Response to Reviewer #3

Manuscript Title: *Landslide susceptibility mapping with explainable AI techniques: Evidence from Bavaria, Germany*

Dear Editor and Reviewer,

We thank the Editor for handling our manuscript and the Reviewer for the thorough and constructive review, which has helped us sharpen the paper's message and improve its rigor.

In revising the manuscript, we made our central contribution more explicit, especially in the introduction, discussion, and conclusion. In response to the reviewer's concerns and suggestions, we included statistics on susceptibility levels within mountainous terrain, clarified the flat-area masking and explained in more detail how the class imbalance was handled. We are also improving the figures and maps. In addition, we ran more experiments, as suggested by the reviewers, such as the spatial cross-validation, and benchmark modeling. The results of these experiments will be incorporated in the revised version of the manuscript.

Our point-by-point responses are provided below in blue, with each comment in italics.

Sincerely,

Veronika Buchauer, Marta Sapena, Christian Geiß, Patrick Aravena Pelizari, and Hannes Taubenböck

RC3: 'Comment on egusphere-2026-1647', Anonymous Referee #3

Comment 1: *I have read your manuscript with great detail. I have the following major feedbacks: I see that the flat areas are not masked out in the slope units, is there a reason for that? If you look at the Figure 9, the compared Susceptibility map has a masked region. This is important because your model shows great variation in susceptibility overall but I am not sure if it has similar variation in the mountainous region only too or not (e.g. Figure 4). If all the mountainous regions have susceptibility above 0.7 it would not be a robust analysis. Therefore, I recommend masking the flat regions out.*

Response: We thank the reviewer for this constructive comment, which raises two related concerns: the masking of flat areas, and whether the estimated susceptibility varies within the mountainous parts of the study area.
(1) Masking flat areas: As described in Section 2.2.2, we exclude slope units that are too flat and highly unlikely to trigger a landslide. This masking removes $11,720$ of the $317,788$ slope units and leaves $306,068$ units in the analysis. Beyond this mask based on the topography and land cover of the slope units, we deliberately did not exclude more flat parts of Bavaria from the susceptibility estimation, because our objective was to provide a comprehensive susceptibility map for the entire study area rather than only for selected regions. The compared susceptibility map from LfU only shows binary labels (susceptible or not susceptible to landslides) while we have a continuous approach where nuances in susceptibility are visible and that we want to share for the whole region of Bavaria.
(2) Susceptibility variation within mountainous terrain: To confirm that the model does not simply assign high susceptibility to all steep or high-elevation terrain, Figure 1 in this response letter shows the distribution of the estimated susceptibility across slope and elevation classes. Susceptibility varies markedly within mountainous terrain and is far from uniformly high. Among steep units (mean slope ≥ 15, n = $17,545$) the median estimated susceptibility is 0.25 (mean 0.30; 95th percentile 0.64): about $70\,\%$ of these units fall below 0.5, and only $2.0\,\%$ reach the high-susceptibility range mentioned by the reviewer (≥ 0.7). The same holds at high elevation: for units above $1,000\,m$ (mostly in the Bavarian Alps; n = $9,632$) the median is 0.25, $71\,\%$ lie below 0.5, and only $0.9\,\%$ exceed 0.7.
Moreover, the estimated susceptibility is not a monotonic function of slope. The median susceptibility increases with slope up to a maximum in the 20–$25°$ class (0.36) and then decreases again for the steepest units ($\geq 30° : 0.22$). The observed landslide frequency follows the same pattern, peaking at $18.3\,\%$ in the 20–$25°$ class and decreasing to $6.2\,\%$ for slopes $\geq 30°$. The model therefore reproduces the geomorpholog-

ical behaviour of the inventory, the steepest, predominantly bedrock slopes are less prone to the landslides recorded here, rather than seing steep terrain directly as high susceptibility. In the revised version of the manuscript we will describe these topographic characteristics of the predicted susceptibility across slope units, and the figure as supplementary material.

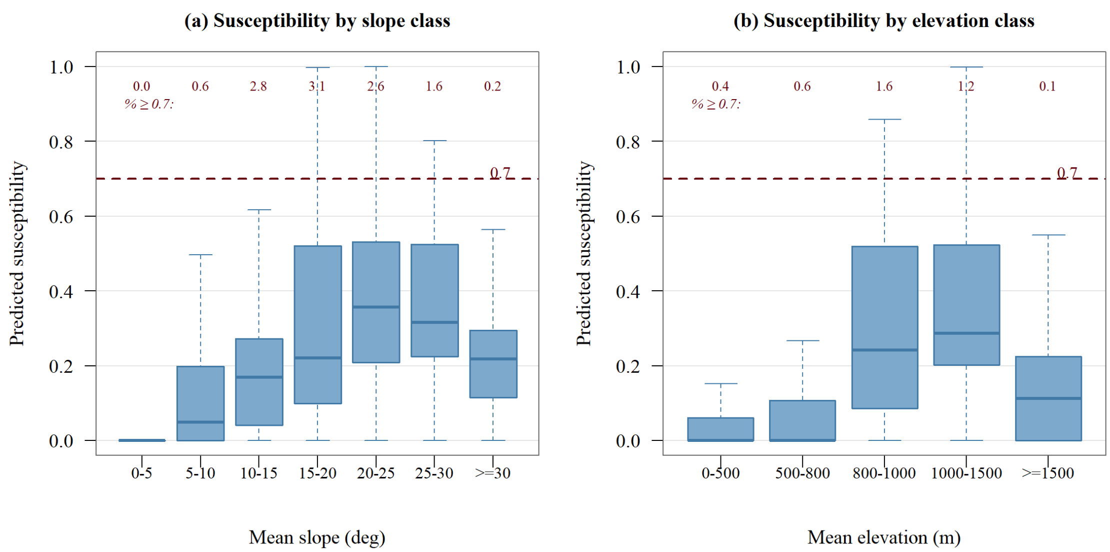

Figure 1: Distribution of the estimated landslide susceptibility within slope (a) and elevation (b) classes for all modeled slope units. Boxes show the interquartile range and median; the dashed line marks the 0.7 high-susceptibility level suggested by the reviewer, and the value above each box gives the percentage of units in that class reaching ≥ 0.7.

Comment 2: *The random train–test split likely introduces spatial leakage, so spatial cross-validation or blocked validation should be implemented to provide more realistic performance estimates.*

Response: We thank the reviewer. We agree that the random hold-out can be optimistic, so we ran a spatial leave-one-region-out cross-validation over the seven Regierungsbezirke, training and validating the same model (same architecture, 84 features, flat-area mask and oversampling) on all but one region and testing on the held-out one (Table 1, Figure 2).

Performance drops relative to the random split, confirming that the previously used metrics are optimistic. The pooled ROC AUC is 0.93 (vs 0.953) and the PR AUC 0.37 (vs 0.53), which is a small drop for the ROC AUC but larger for the minority-class sensitive PR AUC. It also varies strongly by region, as most of the seven Regierungsbezirke transfer well (ROC AUC 0.91–0.95), but Niederbayern collapses (ROC AUC 0.80, PR AUC 0.03). This is not only a prevalence effect, i.e. not only the amount of landslides in the region is causing this effect, as Oberpfalz has the lowest prevalence ($0.6\,\%$) yet transfers well (Figure 2).

The important component is whether a region's landslides lie in settings seen elsewhere in the training set. The model predicts slope units with landslides in Oberpfalz's with a median susceptibility of 0.54 but Niederbayern's at only 0.09 (Figure 3). This is not due to different topographic settings of landslides in Niederbayern but it rather reflects their different geological environment (Table 2). On average, $44\,\%$ of the area of a Niederbayern landslide slope unit lies in the Tertiary hill country of the Alpine foreland, against only $3\,\%$ for the landslides of the other regions, and these units contain virtually none of the Alpine and young-moraine terrain that makes up a large share of the training landslides elsewhere. A model trained on the other regions therefore does not learn this setting well.

Spatial generalization is thus governed by how well the training inventory represents a region's landslide-generating settings, not by the model or by regional prevalence. We will add this analysis to the results and discussion of the revised manuscript.

In addition, while training the cross validation models, we identified an error in the original computation of

the PR AUC. The corrected PR AUC of the random split is 0.53 (the ROC AUC of 0.953 is unaffected, as well as confusion matrix based metrics), and we will correct this in the revised manuscript. We sincerely thank the reviewer to help us discover this miscalculation.
A PR AUC of 0.53 remains a strong result given the severe class imbalance. With a landslide prevalence of only $2.6\,\%$, it is about twenty times the no-skill baseline of 0.026. The model also retains a high sensitivity (0.92) and balanced accuracy (0.89) at the threshold of 0.20 (Table 1 of the manuscript), i.e. it still correctly identifies the large majority of slope units with landslides. The correction therefore does not affect the conclusion that the model discriminates landslide-prone from stable slope units well.

Region (test)	Prevalence	ROC AUC	PR AUC	Sensitivity	Specificity	Bal. acc.
Mittelfranken	2.70%	0.952	0.452	0.862	0.900	0.881
Oberfranken	4.11%	0.937	0.490	0.783	0.914	0.849
Oberbayern	3.82%	0.935	0.416	0.904	0.828	0.866
Oberpfalz	0.60%	0.916	0.258	0.782	0.929	0.855
Schwaben	3.93%	0.915	0.347	0.816	0.855	0.835
Unterfranken	0.76%	0.912	0.082	0.714	0.879	0.796
Niederbayern	0.75%	0.804	0.027	0.249	0.949	0.599
Pooled	2.64%	0.932	0.370	0.824	0.884	0.854

Table 1: Spatial leave-one-region-out cross-validation over the seven Regierungsbezirke (each region is the held-out test set once). Confusion-matrix metrics are at the 0.20 decision threshold.

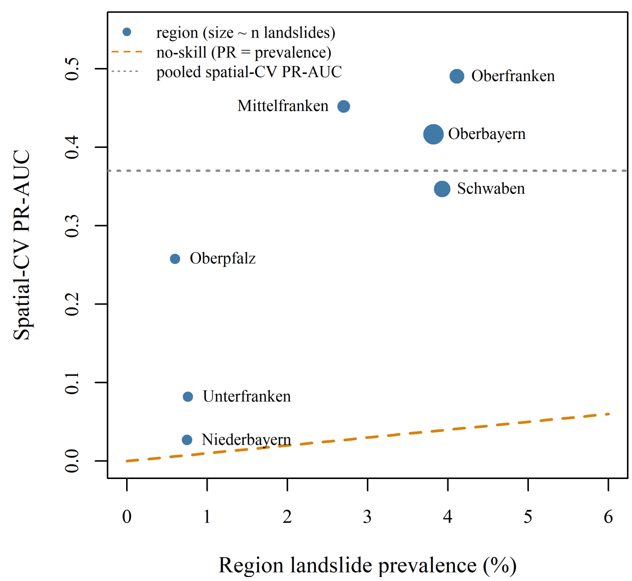

Figure 2: PR AUC from spatial-CV versus landslide prevalence per region. The point size is proportional to the absolute number of landslides. The dashed line shows the no-skill baseline with PR $=$ prevalence, while the dotted line represents the pooled spatial-CV PR AUC.

Setting of the landslides	Niederbayern	Other regions
Tertiary hills / Alpine-foreland cover gravels (soil region)	44%	3%
Low to very low permeability	42%	3%
Silicate parent rock	48%	11%
Alpine soils	0%	24%
Young-moraine soils	0%	24%

Table 2: Geological and soil settings of the landslides in Niederbayern compared with the landslides of the other six regions. Values give the mean areal cover of each setting across the landslide slope units, i.e. the average fraction of a landslide slope unit's area that falls in that setting.

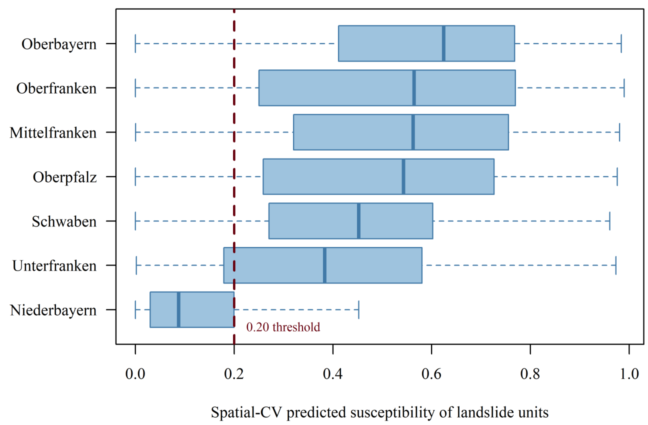

Figure 3: Cross-validated predicted susceptibility of the actual landslide units in each region. Niederbayern's landslides are almost all scored below the 0.20 threshold (median 0.09), unlike the other regions.

Comment 3: *The modeling framework incorrectly treats all non-landslide slope units as true negatives, whilst you remove 11k SU out of 317K and justify accounting for class imbalance, does it really solve your problem of strong class imbalance of positive 2.6 %? In my calculation it makes it 2.7 %. I would expect a better way to handle such a strong class imbalance.*

Response: We thank the reviewer for this comment. We think the purpose of masking some units has been misunderstood, therefore we will clarify this first and then explain how we actually handle the class imbalance.
First, we remove some slope units not only to balance the classes. The $11,720$ excluded slope units are water bodies, sealed/artificial land or (near-)flat terrain that are very unlikely to experience a landslide, so we discarded them from the analysis since the probability of slope failure is very low. This was not meant for handling the class imbalance, and we agree with the reviewer that it leaves the imbalance essentially unchanged. We will rephrase the corresponding sentences in Section 2.3.2 to make clearer that discarding these slope units does not substantially help with class imbalance. We handle that during model training instead.
Second, in the training we used oversampling until we reach 10 % of positives along with the binary cross-entropy (BCE) loss function. However, we also tested imbalance-aware losses that tackle the imbalance problem, such as class-weighted BCE and binary focal loss, on their own and combined with several re-sampling schemes: random over-/under-sampling, SMOTE and Borderline-SMOTE. Nevertheless, none of the combinations provided better results than BCE with random oversampling. That is why we kept this strategy for the final model. We described these tests in Section 2.3

Comment 4: *The ANN architecture is insufficiently justified and should be benchmarked against simpler baseline models such as logistic regression, random forests, or XGBoost and compared in greater detail.*

Response: We thank the reviewer for this suggestion. The reasoning for choosing the ANN is addressed

model	ROC AUC	PR AUC	sensitivity	specificity	balanced accuracy
ANN	0.953	0.53	0.915	0.858	0.887
Logistic regression	0.935 (± 0.0004)	0.338 (± 0.0004)	0.772 (± 0.0007)	0.916 (± 0.0004)	0.844 (± 0.0004)
Random Forest	0.959 (± 0.0003)	0.517 (± 0.0024)	0.783 (± 0.0038)	0.949 (± 0.0001)	0.866 (± 0.0019)
XGBoost	0.963 (± 0.0004)	0.530 (± 0.0053)	0.704 (± 0.0041)	0.970 (± 0.0006)	0.837 (± 0.0020)

Table 3: Benchmark of the ANN against logistic regression, random forest and XGBoost under the same experimental setup. ROC AUC, PR AUC and confusion matrix based metrics (sensitivity, specificity and balanced accuracy at a 0.20 decision threshold) are reported on the test set. For the compared models, values are the mean (± standard deviation) over three runs. The ANN row is the model used for landslide susceptibility mapping with the corrected PR AUC.

more thoroughly in the response to RC2, comment 4 and will be taken into the revised manuscript. As this concern has come up by multiple reviewers, we have benchmarked the ANN against some suggested models: logistic regression (LR), random forest (RF) and XGBoost, following the same experimental setup (flat-area mask, oversampling and an 80/20 random split). To ensure a fair comparison, the hyperparameters of RF and XGBoost were tuned on the validation set (for RF, the number of variables sampled at each split, mtry; for XGBoost, the tree depth, learning rate and number of boosting rounds), whereas LR was kept as a standard baseline, as adding regularisation did not improve its performance. We trained each model with three different seeds and the results can be seen in Table 3 of this response letter. The tree-based models (RF and XGBoost) performed similarly than the ANN, whereas LR shows a poorer performance, particularly in PR AUC, which is consistent with the non-linearity in our data. Apart from LR, therefore, the models show similar performance with only minor differences. We additionally report sensitivity, specificity and balanced accuracy at a 0.20 threshold in Table 3. These threshold-based metrics are, however, only of limited comparability across models, because their predicted susceptibilities are not equally calibrated.
Overall, we interpret the benchmarking results as support for the central message of our study, which we will state explicitly in the revised version of the manuscript. The quality and completeness of the landslide inventory have a greater impact on the resulting susceptibility map than the choice of model. We include the comparison in the appendix, as a detailed model comparison is not the focus of the paper. Finally, as noted in our response to Reviewer 2, we will also extend the justification for choosing the ANN in the revised manuscript.

Comment 5: *I find it really hard to intuitively get what is the novelty in the work, I know it is on the role of inventory but it should be discussed in greater detail and the story must be made more coherant to explain the story. Now the paper looks more like new AI model for landslide susceptibility (which is not the case as such models have been used a lot). I recommend framing the story in similar lines of this paper: https://link.springer.com/article/10.1007/s10064-005-0023-0*

Response: We thank the reviewer for this valuable comment. We agree that the novelty of our study should be clear. If it is not clear enough yet, we are happy to sharpen it.
To restate our contribution clearly: it is not a new AI model, as the reviewer rightly notes, such models are by now well established. The contribution aims on the role of the landslide inventory in susceptibility mapping, in line with the focus of this special issue on inventory quality. Our study demonstrates that the completeness and quality of the inventory, rather than the choice of model, govern the quality of resulting susceptibility map. In addition, we present the first Bavaria-wide landslide susceptibility map, which represents a further novel aspect of this work in terms of its spatial extent and regional relevance. The discussion is already centered on these points, and we will make them equally explicit in the introduction and the conclusion, so that the inventory-centered storyline is clear and the model is presented as a tool rather than the primary contribution.
We are also grateful for the suggested reference, which is a good example of a coherent, inventory-driven narrative. We would, however, note a difference in scope: the suggested paper reviews quantitative landslide hazard and risk zonation, taking temporal probability, run-out modeling, vulnerability, and risk at various planning scales into account, whereas our work concerns susceptibility mapping and, within it, the role of the inventory. While we share their emphasis on the inventory as a central input, reframing the paper along hazard- and risk-zonation lines would extend it beyond its intended scope and beyond the goal of this spe-

cial issue. We will therefore adopt the reviewer's suggestion and clarify the inventory-centred story, while retaining the susceptibility-mapping scope.

Comment 6: *The threshold selection and evaluation framework should be strengthened using more rigorous imbalance-aware and calibration-sensitive performance metrics.*

Response: We fully agree that, given the strong class imbalance, the evaluation must rely on imbalance-aware metrics rather than on overall accuracy, and this is in fact a point we were careful to address in the manuscript.
As described in Section 3.1 and summarized in the manuscript's Table 1, we deliberately do not judge the model on accuracy, which we noted is unsuitable for imbalanced data and highly depends on the selected threshold, thus it is only reported for completeness and for allowing comparison with other studies. Our assessment instead rests on metrics that remain informative under imbalance:
- The Precision–Recall AUC, it reflects the performance on the minority class and is not masked by the dominant majority class;
- Balanced accuracy (the mean of sensitivity and specificity) rather than raw accuracy;
- An operating threshold (0.20) chosen to maximize balanced accuracy while prioritizing sensitivity over specificity, since in landslide susceptibility a false negative is far more costly than a false positive.
The evaluation framework is in our understanding already built around imbalance-aware metrics.
In addition, to evaluate the model with the validation tools most widely used in landslide susceptibility mapping (Lima et al. (2022)) and to demonstrate its generalizability to inventory changes, we add success-rate and prediction-rate curves in Figure 4 of this response. The success-rate curve is computed against the training inventory and summarizes the model fit (AUC = 0.93), whereas the prediction-rate curve is computed against the updated, independent landslides and therefore measures how well the model generalizes to landslides that were not used in training (AUC = 0.73). We will include these results and metrics in the revised manuscript.

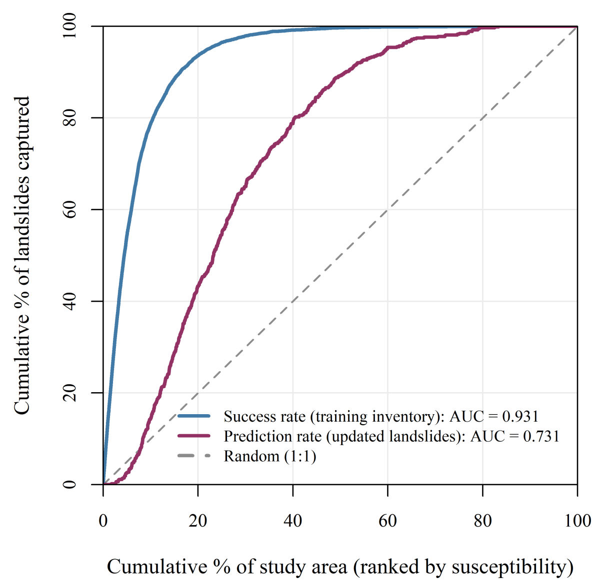

Figure 4: Success-rate and prediction-rate curves for the susceptibility map. Slope units are ranked by predicted susceptibility. The x-axis gives the cumulative proportion of the study area (by slope-unit area) and the y-axis the cumulative proportion of landslides captured. The success-rate curve (blue) uses the training inventory (AUC = 0.93); the prediction-rate curve (purple) uses the updated, independent landslides (AUC = 0.73).

Comment 7: *The figure-1 workflow should be removed in my opinion, things such as "transform python code to R" is not actual method but a mere technical detail and flowcharts with such a long flow and details do not describe and justify the choice of method (which is done in text). Therefore I recommend removing* *or simplifying this figure.*

Response: We agree that the workflow figure includes technical implementation details that are not part of the methodology itself and are better left to the text. We would rather simplify Figure 1 so that it presents only the main methodological steps at a higher level of abstraction, as we think it is in general a valuable figure giving readers an idea of the workflow at one glance. The revised manuscript will contain this updated figure (Figure 5).

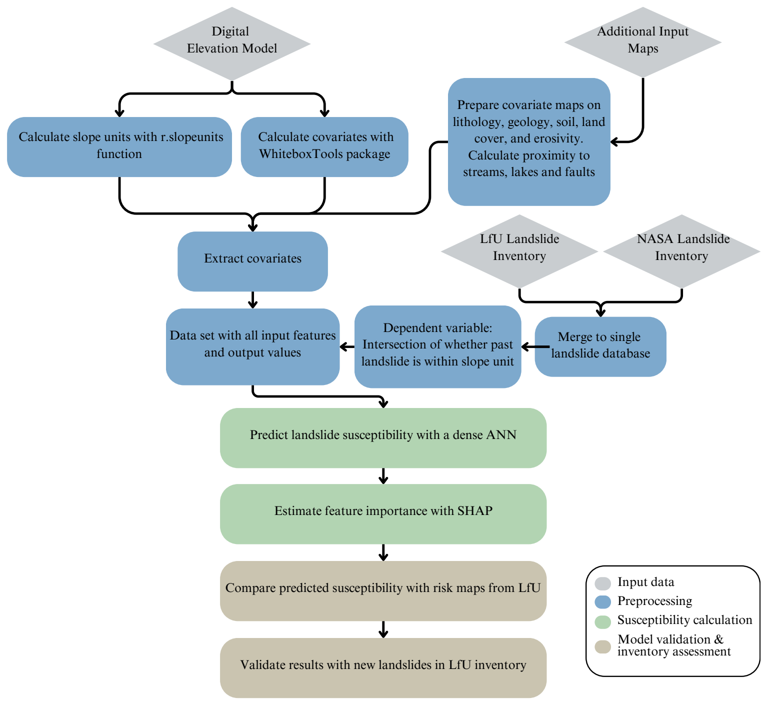

Figure 5: Revised workflow figure

Comment 8: *In general all maps should follow cartographic standards, which are missing overall (like co-ordinates, north arrow, consistency on using capital and small letter (in Digital Elevation Model, , shaded relief, either make everything uppercase or everything lowercase).*

Response: We thank the reviewer for these helpful suggestions, and we have revised the maps accordingly. North arrows have been added to all maps (as also recommended by other reviewers), and scale bars are present on all of them. We have also harmonized the capitalization by correcting the inconsistency in Figure 3, while the remaining maps follow the journal's conventions. As for coordinates, we propose to state the AOI coordinates in the figure captions.

Comment 9: *The manuscript should discuss temporal inconsistencies between landslide inventories and environmental predictor datasets more explicitly.*

Response: We map landslide susceptibility, i.e. the spatial likelihood that a slope unit is prone to landsliding, rather than landslide hazard or risk. Susceptibility expresses where landslides are likely, not when or how frequently they occur. This temporal dimension enters only in hazard (the probability of occurrence within a given period) and, building on it, in risk. Because our target is therefore time-independent by definition, temporal consistency between the inventory and the feature datasets is far less critical here than it would be for a hazard or risk assessment.
This is reinforced by our choice of features, which are all temporally stable. They are essentially time-invariant terrain and ground attributes derived from the digital elevation model and from geological and soil data, none of which change significantly over the period covered by the inventory. The rainfall erosivity is likewise a long-term climatological average rather than a value for a specific year, and the land cover in the

landslide-prone terrain is dominated by stable forest and (semi-)natural cover. The susceptibility mapping is therefore effectively time-insensitive, and the temporal mismatch between the datasets is not expected to affect the results.

Comment 10: *The study should must reproducibility by providing complete implementation details, hyper-parameter ranges, code availability, and workflow documentation.*

Response: We fully agree that reproducibility is essential. The implementation details like the network architecture and the hyperparameters are reported in Section 2.4 of the manuscript, and the complete code will be made publicly available in a documented repository upon acceptance, as stated in the Code and Data Availability section. The repository contains the full workflow documentation, while Figure 1 provides a high-level overview of the same workflow. Together, the manuscript and the repository will allow the analysis to be reproduced.

Comment 11: *I would remove challenges in slope unit segmentation from the discussion as that is not the main focus of your manuscript but impact of inventory incompleteness and structure the manuscript in that line with a one clear message which has better scientific impact.*

Response: We thank the reviewer for this suggestion, and we agree that the manuscript should convey one clear message centered on the role of inventory completeness. However, we would prefer to retain the discussion of slope-unit segmentation because it is directly relevant to our results, rather than being a separate topic. The segmentation of slope units propagates through the entire analysis as it determines how the features are aggregated and extracted for each unit, and it decides which units are labeled as positive or negative. Both of these directly influence the resulting susceptibility estimates, and we therefore consider it important to acknowledge this source of uncertainty.
To address the reviewer's concern about focus, we will reframe this discussion to highlight the importance of slope-unit segmentation and its influence on the quality of the results, so that it reinforces the manuscript's central message rather than detracting from it.

References

Lima, P., Steger, S., Glade, T., and Murillo-García, F. G.: Literature review and bibliometric analysis on data-driven assessment of landslide susceptibility, Journal of Mountain Science, 19, 1670–1698, https://doi.org/10.1007/s11629-021-7254-9, 2022.