Responses to the reviewers July 2024

Christoph Schaller[1,2], Luuk Dorren[1], Massimiliano Schwarz[1], Christine Moos[1], Arie C. Seijmonsbergen[2], and E. Emiel van Loon[2]

[1]Bern University of Applied Sciences - HAFL, Länggasse 85, 3052 Zollikofen, Switzerland
[2]University of Amsterdam UVA - IBED, Sciencepark 904, 1098 XH Amsterdam, The Netherlands

Correspondence: Christoph Schaller (christoph.schaller@bfh.ch)

Date: 29 July 2024

Dear Reviewers, dear Editor,

Thank you very much for your feedback on our manuscript. We have opted for a combined reply with a co-listing of the comments from the reviewers RC1 and RC2. We intend to revise our study by incorporating points from the comments and rework the manuscript accordingly. For data processing we intend to include the NFI forest type as an additional variable to be tested. For the subsequent modelling and analysis we intend to incorporate the MAD as an additional measure and to explore several options for improved visualisations based on the reviewers' suggestions. For the manuscript we plan to partially amend and partially revise the first two sections for better clarity. For the methods we intend to amend text to emphasise the explored variables and methods more clearly and to eliminate unclear parts. We will integrate optimised visualisations and tables based on the analysis in the results section. For the discussion, the focus will primarily lie on the addition of a discussion of uncertainty. Please refer to our individual replies in the sections below for details.

With best regards,
Christoph Schaller

1 Anonymous Referee RC1

Dear Authors,, I apologize for the late upload of my review report. The good news is that my review will be very positive: I liked the manuscript and I think it addresses an important topic, well inside the aims and scopes of NHESS. It has an ambitious goal (soil thickness prediction at national scale is very challenging and, to the best of my knowledge, never attempted before in these terms), which is pursued with a robust and original approach. The manuscript is well structured, well written and very clear. Overall, the manuscript surely deserves publication and wit will be a valuable contribution to the journal. I just have a few comments. None of them is a big issue, and the manuscript can be accepted after minor revisions.

1.1 GENERAL COMMENTS

1 - The thing I miss the most in your study is a better linkage with the geology.

1a- I would add a test site description section, to make clear that you work at national scale (an amazing feature of your work) and briefly describing the main features of Switzerland (geology, climate).

1b- Geology is important in influencing both landsliding and soil thickness distribution. You use geology/lithology in your study by means of bedrock density, which you consider a proxy for bedrock lithology. In my opinion, this strategy needs a better justification. I guess you did this to have a numerical variable instead of a categorical variable (e.g. lithology classes). However, rock/soil cohesion, internal friction angle or hydraulic conductivity (just to name a few) may sometimes have a better relation with soil properties or slope stability. Some more reasoning on your strategy would be welcome.

1c- Building on previous comment, I recommend adding a table in which you list the main lithologies and the density value you assigned to them (or viceversa, as it suits you better). This may be also linked to a figure with main lithologies (see comment about test site description). Does Swisstopo contain thematic layers about other geotechnical properties? If yes, why didn't you consider them as well? When accounting for morphology, you took into account many parameters, so geology could be also accounted for by different parameters.

1d- In section 6.1 you analysed landslide distribution by slope class. However, the first thing that came up in my mind was to evaluate their distribution across lithologies. Is it possible to quickly perform a similar analysis?

>> 1a) Due to the covered extent and the wide range in different geological conditions, we consider a detailed description (especially of the geology) out of scope for of this study. We will add a brief description highlighting the extent of the study and this diversity along with suitable references covering details on geology and climate.

>> 1b) In our study we tested three different datasets on geology. In addition to the finally used dataset on rock density, we also tested data from swisstopo on bedrock from the geological map (Swisstopo, 2023a) and unconsolidated deposits from the Geocover data (Swisstopo, 2023b). For all three datasets we also tested categorical variables. Much of this was omitted from the manuscript in order to keep the text more focused and concise. Parts of this are still evident in the Jupyter notebook accompanying the manuscript. However, we plan to add some more details on these additional tests and their results back to the manuscript. This should better justify the final choice of the numerical variable with the rock density as a proxy. The choice of the numerical variable is ultimately a result of its overall higher variable importance

during the explorative tests of the possible covariates. The lower importance of the categorical variables is mainly due to the high number of categories (especially in the Geocover dataset) and a resulting class imbalance when applied to the reference datasets. In addition, Geocover suffers from inconsistencies in encoding and discontinuities at map sheet borders, since the data is based on digitised paper maps by different authors. The dataset on rock density, on the other hand, is less detailed than Geocover but consistent across Switzerland and appears to represent well the main lithology classes in Switzerland.

>> 1c) We did not assign density values to the lithologies ourselves. We used the values from the existing dataset provided by Swisstopo (2020) that is based on the work by Zappone and Kissling (2021). We are considering to include a table or visualisation with the lithologies and densities but may refer to the existing figures in Zappone and Kissling (2021) for the sake of conciseness since we didn't add anything new to the data. We are not aware of datasets with geotechnical properties by Swisstopo. We tested layers with estimated percentages for sand, silt and clay at depths of 0 cm to 30 cm, 30 cm to 60 cm, and 60 cm to 120 cm provided by the Competence Center for Soils https://ccsols.ch/. However, we ultimately opted not to include these data in the study due to low variable importance values and uncertainties connected to these data.

>> 1d) We will explore the possibility to include a table and/or figure with details on the lithology and rock densities. We are not yet sure, if box plots by lithology would fit well with the argumentation. They could, however, be included as additional material in the accompanying notebook.

1.2 SPECIFIC COMMENTS

L2: I would mention the areal extension of Switzerland. I think the width of your test site is a big constrain to the work; as such, it should be emphasized as an additional point of strength (you may consider to stress it also in the discussion and conclusion).

>> We will add a description with the overall area of Switzerland and the area of the cantons included in the used landslide inventories.

L13: Would you consider adding also https://doi.org/10.1016/j.jeem.2024.102942 ? I think it is pertinent, as it reports on indirect impacts of hydrogeological processes, which are rarely accounted for (most studies focus only on casualties or direct economic damages).

>> Thank you for pointing out this publication. However, after careful consideration we opted not to include it, since its content is not directly linked to our argumentation.

L7 - I suggest to also convert CHF amount to USD. A few European readers may be aware of the value of CHF, but maybe not all the international readers are familiar with this currency.

>> Your argument is understandable. We will convert the amounts to USD.

L21 - My English is not better than yours, but isn't "carried out within" more appropriate than "carried out at within"?

>> You are absolutely correct. We have removed the superfluous "at" in our draft.

L25 - I found odd to read the landslide definition in the middle of the introduction... Isn't it better to move it earlier in the text?

>> We will consider this suggestion along with suggestions of reviewer 2 to restructure the introduction and section 2 on the theoretical background.

L51 and L55 are mentioning three and two landslide inventories, respectively. At this stage, this is confusing.

>> After careful deliberation of the feedback from both reviewers, we decided to exclude the StorMe inventory from the main analysis and only use it as an additional example in the discussion. We will thus not mention it at this point, which should resolve this confusion.

Section 2.1 - Just a comment: another point clearly highlighting the importance of soil thickness is the math formulae used in slope stability models. As scientific knowledge advances, the complexity of models has always been increasing, integrating new parameters and new processes in the stability equation. However, soil thickness has always been there: since the first pioneering equations (e.g. Skempton, A. W., & DeLory, F. A. (1957). Stability of natural slopes in London clay. Thomas Telford Publishing, London, UK, 15, 378-381.), soil thickness has always been there, among the uncontested key parameters!

>> Thank you for this comment. We are considering to include this in section 2.1 when revising the manuscript.

Tab1 - GIST-RF was recently applied to another case of study. If you want, you can add https://doi.org/10.1016/j.catena.2024.108024 along with Xiao's work.

>> Thank you for pointing out this publication. We will include it in the table as well as in the discussion.

Section 3.1 - I suggest to be clearer on one point: the geometry used to map landslides in the inventory. Are they mapped as polygons or points? In case they are points, it would make a big difference if the mapped point is the triggering point or the impact point, especially in case of shallow landslides with large runout distances.

>> Actually, the HMDB inventory only includes points in the triggering area while the KtBE inventory is based on polygons. We will try to make this clearer.

L133 - Geological bedrock instead of geological underground?

>> Thank you for the comment. This will be replaced by "geological substratum".

L148-153 - see comment 1c.

>> Please refer to the reply on comment 1c.

L212 (and elsewhere) - Shouldn't it be CARET, with capital letters? If yes, please adjust all occurrences in the text.

>> The explanation in brackets corresponds to the explanation in the introduction chapter of the caret package. However, the package authors render the name in lower-case in all other instances. Therefore, we will keep the lower-case rendering.

L311 I was surprised in seeing errors higher than the maximum expected value (2.5m) Didn't you applied a upper constrain to the modeled thickness values?

>> We didn't add an upper constraint on the model outputs. It is inherent to linear models that values higher or lower than the values in the training data can result.

L330 I would make reference to Fig B1

>> We will add a reference to figure B1 in the supplementary materials.

L376-379- Could it be that by adding many 0 values in calibration the overall average modeled thickness is lowered?

>> Thank you for the question. So far, we only compared the error but not the overall average of the predicted landslide depth with or without the 0 values. We will look into this during the revision of the study.

2 Anonymous Referee RC2

This relevant and interesting manuscript by Schaller et al. presents a statistically based approach to predict the thickness of shallow landslides in Switzerland.

The authors state four objectives:

(i) To present descriptive statistics of shallow landslide thickness and related explanatory variables based on three Swiss landslide databases.

(ii) Develop and test three different machine learning approaches, ranging from linear regression models to generalised additive models and random forest models, to predict shallow landslide thickness based on different geospatial datasets and their derivatives.

(iii) Evaluate the performance of the models.

(iv) Compare the developed model with three existing models (focusing only on elevation, slope and cumulative slope distribution).

As shallow landslide thickness is a key variable in shallow landslide susceptibility modelling and in a further step of run-out modelling, i.e. shallow landslide hazard indication mapping, the prediction of shallow landslide thickness is of high relevance for further model development, but also for practitioners in natural hazard management. Thus, the authors address a highly relevant topic, especially as they develop models for larger areas and are not limited to smaller catchments. This last aspect in particular highlights the study, as there is no other large-scale approach to predicting shallow landslide thickness over a large area and in high mountains, at least to my knowledge.

Predicting the thickness of shallow landslides is a very difficult task, not least because of the very small-scale heterogeneity of the influencing factors. The authors test many geomorphometric properties derived from a digital elevation model. However, two very important properties are analysed in less detail: The geology/soil and the vegetation/forest.

The latter two in particular can influence the depth of landslides on a very small scale. Although the authors use Zappone and Kisslings rock density dataset, it is questionable to what extent other layers with geological properties could be taken into account to significantly improve the models. In addition, comprehensive information on soil properties is available for Switzerland (forest and agricultural areas; e.g. Baltensweiler et al. 2021 doi:10.1016/j.geodrs.2021.e00437).

>> As mentioned in the reply to 1a to 1d by reviewer 1, we actually did test data on soil properties from the Competence Center for Soils that is very similar the data by Baltensweiler et al. 2021 you mentioned. In the end, we did not include the data in the final manuscript due to several factors. On factor was the only medium to low variable importance during our tests. Another factor was the uncertainty associated with these data, which is also based on machine learning models. We also didn't mention these tests for the sake of a more focused and concise manuscript. We will add some details on these additional tests to the final manuscript but are not yet sure where and to what extent. At the very least they will be mentioned in the discussion.

For vegetation, a vegetation height model by Schaller et al. is used - which obviously also influences the model and the prediction. However, the VHM does not provide any insight into the forest structure, which has a decisive influence on soil stability (the authors cite e.g. Rickli et al. 2019). A spruce-dominated mountain forest will have a different influence on the prediction of the thickness of shallow landslides than a mixed forest in the lowlands - even though both sites may have the same geomorphometric characteristics. As a first step, one could, for example, use the mixed forest layer of the Swiss National Forest Inventory, which is available for the whole of Switzerland. I think, this should definitely be included in the discussion.

>> Thank you for this comment. We agree that the species may have an influence on the stabilising function of the forest and may be a suitable variable to represent the forest structure. For the revision of the study, we will add the NFI forest type (Waser and Ginzler, 2018) as an additional covariate to be tested. Depending on its importance it may be included in the revised models.

As the prediction of shallow landslide thickness is intended to be used to generate raster data, i.e. maps, it would be of general interest to know what these rasters will look like. Therefore, I strongly suggest including one or two case study sites and showing how the approach will work for a spatial prediction (including a critical discussion of uncertainties).

>> We consider the generation and use of the rasters as a next step that is part of a subsequent study. The main focus of the current study is on the modelling and prediction of the landslide depth at the locations recorded in the inventories. We have therefore decided not to include such case studies and rasters. Our main concern is that the addition of such rasters would undermine the focus and conciseness of the paper. The prediction rasters would necessitate the addition of a considerable amount of text to properly explain the generation of the predictions and additional visualisations to allow for their interpretation (e.g., local context and model inputs). This would extend the manuscript beyond the recommended limits while adding limited value to the study.

The manuscript is clearly written. However, with the four objectives in mind, I feel that there is room for shortening and more precise wording in a number of places.

Some of the figures are difficult to read and extended captions could help the reader to follow more quickly (figures and tables are not always self-explanatory).

In the methods section, some important aspects are lost or neglected, which makes it difficult to fully evaluate the results. I think, the discussion chapter lacks a more critical discussion of the methods and data used, especially with regard to uncertainties and sensitivities of the models.

Based on the relevance of the topic (which is clearly within the aims and scope of NHESS) and first promising results, I recommend accepting this manuscript after major revisions.

2.1 General Comments

2.1.1 Introduction and Theoretical Background

I suggest combining Chapter 1 and Chapter 2. Otherwise the four objectives seem a bit lost between the two chapters. The introduction sounds more like an extended motivation, but lacks a description of the existing research and the subsequent research gap. Combining the two chapters could easily solve this. However, I suggest shortening the first part "motivation" and adding a short overview of the ML models chosen and why you chose them for your presented study. In chapter 2.2 you describe the existing soil thickness estimation models for landslide modelling and provide Table 1 for a more detailed overview. I think this is a very good idea. However, is it possible to add one or two sentences about them (common basis/differences; your list is "non-exhaustive" -> on what basis did you select the models/studies presented?).

>> We will revise chapters 1 and 2 trying to address these concerns.

2.1.2 Materials

Most of the records in the HMDB were collected after heavy rainfall events within defined perimeters. I suggest mentioning this in the text as it may influence the choice of statistical models.

>> We will add this to the description.

The thickness of landslides in the KtBE dataset was estimated by experts and orthoimagery. Hählen (2023) estimated an error of up to 50%, which is very high. - How well can you fit an ML model with such a high error in the data? Is this a suitable dataset for training the ML models? I suggest including this in the discussion chapter.

>> We had a short exchange with Nils Hählen on this topic. We should be more precise and write that the possible error is estimated to be 25-50%. While there potentially is a high error, the similarities to other distributions (including HMDB) suggests, that the overall error is small enough to warrant an inclusion in the modelling process. We agree that this should be included with more detail in the discussion.

You mention that the StorMe dataset was excluded from the model development because of doubts about the data quality (L122). However, you use it for descriptive statistics. How robust are the descriptive statistics? I am not familiar with the StorMe dataset in detail. However, I understand that landslides from the HMBD are included in the StorMe dataset. If so, have you removed the 709 HMDB records from the StorMe dataset?

>> We agree with your concern about the statistics based on StorMe. We can't be completely sure about the robustness due to the uncertainties stemming from the data quality issues. After careful deliberation we decided to remove StorMe from the descriptive statistics as well and only include this inventory as an example in the discussion.

Model input data:

What about other geological data? What about soil maps? (Please, see my comments above).

>> Please refer to the reply on points 1a to 1d of reviewer 1.

In the methods chapter you mention the Topographic Landscape Model TLM (ground cover rock / e.g. Fig. 3). If so, I suggest you include it in your list of input data used.

>> We will add an according point to the list.

2.1.3 Methods

I have some unanswered questions in the methods chapter. This is also where my main concerns about the study lie. These concerns mainly relate to:

(i) Sampling of covariates at the failure points of the slides (S2.3). Did you use a buffer around the failure points? For example, the HMBD was recorded with less accurate GPS systems (many records from the 90s/00s). Uncertainties may range from 5 to 20 metres. How did you extract the points - with a grid size of 5m for example? Wouldn't it be necessary to use a buffer and use the values of several grid cells (as has been done in other studies)? As far as I know, the reliability of the older StorMe database records is even more important (even if they were only used for descriptive statistics).

>> We acknowledge that there is a degree of uncertainty in the coordinates of the landslides. However, we decided against the use of a buffer and only sample at one point. On the one hand, the different resolutions used in the tested covariate rasters partially have a similar averaging function as such a buffer. Most of the chosen variables already include an averaging either by means of the resolution (10 m or 25 m) and/or a radius in the calculation of the covariate (e.g. TRI and openness). We actually expect an increase in the uncertainty for some covariates when aggregating over a buffer (e.g. the slope of a failure at the transition from flat to steep terrain).

(ii) You used regression models. How did you build / structure these models? As already mentioned, the HMBD, for example, was recorded in perimeters after defined heavy rainfall events. This results in a spatial and temporal hierarchy in the data. Can simple linear models be used or should more complex models be used (e.g., linear mixed effect models)? (iii) How was the quality of linear and generalised additive models checked? Did you perform diagnostic plots of the residuals? Was

transformation of the data necessary to meet the assumptions of the models? Based on your list of input variables, there may be interactions between some variables. Have you included an interaction term in your analysis? Perhaps you can include a table/chart of the models you fitted?

>> In our study we mostly tested models based on the caret library. For all three tested model types, the same covariates were used to build models based on the facilities of caret. For the linear regression and GAM structure of the models is automatically estimated by the library. There is little leeway to influence the model structure except through the hyper-parameters, which are documented in the manuscript.

>> We did a few initial, manual tests with mixed effect models, but didn't see enough of a difference to warrant further exploration.

>> The limited possibilities to influence the structure of the GAM models is admittedly somewhat of disadvantage of caret. Due to the main focus on random forests, we accept this downside. We did apply some standard diagnostic plots to the fitted GAM models, but the results were rather inconclusive. Therefore, we left out these plots and concentrated on the used indicators for gauging the performance of the models. We will look into whether some general conclusions the visual interpretation of the plots can be integrated into the results or the discussion.

(iv) You mainly use MAE and $R\hat{2}$ for model validation. What about a confusion matrix, ROC curves or AUC values to evaluate your RF model in more detail (particularly in case of an imbalance)?

>> Our study uses machine learning models for regression i.e. we predict numerical values. Confusion matrices, ROC curves and AUC values are only applicable for classification tasks.

(v) Can you provide more information and results on the importance of your covariate selection and fitted models?

>> We already included a heat map with the most important results in the appendices of the manuscript. We deliberately chose not to go into more detail in order to focus the manuscript more on the actual prediction results.

2.1.4 Discussion

The discussion of landslide inventories is very brief (Chapter 6.1). It would be desirable to address the uncertainties and problems of the databases, such as the accuracy of the measured coordinates or the derivation of depth from orthophotos, in the context of the descriptive statistics and model performance presented.

>> We intend to restructure and extend the discussion on the landslide inventories, taking into account the points you mentioned. We are planning to retain the StorMe inventory as an additional example for the possible problems with the data in the inventories.

Chapter 6.3 discusses the selection of covariates. I suggest adding a few lines, i.e. a critical discussion, on the effect of geology / soil characteristics and vegetation / forest effects.

>> We will include a more in-depth discussion on the effects of geology, soil characteristics and vegetation, possibly including some learnings from our tests with the soil characteristic data from the Competence Center for Soils.

I suggest adding a critical discussion of the ML models used (see also my comments above on the methods chapter). Have you considered other models (e.g. logistic regression)? Could the robustness be increased by a bootstrap approach?

>> We are taking this suggestion under consideration.

>> In the early phases of the study, we explored a number of different possible ML models for this study but settled on the three types presented in the manuscript. Logistic regression is not suitable for our study, since it is used for classification.

>> Bootstrap procedures are used to estimate confidence intervals and they are usually more robust in estimating such confidence intervals than model-based estimation. However, bootstrapping does not increase the robustness of the model (the parameter-estimates of the model) or its predictions. Therefore, we will not adapt bootstrapping for this study.

I also think that an extended critical discussion of the uncertainties is missing.

>> As mentioned in the replies above, we intend to add points on the uncertainties. We have yet to decide if there will be a central paragraph or section discussing uncertainties or if we will discuss them mostly in context of the different topics such as inventories, input data, and models.

2.2 Minor Comments

L31: Change meters to m to be consistent.

>> This will be modified.

L35: Can you add a reference for your chosen definition of landslide thickness?

>> This definition was formulated by the authors as the definition to be used in our study.

L46/47: I suggest rewording the sentence and removing the "according to our recent paper".

>> We will take this into consideration when revising chapter 1.

L55: At this stage it is not clear to the reader why you have chosen only two landslide inventories and not all three?

>> The mention of StorMe will be removed at this point (see also the reply to your comment on L122 above).

L60: What do you mean by "landscapes"? - Is that topography? Geomorphometry? L60: Soil types in terms of pedological soil types (e.g. Cambisols) or in terms of a more geotechnical description (e.g. USCS).

>> This is a statement of Hungr et al. (2014). It refers to the topographical variation.

>> From our perspective, this refers to the geotechnical soil types. We will try to make this more clear when revising the manuscript.

L81: I suggest avoiding 'in' or 'see' before a figure reference. This occurs frequently throughout the manuscript, and I suggest it be changed.

>> We will take this into consideration and remove the instances of 'see' where not appropriate.

Fig1: You mention the failure area, transit zone and deposition zone in the figure caption. I suggest you highlight these in the figure. In the figure you highlight the regolith. In the text you write "soil / soil type". What are you referring to? Is regolith = soil? Do you define soil as the entire weathering mantle from the surface to the weathering front / bedrock? Probably you can change the word "regolith" to be consistent with your text.

>> The missing depiction of transit zone and failure area is a result of an intentional simplification of the figure to highlight the parts pertaining to failure, which are the focus of this paper. We will discuss whether to modify the figure or the caption.

>> Mostly in the sense of regolith=soil. We will revise the figure and/or text to eliminate ambiguities and be consistent throughout the text.

Tab1: I suggest adding a semicolon after each reference in column 4 and changing the parenthesis in the fifth line. Can you give a brief explanation of the parameters in the table heading? In the last row you mention 158 remotely sensed covariates. Can you somehow classify and mention them?

>> We will modify the references accordingly.

>> We will add a short specification on the general nature of the covariates based on the abstract by Hengl et al. (2017): "158 remote sensing-based soil covariates (primarily derived from MODIS land products, SRTM DEM derivatives, climatic images and global landform and lithology maps)"

L96: You have filtered your data and kept only entries with a thickness of up to 2.5m. In your introduction you refer to the "Swiss definition" of shallow landslides, where the thickness of the instable mass does not exceed 2m. You are working with Swiss landslide inventories. Why did you finally choose a threshold of 2.5m? Can you give some information on how many landslides (n) occur between 2 and 2.5m? L110: Did you remove the six landslides (from tab 2 it seems that you left them in the dataset)? Why - especially since you then cross-validate.

>> You are correct, that the cited "Swiss definition" and the used limit of 2.5 m are inconsistent. In order to be consistent, we will adopt a 2 m limit for the revision of the study.

>> The landslides mentioned on line 110 were removed based on the filtering due to the 2.5 m limit. We will try to reformulate the description to make this clearer. Adopting the 2 m limit will result in the removal of 2 additional landslides for the KtBE dataset and 66 additional landslides for the StorMe.

L97: Please give the URL for the WSL landslide database.

>> We will add an additional reference for the homepage or the database.

L110: (Hählen 2023).

>> Thank you for pointing this out. We will correct the citation style.

L126: I suggest starting a new paragraph and shortening the paragraph as you provide Tab2 and the maps (Fig. 2).

>> We are considering this change and are deliberating the possibility of an integration with the study site description.

L133: Please replace geological underground (e.g. with bedrock or geological condition).

>> Thank you for the comment. This will be replaced by "geological substratum".

Tab2: Please include the full name of any abbreviations used in the captions (HMBD, KtBE, StorMe).

>> We will take this into consideration.

Figure 2: Change "slides" to "shallow landslides". Remove "showing the locations" in the figure caption. Please add the sources/background maps in the reference list (swissBOUNDARIES, and relief map).

>> We will modify the figure and add corresponding references.

L141: Why do you need the areas outside the Swiss borders if you are using inventory data from Switzerland? This is unclear.

>> Some of the covariates are calculated with a window or radius. To prevent erroneous values at the border, we fill the area outside Switzerland such that all covariates within the borders are calculated without missing values. We will try to amend the description to make this clearer.

L146/147: This sentence is unclear, especially the last part ("assuming that this leads to less soil cover"). Could you please rephrase the sentence?

>> The assumption is, that areas with stronger extreme precipitation will suffer from higher erosion and landsliding thus leading to less soil cover. We will try to rephrase this.

L153: Why do you use catchment areas, since you use them as the basis for tiling / parallel processing? Why not use the Swiss reference grid (e.g. SwissSURFACE divisions)? Do you influence the values of some geomorphometric derivatives by tiling? For example, the TWI, if you do not process the whole grid at once? Later you mention that you create a buffer around the catchments for data processing. Do the intersecting areas have the same values (e.g. for TWI)?

>> We intend to amend the method description to make our choices clearer:

>> – Catchments are used as a tiling unit because some of the indicators (e.g. TWI, catchment area) have a hydrological background. Processing the DEM in a regular grid like the SwissSURFACE division would result in erroneous values for these indicators.

- The geomorphometric derivatives should not be influenced since they are processed with the catchment plus a sufficient buffer. This buffer is chosen such that the derivatives calculated with a window or radius have input values for the full needed area for each cell within the catchment. The potentially erroneous values outside the catchment border are not used for further processing.
- Processing the hydrological indicators such as TWI and catchment area over the whole of Switzerland would be ideal but takes too long on the hardware available to us. Processing across the entire grid would result in higher values for the catchment area/flow accumulation in the areas of rivers and lakes (which would also influence the TWI in these areas). However, since we don't intend to predict the failure depth within water bodies and since the calculated values on the slopes to the rivers and lakes are still correct, we accepted the too low values in these parts in exchange for significantly faster data processing through tiling.

L158: The covariates such as terrain variables or geology were derived from the DEM. What do you mean by that? What geology do you derive from the DEM? I thought you used the rock density layer?

>> The terrain variables are derived from the DEM while the geology is based on the "other inputs". To avoid confusion, we will reformulate the text to say "The covariates, such terrain variables or geology, were derived based on the DEM as well as other input data."

L159: By reference data you mean the inventory data?

>> Yes. We will make this explicit.

L160: You mention Stage 4. However, in L156 and Fig3 you highlight three stages.

>> Thank you for pointing this out. This will be changed to "(stage 3)".

Fig3: (i) Please give the full spelling of the abbreviations S and O used. (ii) In Stage 2 you refer to geological maps. This is the density map? What about the geomorphometric derivatives? (iii) I assume there is a typing error with O3.1 and O3.2 in Stage 2? (iv) S1.4: You average the average? Does this make sense? Why don't you use the original data to do the averaging? Also, if I understand correctly, you resample the EUDEM to 5m that you aggregate/average again to 25m (S1.5)?

>> (i) We will add a small legend with the spellings to the figure.

>> (ii) The geomorphometric derivatives are derived/calculated from the DEM as part of the "Generate covariate rasters" step. For compactness' sake we will not show every calculation as a separate step and only show "raw" inputs used by the calculations in this step.

>> (iii) Thank you for pointing out this mistake. It is indeed a typo. We will correct this.

>> (iv) Averaging the average is indeed usually wrong. The exception are cases where the n within all averaged groups is identical. This is ensured in our case because we work on an aligned grid and the lower resolutions being multiples of the

higher ones. We thereby accept the small error due to rounding, since the preparation process is considerably simplified because only one filling step is necessary. This also means that all covariates are ultimately derived from the same filled 5 m raster. The resampling of the EUDEM an subsequent averaging admittedly introduces a slightly bigger error in height. But again, we are willing to accept in exchange for a simplified process since these height values are only used for cells at the border for a few covariates which are calculated with a larger radius (mainly position based indicators like TPI or openness).

L162: Please use the correct reference for software R.

>> We will modify the entry in the bibliography such that the reference will be rendered as (R Core Team, 2022) by the Copernicus LaTex Template.

L169ff: See my comment on Fig 3(iv).

>> We will try to make the motivation behind the process clearer in the text.

L180: Please delete the first sentence of chapter 4.2.1 as it is repetitive.

>> We will take this into consideration.

Tab3: (i) You use a number of explored covariates. However, you have somehow forgotten to provide the references. Could you please add a column with the references and the input datasets used? For example, there are several ways to calculate flow accumulation and/or TWI (based on multiple flow directions / single flow directions / weighted / ....). (ii) You highlight the variables and cell size used in the final ML models. Following this description, you use the TPI for 15m, 50m, 200m, ..., 4km with a cell size of 25m. Is this correct? However, I am confused as you only mention tpi_500m_25 and tpi_4km_25 in L224. (iii) I suggest adding the other covariates examined as VHM and rhob_m to the table as well.

>> (i) We will extend the table with information on the input data and the exact tool used to calculate the respective covariate.

>> (ii) The values 15m ... 4km pertain to the radius the TPI is calculated at. And not all radii are calculated for all resolutions. We will make this clear in the table.

>> (iii) Strictly speaking these correspond to the last two rows in the table. We will try to make this clearer.

L200-222: The approximation of the failure point is not clear. Could you please rephrase the sentence?

>> We will rephrase this.

L205: You have randomly generated points. How did you do this? What procedure did you use?

>> The points were generated with a sampling function of the terra package in R. We will add this information to the text.

L208/209: Are the numbers of the points generated correct? (HMBD 29? / KtBE 50?).

>> Yes, the numbers are correct. The number depends on the ratio of rock signature within the catchments where the inventory points are located. Since the locations of the HMDB are more closely clustered and differently located, fewer catchments with less rock signature are affected, thus resulting in less points.

L218: You have chosen a combination of exploratory analysis based on literature and landslide experts. Could you please refer to the literature you used? Who are your experts and how many experts did you contact? On what criteria did they suggest the covariates?

>> We will take this into consideration. Since this mainly overlaps with the references in chapters 2.1 and 2.2, we may refrain from repeating all the references and refer to the relevant parts of the chapter.

>> The mentioned experts are among the co-authors of the paper. We will modify the manuscript such that it states "based on expert knowledge of the authors" instead of "experts on landslides and geomorphology". The expert inputs were based on their backgrounds in geomorphology, hydrology and landslide modelling. We tested covariates they considered to have a potential influence on the landslide failure and retained the most promising ones for the final study.

L219: "both datasets", i.e. HMDB and KtBE?

>> Yes. We will make this explicit.

L263: Why did you fit the models without intercept?

>> For the hyper-parameter tuning of LM models, the caret package only supports the distinction between fitting with or without intercept. The fit without intercept is a result of the tuning procedure. We will try to make this clearer in the text.

L285-306: This paragraph seems rather long. I would suggest shortening it and referring more to the table and figures.

>> We will take this into consideration.

Tab5: (i) Why do you highlight the sampling depth in the forest? What are the consequences of that? Are there differences in the descriptive key figures? (ii) You give the standard deviation for the arithmetic mean. However, you also show the median. Please add the corresponding MAD to be consistent. (iii) In general, it might be useful to replace the table with a figure. My suggestion is a combination of violin plots and box plots, as they contain the robust characteristic values on the one hand, and show the data distribution on the other. As mentioned in the introduction, geology and vegetation/forest are two key factors that could significantly influence the prediction of shallow landslide thickness. So why not split the data set further (see comment (i)).

>> (i) We mainly highlighted this because the KtBE dataset only contains few cases within forests compared to the other datasets.

>> (ii) We will add the MAD.

>> (iii) We will explore whether we can replace the table by plots that can convey the same information. One of the potential problems we see are the different scales and divisions by categories, which could lead to rather small plots. We will also test whether there are further divisions by geology or vegetation that make sense. Depending on the classes, only few cases per class might result, leading to similar problems like the box plots with too few points per box. Depending on the results we may only include them as supplementary material in the accompanying notebook.

Fig4: Is the mean missing from the bottom left figure (or is it the same as the median)?

>> Its not missing but identical to the median (compare Tab. 5).

L308: I cannot find the value 0.24m in Table 6. Is 0.24m correct (or should it be 0.25)? Which is the correct value?

>> Thank you very much for this hint. There was indeed a discrepancy between the value in the text and the table. This instance should indeed have been 0.25 m. We will correct the text.

Fig5 (and all other box plots in the manuscript): (i) Individual boxes contain only one to 10 data points. At least five values are needed just to define the box. Perhaps you could point this out or simply replace the box with a median and MAD? (ii) In some cases, your data have a wide range (outliers and long whiskers). Could you try to transform your data (e.g. symlog) so that the reader gets more insight into the differences/major part of the data (boxes). This is especially the case for Fig7).

>> You are correct, that most sources recommend at least 5 samples for constructing a box. We will modify the plots to show the individual points together with a median or mean value.

>> We will explore if the figures could be further improved by applying transformations to the data.

L318: Compared to?

>> We will clarify this by adding "when compared to the other three models" to the sentence

L328: Please show/refer to a figure.

>> We will add a reference to figure B1 in the supplementary materials.

Fig6: (i) It is very difficult to read this figure as the plotted points are very small. The same applies to the 2D kernel density contours. (ii) Figure caption, last line: How many outliers are outside the display range (n=?)?

>> We will try to optimise the raster file for the figure as far as possible and plan to provide a high-resolution variant for download in the final version.

>> We will either specify the number of outliers in the caption or as an overlay in the plots themselves.

L382: What variables are you referring to (overlap with Zweifel et al. 2021)?

>> The notable overlaps are elevation, slope, TWI, curvature, and aspect (though transformed in our case) are found in both studies. We will reformulate the sentence to make this explicit.

References

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLOS ONE, 12, 0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hungr, O., Leroueil, S., and Picarelli, L.: The Varnes classification of landslide types, an update, Landslides, 11, 167–194, https://doi.org/10.1007/s10346-013-0436-y, 2014.

Swisstopo: Geologie Gesteinsdichte, https://data.geo.admin.ch/ch.swisstopo.geologie-gesteinsdichte/, 2020.

Swisstopo: GeoMaps 500 - Vector, https://www.swisstopo.admin.ch/en/geodata/geology/maps/gk500/vector.html, 2023a.

Swisstopo, B. f. L.: GeoCover v2, https://www.swisstopo.admin.ch/en/geodata/geology/maps/geocover.html, 2023b.

Waser, L. and Ginzler, C.: Forest Type NFI, https://doi.org/https://doi.org/10.16904/1000001.7, 2018.

Zappone, A. and Kissling, E.: SAPHYR: Swiss Atlas of Physical Properties of Rocks: the continental crust in a database, Swiss J. Geosci., 114, 13, https://doi.org/10.1186/s00015-021-00389-3, 2021.