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).

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.

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).

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.

General Comments

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?).

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.

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.

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?

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

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).
(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?
(iv) You mainly use MAE and R^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)?
(v) Can you provide more information and results on the importance of your covariate selection and fitted models?

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.

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.

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?

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

Minor Comments

L31: Change meters to m to be consistent.

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

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

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

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).

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

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.

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?

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?

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

L110: (Hählen 2023).

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.

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

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

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

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).

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

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

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)?

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?

L159: By reference data you mean the inventory data?

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

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 EU-DEM to 5m that you aggregate/average again to 25m (S1.5)?

L162: Please use the correct reference for software R.

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

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

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.

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

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

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

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?

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

L263: Why did you fit the models without intercept?

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

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)).

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

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?

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).

L318: Compared to?

L328: Please show/refer to a figure.

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=?)?

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