Answers to Reviewer RC1: Thank you for your interest in the paper and the valuable feedback. We considered all your comments (in blue), gave our answers (in black) alongside the revised text from the manuscript (in green).

This study presents an analysis of how certain variables may limit the performance of shallow landslide susceptibility modelling in Switzerland. I find the topic relevant and promising; however, improvements are necessary, particularly in the introduction and methodology sections, which require a more detailed description. Additional comments and suggestions are outlined below:

1. Lines 44-48: The argument is weak. I do not agree with this statement. Landslides do not always occur in unexpected locations. In most cases, susceptible areas are already known, but other factors and competing interests often take precedence, allowing these areas to be occupied.

   We agree with the reviewer that our argument was misleading. Susceptible areas are often well known, but rather the precise location of where the landslides may initiate within susceptible areas remains difficult to predict. We therefore corrected the sentence to the simple argument "that despite improvements, discrepancies between predicted susceptibility and observed landslide occurrence remain"; which in our opinion applies generally to landslide susceptibility studies. The sentence further introduces the necessity to study limitations in detail (which is our aim with this study).

   "Despite substantial progress in improving the quality of susceptibility predictions across various fields of research, discrepancies between predicted susceptibility and observed landslide occurrence remain, highlighting the need to examine the common limitations of landslide susceptibility models."

2. Lines 62-64: The argument is weak. The authors mention the availability of "large landslide inventories", but this is not supported or properly referenced. In which regions or studies is this the case? In my view, this is not representative of most countries and should be revised and properly justified with references.

   We revised this argument and specifically refer to the great improvement in spatial geodata (what the entire paragraph is about) and the development of novel tools (machine learning algorithms) to process the large amount of data available. The argument than concludes that despite these improvements, landslide model can still not perfectly predict landslide locations. Which refers to the previous argument that even though susceptible areas are often known, the precise locations are difficult to predict.

   "Despite improvements in model performance resulting from the increasing availability of spatial geodata and advances in large-scale data processing, even state-of-the-art machine learning algorithms (Merghadi et al., 2020; Ado et al., 2022) cannot perfectly predict the locations of landslide occurrences."

3. Introduction: Poorly developed with week arguments and general phrases, needs to be rewritten.

   We revised the introduction to give more credit to research within the field of landslide susceptibility modelling. It was not our aim to criticise previous work, but rather to highlight that despite great improvement, models can still not perfectly predict landslide susceptibility. Finding the causes constraining models, therefore, is relevant to the scientific community in order to provide practitioners with even better solutions to establish accurate susceptibility maps.

It would be helpful to provide more background information about the study area. The manuscript barely discusses Switzerland, yet it is important to explain why this country was selected, its historical context regarding landslides, and the current state of landslide inventory mapping and management. This additional context would help readers better understand the relevance and applicability of the study.

To provide the readers with background information about the study areas, we included following paragraph introducing the inventories (section 3.1):

"In Switzerland's mountainous terrain, shallow landslides can occur throughout the entire country (Halter et al., 2025a). Landslides have been reported by individual cantonal authorities with the earliest documentation dating back to 1739. The Federal Office for the Environment (FOEN) has combined these records in a national dataset, which provides valuable information on the temporal occurrences of landslides (FOEN, 2024). However, the locations of these events were often not mapped accurately, making it difficult to integrate the data into landslide susceptibility models. To better understand the processes and conditions leading to shallow landslide triggering, the FOEN mandated the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) to document and analyse all landslides that occurred within certain perimeters during individual rainfall events. In total, landslides from eight rainfall events between 1997 and 2012 were documented and combined into the SLD (Rickli et al., 2016). Independently, the Cantonal Office of Bern compiled an inventory comprising historical landslide records and landslides visually identified from orthophotos, forming a basis for the development of a cantonal landslide susceptibility model (Hählen 2024)."

Lines 188-89: Please provide the version of the software used.

We provided the version of software used in the main text:

[...], variables were derived using SAGA v2.1.4 (Conrad et al., 2015), GDAL v3.12.1 (GDAL/OGR contributors, 2025), QGIS v3.40.11 (Dawson et al., 2025) and GRASS v8.4 (GRASS Development Team, 2024) tools within QGIS (Table 1).

1. Section 3: Did the authors assess the correlation among the predictor variables? If so, please describe the methodology used and report the results. If not, I recommend including a correlation or multicollinearity analysis to ensure that highly correlated variables do not affect the interpretation of the results.

   Yes, we assessed the correlation of predictor variables used. Pearson Correlation Coefficients between predictor variables are added in the supplementary material and their potential influence on the model is now discussed within the section 3.2:

   Pearson correlation coefficients (Benesty et al., 2009) were computed for the training area (Canton Bern) to assess correlations among the numeric explanatory variables (Supplementary Material Fig. S1). Although some variables showed significant correlation — particularly among the climatic and topographic variables — Random Forest's predictive performance is generally robust to correlated predictors (Genuer et al., 2010). However, we note, that correlated predictors can bias variable importance rankings, typically overestimating the permutation importance attributed to correlated variables (Strobl et al., 2008).

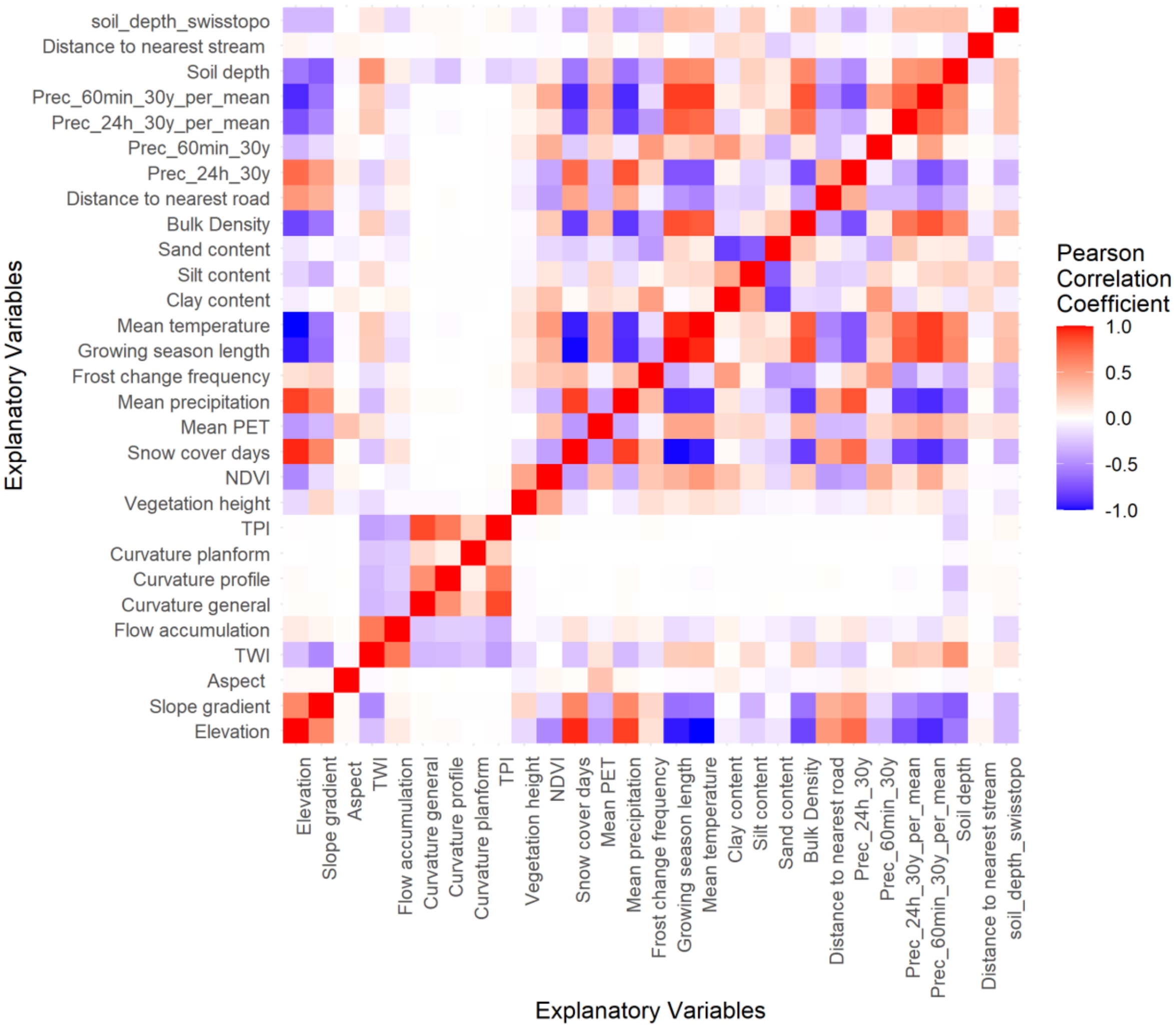

Figure S1. Pearson Correlation Coefficients between the explanatory variables of the Canton Bern region used to train the Random Forest model

2. Table 1: Please provide the scale/resolution of all thematic variables used.

We provided for all explanatory variables the resolution used in Table 1. Furthermore, we state in section 3.4 that all data was resampled to 10 m raster grid:

A raster grid with a spatial resolution of 10 m × 10 m was applied across the entire training and test study areas (Canton Bern and SLD study area perimeters), to which all explanatory variables were resampled.

The manuscript uses predictor variables with spatial resolutions ranging from 2 m to 1000 m. This considerable difference in spatial resolution raises concerns about the consistency of the input data and its potential impact on the model's performance. The authors should clarify how these datasets were harmonized prior to the analysis (e.g., resampling method and target resolution) and discuss the possible implications of combining variables at such different spatial scales.

We agree that using predictor variables with different spatial resolution as well as the method used for harmonization may impact on the model outcome. For clarification, we revised the methodological section 3.4 accordingly:

"Variables with finer resolution were resampled to 10 m using bilinear interpolation, whereas variables with lower resolution were disaggregated to a 10 m resolution (without interpolation). The influence of harmonizing spatial variables with largely different original resolution (Table 1) and the resampling methods used were not investigated in this study. Explanatory variables with a coarse native resolution (e.g., extreme precipitation statistics, available at a 1000 m resolution; Table 1) should be interpreted at the scale of their original data rather than at the downscaled 10 m resolution used for model input. Variables with such coarse spatial resolution are inherently unable to capture the fine-scale spatial heterogeneity that influences shallow landslide initiation and instead provide information on broader regional trends relevant to susceptibility modelling."

1. Table 2: There appears to be an inconsistency between the title and the landslide classification used in the manuscript. While the title refers to "shallow landslides," the inventory includes different movement types (flows, slides, rotational and translational landslides). This is somewhat confusing, as it is not clear whether the study focuses exclusively on shallow landslides or on a broader set of mass movement processes. The authors should clarify the definition of "shallow landslide" used in this context and justify the inclusion of the different movement types in the inventory.

   From our understanding the term "shallow landslides" refers both to translational and rotational, as well as movement processes described as flow and slide. To clarify what we refer to, the very first sentence in the revised manuscript introduces shallow landslides as follows:

   "Shallow landslides, defined here as translational or rotational slope failures with failure surfaces at depths of less than 2 m, typically evolve into either slide or flow movements (Hungr et al., 2014)."

2. Section 5: More discussion with other studies developed in Switzerland are required and/or similar Alpine environments.

   Thank you for your suggestion, a detailed comparison to two landslide susceptibility studies from Switzerland was added into the discussion (section 5.3 Transferability limitations):
   The primary cause of reduced test performance of the test compared to the training (AUC = 0.74, compared to 0.89) is related to limited model transferability to distinct geographic regions of Switzerland with contrasting environmental

conditions. Von Ruette et al. (2011) analysed model transferability using a subset of the SLD by training a logistic regression susceptibility model on the Napf_05 perimeter and applying it to the independent perimeters Napf_02 (AUC = 0.72), Entlebuch_05 (AUC = 0.69), and Appenzell_02 (AUC = 0.82). These performance values are comparable to those obtained in this study for the independent model using a Random Forest model trained on the Bern inventory (AUC = 0.75, 0.70, and 0.76, respectively).
[…]
This result is consistent with the findings of Zweifel et al. (2021), who developed landslide susceptibility models for ten study areas across Switzerland and obtained an overall AUC of 0.79 when evaluating model performance across all regions. They further reported that regional-scale explanatory variables such as climate-related datasets contributed little to improving model performance.

1 The validation strategy appears limited, as it relies solely on AUC to evaluate the different susceptibility scenarios. While AUC is a useful measure of overall model performance, it does not provide information about the spatial agreement between predicted susceptibility and observed landslides. I recommend including at least one spatially explicit evaluation metric to better assess model performance in a spatial context, especially given that different predictor sets and inventory configurations are compared.

Evaluating the performance of landslide susceptibility maps remains a challenging task (e.g. Guzzetti et al., 2006). Raster-based susceptibility maps, such as those presented in this study, typically exhibit complex spatial patterns, with areas of high and low predicted susceptibility occurring in close proximity. Consequently, directly quantifying the spatial agreement between observed landslides and individual raster cells with high predicted susceptibility is not straightforward, as a single landslide may correspond to multiple adjacent cells predicted as highly susceptible. One possible approach would be to aggregate neighbouring cells into susceptibility zones and assess their spatial correspondence with observed landslides. However, developing and validating such a framework would require a dedicated methodological study and is beyond the scope of the present work.
Moreover, the *spatial agreement* metrics such as Cohen's kappa are commonly used in landslide research to compare the spatial similarity between different susceptibility maps rather than to quantify the geographic correspondence between predicted susceptibility and observed landslide locations (e.g. Adnan et al., 2020; Sterlacchini et al., 2011). Although this study produced two susceptibility maps (the independent and dependent models), our objective was not to compare their spatial patterns, but rather to investigate how different sets of explanatory variables influence model performance.

Nevertheless, we agree with the reviewer that relying solely on the Area Under the Receiver Operating Characteristic Curve (AUC) provides only one perspective on model performance. We have therefore included the True Skill Statistic (TSS) as an additional evaluation metric. Unlike AUC, TSS evaluates classification performance at a selected threshold (here defined as the optimal threshold in the training set) by combining sensitivity and specificity (TSS = sensitivity – specificity). As a result, TSS accounts equally for the true positive and true negative rates and is well suited for highly imbalanced datasets, such as landslide inventories where non-landslide locations greatly outnumber landslide locations.

In the manuscript we introduced TSS alongside the AUC in method section 3.4:

Model performance was evaluated using the Area Under the receiver operating characteristic Curve (AUC) and the True Skill Statistic (TSS). The AUC provides a threshold-independent measure of the model's ability to discriminate between landslide and non-landslide locations across all possible probability thresholds. AUC values range from 0.5, indicating random discrimination, to 1, representing perfect discrimination. As one of the most widely used evaluation metrics for landslide susceptibility models, reported AUC values typically range considerably from approximately 0.7 to values close to 1 (Ado et al., 2022; Frattini et al., 2010). The TSS score provides a complementary, threshold-dependent assessment of model performance and is calculated as the sum of sensitivity and specificity minus one. It ranges from 0, indicating performance no better than random, to 1, indicating perfect classification. The probability threshold used to compute the TSS was determined from the training dataset as the value that maximized the TSS. This optimal threshold was subsequently applied to the validation and test datasets to evaluate model performance."

To further clarify that AUC and TSS cannot assess spatial agreement, we added following sentence in the discussion (section 5.1):

"While AUC and TSS quantify the model's ability to discriminate between landslide and non-landslide locations, they cannot assess the spatial agreement between the spatial pattern of predicted susceptibility and observed landslide locations (Sterlacchini et al., 2011). Consequently, susceptibility maps with similar AUC or TSS values may differ substantially in the spatial distribution of predicted susceptible areas (Vakhshoori and Zare, 2018)."

Bibliography

Adnan, M. S. G., Rahman, M. S., Ahmed, N., Ahmed, B., Rabbi, M. F., & Rahman, R. M. (2020). Improving Spatial Agreement in Machine Learning-Based Landslide Susceptibility Mapping. *Remote Sensing*, *12*(20). https://doi.org/10.3390/rs12203347

Ado, M., Amitab, K., Maji, A. K., Jasińska, E., Gono, R., Leonowicz, Z., & Jasiński, M. (2022). Landslide susceptibility mapping using machine learning: A literature survey. *Remote Sensing*, *14*(13), 3029. https://doi.org/10.3390/rs14133029

FOEN. (2024). *Federal Office for the Environment, StorMe Landslide Inventory*. https://www.bafu.admin.ch/bafu/de/home/themen/naturgefahren/fachinformationen/naturgefahrensituation-und-raumnutzung/gefahrengrundlagen/naturereigniskataster-storme.html

Frattini, P., Crosta, G., & Carrara, A. (2010). Techniques for evaluating the performance of landslide susceptibility models. *Engineering Geology*, *111*(1), 62–72. https://doi.org/10.1016/j.enggeo.2009.12.004

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., & Galli, M. (2006). Estimating the quality of landslide susceptibility models. *Geomorphology*, *81*(1), 166–184. https://doi.org/10.1016/j.geomorph.2006.04.007

Hählen, N. (2024). *Orthofotobasiertes Inventar spontaner, flachgründiger Rutschungen im Kanton Bern* [Dataset]. https://www.researchgate.net/publication/384051894_Orthofotobasiertes_Inventar_spontaner_flachgrundiger_Rutschungen_im_Kanton_Bern

Halter, T., Lehmann, P., Bast, A., Aaron, J., & Stähli, M. (2025). In situ soil moisture data improve precipitation-based shallow landslide early warning through innovative machine learning methods. *Landslides*, *22*(11), 3599–3614. https://doi.org/10.1007/s10346-025-02599-4

Rickli, C., McArdell, B., Badoux, A., & Loup, B. (2016). Database shallow landslides and hillslope debris flows. *Proceedings of the 13th Congress INTERPRAEVENT*, *30*, 242–243. https://hangmuren.wsl.ch/

Sterlacchini, S., Ballabio, C., Blahut, J., Masetti, M., & Sorichetta, A. (2011). Spatial agreement of predicted patterns in landslide susceptibility maps. *Geomorphology*, *125*(1), 51–61. https://doi.org/10.1016/j.geomorph.2010.09.004

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Vakhshoori, V., & Zare, M. (2018). Is the ROC curve a reliable tool to compare the validity of landslide susceptibility maps? *Geomatics, Natural Hazards and Risk*, *9*(1), 249–266. https://doi.org/10.1080/19475705.2018.1424043

von Ruette, J., Papritz, A., Lehmann, P., Rickli, C., & Or, D. (2011). Spatial statistical modeling of shallow landslides—Validating predictions for different landslide inventories and rainfall events. *Geomorphology*, *133*(1), 11–22. https://doi.org/10.1016/j.geomorph.2011.06.010

Zweifel, L., Samarin, M., Meusburger, K., & Alewell, C. (2021). Investigating causal factors of shallow landslides in grassland regions of Switzerland. *Natural Hazards and Earth System Sciences*, *21*(11), 3421–3437. https://doi.org/10.5194/nhess-21-3421-2021