Response to Reviewer #2

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

Dear Editor and Reviewer,

We would like to thank the reviewer for the careful and constructive evaluation of our manuscript. We are pleased that the research question, methodology, and results section were found to be sound. We took the technical suggestions seriously, as we believe they add meaningful value to the work.

We hope the revisions are satisfactory.

Sincerely,

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

RC2: 'Comment on egusphere-2026-1647', Anonymous Referee #2

Comment 1: *Dear authors and editor,*
the article is well written, the research question is clear, methodology appropriate, results are well presented and the discussion is thorough. The manuscript is fit for publication after minor changes as there are no significant flaws in the presented research but rather "food for thought" questions and/or suggestions listed below which can help increase the manuscript quality.
I have a few technical suggestions:

- *the figures which represent maps should have a north arrow*
- *I failed to see study area size i.e. in terms how many square kilometres*
- *maybe it would be nice to state which seven land cover classes are used as "land cover" variable*
- *it would be beneficial to add a geological map, e.g. alongside figure 2: make it 2b and keep the current figure 2 as 2a*

Response: We agree that the suggestions will improve the clarity of the study area and data used, thus we will add the corresponding information in the revised manuscript. North arrows will be added to the Figures 2, 5, and 9. The information on study area size of around $70,550\,\mathrm{km}^2$ will be included in Section 2.1 Study area. The enumeration of the land cover classes is added in Section 2.2.2. As also suggested by reviewer 1, the geological map will be included in Figure 2 (b) with the aggregated categories used in our study. The original 33 categories can be consulted in the Umweltatlas.

Comment 2: *if possible, try to quantify information about landslide inventory which is key in this article. E.g. the data in lines 120-123. Were there some mappings done via HR LiDAR derivatives, what amount of the inventory is field surveyed and verified, time spans etc. For an article focusing on the inventory, some more numbers should be present in my opinion, considering >11 000 landslides*

Response: Thank you very much for this suggestion. We will add information on the timespan and survey detail level of the LfU inventory to Section 2.2.1. In the revised version of the manuscript, we include the following sentences marked in red:
The two inventories differ in their compilation strategies and data sources. The LfU GEORISK inventory is maintained by the Bavarian Environment Agency and combines historical records, verified civilian reports, and results from systematic site inspections conducted by experts. These inspections apply various geodetic and geotechnical measurement techniques, such as extensometers, inclinometers, repeated geodetic surveys, and terrain change detection using laser scanning or photogrammetry. The inventory spans several centuries, with 301 events before 1800, 13 events between 1800 and 1900, and 507 events between

1901 and 2000, with the remainder occurred after 2000. Regarding the level of survey detail, 87.5 % of the recorded events were captured through overview field inspection, 9.9 % have only basic data recorded, and 2.5 % were subject to detailed field investigation. Historically, mapping activities focused on areas with higher susceptibility, particularly the Alps and the Franconian Jura, resulting in spatial differences in inspection intensity across Bavaria.

Comment 3: *A few questions for possible discussion, consider adding to the manuscript some of it (and reply in the discussion to the rest of them):*
you mentioned 84 input features (LCFs) for the modelling. It seems rather uncommonly much – what was the motivation for this and were there any collinearity tests done preliminary or during the modelling

Response: The motivation for using 84 input features was intentional. In data-driven studies, it is common practice to provide as much data as possible representing the study area and letting the model decide which features are the most important or relevant. It is, of course, important to evaluate these features later and check the plausibility of the results. This enables the model to identify features that may be relevant in Bavaria but not in other places, helping to create a more locally adjusted map. Rather than pre-selecting a reduced feature set based on prior assumptions, we used a broad set of landslide-influencing factors to let the SHAP analysis empirically identify which variables contribute to model performance in Bavaria. As stated in the manuscript in Section 2.5, this approach avoids the risk of overemphasizing certain features while obscuring the relevance of others, which can occur when feature selection is performed beforehand.
Regarding collinearity tests: These checks are most relevant for linear and logistic regression models, where correlated predictors can distort coefficient estimates. Neural networks handle this differently. Rather than estimating a single coefficient per feature, they distribute information across many weights, which means correlated inputs are naturally absorbed into the learned representations without destabilizing the model (Bishop, 1995; Goodfellow et al., 2016). However, highly correlated features may still affect training efficiency and interpretability. Regularization in the ANN through dropout of, in our models case, 45 % of the connections at each step in the training can also mitigate any potential negative effects of correlated inputs on model generalization. We will add a clarification in the revised manuscript that feature correlation is not a major issue in ANNs.

Comment 4: *how come you decided "only" for ANN method? were there some preliminary or similar cases done which made you prioritize "only" this method? maybe another (quite different) method would show differently some results which you presented*

Response: The ANN was chosen primarily because of the nature of our dataset: 84 input features with complex nonlinear interactions, a large training set (ca. 220,000 slope units), and a heavily imbalanced class distribution. As stated in lines 68-72, a key advantage of ANNs is that they make no assumptions about the functional form of relationships between inputs and the predicted outcome (Fischer, 2006), which makes them well-suited for susceptibility modeling where these relationships are hard to specify in advance. This is supported by the SHAP scatterplots in Section 3.2, which show nonlinear relationships that simpler models like logistic regression would struggle to capture.
Regarding alternative and different methods: Support Vector Machines (SVMs) are a natural alternative, because it is widely used in landslide susceptibility mapping (Lima et al., 2022; Merghadi et al., 2020), but have practical limitations for our specific application. Kernel SVMs scale quadratically to cubically with training set size, which makes them computationally impractical for our study. They also require a pre-specified kernel function, which assumes a particular structure in the feature relationships. This seems hard to justify given the nonlinearities in our SHAP results. Another method that is common in landslide susceptibility analysis is logistic regression (Lima et al., 2022; Reichenbach et al., 2018), which also face the problem of assuming a linear relationship between the input and the logits. In addition, it is less computationally efficient than ANNs and have problems with multicollinearity when calculated with many input features. ANNs, in contrast, learn relationships directly from the data without functional assumptions (Fischer, 2006; Goodfellow et al., 2016), and have been shown to be particularly well-suited for problems where input features exhibit strong nonlinear interactions (Merghadi et al., 2020).
We would like to add that despite comparative studies being interesting themselves to find model uncertainties, since it is common to have different performances and results, this study is focused on the impact of landslide inventories in the model results, rather than comparing several models. We agree that comparing

different modeling approaches, and even creating an ensembled map to account for model uncertainties (like it was done in a previous study for flood susceptibility (Montien Tique et al., 2026)) is interesting itself, but out of the scope. Still, we will extend the justification of the selected modeling method in the revised manuscript.

Comment 5: *how were the exemplary slope units for figure 8 selected?*

Response: The two slope units in Figure 8 were randomly selected from the predictions, with the constraint that one had high and one low predicted susceptibility, to illustrate SHAP force plots at opposite ends of the susceptibility range. A short explanation on this will be added in the corresponding Section 3.2 Model explainability.

Comment 6: *did you consider taking LfU polygons as a LCF? (i.e. categorical LCF: present deep, new deep etc, presented in figure 9 left)*

Response: We did not consider incorporating the *LfU* susceptibility polygons as an input feature for two main reasons. First, the *LfU* maps cover only $51\,\%$ of Bavaria, so including them as a feature would introduce a large spatial gap in the input data, effectively suggesting stability in unmapped areas. This again relates to the problem of interpreting absence of landslide susceptibility polygons as stability of the slopes, like discussed in Section 4.1. Second, and more fundamentally, the *LfU* maps are closely related to the ground truth used to train and evaluate the model, as both are derived from the same *LfU* landslide inventory. Using them as an input would therefore risk predicting in a loop, meaning the model is trained on information derived from the same source as its labels. For these reasons, we chose to use the *LfU* maps only as an independent comparison and validation source.

Comment 7: *could your model point out the locations where more inventory adjustment/corrections/upgrades are needed? can you elaborate on the uncertainty aspect of the modelling and why it wasn't quantified*

Response: This is a good point, but unfortunately our model and data are not well-suited to identify where inventory updates are most needed. The core problem is that slope units without recorded landslides were treated as stable (negative) during training, so the model learned to predict low susceptibility in areas where there were no reported landslides, i.e. under-sampled regions in the landslide inventory. Therefore, low predicted values show areas with a low probability of being landslide-prone, but also underrepresented areas. To disentangle areas with stable slopes from inventory incompleteness, one would need additional information that is currently unavailable, such as an explicit map of survey coverage or reliable absence data confirming stable conditions. The spatial distribution of *LfU* fieldwork effort provides some indirect guidance on where the inventory is more or less complete, but this is not available in a reliable form that can be directly incorporated into the model. We agree that explicit uncertainty quantification would be a valuable addition here, and is stated as a direction for future work in Section 4.1.

Comment 8: *it would be interesting to see spatial K-fold validation in this type of modelling, if you could elaborate on the topic*

Response: We thank the reviewer for this suggestion. Spatial k-fold cross-validation would indeed have been a valuable addition, as it gives a better picture of how well the model generalizes across different areas, compared to a standard random train-test split. In our case, the evaluation against the updated landslide inventory in Section 3.4 partly serves this purpose, as it tests the model on landslides from previously underrepresented regions and reveals its generalization limitations.
Spatial cross-validation could potentially reduce some of the generalization problems observed in our results by encouraging the model to learn more transferable and spatially robust features instead of region-specific patterns. Training and validating the model on spatially separated subsets may therefore improve predictive performance in areas of Bavaria that are poorly represented in the inventory. We have added this aspect as a recommended direction for future work in Section 4.1.

Comment 9: *you mentioned that stable slope units are in fact unknown, i.e. possibly stable or unstable which is correct (not verified as negatives). If you could declare some slope units as 100% stable, would you prioritize them in the modelling? was this done in some research before and if yes what were the reported results?*

Response: In our study, we partially address this by removing slope units that can be considered inherently stable with high confidence, specifically units dominated by water bodies or artificial surfaces ($> 90\,\%$ coverage) and very flat units (slope mean < 1 and standard deviation < 0.5), as described in Section 2.3.2. Beyond these cases, definitively classifying a slope unit as stable without field verification by experts is rarely possible.
Even if reliable stable slope information were available, we would be cautious about increasing their weight during training. The dataset already contains roughly $97\,\%$ negative observations, and further emphasizing the negative class would push the model toward predicting stability more aggressively. This reduces sensitivity, which is a highly relevant performance criterion in hazard mapping, where missing a susceptible area carries far greater consequences than a false positive.
Some studies have attempted to identify stable slopes using geomorphological criteria and use them as verified negatives in training (Zhu et al., 2018; Wang et al., 2024; Yuzhong et al., 2025). However, results from Wang et al. (2024) suggest that model performance is relatively insensitive to the ratio of positive to negative samples, with AUC remaining largely stable across ratios ranging from 1:1 to 1:10. This indicates that improving the spatial completeness of the landslide inventory is likely to have a greater impact on model performance than refining the weighting of negative samples.

Comment 10: *your main conclusion is stated in lines 615-618, can your research propose some novelty to this topic? if so, specify some ideas for future work and perspectives to mitigate this issue with the "unseen locations"*

Response: Thank you for this remark. Future perspectives were already discussed in the discussion section, but we agree that the conclusion benefits from a more explicit summary of how the generalizability limitations from spatially incomplete inventories could be addressed in future work. We will therefore add the following paragraph at line 622:
"Addressing these generalizability problems represents an important direction for future work. A promising addition is the application of stratified sampling during model training, either by balancing samples across geographic regions or by stratifying along feature gradients such as slope, elevation, or geology. This helps to reduce the model's tendency to learn associations specific to well-surveyed terrain types. In addition, spatial cross-validation could serve as a diagnostic tool to systematically identify which regions or feature combinations generalize poorly."

References

Bishop, C. M.: Neural Networks for Pattern Recognition, Oxford University Press, Inc., USA, ISBN 0198538642, https://doi.org/10.1093/oso/9780198538493.001.0001, 1995.

Fischer, M. M.: Neural Networks: A General Framework for Non-Linear Function Approximation, Transactions in GIS, 10, 521–533, https://doi.org/https://doi.org/10.1111/j.1467-9671.2006.01010.x, 2006.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, http://www.deeplearningbook.org, 2016.

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.

Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., Avtar, R., and Abderrahmane, B.: Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance, Earth-Science Reviews, 207, 103 225, https://doi.org/10.1016/j.earscirev.2020.103225, 2020.

Montien Tique, W. F., Sapena, M., Weigand, M., Groth, S., Geiß, C., and Taubenböck, H.: A comparative assessment of data-driven flood susceptibility mapping in Nigeria, Natural Hazards, 122, 139, 2026.

Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M., and Guzzetti, F.: A review of statistically-based landslide susceptibility models, Earth-Science Reviews, 180, 60–91, https://doi.org/10.1016/j.earscirev.2018.03.001, 2018.

Wang, Y., Wang, L., Zhang, W., Liu, S., Sun, W., Hong, L., and Zhu, Z.: A physics-informed machine learning solution for landslide susceptibility mapping based on three-dimensional slope stability evaluation, Journal of Central South University, 31, 3838–3853, https://doi.org/10.1007/s11771-024-5687-3, 2024.

Yuzhong, K., Hua, W., Chong, X., Jingjing, S., Kangcheng, Z., Chenguang, Z., Jianwei, Z., Tong, X., Taijin, S., Zelin, Z., and Hui, K.: Landslide susceptibility mapping using an entropy index-based negative sample selection strategy: A case study of Luolong county, PLOS ONE, 20, 1–27, https://doi.org/10.1371/journal.pone.0322566, 2025.

Zhu, A.-X., Miao, Y., Yang, L., Bai, S., Liu, J., and Hong, H.: Comparison of the presence-only method and presence-absence method in landslide susceptibility mapping, CATENA, 171, 222–233, https://doi.org/https://doi.org/10.1016/j.catena.2018.07.012, 2018.