Answers to Reviewer RC2: 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).

Halter et al. explore three components of data-driven statistical models for assessing landslide susceptibility: 1. the effects of incomplete inventories on model development; 2. the transferability of these models; 3. the effects of hidden variables on model outputs. The first two components have already been extensively covered in previous studies, as the authors recognize. The unique aspect of this study is the third component. The authors try to assess the influence of hidden variables generally not included in data-driven statistical models, due to the data being highly heterogeneous and difficult to collect, on susceptibility model outputs. These hidden variables include soil and vegetation characteristics. It is well known that hidden variables can exert a strong control on susceptibility model results. This paper attempts to provide a more detailed understanding of these hidden variables using a database developed from field observations at known landslide locations. I found the manuscript to be well-written and the topic to be of interest to the landslide community. However, I have some serious concerns about the limits of the data used in the analysis. Below, I provide some general comments on the manuscript, followed by more detailed comments.

General Comments:

I have some major concerns about the hidden variable aspect of this study. Most of them are tied to the fact that the authors only have information about the hidden variables at landslide locations, not non-landslide locations. Because this is a data shortage issue, it is not clear to me if it can be overcome to allow this manuscript to be published.

1. The analysis only analyzes the influence of hidden variables at landslide locations and not at non-landslide locations. I understand that this is a limitation of the dataset, but I don't think this gives you enough information to understand the influence of the hidden variables on the susceptibility models. All it can tell you is if your model is underpredicting when a given hidden variable is present. It cannot tell you anything about where the model overpredicted when the same hidden variable is present, which is just as important as assessing underprediction. It also doesn't directly tell you anything about the relative abundance of these hidden variables where you do not have data, which is most of your study area. It could be that the hidden variables with low susceptibility values at landslide locations (e.g., the rendzina soil type) are prolific at non-landslide locations. In which case, we may not want landslides with these hidden variables to be modeled with high susceptibility.

2. Building on the previous concern, there is no indication of the effect of hidden variables on susceptibility model performance. This again relates to the hidden variables only being evaluated at known landslide locations. We still don't know if

including detailed soil observations, or any other of the hidden variables, would have a notable influence on model performance. Put another way, the analysis doesn't tell us anything about the effects of hidden variables on modelled landslide susceptibility. Despite what the manuscript currently says. It only provides information about the variance of *susceptibility values of landslide locations*. These are not the same thing! It is already well understood that data-driven statistical models omit important processes for assessing slope stability. I'm not convinced that this manuscript does anything more than provide more discussion of this fact without any real data to support it.

Many thanks for the thorough review. We recognise and acknowledge the key point raised by the Reviewer, namely that our dataset contains information only for landslide sites (and not for sites without landslides). We agree with the reviewer that with the available data and analysis we cannot draw clear conclusions as to whether an ML model with 'hidden variables' (HV) leads to higher model quality or not. Furthermore, the reviewer correctly identified another limitation of this study: our analysis focuses exclusively on locations where the model underpredicted landslide susceptibility by examining residual model errors. Consequently, it does not consider conditions associated with overprediction and therefore provides only a partial assessment of the factors contributing to model prediction errors. For these reasons, parts of the manuscript and its title ("*How hidden variables limit the performance of shallow landslide susceptibility models")* were misleading; we will therefore change the title to '*Limitations of Shallow Landslide Susceptibility Modelling*' (discussed in subsequent comments).

Despite these limitations, we argue that the analysis of hidden variables provides important insights into the shortcomings of landslide susceptibility models. The concept that omitted or unmeasured variables can lead to systematic deficiencies in model predictions is well established across several research fields, including applied statistics, ecology, and medicine. Rather than focusing solely on the predictive performance of a model, these disciplines analyse systematic prediction errors or residual structure to identify missing environmental controls or unmeasured covariates that are not represented in the predictor set (e.g. Yackulic et al., 2013; Sperrin et al., 2020; Clarke, 2005; Rinella et al., 2020). Although the statistical frameworks differ from our landslide susceptibility modelling analysis, they share the common objective of using systematic discrepancies between model predictions and observations to generate hypotheses about missing processes.

In our context, we found that residual model errors are systematically associated with different categories of hidden variables. In other words, landslides characterized by certain hidden-variable categories are more frequently underpredicted than others. This raises the question of why the model systematically underpredicts susceptibility for these landslides:

1. The fact that we observed significant differences in susceptibility between individual categories of HV within the data suggests that randomness cannot explain the differences. Furthermore, if underpredicted categories of hidden variables systematically appear in specific terrain settings that are correlated with one or more of the spatial predictors included in the model, the Random Forest algorithm is expected to implicitly learn these relationships and improve their representation. As a result, the differences in predicted susceptibility among categories of hidden variables would be expected to diminish.
2. The fact that **systematic differences (model residuals) remain** suggests that the model cannot capture these conditions with the available spatial predictors. Thereby, **HV (or specific categories thereof)** may serve as **proxies** for the conditions or processes controlling hydromechanical effects relevant to slope stability that contribute to the model's systematic underprediction of susceptibility. To explore these controlling effects, we seek for explanation that mechanistically explain how HV relate to hydromechanical processes relevant for landslide initiation. These explanations are purely based on mechanistic considerations; they are thus plausible, but do not constitute definitive proof. Consequently, we highlight the **need for in-depth process-based investigations** (last sentence in the abstract and in the conclusions).

For transparency about the study limitations, we add the following sentences within the discussion section 5.4, clearly highlighting that HV may only serve as proxies for hydro-mechanical processes relevant to landslide initiation:

> "However, we note, that because hidden variables are available exclusively for landslide locations and no corresponding information exists for non-landslide locations, they could not be incorporated into the susceptibility model. As a result, their potential contribution to improving overall model performance could not be quantified. Furthermore, the analysis focuses exclusively on locations where the model underpredicted landslide susceptibility by examining residual model errors. Consequently, it does not account for conditions associated with overprediction and therefore provides only a partial assessment of the factors contributing to model prediction errors. Nevertheless, the non-random distribution of residual model errors suggests that hidden variables may act as proxies for hydromechanical controls that are not captured by the classical spatial predictors commonly used to assess landslide susceptibility. To explore these controlling effects, we seek for explanation that mechanistically explain how hidden variables are related to hydromechanical processes relevant for landslide initiation. These explanations are purely based on mechanistic considerations; they are thus plausible, but do not constitute definitive proof."

In addition, we reworked the method section 3.5 to clarify that the analysis aims to quantify the variance of hidden variables associated with predicted susceptibly. In

accordance with the line-by-line comments by the reviewer, we changed the name of the statistical measure “effect size” to “associated variance”:

> “To assess how hidden variables relate to predicted susceptibility, we computed the “associated variance”, a statistical measure of association between a hidden variable and the model's predicted susceptibility. High associated variance indicated that the model predicts systematically different susceptibility values across the categories of a hidden variable (e.g., higher median susceptibility for landslides occurring in Brown Earth than in Renzina soils). Hidden variables with high associated variance therefore warrant closer examination, as they may reveal controlling effects on landslide initiation that are not captured by the available explanatory variables. Specifically, categories of hidden variables with significantly lower predicted susceptibility that other categories of the same hidden variable may provide insight into the occurrence of false-negative predictions. […]“

As the authors acknowledge in the introduction and background sections of the manuscript, the real novelty of this manuscript is the exploration of the hidden variables. However, about half of the results and discussion are devoted to well-investigated issues with data-driven statistical susceptibility models. I think the paper could cut back on these other sections to better focus on the more novel parts of this study. This only works if the authors can fix the issues I raised with the hidden variable component of the study.

Reviewer 2 focuses almost exclusively on the fourth component (d) of our study in the review and gives little consideration to the remaining discussion points and findings, arguing that the other limitations of susceptibility models (e.g., transferability and inventory incompleteness) are already well established in literature. While it is true that these topics have been extensively studied, to the best of our knowledge, no previous study has comprehensively demonstrated and discussed all four major limitations of susceptibility models using a single, comprehensive dataset. We consider this to be a relevant novelty in the present manuscript. With the article it was our objective to provide a comprehensive assessment of these limitations, supported by our unique database. This aim is explicitly stated on Line 81-82 (“Although many studies have investigated individual sources of model imperfection […], no study has systematically summarized and quantified the effect of these limitations in a single, comprehensive study framework”). Overall, we are therefore convinced that the novelty and scientific value of this publication should not be reduced solely to the hidden-variable component but also lies in its comprehensive discussion of the principal limitations of landslide susceptibility models. To further highlight the paper’s novelty, we suggest renaming the title to ‘*Limitations of Shallow Landslide Susceptibility Modelling*’. Accordingly, we revised the abstract and introduction of the manuscript to focus more broadly about general limitation so landslide susceptibility modelling.

There were several references to soil depth data, but I did not see this data anywhere in the manuscript or supplemental. Please include this data somewhere.

We initially intended to include the following figure illustrating the relationship between soil development and landslide thickness in the main manuscript. However, we removed it because we considered it sufficient to summarize the main finding in one sentence, namely that landslides tend to initiate at greater depths in well-developed than in poorly developed soils. But we agree with the reviewer that the data should be shown, and therefore we have included the figure as Figure S5 in the Supplementary Material.

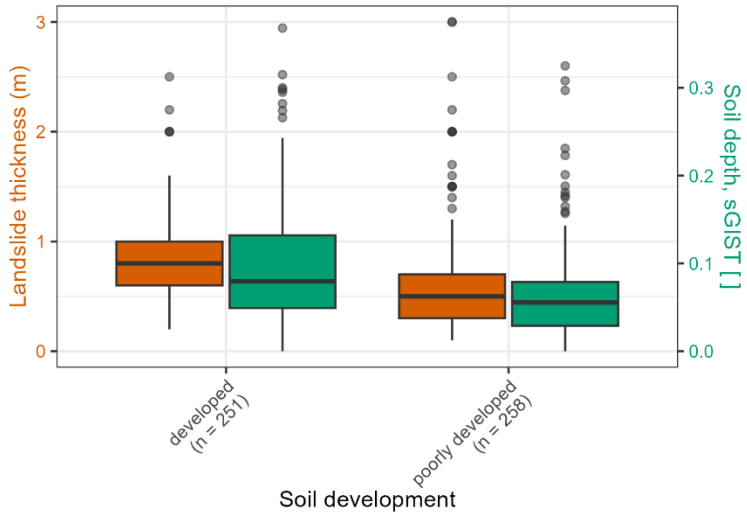

Figure S5. Distribution of landslide thickness (orange boxplots) and soil depth index (green boxplots) modelled with the sGIST model (Catani et al., 2010) for the categories developed and poorly developed soils.

Line-by-lines:

66: "no study has systematically summarized and quantified the effect of these limitations on model performance"- I don't think this is accurate. Lots of studies have quantified the limits of model performance. I think you need to be more specific in what exactly you are referring to.

> To explicitly highlight that we refer to the novelty that a single framework was used to study limitations of shallow landslide models, we rewrote the sentence to: "Although many studies have investigated individual sources of model imperfection (described in detail in the next section), to our knowledge, no study has systematically summarized and quantified these limitations within a single study framework."

73: I think you are limiting this discussion to data-driven statistical landslide susceptibility models rather than other types of models. Please specify.

We agree and specified the sentence: In landslide research, individual sources of data-driven landslide susceptibility model imperfection have been studied in detail [...].

250: Please define KOBO.

KOBO stands for the "competence centre of soils"

257: Please explain how you downscaled this data.

We checked our workflow documentation and realized that we mixed up the terms "downscaling" and "upscaling" (the terms are used differently in image processing and environmental sciences/remote sensing). As the original DEM was at 2 m resolution, we had to **upscale** the DEM to 10 m using bilinear interpolation. We corrected the sentence to:

Topographic inputs for the soil depth mapping were derived from SwissALTI3D DEM (Swisstopo 2025a) which was upscaled from 2 m to 10 m resolution using bilinear interpolation.

262: Please clarify what is meant by 'paths'.

All man-mad -trails were considered. To clarify we add in parathesis: [...] paths (including all types of man-made trails, such as hiking trails) [...]

Table 1: I suggest being consistent with the model names. Choose Model 1 or Independent.

Thank you for your suggestion. We changed the name Model 1 to "independent model" and Model 2 to "dependent model"

285: Please specify how you determined a 'statistically sufficient sample size'.

We defined the numbers to represent statistically sufficient sample size' based on the general rule of using a minimum of 10 events per variable. For clarification we changed the sentence to:

"To ensure statistically sufficient sample sizes, we excluded variables documented for fewer than 100 landslides, as well as individual categories of these variables represented by fewer than 10 landslides. These thresholds were defined based on general recommendations for minimum sample sizes in standard statistical modeling, such as the widely used "10 events per variable" heuristic (Peduzzi et al., 1996).

363: You tuned the same hyperparameters; you did not use the same hyperparameter values. Right? Please clarify your meaning.

We did used the same hyperparameter values in both the dependent and independent model to isolate the effect on model performance using different sets of explanatory variables. To clarify this, we changed the sentence accordingly:

"To isolate the effect of using different sets of explanatory variables, the dependent model was trained using the same hyperparameter values as the independent model."

367: Please clarify your meaning here. Some of the SLD data is outside of the Canton Bern area. And you said that you removed some areas of the Canton that were in the SLD data.

Thank you for pointing this out. We removed the SLD perimeter areas located within the canton of Bern from the training dataset. To clarify this point, we have rewritten the sentence as follows: "Ultimately, both models were separately trained with the optimal selection of explanatory variables based on the Canton Bern training area (excluding SLD perimeters areas located within the canton Bern, section **Error! Reference source not found.**)."

383: What about the occurrence of false positives? These are important too.

We agree that the occurrence of false positives is important too. As we do not have any information about hidden variables from non-landslide location, we cannot study the occurrence of false positives and instead focus on the conditions were the model underpredicted susceptibility. To adequately highlight this as a study limitation we added following sentences in the discussion section:

Furthermore, the analysis focuses exclusively on locations where the model underpredicted landslide susceptibility by examining residual model errors. Consequently, it does not account for conditions associated with overprediction and therefore provides only a partial assessment of the factors contributing to model prediction errors.

384-398: It took me a while to understand how this statistical workflow worked. I'll try and outline what I found confusing with suggestions to help you rewrite this section:

1. I did not understand where the paragraph was going and why you were explaining these things. Provide a brief overview of what this paragraph explains before diving into it.
2. Explaining your workflow in general terms before diving into the details will also help orient the readers. Also, in this general description, explain the motivation for each comparison.
3. More details are needed on the Kruskal-Wallis test. You write it as if the readers should already know what this test is for and how it is used. Most will not.

4. I found myself getting confused on whether you were comparing the variables or the variable categories (i.e., subclasses). You use the eta-squared statistic for the variables and the boxplots of susceptibility values and the Dunn test for the categories.

5. What exactly is the eta-squared measuring because we only have susceptibility of known landslide locations? Not non-landslide locations.

Thank you for your suggestions. We have completely revised this section to provide a clearer description of the statistical workflow. The paragraph now begins with an overview of the general workflow before introducing the statistical methods and their motivation. In accordance with the reviewer's general comments and the overall revision of the manuscript, we replaced the term *"effect size"* with *"associated variance"* to clarify that our analysis does not quantify the effect of the hidden variables (HV) on landslide susceptibility predictions. Instead, associated variance describes the proportion of variance in model-predicted susceptibility associated with the categories of each hidden variable. In addition, we expanded the description of the Kruskal–Wallis test, as most readers will not be familiar with the method:

"To assess how hidden variables relate to predicted susceptibility, we computed the "associated variance", a statistical measure of association between a hidden variable and the model's predicted susceptibility. High associated variance indicated that the model predicted systematically different susceptibility values for the individual categories of a hidden variable (e.g., higher median susceptibility for landslides occurring in Brown Earth than in Renzina soils). These variables warrant closer examination, as they may reveal controlling effects on landslide initiation that are not captured by the available explanatory variables. Specifically, categories of hidden variables with significantly lower predicted susceptibility may provide insight into the occurrence of false-negative predictions.

To quantify the associated variance, each SLD landslide was first assigned a susceptibility value from each model (dependent and independent) by calculating the area-weighted mean susceptibility across the estimated release area (Sect. **Error! Reference source not found.**), rather than using only the raster cell at the landslide centroid (Pourghasemi et al., 2020). Landslides were then ranked according to increasing susceptibility values, and the Kruskal-Wallis test (1952) was performed for each hidden variable. The Kruskal–Wallis test compared independent groups by ranking all observations and testing whether the mean ranks differ significantly across groups, providing a rank-based alternative to one-way ANOVA for non-normally distributed data; it produces an H-statistic indicating whether at least one group's distribution differs from the others. From the H-statistic, the associated variance (η^2) was computed as:

$$\eta^2 = \frac{H - k + 1}{n - k}$$

where k is the number of categories of the hidden variable and n is the total number of samples. Associated variance (commonly referred to as 'effect size' in statistical literature) ranges from 0 to 1 and quantifies the fraction of variance in predicted susceptibility associated with a given hidden variable (Tomczak & Tomczak, 2014). For the variables with the highest associated variance, boxplots illustrate the susceptibility distribution across categories, showing the interquartile range (Q1–Q3), the median, and whiskers extending to 1.5 times the interquartile range, with values beyond this range shown as outliers. To further test whether the individual categories of the hidden variables differ significantly from one another, the Dunn's test with the conservative Bonferroni p-value adjustment (Dunn, 1964) was applied.

All statistical analyses were performed for both the independent and dependent model. Hidden variables with consistently high associated variance across both models provide evidence that the observed trends are systemic, rather than artifacts of a single model with limited spatial transferability."

391-392: It is unclear to me how to interpret these effect values. Also, why are these thresholds meaningful?

As mentioned in the previous answers, we reworked the section 3.5 and thereby removed the sentence related to the thresholds of effect sizes.

395: Did you mean 'computed' instead of 'competed'?

Thank you for catching this typo. Yes, we meant "computed".

410: Please show the AUC values of these perimeters or refer the reader to where they are. I could not find them.

The AUC values are shown in (**Error! Reference source not found.**b). We added the following reference to the plot:

Despite this, when the independent model was tested against the eight SLD perimeters, the overall average performance was considerably lower (AUC = 0.74) and varied strongly among perimeters (Fig. 2b).

448: The term 'effect sizes' is misleading because it is usually used as an indicator of how the model output changes based on the input parameters. The variables you are comparing were not in the model and we don't really know how they affect model performance. We only know their values and your model's susceptibility values at landslide locations. That does not provide enough information to get indications of the true effects of these variables on the susceptibility model.

As mentioned in previous answers, we renamed the term "effect size" to "associated variance" and it redefined its meaning. In addition, we changed the sentence to describe what the "associated variance was used for (line 549-551):

"To assess whether predicted susceptibility differs among categories of the hidden variables, the variance in susceptibility predictions associated with the hidden variables was quantified for the SLD dataset considering both the dependent and independent models (section 3.5)."

451-452: Can you explain why the effect sizes were smaller?

The explanation why the associated variance is smaller is now stated in the following sentence on lines 545 - 557:

"Overall, the associated variances in the dependent model (Fig. 4a) were smaller than those of the independent model (Fig. 4b), which can be related to the higher performance of the dependent model, leading to smaller variation among susceptibility prediction of the SLD landslides."

464-465: Rendzina and the Brown earth have statistically comparable susceptibility values (group 'a').

Thank you for catching this. While susceptibilities of Regosol and Rendzina soils are both lower than Brown Earths, only Rendzina was predicted with **significantly** lower susceptibility than Brown Earth. Thus, we correct the sentence to:

Landslides occurring in Rendzina soils were predicted to have significantly lower susceptibility than those triggered in Brown Earth soils (**Error! Reference source not found.**, first column).

468: Please explain why having at least 10 landslides matters.

We refer to the defined minimum number of samples considered per categories (see reviewer comment for Line 285) and clarify it with following sentence:

"This trend was consistent across individual perimeters with at least 10 landslides (considered as the minimum sufficient sample size per category; section **Error! Reference source not found.**)."

470: I don't see any data about soil depth anywhere. Please show it.

We agree with the reviewer and have added a figure to support the argument in the supplementary material as Figure S4.

478: How do you know that that is the only reason?

This statement was written unclearly. The sentences is rewritten to: This variation can be related to susceptibility differences between the perimeters and is therefore not further discussed in subsequent sections.

484: Please reword as follows – predicted susceptibility at known landslide locations. This type of clarification needs to be implemented throughout the manuscript.

We reworded “predicted landslide susceptibility” to “predicted susceptibility at known landslide locations” on the specific line and reworded predicted susceptibility to refer only to known landslide locations throughout the manuscript.

524: Because this paper relies on statistics, I wouldn't use this word unless you mean it in a statistical sense. You use it in a non-statistical sense throughout the manuscript.

We removed the word “significantly”, as it was not mean in a statistical sense here

554: I found this sentence to be difficult to follow. What is ‘they’ referring to?

We rewrote the sentence to: “Even though imperfect predictions can partly be explained by previously discussed limitations in transferability; such prediction errors are evident in both the training and testing datasets and were often observed in close proximity to correctly predicted landslides.”

565: Please clarify what is meant by ‘less well predicted’.

We rewrote the sentence to: “Poorly developed soils were not only less well predicted (i.e. lower susceptibility was predicted) by our model”

Bibliography

Catani, F., Segoni, S., & Falorni, G. (2010). An empirical geomorphology-based approach to the spatial prediction of soil thickness at catchment scale. *Water Resources Research*, *46*(5), W05508. https://doi.org/10.1029/2008WR007450

Clarke, K. A. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science*, *22*(4), 341–352. https://doi.org/10.1080/07388940500339183

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Pourghasemi, H. R., Kornejady, A., Kerle, N., & Shabani, F. (2020). Investigating the effects of different landslide positioning techniques, landslide partitioning approaches, and presence-absence balances on landslide susceptibility mapping. *CATENA*, *187*, 104364. https://doi.org/10.1016/j.catena.2019.104364

Rinella, M. J., Strong, D. J., & Vermeire, L. T. (2020). Omitted variable bias in studies of plant interactions. *Ecology*, *101*(6), e03020. https://doi.org/10.1002/ecy.3020

Sperrin, M., Martin, G. P., Sisk, R., & Peek, N. (2020). Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology*, *125*, 183–187. https://doi.org/10.1016/j.jclinepi.2020.03.028

Swisstopo. (2022). *swissALTI3D -Das hoch aufgelöste Terrainmodell der Schweiz*. https://www.swisstopo.admin.ch/en/height-model-swissalti3d

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *TRENDS in Sport Sciences*, *1*(21), 19–25. https://www.wbc.poznan.pl/dlibra/publication/413565

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2013). Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, *4*(3), 236–243. https://doi.org/10.1111/2041-210x.12004