

Dear Editors:

Thank you for your letter and for reviewers' careful review concerning our manuscript entitled “A Statistical-Physical Prior-Guided Framework for Multi-Class Landslide Susceptibility Mapping” ([Submission ID: egusphere-2026-4550](#)). Those comments are valuable and very helpful for revising and improving our paper. We have studied comments carefully and have made corrections which we hope meet with approval.

Looking forward to hearing from you soon.

With kindest regards,
Yours Sincerely

School of Remote Sensing and Information Engineering
Wuhan University, Wuhan, China, 430079

Replies to Reviewer 1

The authors would like to express their gratitude to the reviewer for his/her constructive and helpful comments for substantial improvement of this paper. The manuscript has been carefully revised according to your comments and suggestions, which are very valuable to the improvement of this manuscript. We hope our revision has improved the paper to a level of your satisfaction. The replies to your specific comments/suggestions are as follows.

Comments to the Author:

The study presents a promising direct multi-class susceptibility framework that combines certainty factor analysis, SHALSTAB constraints, six learning algorithms, and SHAP interpretation. The following technically focused comments are intended to strengthen its methodological validity and reproducibility.

Point 1:

Please demonstrate that the training and test partitions are spatially independent through spatial block cross-validation rather than random sampling.

Reply: Thank you for this valuable comment. We sincerely apologize that the original manuscript used a random data split and did not sufficiently ensure the spatial independence between the training and test sets. Following your suggestion, we have revised the data partitioning strategy by adopting spatial-block cross-validation and dividing the study area into spatially separated regions according to geographic location, thereby ensuring spatial independence between the training and test sets. We hope that this revision can reduce information leakage caused by spatial autocorrelation and improve the reliability of the model validation results.

Revision in our paper: (the revisions are marked in red)

L287:

During sample construction, stratified sampling was performed based on the five-class landslide susceptibility zoning obtained from the CF results constrained by SHALSTAB. To ensure spatial independence between the training and test sets, the study area was divided into a 3×4 spatial grid according to geographic location, and 10-fold spatial-block cross-validation was adopted. In each fold, eight spatial blocks were used as the training set and the remaining four blocks as the test set. In each fold, the CF statistical prior was recalculated using only the historical landslide samples within the training area and landslide records within the test area were not involved in CF construction. The SHALSTAB results were independently calculated based on the DEM and parameter inputs, and no landslide records from either the training or test sets were used in its construction. Fig. 4 shows the data distribution for one fold. To ensure that all susceptibility levels were represented in model training, 18 landslide points located in the very-low-susceptibility zone and one invalid landslide point were first

excluded in this fold. The grid cells containing the remaining 1,388 historical landslide points were assigned as very-high-susceptibility samples, while samples from the very low, low, moderate, and high susceptibility zones were selected according to the predefined sampling proportions. A five-class dataset containing 15,116 samples was ultimately constructed. The five susceptibility classes were very low, low, moderate, high, and very high, corresponding to label values of 1 to 5. Model input variables were extracted from the raster layers of landslide conditioning factors, with each sample containing a set of conditioning-factor values and its corresponding susceptibility class. To further evaluate the predictive capability of the models for newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 were used as a temporally independent validation dataset. These landslide records were not involved in label construction or model training, and were used only for independent validation.

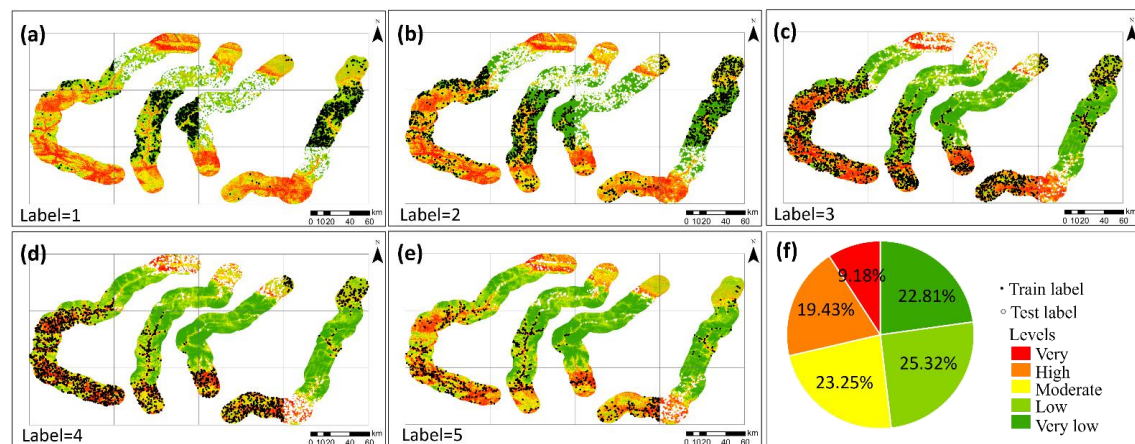


Fig. 5. Distribution of landslide susceptibility samples in one fold. (a) Very low susceptibility samples; (b) Low susceptibility samples; (c) Moderate susceptibility samples; (d) High susceptibility samples; (e) Very high susceptibility samples; (f) Sample composition.

Point 2:

Please quantify spatial autocorrelation in landslide occurrences and conditioning factors, then show how it is controlled during validation.

Reply: Thank you for this valuable comment. We sincerely apologize that the spatial autocorrelation of landslide occurrence and the conditioning factors was not quantitatively assessed in the original manuscript. Following your suggestion, we added a Global Moran's I analysis to quantify the spatial autocorrelation of landslide occurrence and the conditioning factors. To account for this spatial dependence, spatial-block validation was adopted to spatially separate the training and test regions, thereby reducing information leakage and potential overestimation of model performance caused by spatially adjacent samples appearing in both the training and test sets. We hope that these additions more clearly demonstrate the spatial dependence of the data and how it was controlled during model validation.

Revision in our paper: (the revisions are marked in red)

L257:

To quantitatively evaluate the spatial dependence of landslide occurrence and the conditioning factors, this study employed the Global Moran's I statistic for spatial autocorrelation analysis. The spatial weight matrix was constructed using the k-nearest-neighbor method with k=8 and was standardized. Moran's I was calculated as follows:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})^2} \quad (7)$$

where n is the number of samples, x_i and x_j are the observed values at locations i and j , respectively, $\bar{x}$ is the mean of the observed values, w_{ij} is the spatial weight, and S_0 is the sum of the spatial weights. Moran's I generally ranges from -1 to 1 . A positive value indicates positive spatial autocorrelation, meaning that similar attributes tend to cluster spatially; a value close to 0 indicates a random spatial distribution; and a negative value indicates spatial dispersion.

As shown in Table 3, the Moran's I for landslide occurrence is 0.249 , indicating positive spatial autocorrelation in the landslide samples. This suggests that landslides are not completely randomly distributed but exhibit a certain degree of spatial clustering. The strength of spatial autocorrelation varies considerably among the conditioning factors. Distance to epicenters, rainfall, distance to rivers, and distance to faults exhibit the strongest spatial autocorrelation, while DEM, distance to roads, NDVI, soil type, and rock groups also show strong spatial clustering. Therefore, spatial-block validation was adopted in the subsequent model evaluation to spatially separate the training and test sets, thereby reducing information leakage and potential overestimation of model performance caused by spatially adjacent samples appearing in both sets.

Table 3

Moran's I results for landslide occurrence and conditioning factors.

Variable	Moran's I	p-value	Variable	Moran's I	p-value
DEM	0.936	0.005	Soil type	0.826	0.005
Aspect	0.423	0.005	NDVI	0.875	0.005
Slope	0.053	0.005	Distance to rivers	0.995	0.005
Relief amplitude	0.394	0.005	TWI	0.215	0.005
Rock groups	0.800	0.005	Rainfall	0.997	0.005
Distance to faults	0.989	0.005	LULC	0.482	0.005
Distance to epicenters	1	0.005	Distance to roads	0.894	0.005
			Landslide occurrence	0.249	0.005

Point 3:

Please justify the five susceptibility classes with an objective, reproducible rule and assess sensitivity to alternative class boundaries.

Reply: Thank you Thank you for this valuable comment. We sincerely apologize that the basis for the five susceptibility classes and their sensitivity were not sufficiently explained in the original manuscript. Following your suggestion, we further clarified the rule used to construct the five susceptibility classes. We hope that these revisions

more clearly demonstrate the objectivity, reproducibility, and stability of the five-class classification scheme.

Revision in our paper: (the revisions are marked in red)

L400:

To evaluate the historical-landslide coverage and spatial concentration of the susceptibility maps, the areal proportion of susceptibility class (a), the historical-landslide hit rate (h), and HOI were calculated. Binary results were summarized in two ways: Table 7 reports statistics after dividing binary probability maps into five classes by natural breaks, and Table 8 reports the landslide-prone zones defined by the binary threshold ($p > 0.5$). As shown in Tables 7 and 8, TabPFN achieved the highest threshold-based binary HOI (1.70); its landslide-prone zone occupied 16.9% of the study area and contained 86.7% of the historical landslides. Its HOI based on the five-class natural-breaks result was 1.68, lower than the threshold-based HOI and confirming that evaluating a binary model solely by its very high susceptibility class underestimates model capability. Table 9 shows that TabPFN also performed best in the multi-class setting. Its very high susceptibility zone occupied only 10.8% of the study area but contained 86.0% of the historical landslides, producing an HOI of 1.75. The comparison of HOI between binary and multi-class classification is shown in Fig. 6, indicating that the multi-class classification approach can produce a more spatially concentrated high-susceptibility zone.

Table 7

Statistical results of landslide susceptibility maps based on binary classification using natural breaks (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	50.6	11.0	16.3	16.9	21.4	40.2	7.3	24.7	4.4	7.1	1.03
FT-Transformer	62.7	25.4	17.2	20.0	8.8	20.6	7.7	19.0	3.6	15.1	1.11
SVM	66.4	8.4	6.9	6.8	9.1	16.0	12.0	40.5	5.7	28.3	1.23
XGBoost	61.0	3.4	9.0	4.3	7.1	7.0	6.4	5.2	16.5	80.1	1.64
RF	61.1	3.3	9.9	3.3	7.0	6.3	7.1	10.7	14.9	76.4	1.61
TabPFN	56.0	1.5	11.5	1.5	7.3	2.7	7.4	5.0	17.8	85.3	1.68

Table 8

Statistical results of landslide susceptibility maps based on binary classification using threshold (best values in bold).

Model	Non-landslides		Landslides		HOI
	a_{neg}	h_{neg}	a_{pos}	h_{pos}	
LR	92.6	81.0	7.4	19.0	1.12
FT-Transformer	91.8	72.2	8.2	27.8	1.20
SVM	85.0	40.6	15.0	59.4	1.44
XGBoost	87.2	29.0	12.8	71.0	1.58
RF	82.2	18.8	17.8	81.2	1.63

TabPFN	83.1	13.3	16.9	86.7	1.70
---------------	-------------	-------------	-------------	-------------	-------------

Table 9

Statistical results of landslide susceptibility maps based on multi-class classification (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	16.2	1.2	30.2	8.0	15.2	13.2	22.7	20.1	15.6	57.4	1.42
FT-Transformer	18.2	0.2	26.6	7.6	22.4	8.7	21.5	22.5	11.3	61.0	1.50
SVM	17.1	0.7	26	4.7	22.6	5.4	20.8	11.3	13.5	77.8	1.64
XGBoost	16.8	0.2	28.1	4	20.8	7.8	21.6	11.8	12.7	76.1	1.63
RF	16.8	0.2	28.3	3.1	21.9	7.1	21.5	6.9	11.5	82.7	1.71
TabPFN	18.2	0.1	27.6	3.5	22.4	4.8	21.0	5.6	10.8	86.0	1.75

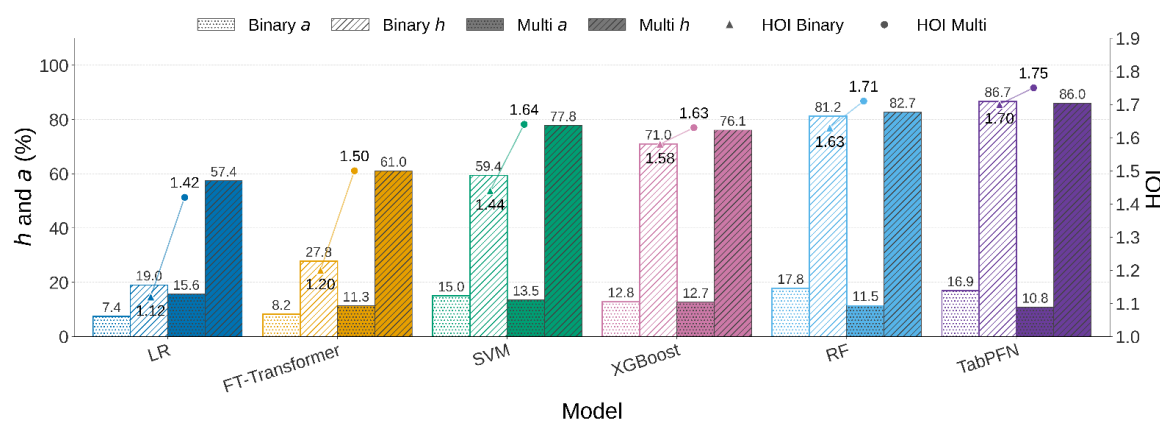


Fig. 6. Comparison of area ratio (a), hit rate (h), and HOI for binary and multi-class susceptibility maps.

The spatial susceptibility patterns produced by the six models under the binary and multi-class settings are shown in Fig. 7. Overall, very high susceptibility zones formed belts or patches concentrated mainly along valleys. In the binary results, most historical landslides occurred within the moderate, high, and very high susceptibility classes. Thus, binary classification identified broadly landslide-prone areas, but the resulting high-risk regions were relatively extensive and increased the area requiring attention. In the multi-class results, historical landslides were more concentrated within the very high susceptibility class, producing a more focused high-risk area. Enlarged views of the TabPFN results show that binary classification generated a broad transition zone of moderate to high susceptibility around landslide locations, whereas multi-class classification assigned the landslide locations more directly to the very high susceptibility class and reduced overestimation in surrounding areas. This further demonstrates that multi-class classification improves the spatial concentration of the very high susceptibility zone.

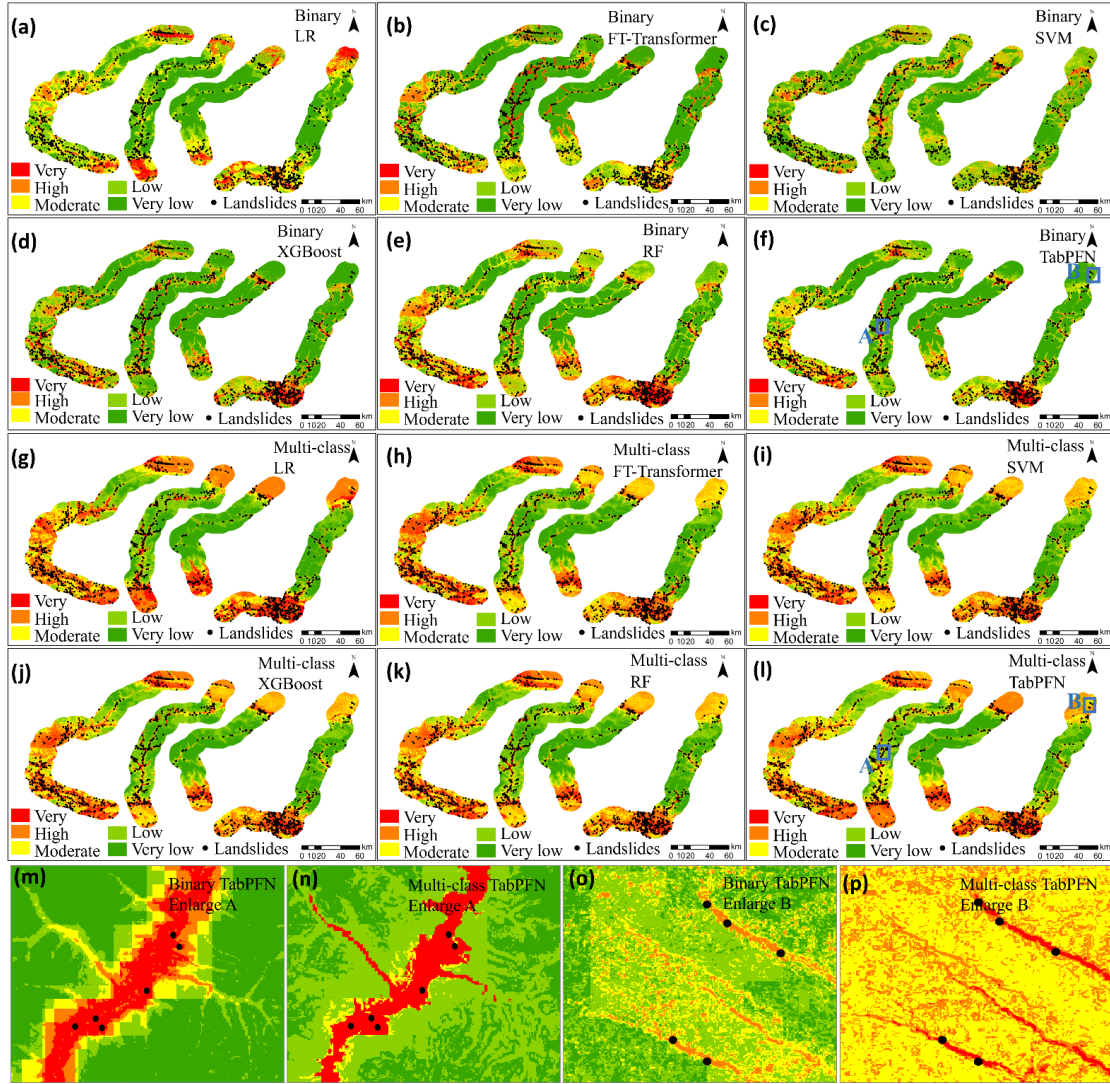


Fig. 7. Landslide susceptibility maps produced by six models under binary and multi-class classification settings. (a–f) Binary classification results of LR, FT-Transformer, SVM, XGBoost, RF and TabPFN; (g–l) multi-class classification results of LR, FT-Transformer, SVM, XGBoost, RF and TabPFN; (m–p) enlarged views of selected TabPFN results.

Point 4:

Please evaluate uncertainty in the certainty factor and SHALSTAB-derived labels, since these labels are treated as reference truth for multi-class learning.

Reply: Thank you for this valuable comment. We sincerely apologize that the uncertainty of the labels derived from the CF statistical prior and SHALSTAB physical constraint was not sufficiently evaluated in the original manuscript. Following your suggestion, we have added an analysis of label uncertainty. We hope that these additions can improve the completeness of the proposed methodology.

Revision in our paper: (the revisions are marked in red)

L454:

5.2. Label uncertainty analysis

To evaluate the uncertainty introduced by the CF statistical prior and SHALSTAB physical constraints into the construction of multi-class labels, landslide samples from eight spatial blocks of the study area were randomly selected each time to construct the CF model, and this procedure was repeated 10 times to characterize the influence of spatial variations in landslide samples on the CF results. Subsequently, the five key SHALSTAB parameters were individually set to 0.8, 1.0, and 1.2 times their baseline values, yielding 15 sets of results. The 10 CF results were then combined with the 15 SHALSTAB results, generating a total of 150 initial landslide susceptibility classification results, which were used to evaluate the uncertainty of the five-class susceptibility classification. The 150 classification results were integrated to obtain the susceptibility class distribution for each grid cell, as shown in Fig. 9(a). Meanwhile, the standard deviation of susceptibility classes across different combined results was used to characterize the variation around the mean class and to quantify classification uncertainty, as shown in Fig. 9(b). A smaller standard deviation indicates less variation in the susceptibility class of a given grid cell across different combined results and, consequently, greater classification stability. The standard deviation is close to 0 across most of the study area, whereas locally higher values, approximately 0.4, occur mainly in areas where susceptibility classes differ markedly and are spatially dispersed.

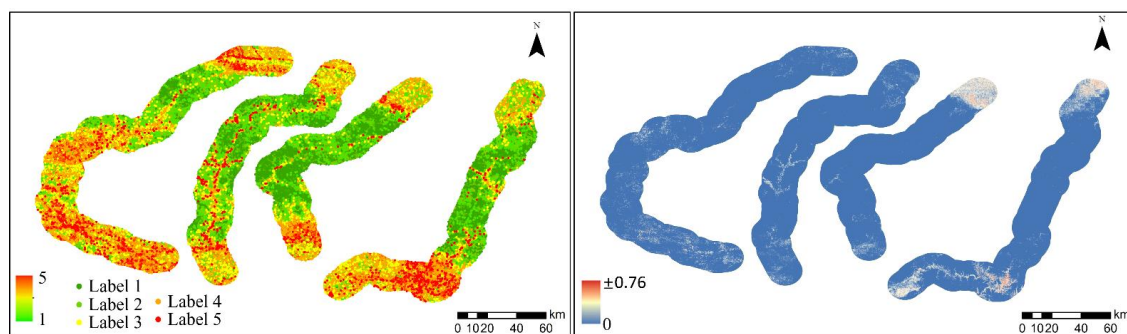


Fig. 9. Spatial distribution of landslide susceptibility classes and their uncertainty. (a) Mean landslide susceptibility classes; (b) Standard deviation of susceptibility classes.

Point 5:

Please clarify whether SHALSTAB parameters were locally calibrated and provide a sensitivity analysis for soil depth, transmissivity, cohesion, and rainfall inputs.

Reply: Thank you for this valuable comment. We apologize for the insufficient clarification in the original manuscript. We have further clarified the SHALSTAB model and parameter settings and added the corresponding sensitivity analysis. We hope these revisions improve the completeness and clarity of the manuscript.

Revision in our paper: (the revisions are marked in red)

L210:

SHALSTAB is a representative physically based model that generally uses a DEM and integrates topographic conditions, geotechnical parameters, and stability criteria to identify spatial units with a clear tendency toward failure (Montgomery and Dietrich, 1994; Ye et al., 2025). The model jointly considers slope angle, upslope contributing area, steady-state hydrological processes, and geotechnical parameters.

Soil thickness is estimated from the local slope using:

$$z = h_{max} [1 - \frac{\tan \theta - \tan \theta_{min}}{\tan \theta_{max} - \tan \theta_{min}} (1 - \frac{h_{min}}{h_{max}})] \quad (5)$$

where θ is the slope angle, and h_{min} and h_{max} are set to 0.2 m and 4.5 m, respectively. Thus, steeper slopes are assigned smaller soil thicknesses. Soil transmissivity T is calculated from the saturated hydraulic conductivity K_s , soil thickness z , and slope angle θ , as $T = K_s z \cos \theta$. The SHALSTAB stability criterion can be expressed as:

$$\frac{a}{b} > \frac{T \sin \theta}{q} \left[\frac{\gamma_w}{\gamma} \left(1 - \frac{\tan \theta}{\tan \varphi} \right) + \frac{C}{\gamma_w z \cos^2 \theta \tan \varphi} \right] \quad (6)$$

where $\frac{a}{b}$ denotes the upslope contributing area per unit contour length, and q is the prescribed rainfall input, set to 100 mm/day in this study. C is cohesion, γ is the natural unit weight of the soil or rock mass, φ is the internal friction angle, and γ_w is the unit weight of water. When the left-hand side exceeds the right-hand side, the grid cell is classified as unstable; otherwise, it is classified as stable.

Based on parameter values reported in previous studies and related information on landslides in the Qinling–Daba Mountains (Qiang et al., 2015; Song et al., 2022), initial values of C , θ , γ , and K_s were assigned to the different engineering geological types. The corresponding parameter settings are listed in Table 2. Because it is difficult to determine spatially distributed geotechnical parameters over such a large regional study area, no local calibration based on site-specific measurements was performed. Most parameters are inherently subject to spatial uncertainty. Therefore, the parameter values adopted for the study area should be regarded as representative relative values rather than uniquely determined absolute values.

Table 2

Parameter settings of different lithologies in the SHALSTAB model.

Engineering geological type	C (kPa)	φ (°)	γ (kN/m ³)	Ks (m/s)
Loose gravelly soil deposits	7.0	6.3	41.4	5.00E-05
Massive hard metamorphic rock group	13.0	15.0	43.0	5.00E-07
Massive hard intrusive rock group	14.0	16.0	44.0	2.50E-07
Schistose–bedded weak metamorphic rock group	8.0	9.0	41.6	5.00E-09
Schistose–massive moderately hard metamorphic rock group	11.0	13.0	42.4	2.50E-08
Medium- to thick-bedded interbedded hard–soft metamorphic rock group	9.0	11.0	42.0	5.00E-08

L433:

5.1. Sensitivity analysis of SHALSTAB parameters

To evaluate the response of the SHALSTAB model to variations in key input parameters, a one-at-a-time (OAT) sensitivity analysis was conducted, in which only one parameter was varied at a time while all other parameters were held at their baseline values. Each parameter was varied from 0.1 to 2.0 times its baseline value, and the SHALSTAB model was rerun for each scenario. The proportion of historical landslides located within unstable areas was used as the model response indicator. A sensitivity coefficient S was used to quantify the relative change in model output caused by the relative change in each parameter, as follows:

$$S = \frac{\Delta Y / Y_0}{\Delta X / X_0} = \frac{(Y - Y_0) / Y_0}{f - 1} \quad (14)$$

where X is the perturbed parameter value, X_0 is the baseline parameter value, Y is the proportion of historical landslides within unstable areas under the perturbed scenario, Y_0 is the corresponding proportion under the baseline scenario, and f is the parameter multiplier relative to the baseline value, with $f = 1$ representing the baseline condition. The sign of S indicates the direction of the relationship between parameter variation and model response, whereas a larger S indicates greater model sensitivity to the corresponding parameter. Fig. 8 shows clear differences in the sensitivity of SHALSTAB to different parameters. Increasing the φ and C generally reduced the historical-landslide coverage within unstable areas, with φ showing the stronger effect. γ was sensitive mainly at very low parameter values, while its influence became small when $f \geq 0.3$. In contrast, K_s and rainfall q caused only minor changes, with sensitivity coefficients close to zero.

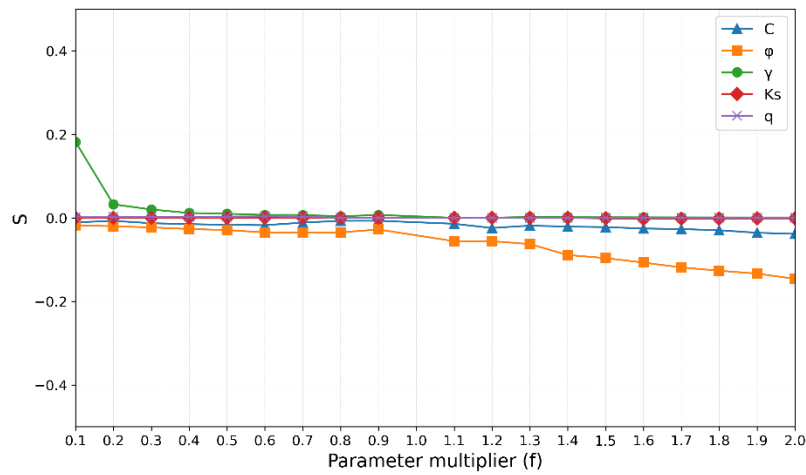


Fig. 8. Sensitivity coefficients of SHALSTAB parameters under different parameter multipliers.

Table 10 further shows that the baseline scenario ($f=1$) covered 98.4% of historical landslides, and in most perturbation scenarios the coverage remained above 95%, indicating that the SHALSTAB results were generally stable despite differences in parameter sensitivity.

Table 10

Historical-landslide coverage within unstable areas under different SHALSTAB parameter multipliers.

f	0.1	0.3	0.5	0.7	0.9	1	1.1	1.3	1.5	1.7	1.9	2
C	99.4	99.3	99.2	98.8	98.5	98.4	98.3	97.9	97.3	96.6	95.3	94.6
φ	100.0	100.0	99.9	99.4	98.7	98.4	97.9	96.6	93.7	90.2	86.6	84.0
γ	82.3	97.0	97.9	98.2	98.3	98.4	98.4	98.5	98.5	98.5	98.5	98.5
K_s	98.4	98.4	98.4	98.4	98.4	98.4	98.4	98.4	98.3	98.3	98.3	98.3
q	98.3	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4	98.4	98.4	98.4

Point 6:

Please assess multicollinearity, nonlinear dependence, and scale mismatch among the 14 conditioning factors before interpreting model performance or SHAP rankings.

Reply: Thank you for this valuable comment. We sincerely apologize that

multicollinearity, potential nonlinear dependence, and scale mismatch among the 14 landslide conditioning factors were not sufficiently assessed in the original manuscript. Following your suggestion, we added a correlation analysis to evaluate potential information redundancy and multicollinearity among the conditioning factors. We hope that these additions make the factor selection and model interpretation more rigorous and complete.

Revision in our paper: (the revisions are marked in red)

L257:

To quantitatively evaluate the spatial dependence of landslide occurrence and the conditioning factors, this study employed the Global Moran's I statistic for spatial autocorrelation analysis. The spatial weight matrix was constructed using the k-nearest-neighbor method with k=8 and was standardized. Moran's I was calculated as follows:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})^2} \quad (7)$$

where n is the number of samples, x_i and x_j are the observed values at locations i and j , respectively, $\bar{x}$ is the mean of the observed values, w_{ij} is the spatial weight, and S_0 is the sum of the spatial weights. Moran's I generally ranges from -1 to 1 . A positive value indicates positive spatial autocorrelation, meaning that similar attributes tend to cluster spatially; a value close to 0 indicates a random spatial distribution; and a negative value indicates spatial dispersion.

As shown in Table 3, the Moran's I for landslide occurrence is 0.249 , indicating positive spatial autocorrelation in the landslide samples. This suggests that landslides are not completely randomly distributed but exhibit a certain degree of spatial clustering. The strength of spatial autocorrelation varies considerably among the conditioning factors. Distance to epicenters, rainfall, distance to rivers, and distance to faults exhibit the strongest spatial autocorrelation, while DEM, distance to roads, NDVI, soil type, and rock groups also show strong spatial clustering. Therefore, spatial-block validation was adopted in the subsequent model evaluation to spatially separate the training and test sets, thereby reducing information leakage and potential overestimation of model performance caused by spatially adjacent samples appearing in both sets.

Table 3

Moran's I results for landslide occurrence and conditioning factors.

Variable	Moran's I	p-value	Variable	Moran's I	p-value
DEM	0.936	0.005	Soil type	0.826	0.005
Aspect	0.423	0.005	NDVI	0.875	0.005
Slope	0.053	0.005	Distance to rivers	0.995	0.005
Relief amplitude	0.394	0.005	TWI	0.215	0.005
Rock groups	0.800	0.005	Rainfall	0.997	0.005
Distance to faults	0.989	0.005	LULC	0.482	0.005
Distance to epicenters	1	0.005	Distance to roads	0.894	0.005
			Landslide occurrence	0.249	0.005

To assess potential correlations and information redundancy among the landslide conditioning factors, pairwise correlation analysis was conducted, and a correlation coefficient matrix was used to evaluate the degree of association among the factors. The closer the absolute value of the correlation coefficient is to 1, the stronger the correlation between two factors; values closer to 0 indicate weaker correlations. In this study, $|r|=0.7$ was used as a reference threshold for strong correlation. When the absolute correlation coefficient between two factors exceeded 0.7, relatively substantial information overlap or multicollinearity was considered possible. As shown in Fig. 4, the overall correlations among the 14 landslide conditioning factors are relatively low, with all pairwise absolute correlation coefficients below 0.7, indicating no obvious strong correlations. Among them, distance to faults and distance to rivers show a relatively high positive correlation, with a correlation coefficient of 0.60. DEM and LULC exhibit a relatively strong negative correlation, with a correlation coefficient of -0.53, while the correlation coefficient between DEM and distance to rivers is 0.51. The absolute correlation coefficients of most other factor pairs are below 0.5. Overall, no severe correlations or obvious information redundancy are observed among the conditioning factors; therefore, all 14 factors were retained for subsequent landslide susceptibility modeling.

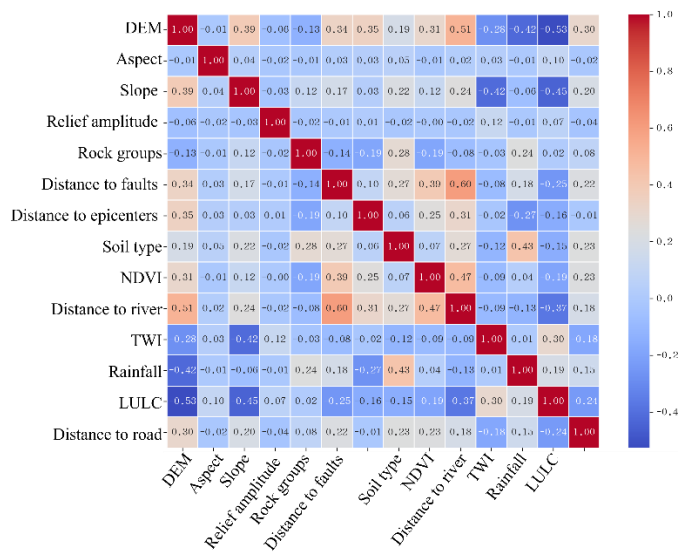


Fig. 4. Correlation matrix of the landslide conditioning factors.

Point 7:

Please report the strategy used to generate non-landslide samples and test whether conclusions remain stable under alternative pseudo-absence designs.

Reply: Thank you for this valuable comment. We sincerely apologize that the strategy for generating non-landslide samples and its potential influence on the model results were not sufficiently described in the original manuscript. Following your suggestion, we further clarified the method used to select non-landslide samples. We hope that these additions further demonstrate the rationality of the non-landslide sample construction and the robustness of the study conclusions.

Revision in our paper: (the revisions are marked in red)

L385:

To systematically evaluate the suitability of different models for multi-class landslide susceptibility mapping, LR, FT-Transformer, SVM, XGBoost, RF, and TabPFN were compared. Experiments were conducted using both binary landslide and multi-class susceptibility datasets. For the binary dataset, non-landslide samples were randomly selected from non-landslide areas at a 1:1 ratio to landslide samples, following commonly used sampling strategies in landslide susceptibility modeling (Chang et al., 2023b; Jia et al., 2026). The binary task evaluated discrimination between landslide and non-landslide samples, whereas the multi-class task focused on distinctions among susceptibility classes. The model performance results are presented in Tables 5 and 6. TabPFN performed best in the binary task, with F1, accuracy, and AUC values of 82.62%, 82.14%, and 91.00%, respectively. TabPFN also achieved the best overall multi-class performance, with corresponding values of 89.26%, 89.04%, and 95.26%. Overall, TabPFN achieved the best comprehensive performance, while RF and XGBoost also showed similarly good performance across both tasks.

Table 5

Performance of landslide susceptibility models on the binary classification dataset (best values in bold).

Model	Non-landslide	Landslide	Overall					
	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance ($\times 10^{-5}$)
LR	71.35	73.20	72.51	72.27	72.51	72.27	81.63	29.04
FT-Transformer	76.31	81.42	79.35	78.86	79.35	78.86	87.71	21.25
SVM	77.71	81.16	79.76	79.43	79.76	79.43	87.80	26.35
XGBoost	77.35	83.59	81.05	80.47	81.05	80.47	89.56	18.93
RF	78.16	84.23	81.73	81.19	81.73	81.19	90.02	16.75
TabPFN	79.28	85.00	82.62	82.14	82.62	82.14	91.00	19.27

Table 6

Performance of landslide susceptibility models on the multi-class dataset (best values in bold).

Model	Very low	Low	Mode rate	High	Very high	Overall					
	Recall	Recall	Recall	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance ($\times 10^{-5}$)
LR	69.05	66.90	31.90	52.38	57.86	55.63	56.13	55.62	55.73	81.15	66.88
FT-Transformer	89.52	76.19	72.14	73.57	61.43	74.76	67.41	74.57	74.12	83.60	60.28
SVM	89.52	82.86	82.14	79.76	78.33	82.53	80.51	82.52	82.51	86.05	53.68
XGBoost	92.86	84.05	74.05	79.76	76.67	81.52	79.43	81.48	80.91	91.12	23.91
RF	94.52	88.81	84.52	87.14	83.33	87.78	85.67	87.67	87.32	90.10	30.67
TabPFN	95.26	90.44	86.21	88.73	85.67	89.35	87.47	89.26	89.04	92.35	95.26

- Chang, Z., Huang, J., Huang, F., Bhuyan, K., Meena, S.R., Catani, F., 2023b. Uncertainty analysis of non-landslide sample selection in landslide susceptibility prediction using slope unit-based machine learning models. *Gondwana Research* 117, 307-320. <https://doi.org/10.1016/j.gr.2023.02.007>.
- Jia, Z., Cheng, Z., Chang, Z., Li, Q., Huang, F., Peng, Y., 2026. Non-landslide sample for landslide susceptibility prediction modeling: A review of selection strategies and their influence rules. *Journal of Rock Mechanics and Geotechnical Engineering* 18, 2859-2880. <https://doi.org/10.1016/j.jrmge.2025.12.011>.

Point 8:

Please compare models using class-specific precision, recall, F1 score, calibration measures, and confusion matrices, not AUC and overall accuracy alone.

Reply: Thank you for this valuable comment. We sincerely apologize that the original manuscript relied mainly on AUC and overall accuracy. Following your suggestion, we added class-specific precision, recall, and F1-score to provide a more comprehensive evaluation of model performance across different susceptibility classes. We hope that these additions make the model evaluation more complete and reliable.

Revision in our paper: (the revisions are marked in red)

L356:

Model performance was evaluated using class-specific recall together with F1, accuracy, and AUC for the multi-class task, thereby preventing overall accuracy from obscuring poor recognition of an individual class (Hand and Till, 2001; Sokolova and Lapalme, 2009). The metrics were calculated as follows:

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (8)$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (9)$$

$$F1 = \frac{1}{K} \sum_{i=1}^K F1_i$$

$$Accuracy = \frac{N_{correct}}{N_{total}} \quad (10)$$

$$AUC = \frac{1}{K} \sum_{i=1}^K AUC_i \quad (11)$$

where TP_i , FP_i , and FN_i denote the true positives, false positives, and false negatives for class i , respectively; K is the number of classes ($K = 5$); $N_{correct}$ is the number of correctly classified samples; N_{total} is the total number of samples; and AUC_i is the one-versus-rest AUC for class i .

To ensure that the final evaluation uses landslide observations that were not involved in any previous modeling or label-construction process, the spatial mapping results were further evaluated using the HOI to measure the ability of the very-high-susceptibility zone to capture historical landslide points within a relatively small area (Zhang et al., 2026). HOI jointly considers the areal proportion of the very high susceptibility zone and the hit rate of historical landslides within that zone and is calculated as follows:

$$HOI = 1 - a_5 + h_5 \tag{12}$$

L385:

To systematically evaluate the suitability of different models for multi-class landslide susceptibility mapping, LR, FT-Transformer, SVM, XGBoost, RF, and TabPFN were compared. Experiments were conducted using both binary landslide and multi-class susceptibility datasets. For the binary dataset, non-landslide samples were randomly selected from non-landslide areas at a 1:1 ratio to landslide samples, following commonly used sampling strategies in landslide susceptibility modeling (Chang et al., 2023b; Jia et al., 2026). The binary task evaluated discrimination between landslide and non-landslide samples, whereas the multi-class task focused on distinctions among susceptibility classes. The model performance results are presented in Tables 5 and 6. TabPFN performed best in the binary task, with F1, accuracy, and AUC values of 82.62%, 82.14%, and 91.00%, respectively. TabPFN also achieved the best overall multi-class performance, with corresponding values of 89.26%, 89.04%, and 95.26%. Overall, TabPFN achieved the best comprehensive performance, while RF and XGBoost also showed similarly good performance across both tasks.

Table 5

Performance of landslide susceptibility models on the binary classification dataset (best values in bold).

Model	Non-landslide	Landslide	Overall					
	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance (×10 ⁻⁵)
LR	71.35	73.20	72.51	72.27	72.51	72.27	81.63	29.04
FT-Transformer	76.31	81.42	79.35	78.86	79.35	78.86	87.71	21.25
SVM	77.71	81.16	79.76	79.43	79.76	79.43	87.80	26.35
XGBoost	77.35	83.59	81.05	80.47	81.05	80.47	89.56	18.93
RF	78.16	84.23	81.73	81.19	81.73	81.19	90.02	16.75
TabPFN	79.28	85.00	82.62	82.14	82.62	82.14	91.00	19.27

Table 6

Performance of landslide susceptibility models on the multi-class dataset (best values in bold).

Model	Very low	Low	Mode rate	High	Very high	Overall					
	Recall	Recall	Recall	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance (×10 ⁻⁵)
LR	71.35	73.20	72.51	72.27	72.51	72.27	81.63	29.04	71.35	73.20	72.51
FT-Transformer	76.31	81.42	79.35	78.86	79.35	78.86	87.71	21.25	76.31	81.42	79.35
SVM	77.71	81.16	79.76	79.43	79.76	79.43	87.80	26.35	77.71	81.16	79.76
XGBoost	77.35	83.59	81.05	80.47	81.05	80.47	89.56	18.93	77.35	83.59	81.05
RF	78.16	84.23	81.73	81.19	81.73	81.19	90.02	16.75	78.16	84.23	81.73
TabPFN	79.28	85.00	82.62	82.14	82.62	82.14	91.00	19.27	79.28	85.00	82.62

LR	69.05	66.90	31.90	52.38	57.86	55.63	56.13	55.62	55.73	81.15	66.88
FT-Transformer	89.52	76.19	72.14	73.57	61.43	74.76	67.41	74.57	74.12	83.60	60.28
SVM	89.52	82.86	82.14	79.76	78.33	82.53	80.51	82.52	82.51	86.05	53.68
XGBoost	92.86	84.05	74.05	79.76	76.67	81.52	79.43	81.48	80.91	91.12	23.91
RF	94.52	88.81	84.52	87.14	83.33	87.78	85.67	87.67	87.32	90.10	30.67
TabPFN	95.26	90.44	86.21	88.73	85.67	89.35	87.47	89.26	89.04	92.35	95.26

4.2. Evaluation of landslide susceptibility maps

To evaluate the historical-landslide coverage and spatial concentration of the susceptibility maps, the areal proportion of susceptibility class (a), the historical-landslide hit rate (h), and HOI were calculated. Binary results were summarized in two ways: Table 7 reports statistics after dividing binary probability maps into five classes by natural breaks, and Table 8 reports the landslide-prone zones defined by the binary threshold ($p > 0.5$). As shown in Tables 7 and 8, TabPFN achieved the highest threshold-based binary HOI (1.70); its landslide-prone zone occupied 16.9% of the study area and contained 86.7% of the historical landslides. Its HOI based on the five-class natural-breaks result was 1.68, lower than the threshold-based HOI and confirming that evaluating a binary model solely by its very high susceptibility class underestimates model capability. Table 9 shows that TabPFN also performed best in the multi-class setting. Its very high susceptibility zone occupied only 10.8% of the study area but contained 86.0% of the historical landslides, producing an HOI of 1.75. The comparison of HOI between binary and multi-class classification is shown in Fig. 6, indicating that the multi-class classification approach can produce a more spatially concentrated high-susceptibility zone.

Table 7

Statistical results of landslide susceptibility maps based on binary classification using natural breaks (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	50.6	11.0	16.3	16.9	21.4	40.2	7.3	24.7	4.4	7.1	1.03
FT-Transformer	62.7	25.4	17.2	20.0	8.8	20.6	7.7	19.0	3.6	15.1	1.11
SVM	66.4	8.4	6.9	6.8	9.1	16.0	12.0	40.5	5.7	28.3	1.23
XGBoost	61.0	3.4	9.0	4.3	7.1	7.0	6.4	5.2	16.5	80.1	1.64
RF	61.1	3.3	9.9	3.3	7.0	6.3	7.1	10.7	14.9	76.4	1.61
TabPFN	56.0	1.5	11.5	1.5	7.3	2.7	7.4	5.0	17.8	85.3	1.68

Table 8

Statistical results of landslide susceptibility maps based on binary classification using threshold (best values in bold).

Model	Non-landslides		Landslides		HOI
	a_{neg}	h_{neg}	a_{pos}	h_{pos}	
LR	92.6	81.0	7.4	19.0	1.12

FT-Transformer	91.8	72.2	8.2	27.8	1.20
SVM	85.0	40.6	15.0	59.4	1.44
XGBoost	87.2	29.0	12.8	71.0	1.58
RF	82.2	18.8	17.8	81.2	1.63
TabPFN	83.1	13.3	16.9	86.7	1.70

Table 9

Statistical results of landslide susceptibility maps based on multi-class classification (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	16.2	1.2	30.2	8.0	15.2	13.2	22.7	20.1	15.6	57.4	1.42
FT-Transformer	18.2	0.2	26.6	7.6	22.4	8.7	21.5	22.5	11.3	61.0	1.50
SVM	17.1	0.7	26	4.7	22.6	5.4	20.8	11.3	13.5	77.8	1.64
XGBoost	16.8	0.2	28.1	4	20.8	7.8	21.6	11.8	12.7	76.1	1.63
RF	16.8	0.2	28.3	3.1	21.9	7.1	21.5	6.9	11.5	82.7	1.71
TabPFN	18.2	0.1	27.6	3.5	22.4	4.8	21.0	5.6	10.8	86.0	1.75

Point 9:

I will also encourage the authors to discuss the incorporation of physics-guided frameworks in other environmental applications, for instance, the following studies are useful among others, i.e., “ <https://doi.org/10.21015/vtse.v13i3.2230> ” , “ <https://doi.org/10.21015/vtcs.v13i2.2229> ” , “ <https://doi.org/10.21015/vtse.v13i4.2231> ” , and “ <https://doi.org/10.1007/s44163-026-00833-z> ”.

Reply: Thank you Thank you for this helpful suggestion. We have added a brief discussion in the Introduction on the application of physics-guided learning in other environmental fields. We hope that this addition makes the manuscript more complete and comprehensive.

Revision in our paper: (the revisions are marked in red)

L83:

Physics-guided learning has also been increasingly explored in other environmental applications. Inam et al. (2025a, 2025b, 2025c) incorporated meteorological and thermodynamic constraints into neural networks for wind-speed, humidity, and temperature forecasting, while Inam et al. (2026) introduced a humidity-based physical constraint into rainfall prediction. In hydrology, Deng et al. (2024) integrated conceptual water-balance models with machine learning and examined the role of physically meaningful intermediate variables in runoff simulation. These studies demonstrate the broader potential of incorporating domain knowledge and physical constraints into data-driven environmental modeling to improve physical consistency and interpretability.

- Deng, C., Sun, P., Yin, X., Zou, J., Wang, W., 2024. Assessment of monthly runoff simulations based on a physics-informed machine learning framework: The effect of intermediate variables in its construction. *Journal of Environmental Management* 362, 121299. <https://doi.org/10.1016/j.jenvman.2024.121299>.
- Inam, S.A., Hashim, H., Ashraf, S.N., Naim, S.W., Rajput, H., Umer, S., 2025a. AI-driven Physics Informed Neural Network for Daily Temperature Forecasting with Constraint Aware Learning and Explainable Feature Attribution. *VFAST Transactions on Software Engineering* 13, 53-67. <https://doi.org/10.21015/vtse.v13i4.2231>.
- Inam, S.A., Hashim, H., Awan, A.M., Rajput, H., Umer, S., 2025b. A Novel Approach toward Windspeed Forecasting using an Advanced Deep Learning Framework with Explainable AI. *VFAST Transactions on Software Engineering* 13, 198-210. <https://doi.org/10.21015/vtse.v13i3.2230>.
- Inam, S.A., Hashim, H., Awan, A.M., Rajput, H., Umer, S., 2025c. Interpretable Physics-Informed Neural Network for Reliable Humidity Forecasting in Contrasting Climates of Sindh. *VAWKUM Transactions on Computer Sciences* 13, 135-148. <https://doi.org/10.21015/vtcs.v13i2.2229>.
- Inam, S.A., Umer, S., Rajput, H., 2026. A physics informed deep learning framework for rainfall forecasting in diverse climatic regions. *Discover Artificial Intelligence* 6, 104. <https://doi.org/10.1007/s44163-026-00833-z>.

Point 10:

Please test whether the reported improvements in AUC, accuracy, and HOI are statistically significant across repeated spatial resampling folds.

Reply: Thank you for this valuable comment. We sincerely apologize that the variability in model performance caused by different spatial partitions and repeated experiments was not sufficiently considered in the original manuscript. Following your suggestion, we adopted 10 spatial-block partitions and conducted multiple repeated experiments under the same spatial partition in each fold to reduce the influence of a single data split and random variation on model evaluation. We hope that these additions make the comparison among models more robust and reliable.

Revision in our paper: (the revisions are marked in red)

L327:

To ensure that the final test data were not involved in model selection, the test region remained completely independent during each spatial-block validation, and hyperparameter optimization was performed using only data from the training region. The training data were further divided into model-fitting and validation subsets, with 20% of the training-region samples used for validation. Macro-F1 on the validation set was used as the optimization objective to account for classification performance across susceptibility classes. The previously generated 10 spatial-block partitions were used in the experiments. For each spatial partition, the test region remained unchanged, while the training-validation split, model training, and hyperparameter search were repeated three times using different random seeds. Therefore, under the complete experimental design, each model produced 30 independent test results. The mean and sample variance of class-wise recall, F1-score, overall accuracy, and AUC were calculated for each model.

For each model, multiple TPE trials were conducted under predefined search budgets and random seeds, and the hyperparameter combination yielding the highest validation F1 was selected as the optimal configuration. The maximum search budget was 50 trials for LR, SVM, and RF, and 30 trials for XGBoost, FT-Transformer, and TabPFN. Optimization was terminated early if no improvement in validation Macro-F1 was observed for 15 consecutive trials. Hyperparameter optimization was conducted independently for each repeated experiment. The parameter combination with the highest validation F1 was then used to retrain the model on the complete training-region dataset, and final evaluation was performed only on the independent test region. The hyperparameter search spaces and the selected parameters from the first repeated experiment of the first spatial partition are presented in Table 4. For LR, C controls the strength of regularization, while class weight adjusts the relative importance of different classes. For SVM, C balances the margin width and misclassification penalty, whereas gamma controls the influence range of the RBF kernel. For RF, the number of trees, maximum depth, minimum number of samples required for splitting, minimum number of samples per leaf, and maximum number of features control the ensemble size and complexity of individual trees. For XGBoost, the number of estimators, maximum depth, and learning rate regulate model complexity and the boosting process; subsample and colsample ratio control the proportions of sampled observations and features, respectively; and min_child_weight, gamma, and L₁ and L₂ regularization are used to constrain model complexity and reduce overfitting. For FT-Transformer, token dimension, number of layers, and number of attention heads determine the network architecture, while dropout, learning rate, weight decay, and batch size control the training process. For TabPFN, the number of estimators and the maximum number of training samples per class control the ensemble size and the amount of input training data.

L395:

Table 5

Performance of landslide susceptibility models on the binary classification dataset (best values in bold).

Model	Non-landslide	Landslide	Overall					
			Precision	F1	F1	Accuracy	AUC	AUC variance ($\times 10^{-5}$)
LR	71.35	73.20	72.51	72.27	72.51	72.27	81.63	29.04
FT-Transformer	76.31	81.42	79.35	78.86	79.35	78.86	87.71	21.25
SVM	77.71	81.16	79.76	79.43	79.76	79.43	87.80	26.35
XGBoost	77.35	83.59	81.05	80.47	81.05	80.47	89.56	18.93
RF	78.16	84.23	81.73	81.19	81.73	81.19	90.02	16.75
TabPFN	79.28	85.00	82.62	82.14	82.62	82.14	91.00	19.27

Table 6

Performance of landslide susceptibility models on the multi-class dataset (best values in bold).

Model	Very low	Low	Mode rate	High	Very high	Overall
-------	----------	-----	-----------	------	-----------	---------

	Recall	Recall	Recall	Recall	Recall	Precis ion	F1	F1	Accur acy	AUC	AUC variance ($\times 10^{-5}$)
LR	69.05	66.90	31.90	52.38	57.86	55.63	56.13	55.62	55.73	81.15	66.88
FT- Transformer	89.52	76.19	72.14	73.57	61.43	74.76	67.41	74.57	74.12	83.60	60.28
SVM	89.52	82.86	82.14	79.76	78.33	82.53	80.51	82.52	82.51	86.05	53.68
XGBoost	92.86	84.05	74.05	79.76	76.67	81.52	79.43	81.48	80.91	91.12	23.91
RF	94.52	88.81	84.52	87.14	83.33	87.78	85.67	87.67	87.32	90.10	30.67
TabPFN	95.26	90.44	86.21	88.73	85.67	89.35	87.47	89.26	89.04	92.35	95.26

Point 11:

Please provide an independent temporal or geographically separated validation to establish transferability beyond the Ancient Qin-Shu Roads study area.

Reply: Thank you for this valuable comment. We sincerely apologize that independent spatial and temporal validation was insufficient in the original manuscript. Following your suggestion, we added temporally independent validation. We hope that these additions further strengthen the assessment of the spatial and temporal generalization and transferability of the models.

Revision in our paper: (the revisions are marked in red)

L300:

To further evaluate the predictive capability of the models for newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 were used as a temporally independent validation dataset. These landslide records were not involved in label construction or model training, and were used only for independent validation.

L487:

5.4. Temporal independent validation

To further evaluate the ability of the models to identify newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 and were not involved in previous CF construction, label generation, or model training were used for temporal independent validation, as shown in Table 12. The results show that all models identified most of the newly recorded landslides within the very-high-susceptibility zone. RF and TabPFN performed best, with 38 landslides located in the very-high-susceptibility zone and a hit rate of 80.85%. Overall, the temporal independent validation results indicate that all models showed good capability in identifying newly occurring landslides that were not involved in model development, with RF and TabPFN providing relatively higher spatial coverage of the new landslides.

Table 12

Temporal independent validation results using landslides recorded during 2021–2025

Model	Total landslides	Very low	Low	Moderate	High	Very high	Very-high hit rate (%)
-------	---------------------	-------------	-----	----------	------	-----------	------------------------

LR	47	2	2	2	7	34	72.34
FT- Transformer	47	0	1	4	7	35	74.47
SVM	47	0	1	5	5	36	76.60
XGBoost	47	0	1	4	5	37	78.72
RF	47	0	1	4	4	38	80.85
TabPFN	47	0	1	4	4	38	80.85

Replies to Reviewer 2

The authors would like to express their gratitude to the reviewer for his/her constructive and helpful comments for substantial improvement of this paper. The manuscript has been carefully revised according to your comments and suggestions, which are very valuable to the improvement of this manuscript. We hope our revision has improved the paper to a level of your satisfaction. The replies to your specific comments/suggestions are as follows.

Point 1:

The comparison among LR, SVM, RF, XGBoost, FT-Transformer, and TabPFN lacks a systematic hyperparameter-tuning procedure. Using one fixed configuration for each model may bias the conclusion that RF performs best. The authors should perform hyperparameter optimization using a validation set separate from the final test data and report the search space, objective, budget, stopping criterion, random seeds, and selected parameters. The study “A hyperparameter tuning strategy for an LSTM model to simulate reservoir outflows: large-scale evaluation across 441 dams in the CONUS” (DOI: 10.1016/j.jhydrol.2026.136204) provides a useful Bayesian optimization/TPE framework that could be adapted here.

Reply: Thank you for this valuable comment. We sincerely apologize that systematic hyperparameter optimization was not performed for the different models in the original manuscript. Following your suggestion and referring to the framework of Basirifard et al. (2026), we adopted Bayesian optimization based on the Tree-structured Parzen Estimator (TPE) to optimize the hyperparameters of LR, SVM, RF, XGBoost, FT-Transformer, and TabPFN. We hope that these revisions reduce potential bias caused by fixed hyperparameter settings and improve the fairness and reliability of the model comparison.

Revision in our paper: (the revisions are marked in red)

L309:

3.4. Landslide susceptibility models

Six representative classifiers were selected for landslide susceptibility modeling: LR, SVM, RF, XGBoost, FT-Transformer, and TabPFN. Together, these methods cover linear statistical, kernel-based, ensemble-learning, and deep tabular-learning models and can characterize relationships between susceptibility and conditioning factors from complementary perspectives. LR served as the linear statistical baseline. An SVM with a radial basis function (RBF) kernel represented nonlinear decision boundaries (Cortes and Vapnik, 1995). RF reduced model variance through bootstrap sampling and random subspaces (Breiman, 2001). XGBoost used regularized gradient boosting to fit complex nonlinear relationships (Chen and Guestrin, 2016). FT-Transformer learned interactions among conditioning factors through feature attention (Gorishniy et al., 2021), whereas TabPFN used pretrained

priors for rapid inference on small and medium-sized tabular datasets (Hollmann et al., 2025). Each model was trained for five-class classification and produced the probability that each sample belonged to each susceptibility class; the class with the highest predicted probability was assigned as the final result.

To reduce potential bias in model comparisons caused by fixed hyperparameter settings, this study followed the optimization strategy of Basirifard et al. (2026) and systematically optimized the hyperparameters of LR, SVM, RF, XGBoost, FT-Transformer, and TabPFN. A Bayesian optimization approach based on the Tree-structured Parzen Estimator (TPE) was adopted to automatically identify the optimal hyperparameter combinations within predefined search spaces. Based on the results of previous trials, TPE constructs separate probability distributions for better- and worse-performing hyperparameter configurations and preferentially explores parameter regions that are more likely to yield improved validation performance, thereby improving search efficiency compared with exhaustive search. This approach has been widely used for systematic hyperparameter optimization in machine-learning and deep-learning models.

To ensure that the final test data were not involved in model selection, the test region remained completely independent during each spatial-block validation, and hyperparameter optimization was performed using only data from the training region. The training data were further divided into model-fitting and validation subsets, with 20% of the training-region samples used for validation. Macro-F1 on the validation set was used as the optimization objective to account for classification performance across susceptibility classes. The previously generated 10 spatial-block partitions were used in the experiments. For each spatial partition, the test region remained unchanged, while the training-validation split, model training, and hyperparameter search were repeated three times using different random seeds. Therefore, under the complete experimental design, each model produced 30 independent test results. The mean and sample variance of class-wise recall, F1-score, overall accuracy, and AUC were calculated for each model.

For each model, multiple TPE trials were conducted under predefined search budgets and random seeds, and the hyperparameter combination yielding the highest validation F1 was selected as the optimal configuration. The maximum search budget was 50 trials for LR, SVM, and RF, and 30 trials for XGBoost, FT-Transformer, and TabPFN. Optimization was terminated early if no improvement in validation Macro-F1 was observed for 15 consecutive trials. Hyperparameter optimization was conducted independently for each repeated experiment. The parameter combination with the highest validation F1 was then used to retrain the model on the complete training-region dataset, and final evaluation was performed only on the independent test region. The hyperparameter search spaces and the selected parameters from the first repeated experiment of the first spatial partition are presented in Table 4. For LR, C controls the strength of regularization, while class weight adjusts the relative importance of different classes. For SVM, C balances the margin width and misclassification penalty, whereas γ controls the influence range of the RBF kernel. For RF, the number of trees, maximum depth, minimum number of samples required for splitting, minimum number of samples per leaf, and maximum number of features control the ensemble size and complexity of individual trees. For XGBoost, the number of estimators, maximum depth, and learning rate regulate model complexity and the boosting process; subsample and colsample ratio control the proportions of sampled observations and features, respectively; and min_child_weight, γ , and L_1 and L_2 regularization are used to constrain model

complexity and reduce overfitting. For FT-Transformer, token dimension, number of layers, and number of attention heads determine the network architecture, while dropout, learning rate, weight decay, and batch size control the training process. For TabPFN, the number of estimators and the maximum number of training samples per class control the ensemble size and the amount of input training data.

Table 4

Models and parameter settings for landslide susceptibility modeling.

Model	Model type	Search space	Search budget	Selected parameters
LR	Linear probabilistic classification model	C: log-uniform, 10^{-3} – 10^3 ; class weight: None or balanced	50 trials	C=23.18; class weight = None
FT-Transformer	Deep tabular Transformer model	Token dimension: 32, 48, 64, 96, or 128; number of layers: 2–6; attention heads: 2, 4, or 8, subject to divisibility by token dimension; dropout: 0–0.5; learning rate: log-uniform, 10^{-5} – 10^{-3} ; weight decay: log-uniform, 10^{-6} – 10^{-2} ; batch size: 64, 128, or 256	30 trials	Token dimension = 64; layers = 4; heads = 2; dropout = 0.05; learning rate = 1.84×10^{-4} ; weight decay = 2.60×10^{-5} ; batch size = 128
XGBoost	Gradient boosting decision tree model	Number of estimators: 100–1000, step 100; maximum depth: 2–10; learning rate: log-uniform, 0.01–0.3; subsample ratio: 0.5–1.0; column sampling ratio: 0.5–1.0; minimum child weight: 1–10; gamma: 0–5; L_1 regularization: log-uniform, 10^{-8} –10; L_2 regularization: log-uniform, 10^{-3} –100	30 trials	Number of estimators = 700; maximum depth = 8; learning rate = 0.043; subsample ratio = 0.620816; column sampling ratio = 0.659; minimum child weight = 5.538; gamma = 2.386; L_1 regularization = 0.437; L_2 regularization = 1.977
SVM	RBF-kernel classification model	C: log-uniform, 10^{-2} – 10^3 ; gamma: scale, auto, or log-uniform 10^{-4} –10; class weight: None or balanced	50 trials	C=10.71; gamma = auto; class weight = balanced
TabPFN	Pretrained tabular foundation model	Number of estimators: 1, 2, 4, 8, or 16; maximum training samples per class: 128, 256, 512, or 1024	30 trials	Number of estimators = 2; maximum training samples per class = 1024
RF	Ensemble learning model	Number of trees: 200–1000, step 100; maximum depth: None, 5, 10, 20, or 30; minimum samples for splitting: 2–20; minimum samples per leaf: 1–10; maximum features: sqrt, log2, or None; class weight: None or balanced	50 trials	Number of trees = 200; maximum depth = None; minimum samples for splitting = 8; minimum samples per leaf = 7; maximum features = log2; class weight = balanced

Point 2:

The random 70/30 split is not sufficient for a spatial landslide dataset because nearby

samples may appear in both training and testing sets, leading to overly optimistic performance. Spatial-block cross-validation, group-wise spatial validation, or leave-one-corridor/region-out testing should be included. Random-split results can be retained, but spatially independent validation is necessary.

Reply: Thank you for this valuable comment. We sincerely apologize that the original manuscript adopted a random 70/30 split, which may have allowed spatially adjacent samples to appear in both the training and test sets. Following your suggestion, we have adopted spatial-block cross-validation by dividing the study area into spatially independent regions according to geographic location and conducting training and testing across different spatial regions. This approach reduces the potential overestimation of model performance caused by spatial autocorrelation. We hope that this revision further improves the independence and reliability of the model validation.

Revision in our paper: (the revisions are marked in red)

L287:

During sample construction, stratified sampling was performed based on the five-class landslide susceptibility zoning obtained from the CF results constrained by SHALSTAB. To ensure spatial independence between the training and test sets, the study area was divided into a 3×4 spatial grid according to geographic location, and 10-fold spatial-block cross-validation was adopted. In each fold, eight spatial blocks were used as the training set and the remaining four blocks as the test set. Fig. 4 shows the data distribution for one fold. In each fold, the CF statistical prior was recalculated using only the historical landslide samples within the training area, and the resulting CF parameters were then applied to the entire study area to generate the corresponding CF-based susceptibility map and five-class classification result; landslide records within the test area were not involved in CF construction. The SHALSTAB results were independently calculated based on the DEM and parameter inputs, and no landslide records from either the training or test sets were used in its construction. To ensure that all susceptibility levels were represented in model training, 18 landslide points located in the very-low-susceptibility zone and one invalid landslide point were first excluded in this fold. The grid cells containing the remaining 1,388 historical landslide points were assigned as very-high-susceptibility samples, while samples from the very low, low, moderate, and high susceptibility zones were selected according to the predefined sampling proportions. A five-class dataset containing 15,116 samples was ultimately constructed. The five susceptibility classes were very low, low, moderate, high, and very high, corresponding to label values of 1–5. Model input variables were extracted from the raster layers of landslide conditioning factors, with each sample containing a set of conditioning-factor values and its corresponding susceptibility class. To further evaluate the predictive capability of the models for newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 were used as a temporally independent validation dataset. These landslide records were not involved in CF construction, model training, or validation, and were used only for independent evaluation of the final susceptibility results.

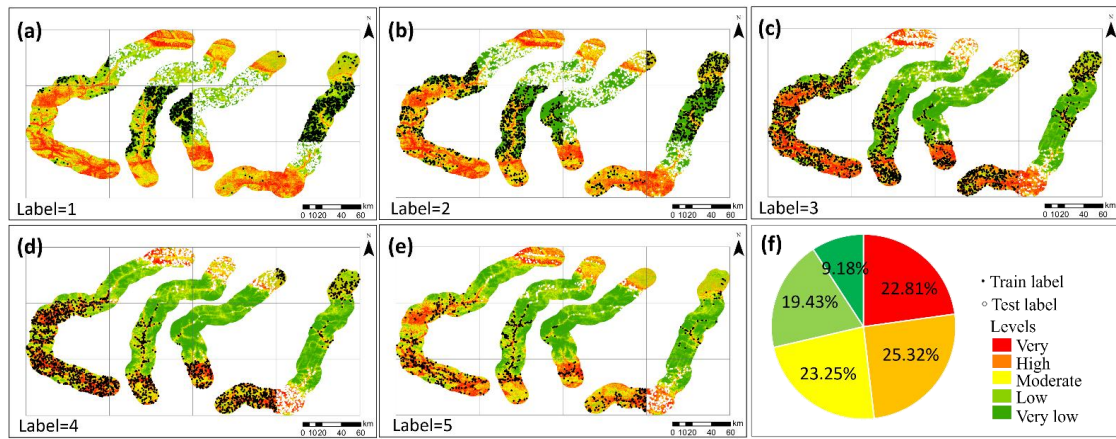


Fig. 4. Distribution of landslide susceptibility samples in one fold. (a) Very low susceptibility samples; (b) Low susceptibility samples; (c) Moderate susceptibility samples; (d) High susceptibility samples; (e) Very high susceptibility samples; (f) Sample composition.

L400:

To evaluate the historical-landslide coverage and spatial concentration of the susceptibility maps, the areal proportion of susceptibility class (a), the historical-landslide hit rate (h), and HOI were calculated. Binary results were summarized in two ways: Table 7 reports statistics after dividing binary probability maps into five classes by natural breaks, and Table 8 reports the landslide-prone zones defined by the binary threshold ($p > 0.5$). As shown in Tables 7 and 8, TabPFN achieved the highest threshold-based binary HOI (1.70); its landslide-prone zone occupied 16.9% of the study area and contained 86.7% of the historical landslides. Its HOI based on the five-class natural-breaks result was 1.68, lower than the threshold-based HOI and confirming that evaluating a binary model solely by its very high susceptibility class underestimates model capability. Table 9 shows that TabPFN also performed best in the multi-class setting. Its very high susceptibility zone occupied only 10.8% of the study area but contained 86.0% of the historical landslides, producing an HOI of 1.75. The comparison of HOI between binary and multi-class classification is shown in Fig. 6, indicating that the multi-class classification approach can produce a more spatially concentrated high-susceptibility zone.

Table 7

Statistical results of landslide susceptibility maps based on binary classification using natural breaks (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	50.6	11.0	16.3	16.9	21.4	40.2	7.3	24.7	4.4	7.1	1.03
FT-Transformer	62.7	25.4	17.2	20.0	8.8	20.6	7.7	19.0	3.6	15.1	1.11
SVM	66.4	8.4	6.9	6.8	9.1	16.0	12.0	40.5	5.7	28.3	1.23
XGBoost	61.0	3.4	9.0	4.3	7.1	7.0	6.4	5.2	16.5	80.1	1.64
RF	61.1	3.3	9.9	3.3	7.0	6.3	7.1	10.7	14.9	76.4	1.61
TabPFN	56.0	1.5	11.5	1.5	7.3	2.7	7.4	5.0	17.8	85.3	1.68

Table 8

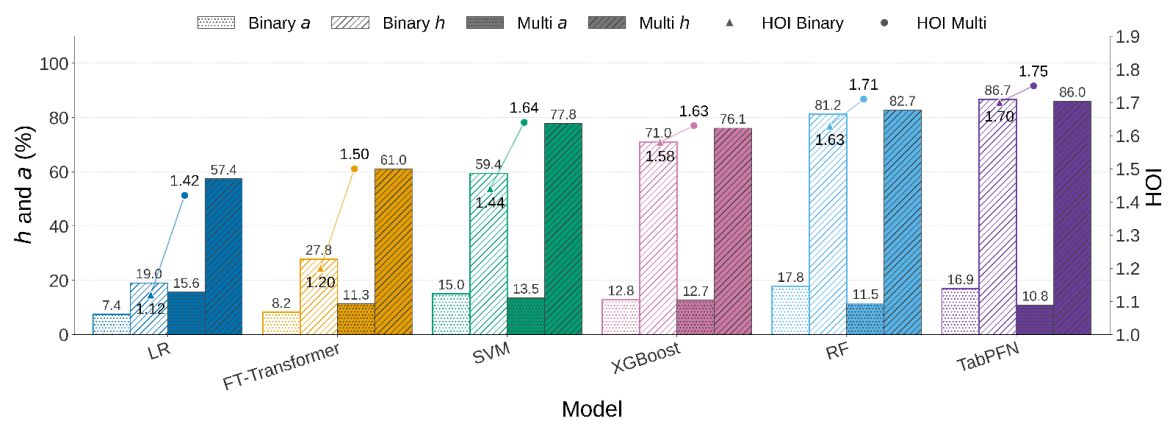
Statistical results of landslide susceptibility maps based on binary classification using threshold (best values in bold).

Model	Non-landslides		Landslides		HOI
	a_{neg}	h_{neg}	a_{pos}	h_{pos}	
LR	92.6	81.0	7.4	19.0	1.12
FT-Transformer	91.8	72.2	8.2	27.8	1.20
SVM	85.0	40.6	15.0	59.4	1.44
XGBoost	87.2	29.0	12.8	71.0	1.58
RF	82.2	18.8	17.8	81.2	1.63
TabPFN	83.1	13.3	16.9	86.7	1.70

Table 9

Statistical results of landslide susceptibility maps based on multi-class classification (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	16.2	1.2	30.2	8.0	15.2	13.2	22.7	20.1	15.6	57.4	1.42
FT-Transformer	18.2	0.2	26.6	7.6	22.4	8.7	21.5	22.5	11.3	61.0	1.50
SVM	17.1	0.7	26	4.7	22.6	5.4	20.8	11.3	13.5	77.8	1.64
XGBoost	16.8	0.2	28.1	4	20.8	7.8	21.6	11.8	12.7	76.1	1.63
RF	16.8	0.2	28.3	3.1	21.9	7.1	21.5	6.9	11.5	82.7	1.71
TabPFN	18.2	0.1	27.6	3.5	22.4	4.8	21.0	5.6	10.8	86.0	1.75

**Fig. 6.** Comparison of area ratio (a), hit rate (h), and HOI for binary and multi-class susceptibility maps.

The spatial susceptibility patterns produced by the six models under the binary and multi-class settings are shown in Fig. 7. Overall, very high susceptibility zones formed belts or patches concentrated mainly along valleys. In the binary results, most historical landslides occurred within the moderate, high, and very high susceptibility classes. Thus, binary classification identified broadly landslide-prone areas, but the resulting high-risk regions were relatively extensive and increased the area requiring attention. In the multi-class results, historical landslides were more concentrated within the very high susceptibility class,

producing a more focused high-risk area. Enlarged views of the TabPFN results show that binary classification generated a broad transition zone of moderate to high susceptibility around landslide locations, whereas multi-class classification assigned the landslide locations more directly to the very high susceptibility class and reduced overestimation in surrounding areas. This further demonstrates that multi-class classification improves the spatial concentration of the very high susceptibility zone.

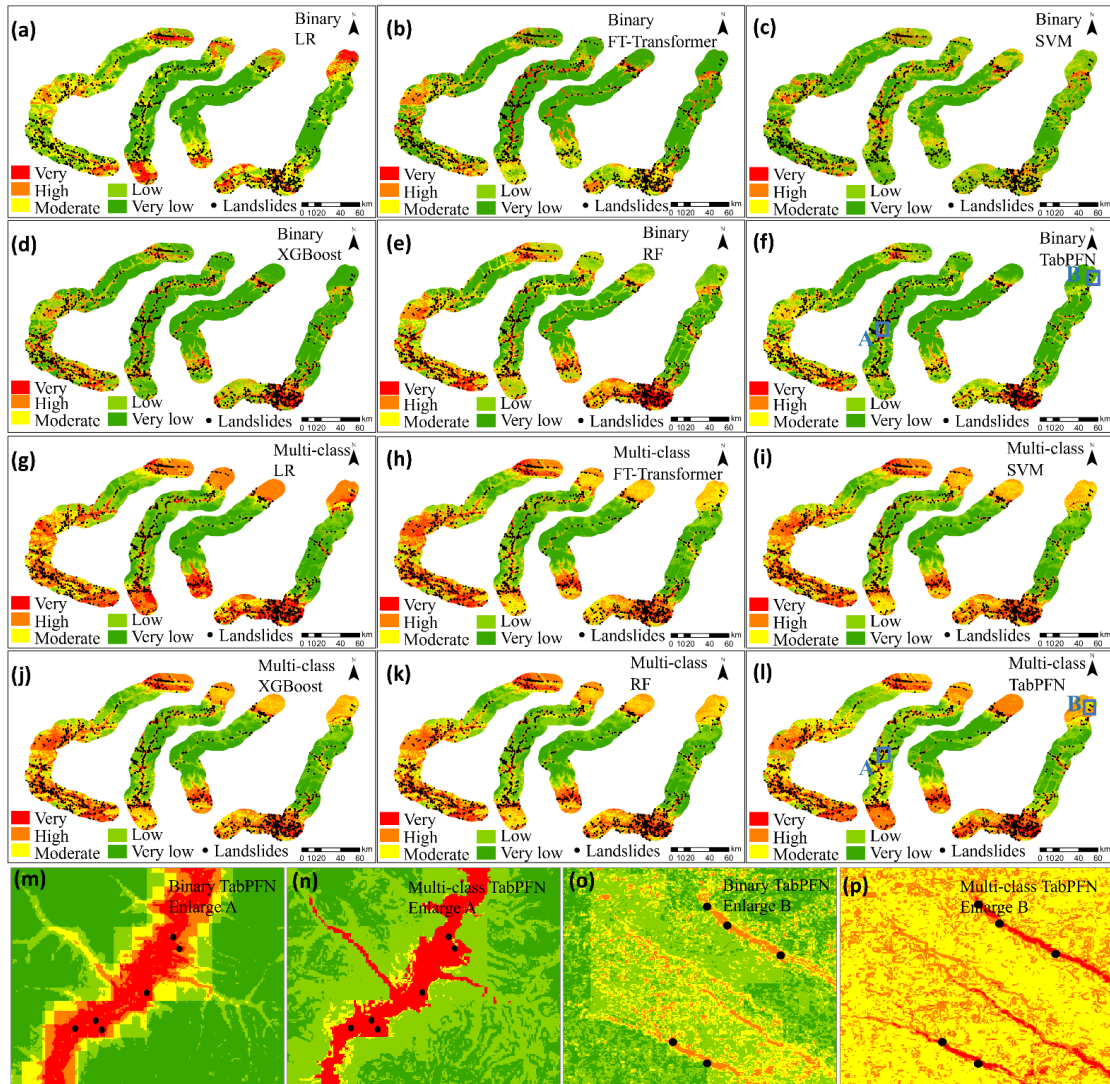


Fig. 7. Landslide susceptibility maps produced by six models under binary and multi-class classification settings. (a–f) Binary classification results of LR, FT-Transformer, SVM, XGBoost, RF and TabPFN; (g–l) multi-class classification results of LR, FT-Transformer, SVM, XGBoost, RF and TabPFN; (m–p) enlarged views of selected TabPFN results.

Point 3:

The construction of the five-class target may introduce circularity. The CF map and SHALSTAB results are derived from the landslide inventory and conditioning factors, while similar factors are later used as ML predictors. Therefore, high accuracy may partly reflect reproduction of the CF/SHALSTAB labeling scheme rather than independent landslide prediction. This concern is stronger if the full landslide inventory

was used before the train-test split.

Reply: Thank you for this valuable comment. We sincerely apologize that the potential circularity associated with the construction of the five-class target was not sufficiently considered in the original manuscript. Following your suggestion, we redesigned the target-generation and validation procedure. In each fold of the spatial-block cross-validation, the CF statistical prior was recalculated using only the historical landslide samples within the training area, while SHALSTAB was calculated independently based only on the DEM and parameter inputs. We hope that these revisions can effectively reduce information leakage and circularity associated with target construction and improve the independence and reliability of the model validation.

Revision in our paper: (the revisions are marked in red)

L287:

During sample construction, stratified sampling was performed based on the five-class landslide susceptibility zoning obtained from the CF results constrained by SHALSTAB. To ensure spatial independence between the training and test sets, the study area was divided into a 3×4 spatial grid according to geographic location, and 10-fold spatial-block cross-validation was adopted. In each fold, eight spatial blocks were used as the training set and the remaining four blocks as the test set. Fig. 4 shows the data distribution for one fold. In each fold, the CF statistical prior was recalculated using only the historical landslide samples within the training area, and the resulting CF parameters were then applied to the entire study area to generate the corresponding CF-based susceptibility map and five-class classification result; landslide records within the test area were not involved in CF construction. The SHALSTAB results were independently calculated based on the DEM and parameter inputs, and no landslide records from either the training or test sets were used in its construction. To ensure that all susceptibility levels were represented in model training, 18 landslide points located in the very-low-susceptibility zone and one invalid landslide point were first excluded in this fold. The grid cells containing the remaining 1,388 historical landslide points were assigned as very-high-susceptibility samples, while samples from the very low, low, moderate, and high susceptibility zones were selected according to the predefined sampling proportions. A five-class dataset containing 15,116 samples was ultimately constructed. The five susceptibility classes were very low, low, moderate, high, and very high, corresponding to label values of 1–5. Model input variables were extracted from the raster layers of landslide conditioning factors, with each sample containing a set of conditioning-factor values and its corresponding susceptibility class. To further evaluate the predictive capability of the models for newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 were used as a temporally independent validation dataset. These landslide records were not involved in CF construction, model training, or validation, and were used only for independent evaluation of the final susceptibility results.

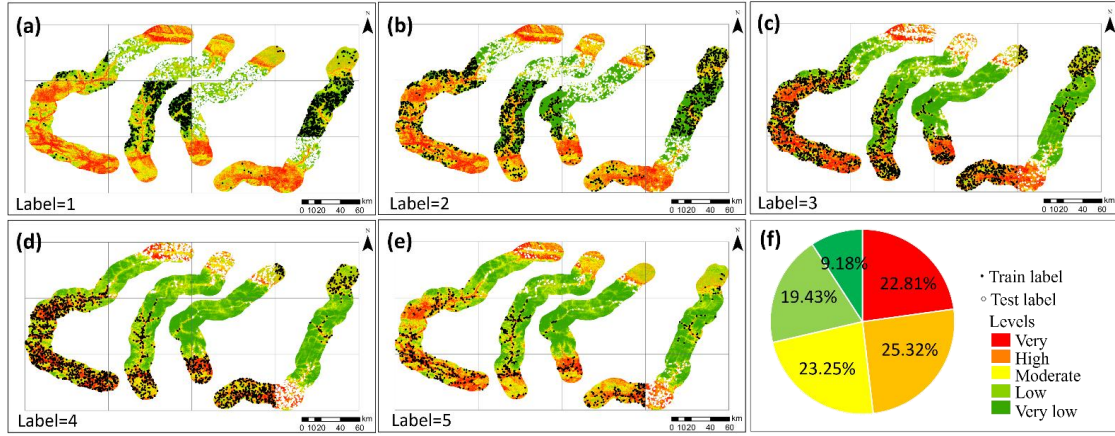


Fig. 4. Distribution of landslide susceptibility samples in one fold. (a) Very low susceptibility samples; (b) Low susceptibility samples; (c) Moderate susceptibility samples; (d) High susceptibility samples; (e) Very high susceptibility samples; (f) Sample composition.

L362:

To ensure that the final evaluation uses landslide observations that were not involved in any previous modeling or label-construction process, the spatial mapping results were further evaluated using the HOI to measure the ability of the very-high-susceptibility zone to capture historical landslide points within a relatively small area (Zhang et al., 2026). HOI jointly considers the areal proportion of the very high susceptibility zone and the hit rate of historical landslides within that zone and is calculated as follows:

$$HOI = 1 - a_5 + h_5 \quad (12)$$

where a_5 is the areal proportion of the very high susceptibility zone within the test region, and h_5 is the proportion of historical landslides falling within that zone in the test region.

5.2. Label uncertainty analysis

To evaluate the uncertainty introduced by the CF statistical prior and SHALSTAB physical constraints into the construction of multi-class labels, landslide samples from eight spatial blocks of the study area were randomly selected each time to construct the CF model, and this procedure was repeated 10 times to characterize the influence of spatial variations in landslide samples on the CF results. Subsequently, the five key SHALSTAB parameters were individually set to 0.8, 1.0, and 1.2 times their baseline values, yielding 15 sets of results. The 10 CF results were then combined with the 15 SHALSTAB results, generating a total of 150 initial landslide susceptibility classification results, which were used to evaluate the uncertainty of the five-class susceptibility classification. The 150 classification results were integrated to obtain the susceptibility class distribution for each grid cell, as shown in Fig. 9(a). Meanwhile, the standard deviation of susceptibility classes across different combined results was used to characterize the variation around the mean class and to quantify classification uncertainty, as shown in Fig. 9(b). A smaller standard deviation indicates less variation in the susceptibility class of a given grid cell across different combined results and, consequently, greater classification stability. The standard deviation is close to 0 across most of the study area, whereas locally higher values, approximately 0.4, occur mainly in areas where susceptibility classes differ markedly and are spatially dispersed.

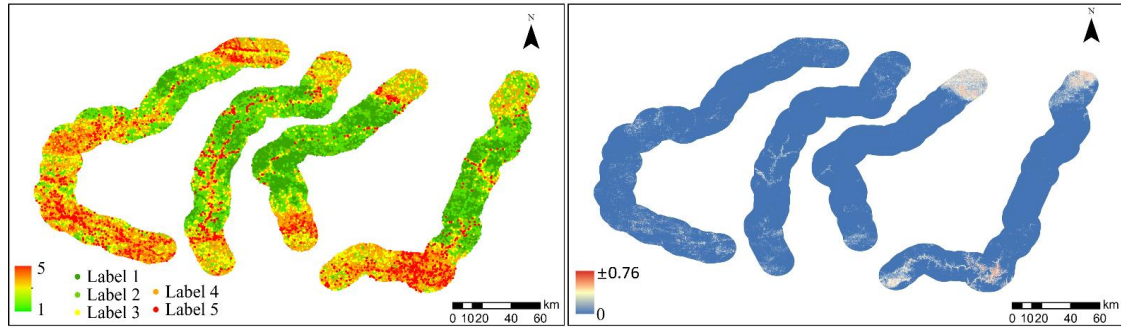


Fig. 9. Spatial distribution of landslide susceptibility classes and their uncertainty. (a) Mean landslide susceptibility classes; (b) Standard deviation of susceptibility classes.

Point 4:

The target-generation and validation procedure should be redesigned or clearly justified. CF generation, SHALSTAB calibration, and sample labeling should rely only on training data, while final evaluation should use landslide observations not involved in any previous modeling or labeling step. Spatially, temporally, or event-based independent validation would provide stronger evidence.

Reply: Thank you for this valuable comment. We sincerely apologize that the target-generation and independent validation procedures were not sufficiently clarified in the original manuscript. Following your suggestion, we redesigned and clarified the target-generation and validation framework. In each fold of the spatial-block cross-validation, the CF statistical prior was recalculated using only the historical landslide samples within the training area, while SHALSTAB was calculated solely based on the DEM and parameter inputs. Meanwhile, we clarified that the HOI was calculated using only actually observed historical landslide points, and added an explanation regarding landslide observations that were not involved in any previous modeling or label-construction process. In addition, landslide points recorded during 2021~2025 were added as a temporally independent validation dataset. We hope that these revisions can further reduce information leakage and circularity in target construction and enhance the independence and reliability of model validation.

Revision in our paper: (the revisions are marked in red)

L136:

Historical hazard data were compiled from expert-verified landslides recorded in national geological-hazard investigations and censuses, together with disaster reports from the Shaanxi Environmental Monitoring Center and relevant county and district authorities. Before 2021, a total of 1,407 landslide records were available, including information on hazard type, occurrence time, location, coordinates, scale, triggering factors, and disaster impacts.

L206:

The study area was divided by natural breaks into five susceptibility classes: very low, low, moderate, high, and very high, producing the initial CF susceptibility map. To reduce local overestimation that might result from a single statistical method, the integrated SHALSTAB result was introduced as a constraint on the initial CF zonation. Given that shallow landslides predominate in the study area, SHALSTAB can provide a regional-scale physical stability constraint, and its results can reflect the stability characteristics across most of the study area. SHALSTAB is a representative physically based model that generally uses a DEM and integrates topographic conditions, geotechnical parameters, and stability criteria to identify spatial units with a clear tendency toward failure (Montgomery and Dietrich, 1994; Ye et al., 2025). The model jointly considers slope angle, upslope contributing area, steady-state hydrological processes, and geotechnical parameters. Soil thickness is estimated from the local slope using:

$$z = h_{max} \left[1 - \frac{\tan \theta - \tan \theta_{min}}{\tan \theta_{max} - \tan \theta_{min}} \left(1 - \frac{h_{min}}{h_{max}} \right) \right] \quad (5)$$

where θ is the slope angle, and h_{min} and h_{max} are set to 0.2 m and 4.5 m, respectively. Thus, steeper slopes are assigned smaller soil thicknesses. Soil transmissivity T is calculated from the saturated hydraulic conductivity K_s , soil thickness z , and slope angle θ , as $T = K_s z \cos \theta$. The SHALSTAB stability criterion can be expressed as:

$$\frac{a}{b} > \frac{T \sin \theta}{q} \left[\frac{\gamma_w}{\gamma} \left(1 - \frac{\tan \theta}{\tan \varphi} \right) + \frac{C}{\gamma_w z \cos^2 \theta \tan \varphi} \right] \quad (6)$$

where $\frac{a}{b}$ denotes the upslope contributing area per unit contour length, and q is the prescribed rainfall input, set to 100 mm/day in this study. C is cohesion, γ is the natural unit weight of the soil or rock mass, φ is the internal friction angle, and γ_w is the unit weight of water. When the left-hand side exceeds the right-hand side, the grid cell is classified as unstable; otherwise, it is classified as stable.

Based on parameter values reported in previous studies and related information on landslides in the Qinling–Daba Mountains (Qiang et al., 2015; Song et al., 2022), initial values of C , θ , γ , and K_s were assigned to the different engineering geological types. The corresponding parameter settings are listed in Table 2. Because it is difficult to determine spatially distributed geotechnical parameters over such a large regional study area, no local calibration based on site-specific measurements was performed. Most parameters are inherently subject to spatial uncertainty. Therefore, the parameter values adopted for the study area should be regarded as representative relative values rather than uniquely determined absolute values.

L287:

During sample construction, stratified sampling was performed based on the five-class landslide susceptibility zoning obtained from the CF results constrained by SHALSTAB. To ensure spatial independence between the training and test sets, the study area was divided into a 3×4 spatial grid according to geographic location, and 10-fold spatial-block cross-validation was adopted. In each fold, eight spatial blocks were used as the training set and the remaining four blocks as the test set. In each fold, the CF statistical prior was recalculated using only the historical landslide samples within the training area and landslide records within the test area were not involved in CF construction. The SHALSTAB results were independently calculated based on the DEM and parameter inputs, and no landslide records from either the training or test sets were used in its construction. Fig. 4 shows the data

distribution for one fold. To ensure that all susceptibility levels were represented in model training, 18 landslide points located in the very-low-susceptibility zone and one invalid landslide point were first excluded in this fold. The grid cells containing the remaining 1,388 historical landslide points were assigned as very-high-susceptibility samples, while samples from the very low, low, moderate, and high susceptibility zones were selected according to the predefined sampling proportions. A five-class dataset containing 15,116 samples was ultimately constructed. The five susceptibility classes were very low, low, moderate, high, and very high, corresponding to label values of 1 to 5. Model input variables were extracted from the raster layers of landslide conditioning factors, with each sample containing a set of conditioning-factor values and its corresponding susceptibility class. To further evaluate the predictive capability of the models for newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 were used as a temporally independent validation dataset. These landslide records were not involved in label construction or model training, and were used only for independent validation.

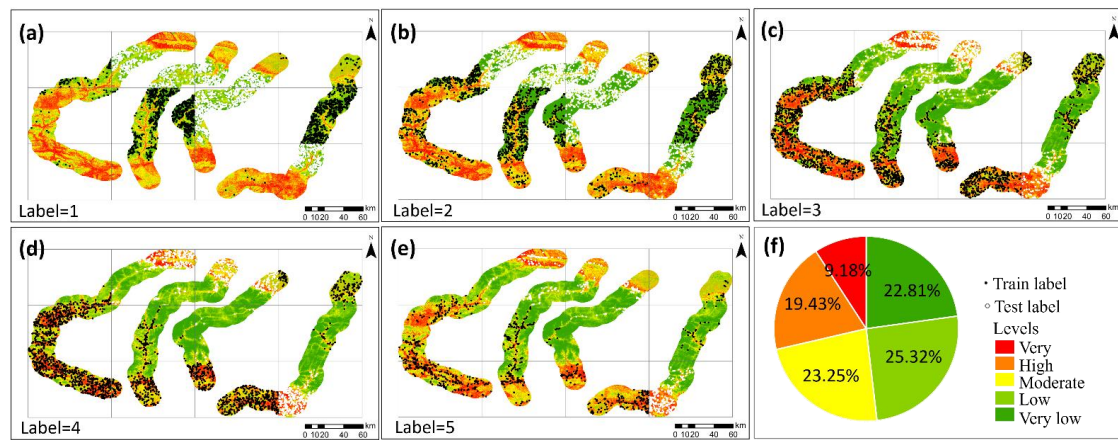


Fig. 5. Distribution of landslide susceptibility samples in one fold. (a) Very low susceptibility samples; (b) Low susceptibility samples; (c) Moderate susceptibility samples; (d) High susceptibility samples; (e) Very high susceptibility samples; (f) Sample composition.

L362:

To ensure that the final evaluation uses landslide observations that were not involved in any previous modeling or label-construction process, the spatial mapping results were further evaluated using the HOI to measure the ability of the very-high-susceptibility zone to capture historical landslide points within a relatively small area (Zhang et al., 2026). HOI jointly considers the areal proportion of the very high susceptibility zone and the hit rate of historical landslides within that zone and is calculated as follows:

$$HOI = 1 - a_5 + h_5 \quad (12)$$

where a_5 is the areal proportion of the very high susceptibility zone within the test region, and h_5 is the proportion of historical landslides falling within that zone in the test region.

L487:

5.4. Temporal independent validation

To further evaluate the ability of the models to identify newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 and were not involved in previous CF construction, label generation, or model training were used for temporal independent validation, as shown in Table 12. The results show that all models identified most of the newly recorded landslides within the very-high-susceptibility zone. RF and TabPFN performed best, with 38 landslides located in the very-high-susceptibility zone and a hit rate of 80.85%. Overall, the temporal independent validation results indicate

that all models showed good capability in identifying newly occurring landslides that were not involved in model development, with RF and TabPFN providing relatively higher spatial coverage of the new landslides.

Table 12

Temporal independent validation results using landslides recorded during 2021–2025

Model	Total landslides	Very low	Low	Moderate	High	Very high	Very-high hit rate (%)
LR	47	2	2	2	7	34	72.34
FT-Transformer	47	0	1	4	7	35	74.47
SVM	47	0	1	5	5	36	76.60
XGBoost	47	0	1	4	5	37	78.72
RF	47	0	1	4	4	38	80.85
TabPFN	47	0	1	4	4	38	80.85

Point 5:

Differences among the best-performing models are small. For example, RF reports AUC = 92.53% and F1 = 68.65%, while TabPFN reports AUC = 92.30% and F1 = 68.25%. These differences may not be statistically meaningful.

Reply: Thank you for this valuable comment. We agree that the performance differences among the best-performing models are small, and conclusions based solely on such limited numerical differences may be ambiguous. Therefore, we revised the relevant statements in the manuscript to avoid potentially overinterpreted expressions. We hope that these revisions make the comparison among different models more objective and rigorous.

Revision in our paper: (the revisions are marked in red)

L385:

To systematically evaluate the suitability of different models for multi-class landslide susceptibility mapping, LR, FT-Transformer, SVM, XGBoost, RF, and TabPFN were compared. Experiments were conducted using both binary landslide and multi-class susceptibility datasets. For the binary dataset, non-landslide samples were randomly selected from non-landslide areas at a 1:1 ratio to landslide samples, following commonly used sampling strategies in landslide susceptibility modeling (Chang et al., 2023b; Jia et al., 2026). The binary task evaluated discrimination between landslide and non-landslide samples, whereas the multi-class task focused on distinctions among susceptibility classes. The model performance results are presented in Tables 5 and 6. TabPFN performed best in the binary task, with F1, accuracy, and AUC values of 82.62%, 82.14%, and 91.00%, respectively. TabPFN also achieved the best overall multi-class performance, with corresponding values of 89.26%, 89.04%, and 95.26%. Overall, TabPFN achieved the best comprehensive performance, while RF and XGBoost also showed similarly good performance across both tasks.

Table 5

Performance of landslide susceptibility models on the binary classification dataset (best values in bold).

Model	Non-landslide	Landslide	Overall					
	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance ($\times 10^{-5}$)
LR	71.35	73.20	72.51	72.27	72.51	72.27	81.63	29.04
FT-Transformer	76.31	81.42	79.35	78.86	79.35	78.86	87.71	21.25
SVM	77.71	81.16	79.76	79.43	79.76	79.43	87.80	26.35
XGBoost	77.35	83.59	81.05	80.47	81.05	80.47	89.56	18.93
RF	78.16	84.23	81.73	81.19	81.73	81.19	90.02	16.75
TabPFN	79.28	85.00	82.62	82.14	82.62	82.14	91.00	19.27

Table 6

Performance of landslide susceptibility models on the multi-class dataset (best values in bold).

Model	Very low	Low	Mode rate	High	Very high	Overall					
	Recall	Recall	Recall	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance ($\times 10^{-5}$)
LR	69.05	66.90	31.90	52.38	57.86	55.63	56.13	55.62	55.73	81.15	66.88
FT-Transformer	89.52	76.19	72.14	73.57	61.43	74.76	67.41	74.57	74.12	83.60	60.28
SVM	89.52	82.86	82.14	79.76	78.33	82.53	80.51	82.52	82.51	86.05	53.68
XGBoost	92.86	84.05	74.05	79.76	76.67	81.52	79.43	81.48	80.91	91.12	23.91
RF	94.52	88.81	84.52	87.14	83.33	87.78	85.67	87.67	87.32	90.10	30.67
TabPFN	95.26	90.44	86.21	88.73	85.67	89.35	87.47	89.26	89.04	92.35	95.26

Point 6:

The authors should report uncertainty intervals and paired statistical comparisons using identical validation folds and seeds. Confidence intervals or paired bootstrap comparisons would help determine whether the differences between models are robust.

Claims that RF “outperformed” the other models should be moderated unless statistical significance is demonstrated.

Reply: Thank you for this valuable comment. To ensure consistency and fairness in the model comparison, all models were evaluated using exactly the same datasets, spatial validation partitions, and random seeds. Thus, the same training, validation, and test data were used for all models in each repeated experiment. We further reported the

mean and sample variance of the evaluation metrics across repeated spatial validation experiments to characterize the variability and stability of model performance. We hope that these revisions make the model comparison more consistent, robust, and cautious.

Revision in our paper: (the revisions are marked in red)

L327:

To ensure that the final test data were not involved in model selection, the test region remained completely independent during each spatial-block validation, and hyperparameter optimization was performed using only data from the training region. The training data were further divided into model-fitting and validation subsets, with 20% of the training-region samples used for validation. Macro-F1 on the validation set was used as the optimization objective to account for classification performance across susceptibility classes. The previously generated 10 spatial-block partitions were used in the experiments. For each spatial partition, the test region remained unchanged, while the training–validation split, model training, and hyperparameter search were repeated three times using different random seeds. Therefore, under the complete experimental design, each model produced 30 independent test results. The mean and sample variance of class-wise recall, F1-score, overall accuracy, and AUC were calculated for each model.

For each model, multiple TPE trials were conducted under predefined search budgets and random seeds, and the hyperparameter combination yielding the highest validation F1 was selected as the optimal configuration. The maximum search budget was 50 trials for LR, SVM, and RF, and 30 trials for XGBoost, FT-Transformer, and TabPFN. Optimization was terminated early if no improvement in validation Macro-F1 was observed for 15 consecutive trials. Hyperparameter optimization was conducted independently for each repeated experiment. The parameter combination with the highest validation F1 was then used to retrain the model on the complete training-region dataset, and final evaluation was performed only on the independent test region. The hyperparameter search spaces and the selected parameters from the first repeated experiment of the first spatial partition are presented in Table 4. For LR, C controls the strength of regularization, while class weight adjusts the relative importance of different classes. For SVM, C balances the margin width and misclassification penalty, whereas gamma controls the influence range of the RBF kernel. For RF, the number of trees, maximum depth, minimum number of samples required for splitting, minimum number of samples per leaf, and maximum number of features control the ensemble size and complexity of individual trees. For XGBoost, the number of estimators, maximum depth, and learning rate regulate model complexity and the boosting process; subsample and colsample ratio control the proportions of sampled observations and features, respectively; and min_child_weight, gamma, and L₁ and L₂ regularization are used to constrain model complexity and reduce overfitting. For FT-Transformer, token dimension, number of layers, and number of attention heads determine the network architecture, while dropout, learning rate, weight decay, and batch size control the training process. For TabPFN, the number of estimators and the maximum number of training samples per class control the ensemble size and the amount of input training data.

L395:

Table 5

Performance of landslide susceptibility models on the binary classification dataset (best values in bold).

Model	Non-	Landsli	Overall
-------	------	---------	---------

	landslide		model performance					
	Very low	Low	Mode rate	High	Very high	Overall		
	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance ($\times 10^{-5}$)
LR	71.35	73.20	72.51	72.27	72.51	72.27	81.63	29.04
FT-Transformer	76.31	81.42	79.35	78.86	79.35	78.86	87.71	21.25
SVM	77.71	81.16	79.76	79.43	79.76	79.43	87.80	26.35
XGBoost	77.35	83.59	81.05	80.47	81.05	80.47	89.56	18.93
RF	78.16	84.23	81.73	81.19	81.73	81.19	90.02	16.75
TabPFN	79.28	85.00	82.62	82.14	82.62	82.14	91.00	19.27

Table 6

Performance of landslide susceptibility models on the multi-class dataset (best values in bold).

Model	Very low	Low	Mode rate	High	Very high	Overall					
	Recall	Recall	Recall	Recall	Recall	Precision	F1	F1	Accuracy	AUC	AUC variance ($\times 10^{-5}$)
LR	69.05	66.90	31.90	52.38	57.86	55.63	56.13	55.62	55.73	81.15	66.88
FT-Transformer	89.52	76.19	72.14	73.57	61.43	74.76	67.41	74.57	74.12	83.60	60.28
SVM	89.52	82.86	82.14	79.76	78.33	82.53	80.51	82.52	82.51	86.05	53.68
XGBoost	92.86	84.05	74.05	79.76	76.67	81.52	79.43	81.48	80.91	91.12	23.91
RF	94.52	88.81	84.52	87.14	83.33	87.78	85.67	87.67	87.32	90.10	30.67
TabPFN	95.26	90.44	86.21	88.73	85.67	89.35	87.47	89.26	89.04	92.35	95.26

Replies to Reviewer 3

The authors would like to express their gratitude to the reviewer for his/her constructive and helpful comments for substantial improvement of this paper. The manuscript has been carefully revised according to your comments and suggestions, which are very valuable to the improvement of this manuscript. We hope our revision has improved the paper to a level of your satisfaction. The replies to your specific comments/suggestions are as follows.

Comments to the Author:

The manuscript presents an interesting statistical-physical prior-guided framework for direct multi-class landslide susceptibility mapping. The integration of a statistical prior derived from the Certainty Factor method with a SHALSTAB-based physical constraint is potentially valuable, and the comparison among conventional machine-learning and deep tabular-learning models is relevant. However, several methodological and conceptual issues should be addressed before the manuscript can be considered for publication.

Point 1:

The manuscript treats the landslide inventory largely as a homogeneous population, although landslide susceptibility may depend strongly on landslide typology. Different landslide mechanisms can respond to different combinations of topographic, geological, hydrological and anthropogenic conditioning factors. Silva et al. (2018)*, in a susceptibility assessment, demonstrated that models developed separately for different landslide typologies performed better than models based on the total inventory, because the statistical relationships between landslides and predisposing factors may become dominated by the most spatially abundant landslide typology. The authors should consider citing this article. The authors should therefore report the landslide typologies represented in the 1,407-event inventory and discuss whether aggregating all landslides into a single class is appropriate. They should also clarify whether all inventoried landslides are compatible with the shallow-landslide assumptions underlying SHALSTAB.

*Silva, R. F., Marques, R., & Gaspar, J. L. (2018). Implications of Landslide Typology and Predisposing Factor Combinations for Probabilistic Landslide Susceptibility Models: A Case Study in Lajedo Parish (Flores Island, Azores -Portugal).

Geosciences, 8(5), 153. <https://doi.org/10.3390/geosciences8050153>

Reply: Thank you for this helpful comment. We apologize for the insufficient clarification in the original manuscript. We have revised the relevant sections to better describe the landslide inventory and clarify the applicability of SHALSTAB to the

landslides in this study. We hope that these revisions make the manuscript more complete and improve its clarity.

Revision in our paper: (the revisions are marked in red)

L136:

Historical hazard data were compiled from expert-verified landslides recorded in national geological-hazard investigations and censuses, together with disaster reports from the Shaanxi Environmental Monitoring Center and relevant county and district authorities. Before 2021, a total of 1,407 landslide records were available, including information on hazard type, occurrence time, location, coordinates, scale, triggering factors, and disaster impacts. According to the landslide-size classification specified in the Specification for Investigation of Landslide, Collapse and Debris Flow Disasters (1:50,000), landslide size is classified on the basis of landslide volume. Small landslides dominate the study area, accounting for 90.24% of the total inventory, with volumes of less than $10 \times 10^4 \text{ m}^3$. Medium-sized landslides, with volumes between 10×10^4 and $100 \times 10^4 \text{ m}^3$, account for 8.29%, whereas large landslides, with volumes greater than $100 \times 10^4 \text{ m}^3$, account for only 1.46%. The same specification defines landslides with a sliding-body thickness of less than 10 m are classified as shallow landslides, which is also consistent with the depth criterion adopted in previous studies (Shou and Chen, 2021; Deng et al., 2022). Statistics from the Shaanxi Environmental Monitoring Center and related studies indicate that landslides account for approximately 86.4% of geological hazards in the study area, and more than 90% of these landslides have a sliding-body thickness of less than 10 m (Xiong et al., 2018; Song et al., 2022). Shallow landslides are also widely developed in the Qinling-Daba Mountains and are strongly associated with regional geological and rainfall conditions. Given that small and shallow landslides account for more than 90% of the landslides along the Ancient Shu Roads, this study focuses primarily on shallow landslides and does not further distinguish their detailed types.

L208:

Given that shallow landslides predominate in the study area, SHALSTAB can provide a regional-scale physical stability constraint, and its results can reflect the stability characteristics across most of the study area.

L530:

Several limitations should be acknowledged. First, SHALSTAB relies on regionally parameterized geotechnical parameters and provides relative stability estimates and therefore has limited ability to represent deep-seated landslides, rockslides, and preferential flow through fractures. It cannot fully capture the spatial heterogeneity of rock and soil parameters at the regional scale. Second, the initial CF map was constructed primarily from statistical relationships between historical landslides and conditioning-factor classes. It is consequently sensitive to inventory completeness, factor-classification schemes, and uneven sample distributions and may locally overestimate or underestimate susceptibility. Third, the landslide inventory is affected by survey accessibility and the completeness of historical records. Under-recording in high-elevation or inaccessible areas may influence both the statistical prior and SHAP interpretation. Fourth, previous studies have shown that different landslide typologies may respond differently to combinations of topographic, geological, hydrological, and anthropogenic conditioning factors, and aggregating different landslide typologies may cause the statistical relationships to be dominated by the most abundant landslide typology (Silva et al., 2018). Although shallow

landslides predominate in the study area, which to some extent reduces the influence of mixing different landslide types, this treatment may still obscure differences among individual typologies. Future work can improve the framework in three directions: (1) incorporate measured geotechnical parameters, geological survey data, and uncertainty analysis to increase the reliability of the physical constraint; (2) develop separate susceptibility models for different landslide typologies and compare their dominant conditioning factors and spatial patterns to evaluate how typology-specific modeling affects the statistical relationships between landslides and conditioning factors; and (3) introduce dynamic factors such as multi-temporal rainfall, LULC change, and road-construction evolution to support event-driven or spatiotemporally dynamic susceptibility assessment. Overall, the proposed statistical-physical prior-guided multi-class framework improves the representation and geomorphic plausibility of susceptibility gradients and increases the spatial concentration of very high susceptibility zones. It can support landslide-risk identification, priority mitigation zoning, and the protection of transportation heritage corridors in complex mountainous regions.

Deng, R., Liu, H., Zheng, X., Zhang, Q., Liu, W., Chen, L., 2022. Towards Establishing Empirical Rainfall Thresholds for Shallow Landslides in Guangzhou, Guangdong Province, China. *Water* 14, 3914. <https://doi.org/10.3390/w14233914>.

Shou, K.-J., Chen, J., 2021. On the rainfall induced deep-seated and shallow landslide hazard in Taiwan. *Engineering Geology* 288, 106156. <https://doi.org/10.1016/j.enggeo.2021.106156>.

Silva, R.F., Marques, R., Gaspar, J.L., 2018. Implications of Landslide Typology and Predisposing Factor Combinations for Probabilistic Landslide Susceptibility Models: A Case Study in Lajedo Parish (Flores Island, Azores-Portugal). *Geosciences* 8, 153. <https://doi.org/10.3390/geosciences8050153>.

Xiong, W., Liu, K., Fan, W., 2018. Analysis on internal dynamic geological genesis of shallow landslide in Qin-Ba Mountain area. *Journal of Geomechanics* 24, 424-431. <https://doi.org/10.12090/j.issn.1006-6616.2018.24.03.044>.

Point 2:

The manuscript reports that the historical landslides are used to construct the very-high-susceptibility class and subsequently evaluates the resulting susceptibility maps using historical-landslide coverage and HOI. The authors should clarify whether the same landslide records used to construct the labels are also included in the evaluation of the final maps. If so, the reported HOI and historical-landslide hit rates should not be interpreted as fully independent predictive validation.

Reply: Thank you for this valuable comment. We further clarified the data used for HOI and historical-landslide hit-rate evaluation. In each fold, only landslides in the training regions were used for CF construction and label generation, while landslides in the test regions were excluded from model development. HOI and historical-landslide hit rates were calculated only using landslides in the corresponding test regions. in the corresponding test regions. We hope that this clarification more clearly demonstrates the independence of the HOI and historical-landslide hit-rate evaluation.

Revision in our paper: (the revisions are marked in red)

L287:

During sample construction, stratified sampling was performed based on the five-class landslide susceptibility zoning obtained from the CF results constrained by SHALSTAB. To ensure spatial independence between the training and test sets, the study area was divided into a 3×4 spatial grid according to geographic location, and 10-fold spatial-block cross-validation was adopted. In each fold, eight spatial blocks were used as the training set and the remaining four blocks as the test set. In each fold, the CF statistical prior was recalculated using only the historical landslide samples within the training area and landslide records within the test area were not involved in CF construction. The SHALSTAB results were independently calculated based on the DEM and parameter inputs, and no landslide records from either the training or test sets were used in its construction. Fig. 4 shows the data distribution for one fold. To ensure that all susceptibility levels were represented in model training, 18 landslide points located in the very-low-susceptibility zone and one invalid landslide point were first excluded in this fold. The grid cells containing the remaining 1,388 historical landslide points were assigned as very-high-susceptibility samples, while samples from the very low, low, moderate, and high susceptibility zones were selected according to the predefined sampling proportions. A five-class dataset containing 15,116 samples was ultimately constructed. The five susceptibility classes were very low, low, moderate, high, and very high, corresponding to label values of 1 to 5. Model input variables were extracted from the raster layers of landslide conditioning factors, with each sample containing a set of conditioning-factor values and its corresponding susceptibility class. To further evaluate the predictive capability of the models for newly occurring landslide events, 47 landslides that occurred between 2021 and 2025 were used as a temporally independent validation dataset. These landslide records were not involved in label construction or model training, and were used only for independent validation.

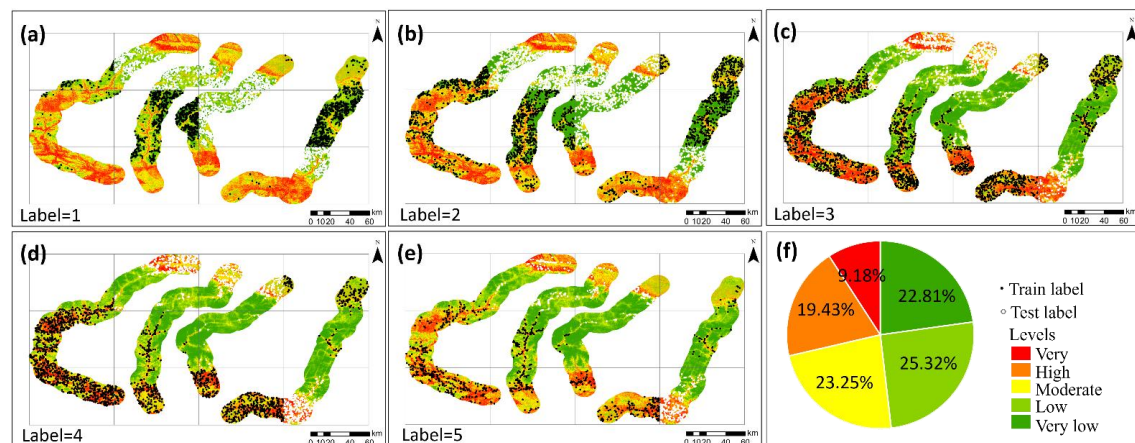


Fig. 5. Distribution of landslide susceptibility samples in one fold. (a) Very low susceptibility samples; (b) Low susceptibility samples; (c) Moderate susceptibility samples; (d) High susceptibility samples; (e) Very high susceptibility samples; (f) Sample composition.

L356:

Model performance was evaluated using class-specific recall together with F1, accuracy, and AUC for the multi-class task, thereby preventing overall accuracy from obscuring poor recognition of an individual class (Hand and Till, 2001; Sokolova and Lapalme, 2009). The metrics were calculated as follows:

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (8)$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (9)$$

$$F1 = \frac{1}{K} \sum_{i=1}^K F1_i$$

$$Accuracy = \frac{N_{correct}}{N_{total}} \quad (10)$$

$$AUC = \frac{1}{K} \sum_{i=1}^K AUC_i \quad (11)$$

where TP_i , FP_i , and FN_i denote the true positives, false positives, and false negatives for class i , respectively; K is the number of classes ($K = 5$); $N_{correct}$ is the number of correctly classified samples; N_{total} is the total number of samples; and AUC_i is the one-versus-rest AUC for class i .

To ensure that the final evaluation uses landslide observations that were not involved in any previous modeling or label-construction process, the spatial mapping results were further evaluated using the HOI to measure the ability of the very-high-susceptibility zone to capture historical landslide points within a relatively small area (Zhang et al., 2026). HOI jointly considers the areal proportion of the very high susceptibility zone and the hit rate of historical landslides within that zone and is calculated as follows:

$$HOI = 1 - a_5 + h_5 \quad (12)$$

L400:

To evaluate the historical-landslide coverage and spatial concentration of the susceptibility maps, the areal proportion of susceptibility class (a), the historical-landslide hit rate (h), and HOI were calculated. Binary results were summarized in two ways: Table 7 reports statistics after dividing binary probability maps into five classes by natural breaks, and Table 8 reports the landslide-prone zones defined by the binary threshold ($p > 0.5$). As shown in Tables 7 and 8, TabPFN achieved the highest threshold-based binary HOI (1.70); its landslide-prone zone occupied 16.9% of the study area and contained 86.7% of the historical landslides. Its HOI based on the five-class natural-breaks result was 1.68, lower than the threshold-based HOI and confirming that evaluating a binary model solely by its very high susceptibility class underestimates model capability. Table 9 shows that TabPFN also performed best in the multi-class setting. Its very high susceptibility zone occupied only 10.8% of the study area but contained 86.0% of the historical landslides, producing an HOI of 1.75. The comparison of HOI between binary and multi-class classification is shown in Fig. 6, indicating that the multi-class classification approach can produce a more spatially concentrated high-susceptibility zone.

Table 7

Statistical results of landslide susceptibility maps based on binary classification using natural breaks (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	50.6	11.0	16.3	16.9	21.4	40.2	7.3	24.7	4.4	7.1	1.03
FT-Transformer	62.7	25.4	17.2	20.0	8.8	20.6	7.7	19.0	3.6	15.1	1.11
SVM	66.4	8.4	6.9	6.8	9.1	16.0	12.0	40.5	5.7	28.3	1.23
XGBoost	61.0	3.4	9.0	4.3	7.1	7.0	6.4	5.2	16.5	80.1	1.64
RF	61.1	3.3	9.9	3.3	7.0	6.3	7.1	10.7	14.9	76.4	1.61
TabPFN	56.0	1.5	11.5	1.5	7.3	2.7	7.4	5.0	17.8	85.3	1.68

Table 8

Statistical results of landslide susceptibility maps based on binary classification using threshold (best values in bold).

Model	Non-landslides		Landslides		HOI
	a_{neg}	h_{neg}	a_{pos}	h_{pos}	
LR	92.6	81.0	7.4	19.0	1.12
FT-Transformer	91.8	72.2	8.2	27.8	1.20
SVM	85.0	40.6	15.0	59.4	1.44
XGBoost	87.2	29.0	12.8	71.0	1.58
RF	82.2	18.8	17.8	81.2	1.63
TabPFN	83.1	13.3	16.9	86.7	1.70

Table 9

Statistical results of landslide susceptibility maps based on multi-class classification (best values in bold).

Model	Very low		Low		Moderate		High		Very high		HOI
	a_1	h_1	a_2	h_2	a_3	h_3	a_4	h_4	a_5	h_5	
LR	16.2	1.2	30.2	8.0	15.2	13.2	22.7	20.1	15.6	57.4	1.42
FT-Transformer	18.2	0.2	26.6	7.6	22.4	8.7	21.5	22.5	11.3	61.0	1.50
SVM	17.1	0.7	26	4.7	22.6	5.4	20.8	11.3	13.5	77.8	1.64
XGBoost	16.8	0.2	28.1	4	20.8	7.8	21.6	11.8	12.7	76.1	1.63
RF	16.8	0.2	28.3	3.1	21.9	7.1	21.5	6.9	11.5	82.7	1.71
TabPFN	18.2	0.1	27.6	3.5	22.4	4.8	21.0	5.6	10.8	86.0	1.75

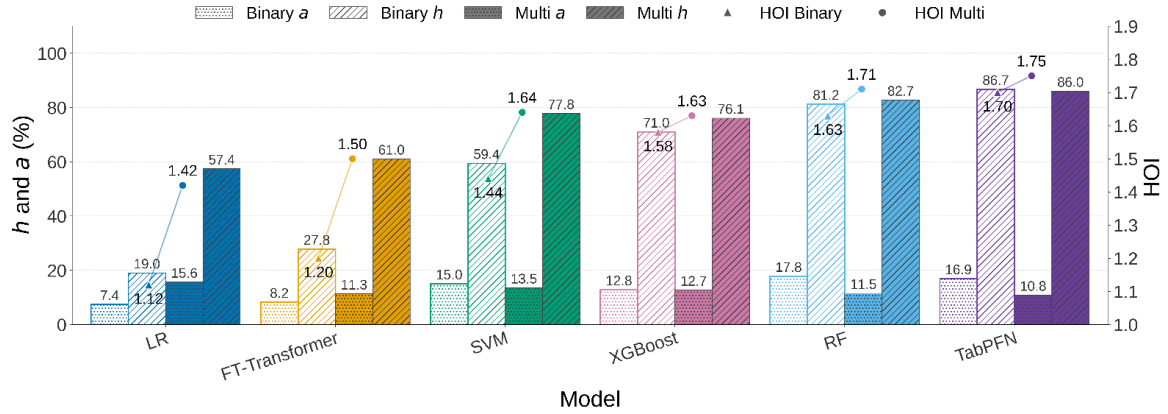


Fig. 6. Comparison of area ratio (a), hit rate (h), and HOI for binary and multi-class susceptibility maps.

Point 3:

The use of 14 conditioning factors should be more strongly justified. Several factors may represent related environmental processes or may be spatially correlated. The authors should provide a correlation/multicollinearity assessment. This would help determine whether the proposed framework benefits from the full set of factors or whether some factors provide limited additional information.

Reply: Thank you for this valuable comment. We sincerely apologize that the correlations and potential multicollinearity among the 14 conditioning factors were not sufficiently assessed in the original manuscript. Following your suggestion, we added a correlation analysis to evaluate potential redundancy among the conditioning factors. We hope that this additional analysis further clarifies the rationale for factor selection and improves the completeness of the methodology.

Revision in our paper: (the revisions are marked in red)

L273:

To assess potential correlations and information redundancy among the landslide conditioning factors, pairwise correlation analysis was conducted, and a correlation coefficient matrix was used to evaluate the degree of association among the factors. The closer the absolute value of the correlation coefficient is to 1, the stronger the correlation between two factors; values closer to 0 indicate weaker correlations. In this study, $|r|=0.7$ was used as a reference threshold for strong correlation. When the absolute correlation coefficient between two factors exceeded 0.7, relatively substantial information overlap or multicollinearity was considered possible. As shown in Fig. 4, the overall correlations among the 14 landslide conditioning factors are relatively low, with all pairwise absolute correlation coefficients below 0.7, indicating no obvious strong correlations. Among them, distance to faults and distance to rivers show a relatively high positive correlation, with a correlation coefficient of 0.60. DEM and LULC exhibit a relatively strong negative correlation, with a correlation coefficient of -0.53, while the correlation coefficient between DEM and distance to rivers is 0.51. The absolute correlation coefficients of most other factor pairs are below 0.5. Overall, no severe correlations or obvious information redundancy are

observed among the conditioning factors; therefore, all 14 factors were retained for subsequent landslide susceptibility modeling.

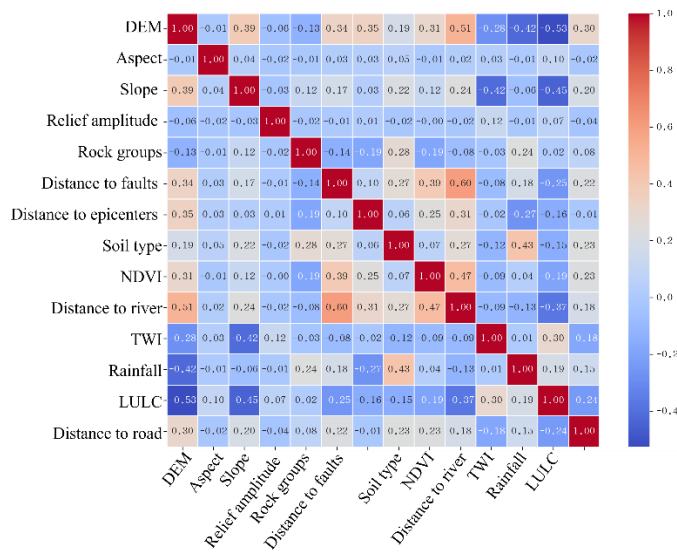


Fig. 4. Correlation matrix of the landslide conditioning factors.

Point 4:

The manuscript should moderate the general statement that multi-class classification provides greater historical-landslide coverage and spatial concentration. The results demonstrate this advantage for the present study area, inventory, class-construction procedure and evaluation framework, but they do not necessarily establish that multi-class susceptibility modelling is generally superior to binary modelling. This distinction should be reflected in the conclusions.

Reply: Thank you for this helpful comment. We apologize for the overly general statement. We have revised the relevant sentence in the manuscript to clarify that the observed advantage of multi-class classification applies to the present study. We hope the revised wording adequately addresses this concern.

Revision in our paper: (the revisions are marked in red)

L27:

These findings indicate that multi-class learning provides greater historical-landslide coverage and spatial concentration in this study.

L511:

5. Conclusions

This study proposed a statistical-physical prior-guided framework for multi-class landslide susceptibility mapping and applied it to direct five-class mapping along the Ancient Qin-Shu Roads in the Qinling-Daba Mountains. The principal conclusions are as follows. (1) A five-class dataset representing very low, low, moderate, high, and very high landslide susceptibility was constructed from the SHALSTAB-constrained initial CF map. The dataset comprised 15,116 sample points, including 1,388 historical landslide records, and 14 conditioning-factor attributes were extracted for each sample

for subsequent multi-class modeling. (2) Using identical samples, TabPFN achieved the best overall performance in both binary and multi-class tasks. In the binary task, TabPFN attained F1, accuracy, and AUC values of 82.62%, 82.14%, and 91.00%, respectively; in the multi-class task, the corresponding values were 89.26%, 82.14%, and 92.35%. (3) Evaluation of the spatial susceptibility patterns showed that multi-class classification identified high-risk areas more compactly in this study. The very high susceptibility zone in the multi-class TabPFN result occupied only 10.8% of the study area but contained 86.0% of the historical landslides, yielding an HOI of 1.75. By contrast, the threshold-defined landslide-prone zone in the binary TabPFN result occupied 16.9% of the area, contained 86.7% of the historical landslides, and had an HOI of 1.70. For the present study area, dataset, class-construction procedure, and evaluation framework, multi-class classification captured more historical landslides within a smaller very-high-susceptibility area, improving spatial concentration and facilitating the delineation of priority mitigation zones. (4) The SHALSTAB stability result produced a physically constrained initial CF susceptibility map. Relative to CF alone, the constraint increased accuracy from 86.6% to 89.04%, AUC from 90.84% to 92.35%, and HOI from 1.70 to 1.75, thereby improving the spatial concentration of landslides. (5) SHAP interpretation identified DEM, LULC, and distance to roads as the principal factors controlling the very high susceptibility class. Low-elevation valleys, cultivated land, and areas close to roads were more likely to be classified as very high susceptibility.

Several limitations should be acknowledged. First, SHALSTAB relies on regionally parameterized geotechnical parameters and provides relative stability estimates and therefore has limited ability to represent deep-seated landslides, rockslides, and preferential flow through fractures. It cannot fully capture the spatial heterogeneity of rock and soil parameters at the regional scale. Second, the initial CF map was constructed primarily from statistical relationships between historical landslides and conditioning-factor classes. It is consequently sensitive to inventory completeness, factor-classification schemes, and uneven sample distributions and may locally overestimate or underestimate susceptibility. Third, the landslide inventory is affected by survey accessibility and the completeness of historical records. Under-recording in high-elevation or inaccessible areas may influence both the statistical prior and SHAP interpretation. Fourth, previous studies have shown that different landslide typologies may respond differently to combinations of topographic, geological, hydrological, and anthropogenic conditioning factors, and aggregating different landslide typologies may cause the statistical relationships to be dominated by the most abundant landslide typology (Silva et al., 2018). Although shallow landslides predominate in the study area, which to some extent reduces the influence of mixing different landslide types, this treatment may still obscure differences among individual typologies. Future work can improve the framework in three directions: (1) incorporate measured geotechnical parameters, geological survey data, and uncertainty analysis to increase the reliability of the physical constraint; (2) develop separate susceptibility models for different landslide typologies and compare their dominant conditioning factors and spatial patterns to evaluate how typology-specific modeling affects the statistical relationships between landslides and conditioning factors; and (3) introduce dynamic factors such as multi-temporal rainfall, LULC change, and road-construction evolution to support event-driven or spatiotemporally dynamic susceptibility assessment. Overall, the proposed statistical-physical prior-guided multi-class framework improves the representation and geomorphic plausibility of susceptibility gradients and increases the spatial concentration of very high susceptibility zones. It can support

landslide-risk identification, priority mitigation zoning, and the protection of transportation heritage corridors in complex mountainous regions.

Point 5:

The manuscript harmonizes all conditioning factors to a 30-m grid, although the original datasets have substantially different spatial resolutions and mapping scales. The authors should discuss the implications of this harmonization, particularly for rainfall, NDVI and geological data, and clarify the effective spatial resolution of the susceptibility model. The representation of individual landslides relative to the 30-m mapping unit should also be discussed.

Reply: Thank you

Revision in our paper: (the revisions are marked in red)

L232:

The landslide conditioning factors are shown in Fig. 3. Topographic factors, including aspect, slope, and relief amplitude, were derived from the DEM. Based on the lithology, the study area were grouped into six classes: unconsolidated gravelly-soil deposits (A), massive hard metamorphic rocks (B), massive hard intrusive rocks (C), schistose-bedded weak metamorphic rocks (D), schistose-massive moderately hard metamorphic rocks (E), and medium- to thick-bedded intercalated hard and soft metamorphic rocks (F). Euclidean distances from each grid cell to the nearest fault, earthquake epicenter, river, and road were calculated in ArcGIS from the corresponding vector datasets. Soil was classified as luvisols, semi-luvisols, primosols, hydromorphic soils, anthroposols, or others. LULC was grouped into cultivated land, forest land, shrubland, grassland, water body, bare land, and impervious surface. All datasets were transformed to the WGS 84 spatial reference and a 30-m grid. Lower-resolution continuous data were resampled by bilinear interpolation, whereas categorical data were resampled using the nearest-neighbor method. After harmonization of projection, resolution, and spatial extent, these factors served as the base data for susceptibility assessment and sample construction. All conditioning factors in this study were harmonized to a 30-m grid to ensure spatial consistency among multi-source datasets and modeling units. Therefore, 30 m represents the unified computational grid of the model rather than implying that all factors inherently contain spatial information at a 30-m resolution. For rainfall and geological data with original spatial scales coarser than 30 m, resampling mainly serves to achieve raster alignment and does not generate new spatial information. Consequently, adjacent 30-m cells may contain identical or interpolated values, while the representation of local rainfall variability and geological boundary details remains constrained by the original data scale. Remote-sensing factors such as NDVI may also undergo a certain degree of spatial smoothing during resampling. The output resolution of the susceptibility maps in this study is 30 m, and the results are more suitable for regional-scale identification of relative landslide susceptibility. For the landslide inventory, each historical landslide was represented by the 30-m grid cell containing its recorded location during sample construction. This representation ensures consistent spatial correspondence between landslide observations and environmental factors but does not describe the complete geometric extent of individual landslides.

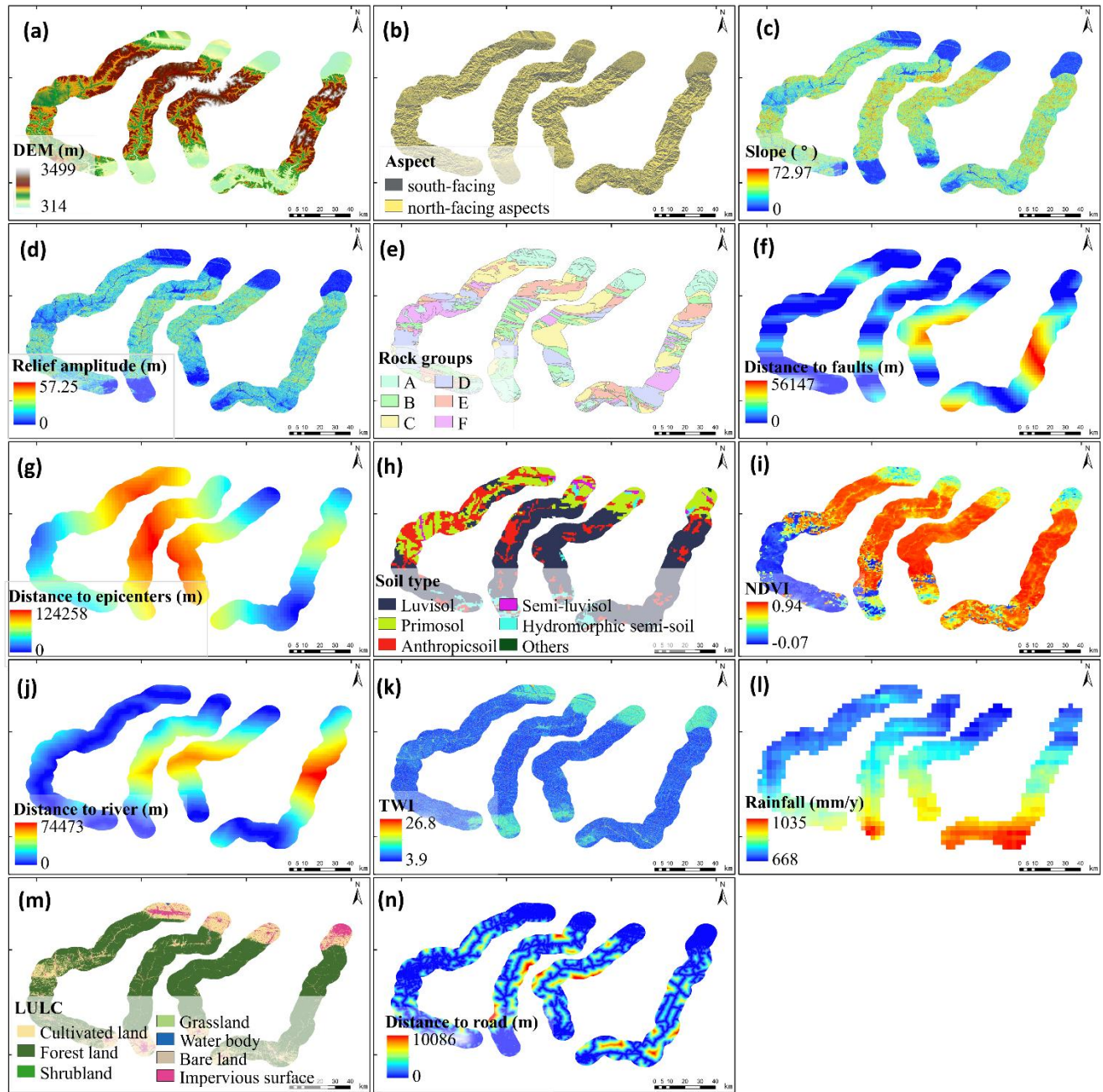


Fig. 3. Landslide conditioning factors. (a) DEM; (b) Aspect; (c) Slope; (d) Relief amplitude; (e) Rock groups; (f) Distance to faults; (g) Distance to earthquake epicenters; (h) Soil type; (i) NDVI; (j) Distance to rivers; (k) TWI; (l) Rainfall; (m) LULC; (n) Distance to roads.