Response to Community Comment #1

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

Dear Editor, dear Dr. Oliver Wigmore,

We appreciate the time and effort spend by the community comment, Dr. Oliver Wigmore, and the feedback on our work. The comments have helped us strengthen both the analysis and its presentation. In response, we are adding a spatial cross-validation to better reflect spatial generalization, we clarify the effective size of the feature set and the rationale for the model, and we frame the SHAP results more cautiously alongside a correlation analysis. Our point-by-point replies follow below and the corresponding changes will be incorporated into the revised manuscript.

Sincerely,

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

CC1: 'Comment on egusphere-2026-1647', Oliver Wigmore

Comment 1: *This is an interesting and ambitious study that addresses an important applied problem. The slope unit framework is well implemented, and the dataset is impressively large. However, I believe there are several important methodological issues that warrant further consideration.*
1) Performance metrics should be interpreted carefully
The ROC-AUC of 0.953 and PR-AUC of 0.844 are calculated on a random holdout, which does not account for the highly uneven spatial distribution of the inventory. Because positive cases are heavily concentrated in a few regions, the test set inevitably over-represents the same terrain already seen in training. These metrics therefore reflect within-inventory discrimination rather than generalisation across Bavaria, and consequently could be optimistic. A blocked spatial cross validation (e.g. holding out entire districts or grids) would be more meaningful. The "out-of-distribution" validation in Section 3.4 is effectively the only real test of generalisation, and the authors highlight poor performance here.

Response: Thank you for your comment. We agree that the random hold-out is optimistic, and we have carried out a spatial leave-one-region-out cross-validation over the seven Regierungsbezirke using the same model. As expected, performance decreases and the pooled ROC AUC drops from 0.953 to 0.93 as well as the PR AUC from 0.53 to 0.37. This decrease varies largely between regions, as most transfer well (ROC AUC 0.91–0.95), whereas Niederbayern clearly does not (ROC AUC 0.80). As this point closely overlaps comment 2 of Reviewer 3, we present the full analysis there. We will also add the results in the revised manuscript.

Comment 2: ***2) Feature set size and model complexity***
84 predictors in a six-layer ANN is a lot with only 8,000 positive cases (2.6% of 300,000 slope units). Many of these predictors are highly intercorrelated (especially the topographic variables) and some are potentially redundant. While ANN are generally less sensitive to multicollinearity it still increases dimensionality of the model and can make generalisation/transferability harder to demonstrate without sensitivity tests. It is unclear how much of the poor "out-of-distribution" performance reflects inventory bias versus model overspecification. The discussion attributes the generalisation failure primarily to inventory incompleteness, but the model architecture may also be contributing. A feature reduction or ablation study would help to disentangle these.

Response: We thank the commenter for this thoughtful point, and we agree that dimensionality relative to the number of positive cases deserves attention.
The figure of 84 features overstates the model's effective dimensionality. About half of these inputs (45 features) are the percentage encodings of only five categorical variables, which we had already aggregated into coarse classes (for example, six geological-unit classes instead of the original 33). The number of

conceptually independent features is therefore considerably smaller than 84. Most of the remaining inputs are continuous DEM derivatives that are correlated because they describe closely related aspects of terrain. This redundancy is inherent to such variables rather than a modeling choice. As the commenter notes, the ANN is comparatively robust to this multicollinearity, and we additionally regularize the network with dropout (rate 0.45) to limit the influence of the added dimensions.

Comment 3: ***3) Negative sampling strategy***
Slope units without a recorded landslide are treated as negatives across all of Bavaria, regardless of whether those areas have ever been systematically surveyed. For an incomplete inventory this is a recognised issue, because it converts "unknowns" into negative class labels, which the model then learns. The resulting underestimation in districts such as Erding and Landshut (on the updated inventory) is consistent with this. The authors identify and acknowledge this issue, but it would strengthen the practical value of the Bavaria-wide map to treat it as a design constraint rather than only a post-hoc explanation. For example by running a sensitivity experiment that restricts negative sampling (or even training) to districts with completed systematic surveys, and treating predictions elsewhere explicitly as extrapolation. This should be paired with an assessment of covariate shift, i.e., whether predictor distributions in application regions fall within the ranges represented in the training data.

Response: We agree that, for an incomplete inventory, treating unsurveyed slope units as negatives effectively converts "unknowns" into negative labels, and we acknowledge this as a genuine limitation of the model.
The proposed sensitivity experiment, however, requires information we do not have, which is a reliable delineation of the districts that have been systematically and completely surveyed. Our inventory is a compilation that does not carry survey-completeness metadata, and completeness cannot be inferred reliably from the data, because a low number of recorded landslides can have two reasons. It may reflect either an under-surveyed area or a genuinely stable one, and the two cannot be distinguished. The inventory update illustrates this: although the newly recorded landslides are concentrated mainly in a few districts (e.g. Erding and Landshut), isolated new detections also appear elsewhere (for instance between the Rhön and the Franconian Jura), and not all these new entries are recent events. There is no defensible basis on which to split Bavaria into "complete" and "incomplete" survey regions, and doing so on an arbitrary criterion could introduce more bias than it removes. For this reason, due to limitations in the available data, we cannot test the proposed experiment, despite acknowledging its relevance.

Comment 4: ***4) Reliability of the SHAP analysis***
With 84 intercorrelated predictors, global SHAP rankings can be difficult to interpret. SHAP importance can be shared among correlated variables and rankings can be sensitive to the correlation structure, not only to genuine predictive signal. The finding that geological and soil variables outrank topographic ones is interesting, but may partly reflect that categorical variables are less intercorrelated amongst themselves than the many DEM-derived metrics. Consequently they receive a high SHAP score, while the DEM derived variables effectively split their SHAP ranking between themselves, appearing less important. A correlation analysis of the feature set and a more cautious framing of the SHAP results as exploratory rather than definitive would considerably strengthen this contribution. Specific mention of how SHAP is affected by multicollinearity and its limitations should also be included for clarity. The use of Group SHAP could also be explored, or at a basic level, simply summing SHAP scores across the variable groups (i.e. sum of SHAP for all DEM derived variables). This article may be useful here: https://doi.org/10.1002/aisy.202400304

Response: We thank the valuable and well-justified comment, with which we agree.
We already stated that the SHAP results should be read as an indication of which features are most informative for the trained model, not as definitive or causal relationships, and we will keep this exploratory framing explicit. We will also add a note on how multicollinearity affects SHAP. When features are strongly correlated, the attributed importance is shared among them, so that several correlated features can each appear less important than a single, less-correlated variable carrying comparable information.
This is the case for part of our feature set: several of the DEM-derived terrain metrics describe closely related aspects of the terrain (e.g. slope, ruggedness and downslope distance gradient) and are therefore correlated. These features were included deliberately, since each carries geomorphologically meaningful information; their correlation simply means that their joint importance is shared between them in the SHAP

ranking, so that individually they appear less prominent. The categorical geological and soil features are comparatively distinct from one another and therefore receive more concentrated SHAP importance. This helps explain why they rank high, in line with the suggested interpretation, and we will add this clarification to Section 3.2.

To make the comparison between thematic groups more transparent, we will additionally report grouped SHAP values in the results, by summing the SHAP values within each feature group (e.g. all DEM-derived variables). We also thank for the suggested reference, which we will cite in this context.

Comment 5: ***Final thoughts***

These comments are not intended to diminish what is clearly a carefully executed, well-written and ambitious piece of work, and I recognise the substantial effort that a study of this scale represents. However, the concerns raised above interact with each other in ways that compound their individual effects, and collectively they affect how confidently the results and interpretations can be presented. That said, they are addressable, and the underlying dataset and slope unit framework remain a genuinely valuable contribution. I hope these comments are useful to the authors as they consider potential revisions.

Response: We sincerely appreciate these encouraging closing remarks and for the careful and constructive reading of our manuscript. We agree that the points raised are addressable, and we have aimed to handle them thoroughly in our point-by-point responses above. We believe the resulting revisions have improved the clarity and robustness of the manuscript. We are grateful for the time and thought invested in the comment.