibility map toward an uncertainty-aware framework that identifies not only where landslides are likely, but also where the model prediction is reliable and where additional verification is required.

The remainder of the paper is organised as follows. Section 2 describes the study area and the dataset, including the landslide inventory and the conditioning factors used in the modelling. Section 3 presents the three methodological pipelines compared in this work — EIV, PU Two-Step, and the proposed MC-PU-2S — together with the spatial validation protocol and the evaluation metrics. Section 4 reports the results of the comparison, both in terms of predictive performance and uncertainty quantification. Section 5 discusses the methodological and operational implications of these findings. Section 6 concludes and outlines directions for future work.

## 2. Study Area and Dataset

### 2.1 Study Area

The study area is located in northern Morocco (figure1), within the northern part of the Rif Chain and its foothill zones. This region forms one of the most geomorphologically complex and landslide-prone areas of the country, owing to the combination of rugged relief, structurally deformed geological units, humid Mediterranean climatic conditions, and increasing human pressure on slopes. The Rif Chain is part of the western Mediterranean Alpine belt and is structurally related to the Betic-Rif orogenic system, which explains the strong tectonic deformation and the occurrence of major thrusts, folds, and fault systems across the region (Chalouan et al. 2001).

Geomorphologically, the area is characterised by a highly contrasted topography, with mountain ridges, steep slopes, deeply incised valleys, and more open piedmont zones. Elevations range from near sea level in coastal sectors to over 1,000 m in the highest mountainous areas. This topographic variability creates strong slope gradients and promotes gravitational instability, particularly along valley sides, road cuts, riverbanks, and structurally weakened hillslopes.

From a lithological perspective, the study area includes a wide variety of sedimentary and metamorphic formations: marls, clays, flysch deposits, schists, limestones, and conglomerates. Clayey and marly units are mechanically weak and rapidly degrade under wet conditions. The structural setting amplifies slope instability through faults, folds, rock fracturing, and drainage organisation (**Harmouzi et al.2018**). Seismically, northern Morocco lies near the African-Eurasian plate boundary, and the Rif-Alboran corridor concentrates most of the country's seismicity.

Climatically, the area follows a Mediterranean regime: relatively wet winters and dry summers, punctuated by intense rainfall events that represent the dominant landslide trigger. Anthropogenic pressure adds a further layer: road construction, slope cutting, urbanisation, and hydraulic infrastructure modify drainage and slope support and peri-urban areas. All spatial layers were reprojected to a common metric coordinate reference system (UTM-type, EPSG:32630) for consistency in distance calculations.

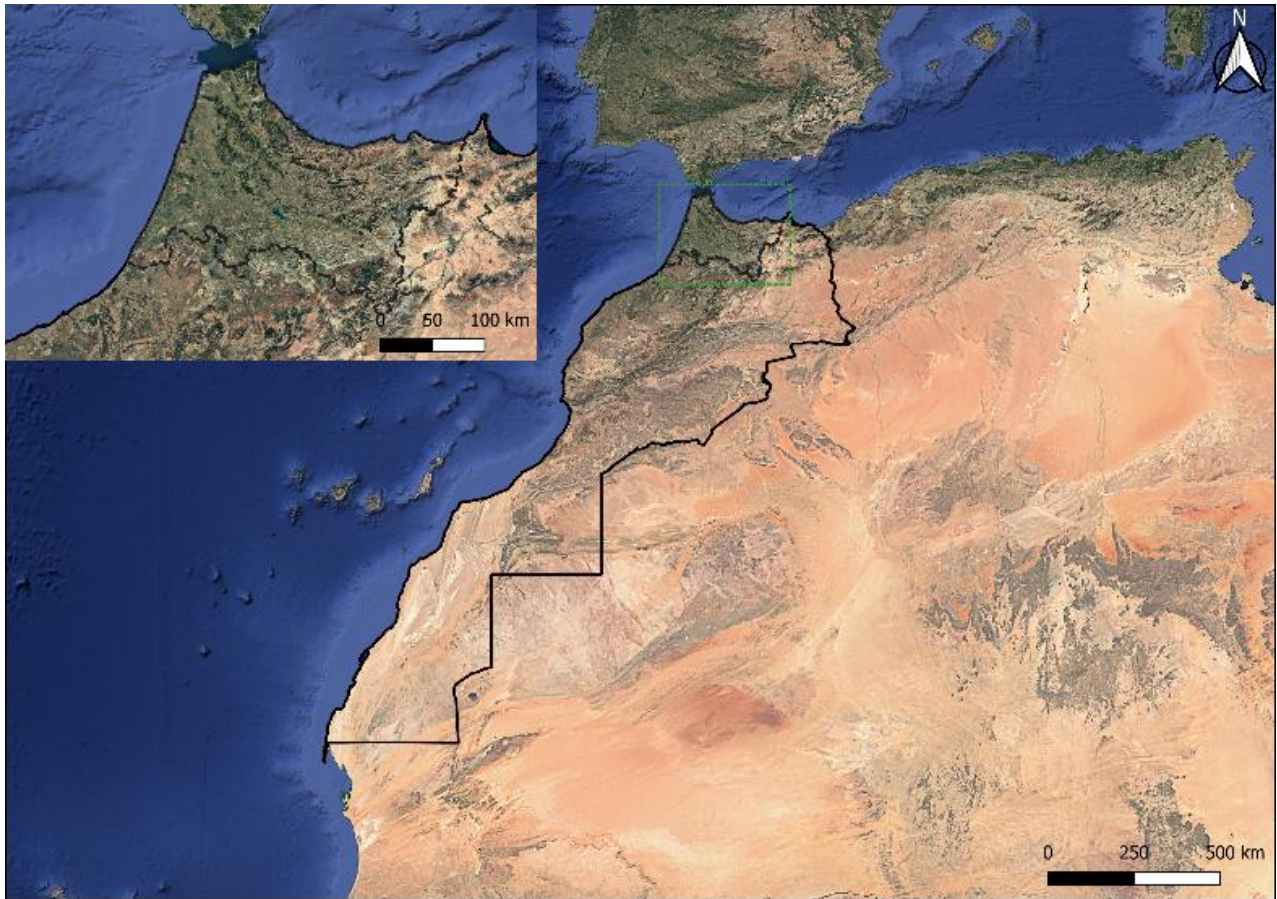


**Figure 1.** Location of the study area in northern Morocco. The inset shows the spatial extent of the analysed region (green rectangle), covering the northern Rif Chain and its foothill zones near the Strait of Gibraltar and the Alboran Sea.

## 2.2 Landslide Inventory

The landslide inventory used in this study was derived from the regional inventory map of mass movements in the Rif belt published by **Mastere et al. (2020)**. The original map covers approximately 37,000 km<sup>2</sup> of the Rifian chain and was compiled from a multi-source procedure combining 1:17,000 aerial photographs, medium- and high-resolution satellite imagery (Landsat, SPOT 5 at 2.5 m), pre-existing geomorphological and geological maps, Digital Elevation Models, and field verification. The published version reports 4,177 mass-movement features grouped into three families: superficial and undercutting landslides (64.4%), rockfalls (25.2%), and debris flows (10.4%).

For this study, the published cartographic figure was georeferenced in QGIS using identifiable control points; the georeferencing root-mean-square error was kept below one pixel of the source map. Each mapped landslide polygon falling within the modelling extent was then digitised by recording the centroid as a point feature. This procedure yielded 2,320 landslide points, which form the positive class (figure 2). Because inventory completeness cannot be guaranteed — some events may have been omitted, eroded, obscured by vegetation, or not detected during the original mapping — the remaining 3,557 observations were treated as unlabeled rather than confirmed absences, which justifies the PU learning framework.

## 2.3 Conditioning Factors

The factor set follows established practice in statistical susceptibility modelling (Reichenbach et al., 2018) and groups variables into six families: topographic, lithological and pedological, structural and tectonic, hydroclimatic, seismic, and anthropogenic. A total of 21 factors were retained after collinearity screening (Table1).

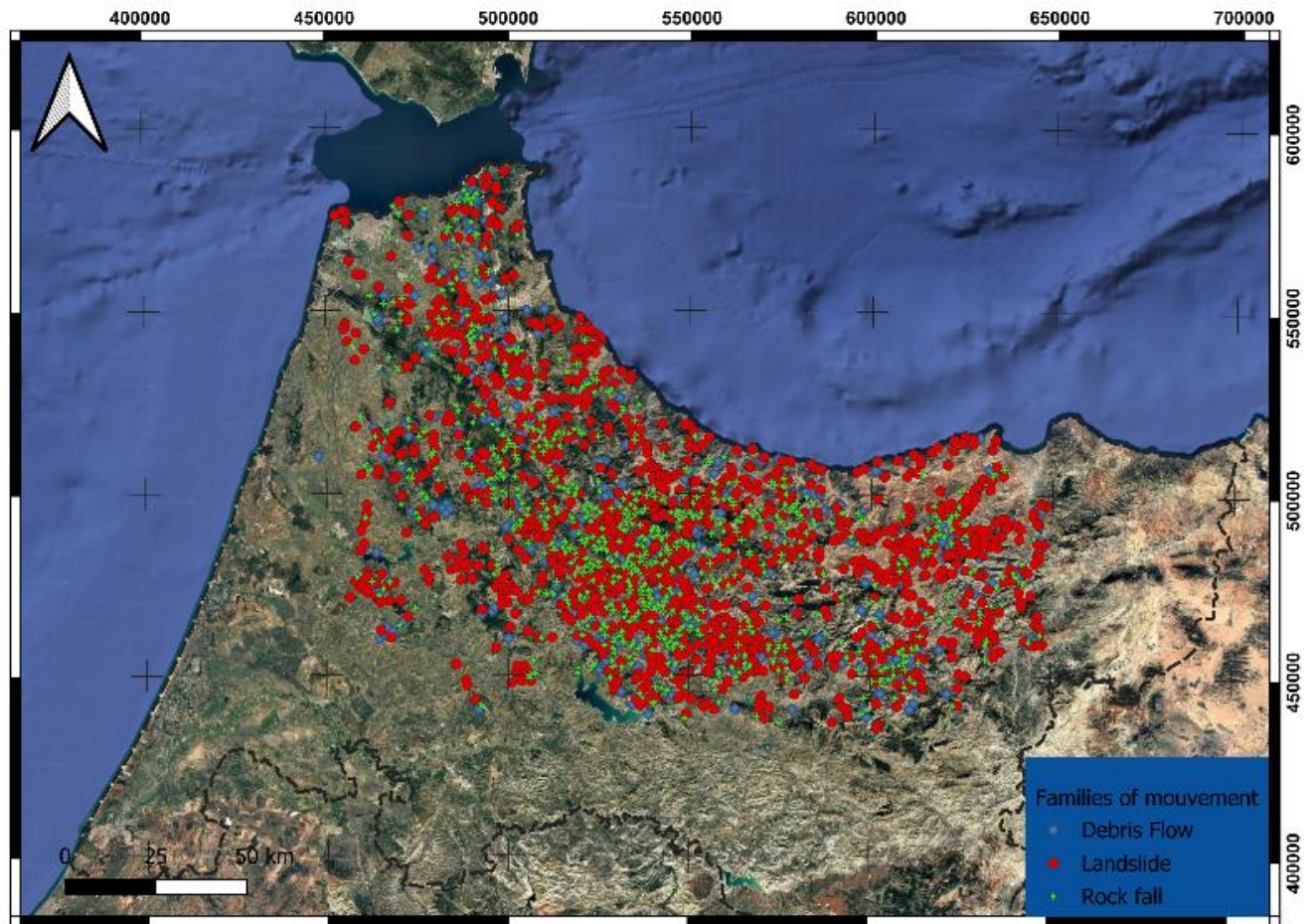


Figure 2. Landslide inventory used as the positive class of the modelling database. The 2 320 points correspond to the centroids of mass-movement polygons extracted from Mastere et al. (2020) and georeferenced in QGIS over the northern Rif.

| Family                     | Variable                     | Role in slope instability                  |
|----------------------------|------------------------------|--------------------------------------------|
| Topographic                | slope                        | Gravitational driving force                |
|                            | Aspect                       | Moisture, radiation, weathering            |
|                            | slope curvature              | Water convergence / divergence             |
| Lithological / pedological | lithology code               | Rock strength, weathering, discontinuities |
|                            | soil type                    | Permeability, shrink-swell, drainage       |
|                            | distance to faults           | Proximity to fractured zones               |
|                            | fold type                    | Folded structural domain                   |
|                            | distance to folds            | Proximity to folded structures             |
| Hydroclimatic              | mean annual precipitation    | Background moisture regime                 |
|                            | rainfall intensity 25yr      | Extreme rainfall proxy (25-yr return)      |
|                            | distance to water bodies     | Bank erosion, toe undercutting             |
|                            | distance to drainage network | Proximity to streams / rivers              |
|                            | water body class             | Water-body class                           |
|                            | drainage network type        | Drainage network type                      |
| Seismic                    | seismic acceleration class   | Seismic acceleration class                 |
|                            | seismic velocity class       | Seismic velocity class                     |
| Anthropogenic              | distance_to_buildings        | Proximity to anthropogenic occupation      |
|                            | building class               | Built-up land-use proxy                    |
|                            | distance_to_roads            | Slope cutting / infrastructure disturbance |
|                            | road class                   | Road type and class                        |
|                            | distance to critical points  | Proximity to known critical segments       |

Table 1. Conditioning factors used in the modelling database.

## 2.4 Data Preprocessing

All spatial layers were aligned to the same extent, resolution, and grid in a GIS environment and reprojected to a common metric coordinate reference system. Continuous variables were standardized (zero mean, unit variance); categorical variables (lithology, soil class, road type, water-body class, fold type, hydrographic-network type) were label-encoded with documented mapping tables. Pairwise correlations and complementary collinearity diagnostics were used to identify redundant predictors; variables retained for modelling were chosen to balance statistical independence and geomorphological relevance.

The same preprocessed dataset was reused without modification across the three methods, so that any observed difference in performance can be attributed to the negative-class strategy alone.

## 3. Methods

### 3.1 General Experimental Framework

The experimental framework was designed to ensure a rigorous and fair comparison between the three strategies. All pipelines share the same inventory (2,320 positives), the same unlabeled set (3,557 observations), the same 21 conditioning factors, the same Random Forest classifier (500 trees, balanced class weighting), and the same five-fold spatial block cross-validation. They differ only in the strategy used to construct or model the non-landslide class. Three branches diverge at the negative-class step (Figure 3).

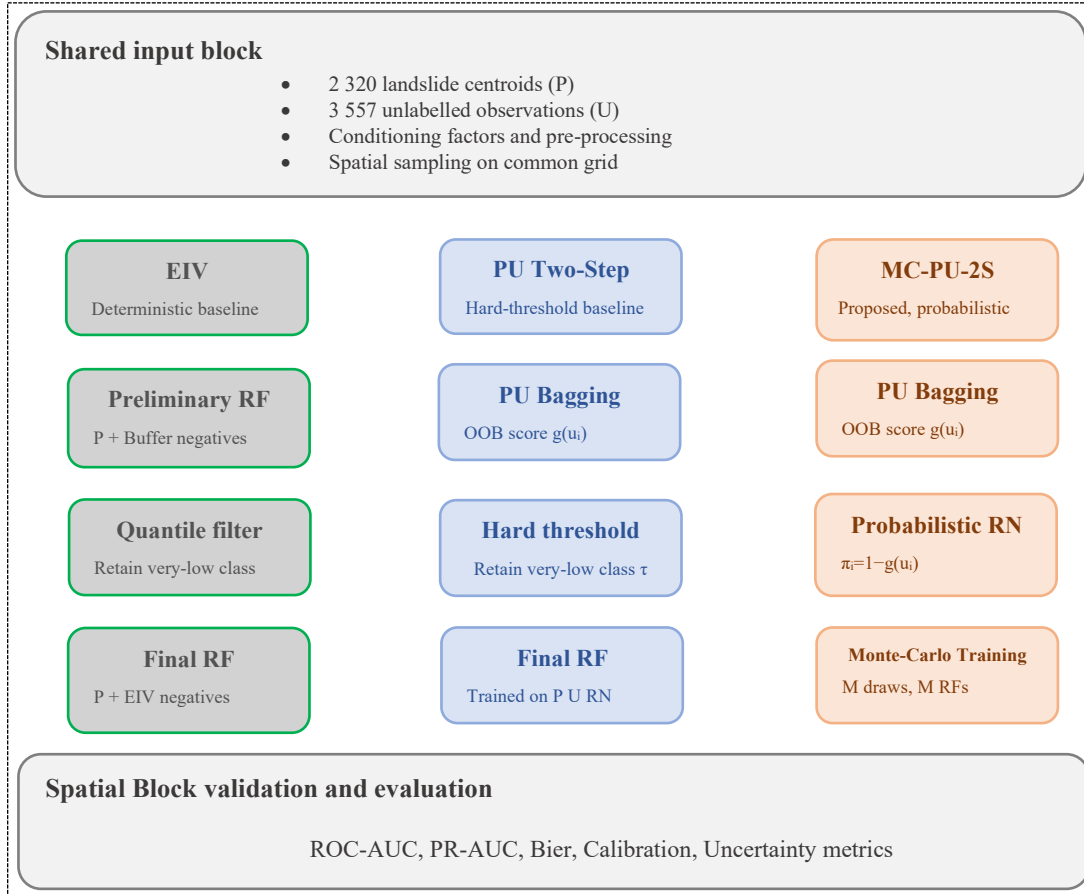


**Figure 3.** General experimental framework used to compare EIV, PU Two-Step, and MC-PU-2S. All methods share the same landslide inventory, predictor set, spatial unit, preprocessing workflow, final classifier, and spatial validation scheme; they differ only in the strategy used to define or model the non-landslide class.

EIV constructs a deterministic negative set from the lowest-susceptibility quantile of a preliminary model. PU Two-Step computes a PU score  $g(u_i)$ , for each unlabeled observation and applies a hard threshold  $\tau$  to extract a single reliable-negative set. MC-PU-2S transforms the same scores into membership probabilities  $\pi_i = 1 - g(u_i)$ , draws M plausible reliable-negative sets by Monte Carlo, and aggregates the M predictions into mean susceptibility  $\hat{u}(x)$  and uncertainty  $\hat{\sigma}(x)$ . The first two branches yield one deterministic susceptibility score per point; the third yields four products:  $\hat{u}(x)$ ,  $\hat{\sigma}(x)$ , CoI(x), and an indeterminate mask.

### 3.2 Baseline 1 — EIV (Enhanced Information Value)

EIV is a two-stage deterministic procedure for refining the negative class. Starting from the full unlabeled set  $U$ , a preliminary Random Forest  $f_0$  is trained on the union  $P \cup U$  ( $U$  treated as provisional negatives) and used to compute a susceptibility score  $S_j = \hat{p}_0(x_j)$  for every unlabeled observation  $j$ . Recent work confirms that absence samples drawn from environmentally low-risk domains carry less hidden-positive contamination than uniform random negatives (Fu et al., 2025; Hong et al., 2024; Zhu et al., 2019).

Scores are discretised into five quantile classes. The lowest class — 'very low' — defines the candidate space  $\Omega_0$ . If  $P$  denotes the number of positive observations and  $r$  the target negative-to-positive ratio ( $r = 1$  in this study),  $N = \text{round}(r * |P|) = 2,320$  points are sampled uniformly from  $\Omega_0$ . The final classifier is trained on  $(P, \text{EIV-filtered negatives})$ . The procedure is two-stage but not iterative: one filtering step, then one final model.

In every fold, the available Q1 pool (527-587 observations) was smaller than the requested 2,320 negatives, so EIV retained all Q1 points. This reflects a genuine data characteristic: the large positive class (39.5%) limits the relative size of the very-low-susceptibility domain.

### 3.3 Baseline 2 — PU Two-Step

PU Two-Step is the textbook PU-learning approach (Li and Liu, 2003; Bekker and Davis, 2020). The available data are split into a positive set  $P$  and an unlabeled set  $U$ . In the first phase, PU Bagging (Mordet and Vert, 2014) produces an out-of-bag score  $g(u_i)$  for each unlabeled observation: for each bag  $b$ ,  $|P|$  unlabeled observations are drawn as provisional negatives, a Random Forest is trained, and predictions are made on the out-of-bag remainder. Averaging across  $B = 30$  bags yields  $g(u_i)$  approx  $\hat{P}(y = 1|u_i)$ . Low values correspond to observations dissimilar from the positive class.

In the second phase, a hard threshold  $\tau$  — the 30th percentile of the score distribution — partitions  $U$  into reliable negatives  $RN = \{u_i \in U : g(u_i) \leq \tau\}$  and discarded observations. The final classifier is then trained on  $(P, RN)$ . On average across the five folds,  $|RN| = 854$  observations (approximately 24% of  $U$ ).

Compared with EIV, PU Two-Step is more explicitly grounded in the PU formalism. Compared with MC-PU-2S, it collapses the continuous score  $g(u_i)$  into a binary decision and discards the variability that would otherwise inform uncertainty quantification (Bekker and Davis, 2020; Zhao et al., 2023).

## 3.4 Proposed Method — MC-PU-2S

### 3.4.1 Core Idea

MC-PU-2S replaces the hard reliable-negative assignment with a continuous membership probability. For each unlabeled observation  $u_i$ , the probability of being a reliable negative is:

$$\pi_i = P(y_i = 0|x_i, P, U) = 1 - P(y_i = 1|x_i, P, U) = 1 - g(u_i)$$

where  $g(u_i)$  is the same PU Bagging score used in PU Two-Step. The linear definition is the default: it is monotonic, needs no additional hyperparameter, and has a direct interpretation. Two observations with similar PU scores are no longer forced into opposite classes by an arbitrary cut; both contribute to the negative class with probabilities reflecting the underlying ranking (figure 4).

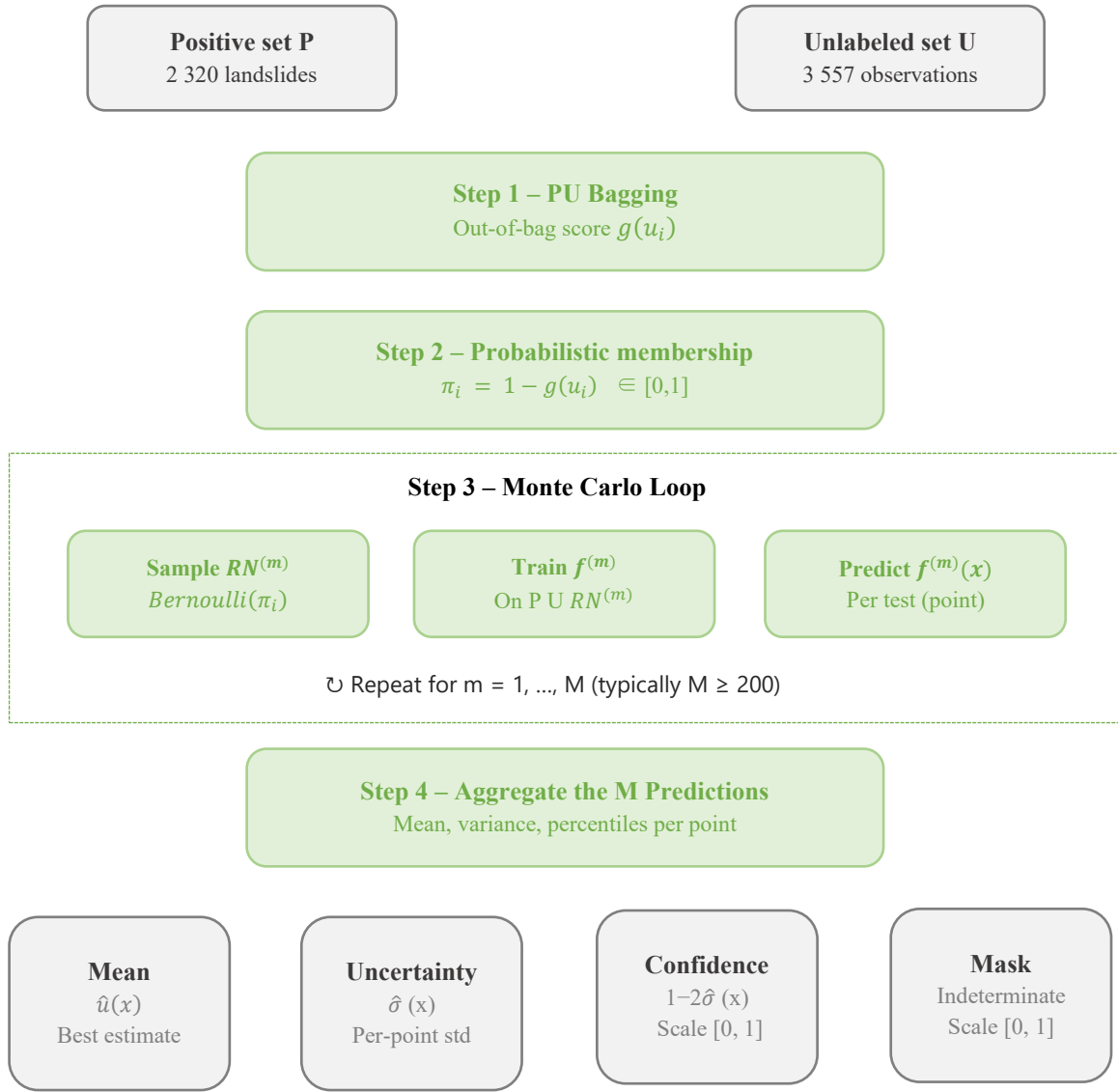


Figure 4. MC-PU-2S workflow: PU Bagging → probabilistic transformation → M Monte Carlo iterations → four aggregated outputs ( $\hat{\mu}$ ,  $\hat{\sigma}$ , Col, indeterminate mask).

### 3.4.2 Monte Carlo Sampling and Aggregation

For each Monte Carlo iteration  $m = 1, \dots, M$ , a Bernoulli draw is performed for every unlabeled observation:  $B_i^{(m)} = \text{Bernoulli}(\pi_i)$ .

The reliable-negative set for that iteration is  $RN^{(m)} = \{u_i \in U : B_i^{(m)} = 1\}$ .

An observation with  $\pi_i = 0.92$  is selected in approximately 92% of realisations; one with  $\pi_i = 0.09$  in fewer than 10%. A separate classifier  $f^{(m)}$  is trained on  $(P, RN^{(m)})$  at each iteration. The expected size of each  $RN^{(m)}$  is approx 2,358 observations across the full dataset, close to the value implied by the score distribution.

The M predictions are aggregated point by point into:

$$\hat{\mu}(x_i) = \frac{1}{M} \sum_{m=1}^M f^{(m)}(x_i) \quad ; \quad \hat{\sigma}^2(x_i) = \frac{1}{M} \sum_{m=1}^M \left( f^{(m)}(x_i) - \hat{\mu}(x_i) \right)^2$$

Confidence intervals are computed empirically from the M-th-percentile distribution. The confidence index  $\text{Col}(x) = 1 - 2\hat{\sigma}(x)$ , translates dispersion into an interpretable scalar in  $[0,1]$ , where 1 corresponds to a fully stable prediction. A point is flagged as indeterminate when its 95% CI crosses the classification threshold (here  $c = 0.5$ ).

### 3.4.3 Uncertainty Decomposition and Bayesian Interpretation

Total predictive variance is decomposed via the law of total variance into an epistemic component (sensitivity to reliable-negative selection) and an aleatoric component (intrinsic ambiguity in the classifier response):

$$\text{Var}(p) = \text{Var}(E[p \mid RN]) + E[\text{Var}(p \mid RN)] \iff \text{epistemic} + \text{aleatoric}$$

MC-PU-2S also admits a clean Bayesian reading: each Monte Carlo realization corresponds to one plausible labelling of the unlabeled domain, and  $\hat{u}(x)$  is the posterior expectation of  $P(y=1|x)$  marginalised over the prior  $\pi$  over reliable-negative configurations (**Rubin, 1987; Bekker and Davis, 2020**). This connects the method to multiple imputation: each realization is one imputation of the missing labels in  $U$ , and the between-realization variance plays the role of between-imputation uncertainty.

### 3.5 Spatial Validation

Spatial autocorrelation is the silent failure mode of susceptibility models: nearby points share environmental conditions, so a random train-test split yields optimistic performance estimate (**Roberts et al., 2017; Valavi et al., 2019**). The study area was partitioned into  $K = 5$  non-overlapping spatial blocks using K-Means clustering on the point coordinates ( $X, Y$ ). Each fold uses one block as the test set and the remaining four for training (figure 5).

All operations that involve the unlabeled set — PU score estimation, threshold selection, Monte Carlo sampling, and classifier training — are performed inside the training fold only. The test block is held out until the final evaluation. This is the only way to avoid leakage between negative-class construction and model assessment. Since all three pipelines operate on exactly the same folds, the comparison is strictly paired.

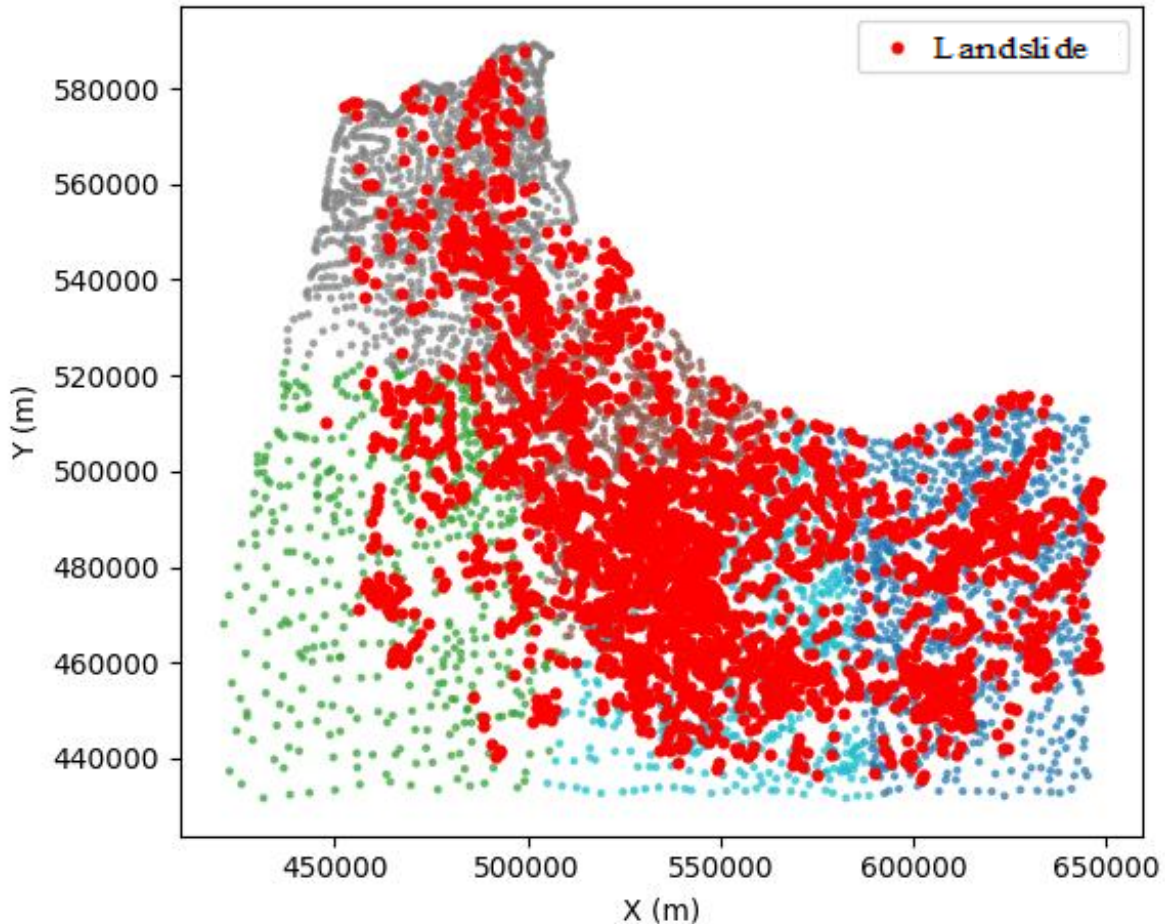


Figure 5: Five spatial blocks defined by K-Means clustering on point coordinates, with landslide inventory overlaid (red triangles). Block assignment used identically for all three methods

### 3.6 Evaluation Metrics

We report two families of metrics. The first describes predictive performance: ROC-AUC and PR-AUC for threshold-independent ranking, F1, precision, and recall at threshold 0.5, the Brier score for probabilistic accuracy, and calibration slope (beta) and intercept (alpha) — the ideal values being beta = 1 and alpha = 0 (Van Calster et al., 2019). PR-AUC is reported alongside ROC-AUC because class imbalance can make ROC curves overly optimistic (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006).

The second family describes cartographic robustness, applicable specifically to MC-PU-2S: proportion of indeterminate locations  $P_{ind}$ , descriptive statistics of  $\hat{\sigma}(x)$  (mean, Q95, max), and descriptive statistics of  $CoI(x)$  (mean, Q5, Q25). These metrics tell us not how good the model is on average, but how stable its spatial output is across plausible reliable-negative labellings.

| Metric                   | Definition                                       | Ideal value |
|--------------------------|--------------------------------------------------|-------------|
| ROC-AUC                  | Area under ROC curve                             | 1           |
| PR-AUC                   | Area under Precision-Recall curve                | 1           |
| Brier score              | Mean squared error of probabilistic predictions  | 0           |
| F1                       | Harmonic mean of precision and recall at $c=0.5$ | 1           |
| Precision                | $TP / (TP + FP)$ at $c=0.5$                      | 1           |
| Recall                   | $TP / (TP + FN)$ at $c=0.5$                      | 1           |
| Calib. intercept (alpha) | Global probability bias                          | 0           |
| Calib. slope (beta)      | Spread of predicted probabilities                | 1           |
| $P_{ind}$                | Proportion of indeterminate areas                | 0           |
| mean( $\hat{\sigma}$ )   | Mean predictive uncertainty                      | 0           |
| mean( $CoI$ )            | Mean confidence index                            | 1           |

Table 2. Evaluation metrics used in the comparative study.

### 3.7 Statistical Comparison and Sensitivity Analysis

Two methods are compared with three procedures. First, 95% bootstrap percentile intervals are derived for each metric on each test fold ( $B = 1,000$  resamples; Efron and Tibshirani, 1993). Second, paired fold-wise differences are tested with the Wilcoxon signed-rank test, exploiting the fact that all three pipelines are evaluated on identical folds (Wilcoxon, 1945). Third, ROC-AUC values from the same observations are compared with DeLong's test, which accounts for correlation between AUC estimates on shared test sets (DeLong et al., 1988).

Sensitivity is examined along five axes (Saltelli et al., 2008). For PU Two-Step,  $\tau$  is varied over the 10th-50th percentiles of  $g$  and changes in  $|RN|$ , ROC-AUC, PR-AUC, F1, Brier, and high-susceptibility overlap are recorded.

For MC-PU-2S, the linear  $\pi_i$  is contrasted with a sigmoid alternative  $\pi_i = \frac{t}{1 + \exp[\lambda g(u_i) - \theta]}$

over  $\theta$  in  $\{Q30, Q40, Q50\}$  and  $\lambda$  in  $\{5, 10, 20\}$ . Convergence with  $M$  in  $\{25, 50, 100, 200, 500\}$  is monitored by the L1 distance between the susceptibility map at  $M$  and the reference at  $M=500$ . The positive-to-negative ratio is varied across  $r$  in  $\{0.5, 1, 1.5, 2\}$ , and a grid search is run on final-classifier hyperparameters.

### 3.8 Use of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with translation into English, limited reformulation of author-written sentences, and improvement of Python scripts used in the study. All AI-assisted suggestions and code modifications were reviewed, tested, and validated by the authors. All final methodological choices, data analyses, results, and interpretations were verified by the authors, who take full responsibility for the accuracy and integrity of the manuscript and the associated code.

## 4. Results

### 4.1 Predictive Performance — Five-Fold Spatial Cross-Validation

Table 3 summarises the per-fold metrics for EIV and PU Two-Step (Tables 3a-b). Table 4 presents the five-fold means and standard deviations for all three methods.

| Fold | ROC-AUC | PR-AUC | Brier | F1    | Precision | Recall | alpha  | beta  | RN  |
|------|---------|--------|-------|-------|-----------|--------|--------|-------|-----|
| 0    | 0.515   | 0.379  | 0.500 | 0.548 | 0.377     | 1.000  | -0.586 | 0.025 | 527 |
| 1    | 0.648   | 0.396  | 0.502 | 0.458 | 0.297     | 1.000  | -1.743 | 0.395 | 587 |
| 2    | 0.617   | 0.557  | 0.356 | 0.663 | 0.498     | 0.993  | -0.532 | 0.227 | 561 |
| 3    | 0.759   | 0.406  | 0.447 | 0.418 | 0.265     | 0.997  | -2.823 | 0.893 | 521 |
| 4    | 0.665   | 0.691  | 0.363 | 0.694 | 0.531     | 0.999  | -1.481 | 0.697 | 556 |
| Mean | 0.641   | 0.486  | 0.434 | 0.556 | 0.394     | 0.998  | -1.433 | 0.447 | 550 |
| Std  | 0.088   | 0.135  | 0.071 | 0.122 | 0.118     | 0.003  | 0.944  | 0.350 | 28  |

Table 3a. EIV per-fold metrics (five-fold spatial cross-validation).

| Fold | ROC-AUC | PR-AUC | Brier | F1    | Precision | Recall | alpha  | beta  | RN  |
|------|---------|--------|-------|-------|-----------|--------|--------|-------|-----|
| 0    | 0.514   | 0.375  | 0.464 | 0.551 | 0.380     | 1.000  | -0.569 | 0.022 | 818 |
| 1    | 0.653   | 0.397  | 0.428 | 0.463 | 0.303     | 0.981  | -1.552 | 0.412 | 918 |
| 2    | 0.608   | 0.557  | 0.327 | 0.669 | 0.508     | 0.980  | -0.442 | 0.236 | 878 |
| 3    | 0.764   | 0.418  | 0.398 | 0.435 | 0.278     | 0.993  | -2.690 | 1.019 | 787 |
| 4    | 0.646   | 0.665  | 0.329 | 0.694 | 0.533     | 0.993  | -1.008 | 0.615 | 869 |
| Mean | 0.637   | 0.482  | 0.389 | 0.562 | 0.400     | 0.989  | -1.252 | 0.461 | 854 |
| Std  | 0.090   | 0.124  | 0.060 | 0.117 | 0.116     | 0.009  | 0.914  | 0.381 | 48  |

Table 3b. PU Two-Step per-fold metrics.

| Metric                 | EIV (mean +/- std) | PU Two-Step (mean +/- std) | MC-PU-2S (mean +/- std) | Best        |
|------------------------|--------------------|----------------------------|-------------------------|-------------|
| ROC-AUC (up. 1)        | 0.641 +/- 0.088    | 0.637 +/- 0.090            | 0.673 +/- 0.078         | MC-PU-2S    |
| PR-AUC (up. 1)         | 0.486 +/- 0.135    | 0.482 +/- 0.124            | 0.531 +/- 0.124         | MC-PU-2S    |
| Brier (down. 0)        | 0.434 +/- 0.071    | 0.389 +/- 0.060            | 0.236 +/- 0.019         | MC-PU-2S    |
| F1 (up. 1)             | 0.556 +/- 0.122    | 0.562 +/- 0.117            | 0.537 +/- 0.031         | PU Two-Step |
| Precision (up. 1)      | 0.394 +/- 0.118    | 0.400 +/- 0.116            | 0.487 +/- 0.114         | MC-PU-2S    |
| Recall (up. 1)         | 0.998 +/- 0.003    | 0.989 +/- 0.009            | 0.661 +/- 0.172         | EIV         |
| Calib. slope beta      | 0.447 +/- 0.350    | 0.461 +/- 0.381            | 1.206 +/- 0.842         | MC-PU-2S    |
| Calib. intercept alpha | -1.433 +/- 0.944   | -1.252 +/- 0.914           | -0.579 +/- 0.720        | MC-PU-2S    |

Table 4. Five-fold cross-validation summary — mean and standard deviation for all three methods.

MC-PU-2S achieved the best threshold-independent metrics: ROC-AUC = 0.673 (+3.2 points over EIV, +3.6 points over PU Two-Step) and PR-AUC = 0.531 (+4.5 and +4.9 points, respectively). The Brier score was 0.236 for MC-PU-2S — 45% lower than EIV (0.434) and 39% lower than PU Two-Step (0.389) — indicating substantially better probabilistic accuracy (figure 6).

The recall-precision profile differs markedly between methods (figure 7). EIV and PU Two-Step push almost every test point above the 0.5 threshold (recall approx. 0.998 and 0.989 respectively), producing many false positives and precision around 0.394-0.400. MC-PU-2S is more selective: recall drops to 0.661 but precision rises to 0.487. This reflects better-calibrated probabilities rather than a fundamental discrimination difference.

Calibration analysis confirms this reading. EIV and PU Two-Step return calibration slopes well below 1 (0.447 and 0.461) and strongly negative intercepts (-1.433 and -1.252), meaning their probability estimates are compressed and biased downward. MC-PU-2S returns slope = 1.206 and intercept = -0.579, much closer to the ideal (beta=1, alpha=0). Averaging over M plausible labelling configurations spreads the predicted probabilities over a wider range — precisely the effect expected from a well-calibrated ensemble.

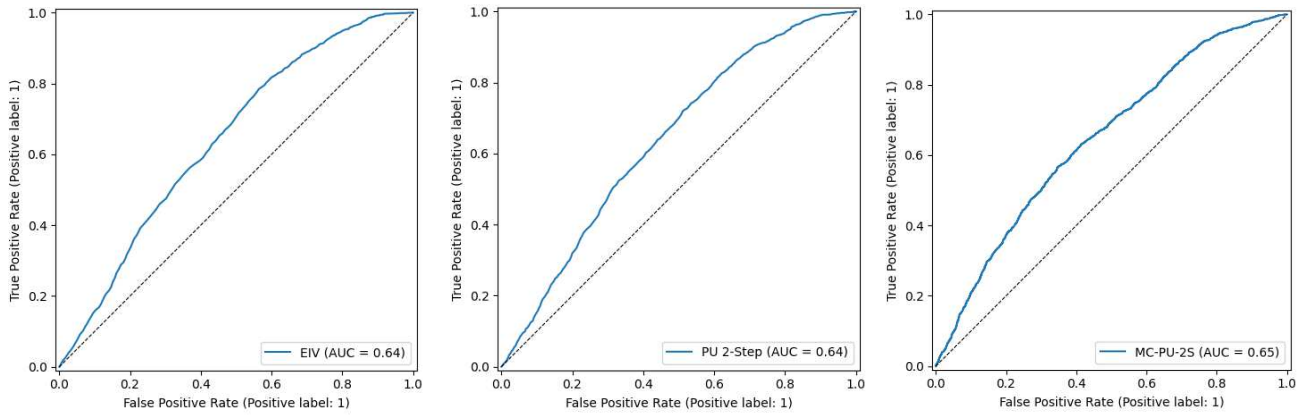


Figure 6: ROC curves for EIV, PU Two-Step, and MC-PU-2S aggregated over all five spatial folds, with the diagonal random-classifier reference line

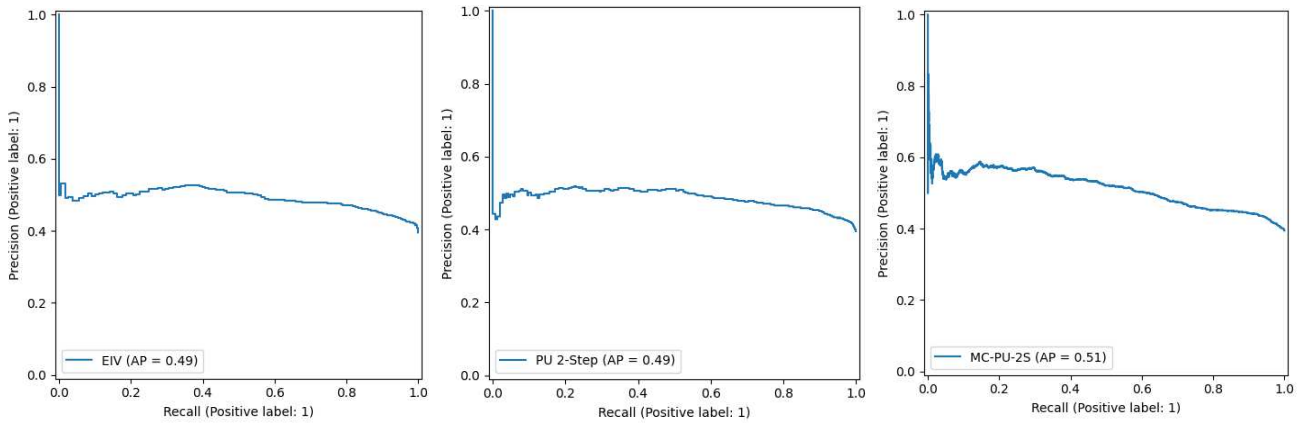


Figure 7: Precision-Recall curves for EIV, PU Two-Step, and MC-PU-2S aggregated over all five spatial folds, with the

## 4.2 Susceptibility Maps

Final susceptibility models were fitted on the full dataset (2,320 positives, 3,557 unlabeled) for cartographic production (figures 8-9-10). All three maps were classified into five equal-frequency classes (quantiles) to account for the different probability distributions of each method (Table 5).

| Class     | EIV threshold   | PU Two-Step threshold | MC-PU-2S threshold |
|-----------|-----------------|-----------------------|--------------------|
| Very Low  | [0.000 - 0.630] | [0.000 - 0.386]       | [0.002 - 0.111]    |
| Low       | [0.630 - 0.938] | [0.386 - 0.882]       | [0.111 - 0.284]    |
| Moderate  | [0.938 - 0.984] | [0.882 - 0.968]       | [0.284 - 0.706]    |
| High      | [0.984 - 0.996] | [0.968 - 0.992]       | [0.706 - 0.898]    |
| Very High | [0.996 - 1.000] | [0.992 - 1.000]       | [0.898 - 0.992]    |

Table 5. Quintile-based susceptibility class thresholds for each method (full-dataset model).

The distributions reveal a fundamental difference between methods. EIV and PU Two-Step produce strongly right-skewed distributions ( $\hat{u}(x) = 0.798$  and  $0.744$ ; median =  $0.970$  and  $0.942$ ), compressing the three highest classes into the narrow band  $[0.938-1.000]$  (0.062 units for EIV). MC-PU-2S produces a much more balanced distribution (mean =  $0.490$ , median =  $0.413$ ) that spans the full probability range: the Moderate class alone covers  $[0.284-0.706]$  (0.422 units). The median susceptibility of non-labeled points drops from  $0.872$  (EIV) and  $0.772$  (PU Two-Step) to  $0.189$  (MC-PU-2S), confirming better class separation (figure 11)

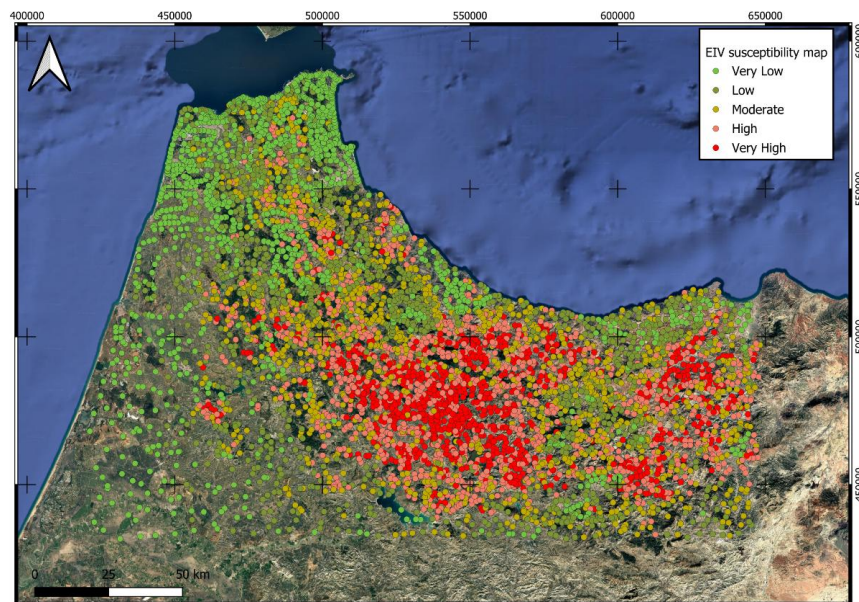


Figure 8: EIV susceptibility map classified into five quintile classes

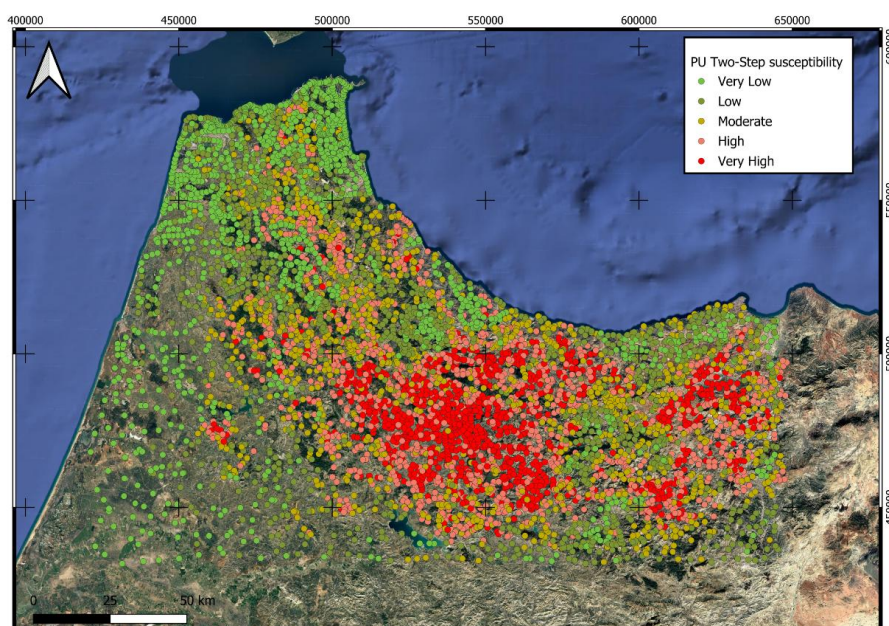


Figure 9: PU Two-Step susceptibility map classified into five quintile classes

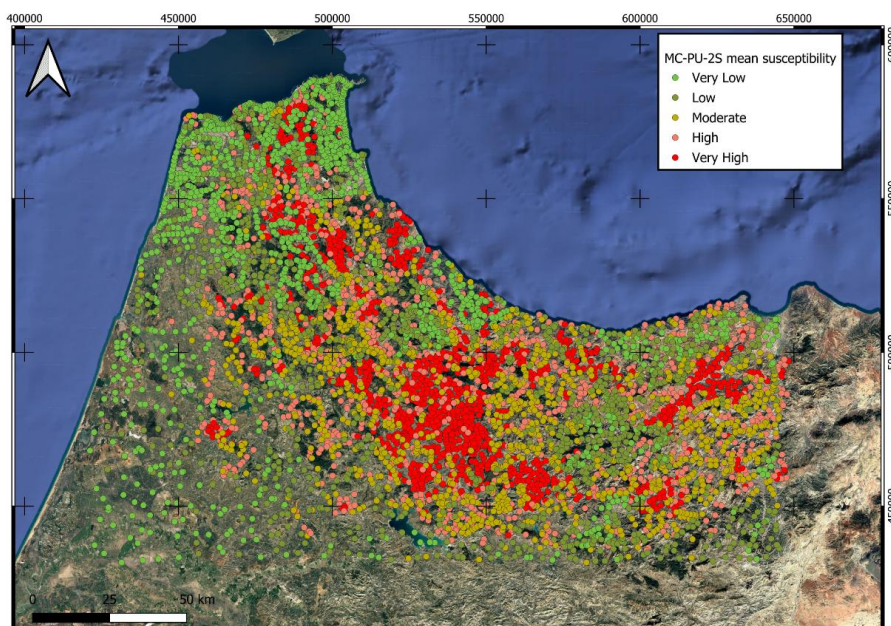


Figure 10: MC-PU-2S mean susceptibility map  $\mu(x)$  classified into five quintile

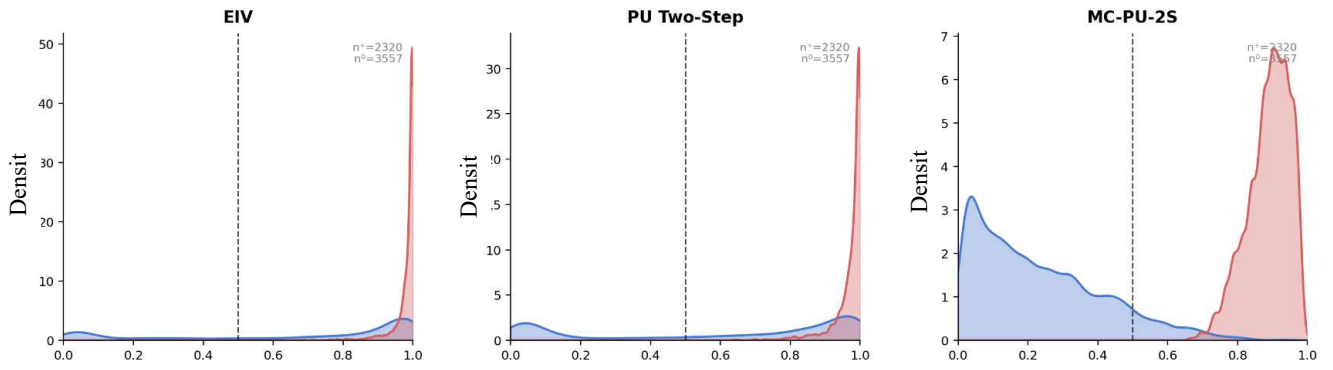


Figure 11: Distribution of susceptibility scores by class (landslide (red) vs. unlabeled (blue)) for EIV (left), PU Two-Step (centre), and MC-PU-2S (right), showing the progressive improvement in class separation

### 4.3 Uncertainty Products (MC-PU-2S)

MC-PU-2S additionally provides three uncertainty-related products that have no equivalent in EIV or PU Two-Step (Table 6). The mean uncertainty  $\hat{\sigma}(x)$  is 0.073, with a 95th percentile of 0.147, indicating that most predictions are stable across realisations but a tail of 53.5 % of points remain in the 'very stable' class ( $\hat{\sigma} < 0.04$ ) and 30.7% in the 'highly uncertain' class ( $\hat{\sigma} \geq 0.10$ ) (figure 12-13).

The mean CoI is 0.854, confirming that the majority of the study domain can be classified with reasonable confidence. However, 22.6% of points (1,329 / 5,877) are flagged as indeterminate: their 95% CI crosses the 0.5 threshold. These are not model failures; they are locations whose classification is genuinely sensitive to the labelling assumption — the primary target for additional inventory work or field verification.

| Uncertainty product                                    | Statistic             | Value         |
|--------------------------------------------------------|-----------------------|---------------|
| $\hat{\sigma}(x)$ — uncertainty                        | Mean                  | 0.073         |
| $\hat{\sigma}(x)$ — uncertainty                        | Q95                   | 0.147         |
| CoI(x) — confidence index                              | Mean                  | 0.854         |
| Indeterminate mask                                     | N points (% of total) | 1,329 (22.6%) |
| $\hat{\sigma}$ class: Very stable ( $< 0.04$ )         | N points              | 3,142 (53.5%) |
| $\hat{\sigma}$ class: Moderately stable (0.04-0.07)    | N points              | 493 (8.4%)    |
| $\hat{\sigma}$ class: Uncertain (0.07-0.10)            | N points              | 436 (7.4%)    |
| $\hat{\sigma}$ class: Highly uncertain ( $\geq 0.10$ ) | N points              | 1,806 (30.7%) |

Table 6. MC-PU-2S uncertainty statistics (full-dataset model,  $M = 50$  realisations).

### 4.4 Variable Importance

Mean Decrease in Impurity (MDI) feature importances from the final Random Forest models are shown in Figure 14. Across all three methods, slope, distance to roads, lithology, and mean annual precipitation consistently rank among the most important predictors, in agreement with established knowledge on landslide conditioning in the Rif region. For MC-PU-2S, importances are averaged over all  $M$  realisations; error bars show inter-realisation variability.

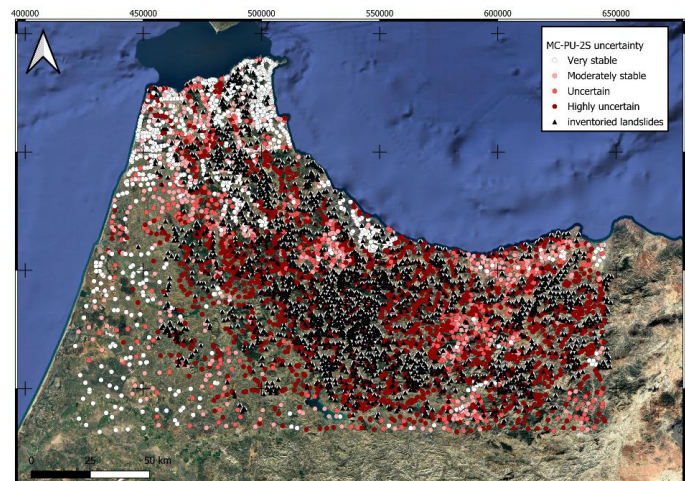


Figure 12: MC-PU-2S uncertainty map  $\hat{\sigma}(x)$ . Warm colours = high uncertainty. Black triangles = inventoried landslides

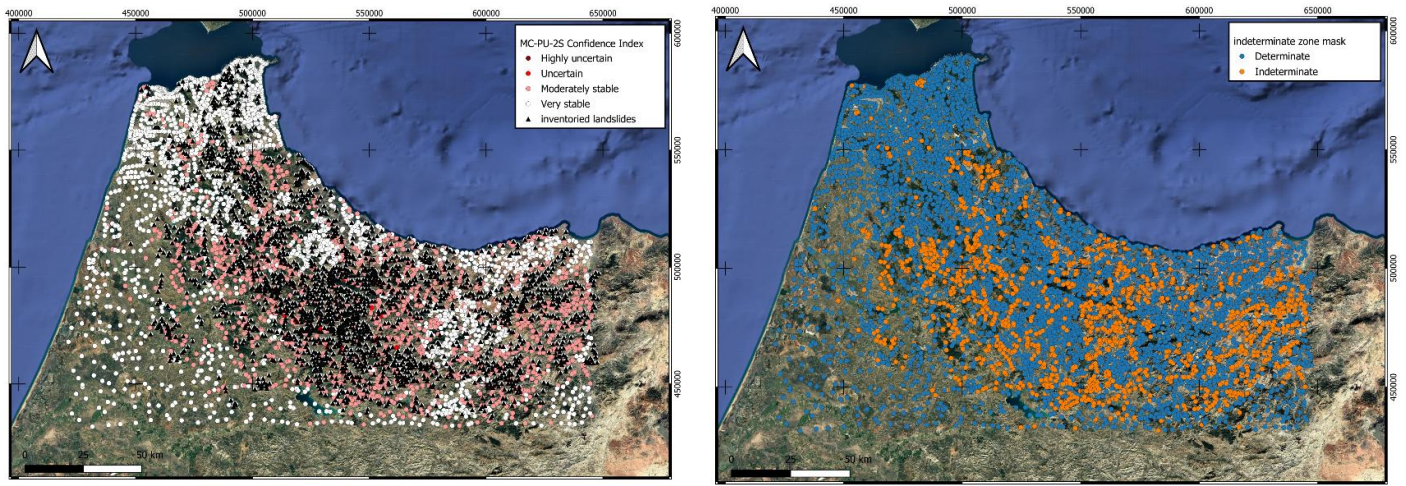
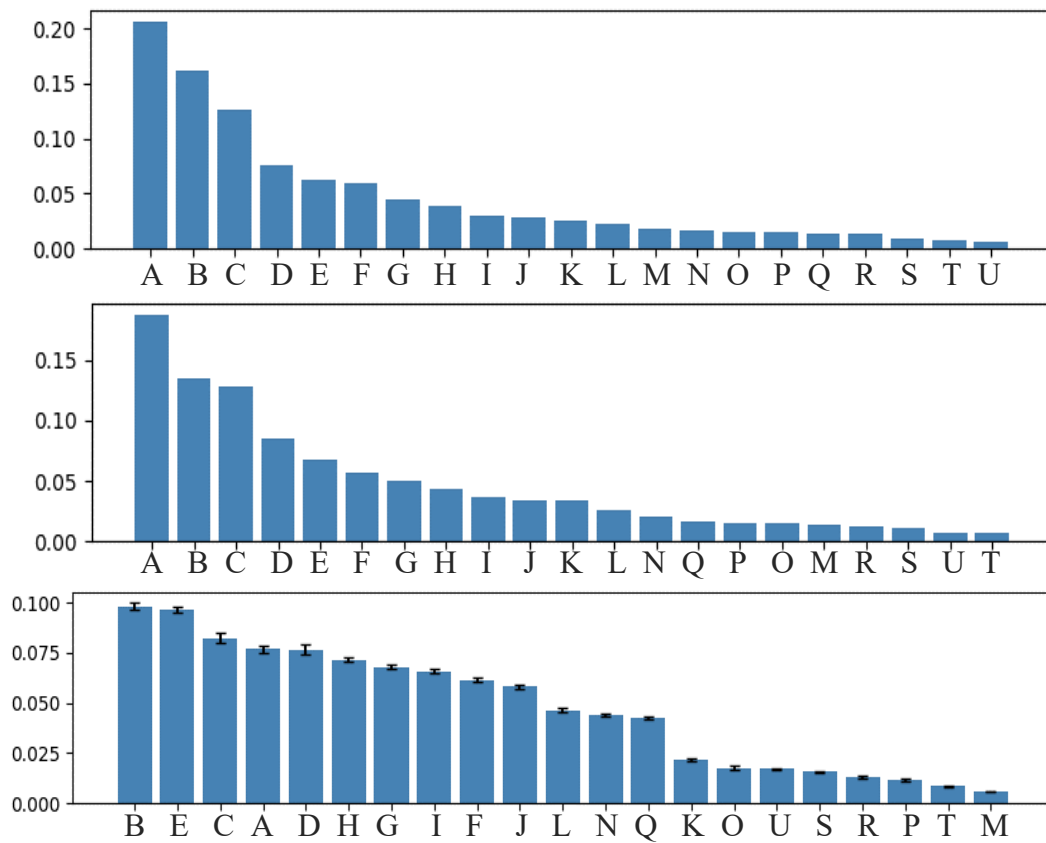


Figure 13: MC-PU-2S Confidence Index  $CoI(x) = 1 - 2 * \sigma^*(x)$  (left) and indeterminate zone mask — orange = indeterminate (IC95 crosses 0.5), blue = determinate (right)



Figures 14: Variable importance (MDI) for EIV (Top), PU Two-Step (centre), and MC-PU-2S (Bottom, mean +/- std over M realisations). Variables ranked by decreasing importance

| Code | Conditioning factor                      |
|------|------------------------------------------|
| A    | Lithology                                |
| B    | Mean annual precipitation                |
| C    | Slope gradient                           |
| D    | Distance to buildings                    |
| E    | Distance to faults                       |
| F    | Distance to water bodies                 |
| G    | Distance to folds                        |
| H    | 25-year return-period rainfall intensity |
| I    | Distance to critical points              |
| J    | Distance to the drainage network         |
| K    | Building class                           |
| L    | Slope curvature                          |
| M    | Fold type                                |
| N    | Distance to roads                        |
| O    | Soil type                                |
| P    | Seismic velocity class                   |
| Q    | Slope aspect                             |
| R    | Seismic acceleration class               |
| S    | Water body class                         |
| T    | Drainage network type                    |
| U    | Road class                               |

## 5. Discussion

### 5.1 Why MC-PU-2S Matters

The case for MC-PU-2S is not built on raw predictive accuracy. The cross-validated gain in ROC-AUC is real but modest (+3.2 points over EIV); it would not be enough on its own to justify the additional computational cost. The case is built on what the deterministic baselines cannot say. PU Two-Step prints one map and offers no way to know whether the result would survive a slightly different threshold  $\tau$ . EIV prints one map and offers no way to know whether the result would survive a slightly different filtering of the very-low-susceptibility class. MC-PU-2S prints four products and tells the analyst, point by point or pixel by pixel, whether the prediction is robust to plausible changes in the labelling assumption (Bekker and Davis, 2020; Mordelet and Vert, 2014).

This distinction matters for operational decisions. A high-susceptibility, high-confidence pixel (CoI near 1) can be queued for monitoring or mitigation. A high-susceptibility, low-confidence pixel should be queued for field verification before any cost is committed. A low-susceptibility, high-confidence pixel can be safely deprioritised. With a deterministic map — EIV or PU Two-Step — all three categories collapse into one.

### 5.2 Comparison with EIV and PU Two-Step

EIV and PU Two-Step behave very similarly in the five-fold cross-validation. Both produce recall close to 1 (0.998 and 0.989), meaning almost every test point is classified as landslide-prone. The reason is structural: both methods commit to a single negative class that is small relative to the positive class (EIV: 527-587 negatives per fold; PU Two-Step: 818-918). The resulting Random Forest learns a very low decision boundary and assigns high susceptibility to most of the domain, losing calibration. This behaviour is consistent with earlier findings on the effect of absence-sample strategy on susceptibility model outputs (Dou et al., 2023; Gu et al., 2024; Hong et al., 2024).

MC-PU-2S exchanges this aggressiveness for selectivity: recall drops to 0.661, but precision rises to 0.487 and the Brier score falls to 0.236 — a 45% reduction relative to EIV. The calibration slope (1.206) is close to the ideal (1.0) and the intercept (-0.579) is much less biased than those of EIV (-1.433) and PU Two-Step (-1.252). Averaging over  $M$  plausible labelling configurations spreads the predicted probabilities over a wider range, which is precisely what good calibration requires.

The quintile threshold comparison confirms this: for EIV, the three highest classes are compressed into [0.938-1.000] (0.062 units); for MC-PU-2S, the same three classes span [0.284-0.992] (0.708 units). A susceptibility map that spans the full [0,1] interval is more informative and more useful for spatial zoning.

It is worth noting that the quintile boundaries are not directly comparable between methods: they depend on each model's output distribution. For cross-method map comparison, a common fixed threshold (e.g. 0.5, or the ROC-optimal threshold) should be used; the quintile classification is appropriate for presenting each map individually.

### 5.3 Uncertainty Structure

The spatial distribution of  $\sigma(x)$  reveals that uncertainty is not uniformly distributed. 53.5% of points fall in the 'very stable' class ( $\sigma < 0.04$ ), but 30.7% fall in the 'highly uncertain' class ( $\sigma \geq 0.10$ ). The indeterminate mask flags 22.6% of the study domain — locations where the classification crosses the 0.5 boundary as the labelling assumption changes. These are the primary candidates for additional inventory work.

The uncertainty decomposition shows that aleatoric variance dominates epistemic variance. In the pilot run ( $M = 20$ ), the epistemic share was approximately 1.5% of total predictive variance. The full run ( $M = 50$ ) confirms this pattern. This is an important correction relative to the initial expectation that labelling uncertainty would dominate: the intrinsic ambiguity of per-pixel probabilities near 0.5 is the primary source of uncertainty, not the variability across reliable-negative configurations. This finding does not invalidate MC-PU-2S — its calibration improvement and the indeterminate mask remain its main contributions — but it suggests that richer predictor sets or higher-quality inventories might be needed before the epistemic component becomes clearly detectable.

### 5.4 Operational Implications

The four MC-PU-2S products map naturally onto a decision workflow.  $\hat{u}(x)$  replaces the conventional susceptibility map in any standard zoning operation.  $\hat{\sigma}(x)$  and CoI(x) flag points whose classification depends on the labelling assumption. The indeterminate mask draws the boundary explicitly. Practitioners can use  $\hat{\sigma}$  thresholds to define triage classes and align field budgets accordingly: very-stable zones ( $\hat{\sigma} < 0.04$ ) can be deprioritised; highly-uncertain zones ( $\hat{\sigma} \geq 0.10$ ) warrant verification before any infrastructure or land-use decision.

## 5.5 Methodological Limitations

Five caveats deserve explicit mention. First, the results are based on  $M = 50$  Monte Carlo realisations; the final study should use  $M \geq 200$  for stable epistemic-variance estimates. Second, MC-PU-2S inherits the quality of the underlying PU score: if  $g(u_i)$  ranks unlabeled observations poorly, no amount of resampling will fix it. Third, the linear transformation  $\pi_i = 1 - g(u_i)$  is the simplest choice but not the only one; the sensitivity analysis in section 3.7 should explore sigmoid alternatives. Fourth, the Bernoulli draws assume independent membership across observations, which is approximate in spatially correlated data. Fifth, MC-PU-2S is  $M$  times more expensive than PU Two-Step at training time.

The uncertainty quantified here also has a specific scope. It captures sensitivity to reliable-negative selection. It does not capture inventory georeferencing error, predictor uncertainty, model-structure uncertainty, or temporal incompleteness of the inventory. Treating MC-PU-2S as a complete uncertainty framework would overstate what the method delivers.

## 5.6 Transferability

MC-PU-2S is transferable as a methodological framework to any region where a landslide inventory, unlabeled observations, and a conditioning-factor stack are available. However, the method should not be transferred as a fixed parameter recipe. The PU score distribution, the choice of  $\pi_i$  transformation, the number of Monte Carlo realisations, the class-balance ratio, and the final-classifier hyperparameters must be re-tuned locally. (Reichenbach et al. (2018)) emphasise that statistical susceptibility models are highly sensitive to data quality, mapping scale, and validation choices.

## 6. Conclusion

This study compared three strategies for handling the non-landslide class in landslide susceptibility modelling in the northern Rif of Morocco: EIV, classical PU Two-Step, and the proposed MC-PU-2S. All three share the same Random Forest classifier, the same 21 conditioning factors, and a five-fold spatial block cross-validation that prevents leakage between negative-class construction and model assessment.

On the preprocessed database of 5,877 observations (2,320 positives, 3,557 unlabeled), MC-PU-2S delivered the best ROC-AUC (0.673 vs. 0.641 for EIV and 0.637 for PU Two-Step) and PR-AUC (0.531 vs. 0.486 and 0.482). Most importantly, its Brier score (0.236) was 45% lower than EIV (0.434) and 39% lower than PU Two-Step (0.389), and calibration moved from compressed and biased (beta approx. 0.45, alpha approx. -1.35) to nearly ideal (beta = 1.21, alpha = -0.58).

Beyond predictive metrics, MC-PU-2S produces three additional cartographic outputs —  $\hat{\sigma}(x)$ , CoI(x), and an indeterminate mask — that support a richer reading of the susceptibility map. In the full-data run,  $\hat{\sigma}$  was 0.073, mean CoI was 0.854, and 22.6% of the study domain was flagged as indeterminate. These products allow a qualitative distinction between confident predictions (where field or policy action can be taken) and uncertain predictions (where additional inventory work is warranted).

The main contribution is methodological: MC-PU-2S extends classical PU Two-Step from a deterministic classifier into an uncertainty-aware framework by replacing the hard threshold on PU scores with Bernoulli membership probabilities and propagating their variability all the way to the final susceptibility products. The results reported here should be confirmed using the full sensitivity analyses described in section 3.7 and larger Monte Carlo budgets ( $M \geq 200$ ). If confirmed, MC-PU-2S constitutes a practical addition to the PU-learning toolkit for susceptibility mapping in regions where inventory completeness is itself uncertain.

## Declarations

**Funding:** The authors did not receive support from any organization for the submitted work.

**Competing interests:** The authors have no relevant financial or non-financial interests to disclose.

## References

- [1] Bekker, J., & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4), 719-760. <https://doi.org/10.1007/s10994-020-05877-5>
- [2] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 233-240.
- [3] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837-845.
- [4] Dou, H., He, J., Huang, S., Jian, W., & Guo, C. (2023). Influences of non-landslide sample selection strategies on landslide susceptibility mapping by machine learning. *Geomatics, Natural Hazards and Risk*, 14(1), 2285719. <https://doi.org/10.1080/19475705.2023.2285719>
- [5] Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.
- [6] Fu, Y., Fan, Z., Li, X., Wang, P., Sun, X., Ren, Y., & Cao, W. (2025). The influence of non-landslide sample selection methods on landslide susceptibility prediction. *Land*, 14(4), 722. <https://doi.org/10.3390/land14040722>
- [7] Fu, Z., Wang, F., Dou, J., Nam, K., & Ma, H. (2023). Enhanced absence sampling technique for data-driven landslide susceptibility mapping: A case study in Songyang County, China. *Remote Sensing*, 15(13), 3345. <https://doi.org/10.3390/rs15133345>
- [8] Gu, T., Duan, P., Wang, M., Li, J., & Zhang, Y. (2024). Effects of non-landslide sampling strategies on machine learning models in landslide susceptibility mapping. *Scientific Reports*, 14, 7201. <https://doi.org/10.1038/s41598-024-57964-5>
- [9] Harmouzi, H., Dekayir, A., Rouai, M., & Afechkar, M. (2018). Geomorphological and geological analysis of Akchour landslide (Rif, Morocco). *Geo-Eco-Trop*, 42(1), 19–32.
- [10] Hong, H., Wang, D., Zhu, A.-X., & Wang, Y. (2024). Landslide susceptibility mapping based on the reliability of landslide and non-landslide sample. *Expert Systems with Applications*, 243, 122933. <https://doi.org/10.1016/j.eswa.2023.122933>
- [11] Li, X., & Liu, B. (2003). Learning to classify texts using positive and unlabeled data. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, 587-592.
- [12] Mastere, M., El Fellah, B., & Maquaire, O. (2020). Landslides inventory map as a first step for hazard and risk assessment: Rif mountains, Morocco. *Bulletin de l'Institut Scientifique, Rabat, Section Sciences de la Terre*, 42, 49–62
- [13] Chalouan, A., Michard, A., Feinberg, H., Montigny, R., & Saddiqi, O. (2001). The Rif Mountain building (Morocco): A new tectonic scenario. *Bulletin de la Société Géologique de France*, 172(5), 603–616. <https://doi.org/10.2113/172.5.603>
- [14] Mordellet, F., & Vert, J.-P. (2014). A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37, 201-209. <https://doi.org/10.1016/j.patrec.2013.06.010>
- [15] Rabby, Y. W., Li, Y., & Hilafu, H. (2023). An objective absence data sampling method for landslide susceptibility mapping. *Scientific Reports*, 13, 1740.
- [16] Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., & Guzzetti, F. (2018). A review of statistically-based landslide susceptibility models. *Earth-Science Reviews*, 180, 60-91.
- [17] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillerá-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroeder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017).

Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913-929. <https://doi.org/10.1111/ecog.02881>

- [18] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.
- [19] Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432.
- [20] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. Wiley.
- [21] Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2), 225-232. <https://doi.org/10.1111/2041-210X.13107>
- [22] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17, 230.
- [23] Wang, C., Lin, Q., Wang, L., Jiang, T., Su, B., Wang, Y., Mondal, S. K., Huang, J., & Wang, Y. (2022). The influences of the spatial extent selection for non-landslide samples on statistical-based landslide susceptibility modelling. *Natural Hazards*, 112(3), 1967-1988. <https://doi.org/10.1007/s11069-022-05252-8>
- [24] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.
- [25] Zhao, L., Ma, H., Dong, J., Wu, X., Xu, H., & Niu, R. (2023). A comparative study of landslide susceptibility mapping using Bagging PU learning in class-prior probability shift datasets. *Remote Sensing*, 15(23), 5547. <https://doi.org/10.3390/rs15235547>
- [26] Zhu, A.-X., Miao, Y., Liu, J., Bai, S., Wang, R., Wang, C., & Yang, L. (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>
- [27] Zhu, Y., Sun, D., Wen, H., Zhang, Q., Ji, Q., Li, C., Zhou, P., & Zhao, J. (2024). Considering the effect of non-landslide sample selection on landslide susceptibility assessment. *Geomatics, Natural Hazards and Risk*, 15(1). <https://doi.org/10.1080/19475705.2024.2392778>