

Dear Editor and Reviewers,

Once again, we would like to thank for thorough and constructive assessments, which have helped us refine the framing and some methodological aspects of the manuscript. In the following, we have copied all comments by Editor, Reviewer #1 and Reviewer #2 in grey. Our responses are shown in black and for quotes and demonstration of changes made in the manuscript we use a smaller Times New Roman font.

Public justification (visible to the public if the article is accepted and published):

Dear Ms Rüther and co-authors,

Thank you for submitting your manuscript to NHESS and for your initial responses to the reviewers' comments.

Both reviewers have recommended that the manuscript be revised, and, based on their reports as well as my own assessment, I concur. The manuscript addresses an interesting topic and is well aligned with the scope of NHESS. At the same time, several aspects would benefit from further refinement.

In particular, I agree with Reviewer 1 that some of the language should be moderated to ensure that the findings are presented in a sufficiently cautious manner and do not over-interpret the results, as you also already acknowledged in your response. The results themselves are valuable and interesting, and I appreciate the manuscript's strong technical and methodological focus. The main opportunity is therefore not to change the substance of the work, but rather to present the findings and their implications in a more balanced way that accurately reflects the strength of the evidence.

Given the strong methodological focus of this work, making the underlying code available would be highly valuable to support reproducibility and facilitate further use of the results. I note that you indicate in the manuscript that the Python code used for data preparation and the R code used for machine learning modelling will be made available upon acceptance. I encourage you to prepare a code repository and provide the corresponding DOI in the revised manuscript. If you prefer, you may apply an embargo or keep the repository private until publication, with the code released publicly thereafter.

The DOI for both codes is provided and referenced in the revised manuscript. Under Code and data availability, we write:

The Python code used for data preparation (Haualand, 2026) and R code and dataset used for machine learning modelling (Gillespie, 2026) are available in code repositories.

Regarding the advocacy and discussion of the importance of spatial cross-validation in landslide susceptibility modelling, you may wish to consider the related study published in Environmental Earth Sciences: <https://doi.org/10.1007/s12665-024-12041-y>. This work discusses aspects of spatial dependence and spatial nested cross-validation in the context of landslide susceptibility modelling, while also highlighting the importance of considering model utility, interpretability, and geomorphic plausibility alongside quantitative performance metrics.

Please note that considering or citing the suggested article is, of course, entirely optional and is not a requirement for the acceptance of your manuscript. I mention it solely because I believe it may be helpful in positioning your work within the existing methodological literature and in framing the novelty of your contribution in relation to previous studies. Whether you find this reference and the associated discussion relevant for your manuscript is entirely at your discretion.

Thank you, the recommended paper was indeed useful both in context of performance metrics vs plausibility discussion and because the paper also advocates for a set of best practice protocol including the masking of absence sampling area.

Other than that I only have minor technical comments regarding the figures, which you may wish to consider addressing:

Figure 1: Since the plotted variables are continuous, a continuous colour scale may be more appropriate. In addition, the contrast between the rainfall raster and the landslide points in subfigure (b) could be increased to improve readability.

Precipitation in Fig 1a is now shown on a continuous colour scale and the legends in 1a and 1b are turned around to show highest values on the top. Landslide points in 1b are shown in a dark brown to create more contrast compared to the relative rainfall raster.

Figure 7: The landslide points and control points are somewhat difficult to distinguish. Please consider improving the figure readability by enhancing the contrast between the basemap, the mask, and the landslide/control points.

We have experimented with the colour for the control points in Figure 7 and conclude that orange provides best overall contrast. Upon request from both Reviewers this figure has been moved the section on control point sampling and is now Figure 2.

Figures 8-11: While I understand the aesthetic motivation for displaying the landslide points using the same colours as the underlying raster, I find that this choice makes the maps somewhat difficult to interpret. In particular, the landslide locations are not always easy to distinguish. Please consider whether increasing the contrast between the points and the background, or slightly increasing the point size, could improve the overall readability of these figures.

In Figures 8, 9 and 11 all landslide points were increased from size 2 to 2.5 points (control points in Fig. 8a are kept at 2 points and in transparent colour to not steal too much attention). In Fig. 10 landslide points were already 3 points, so rather than increasing their size further we simplified the background map such that the landuse layer in green and beige colours no longer interferes with the deposit map.

You may of course retain the current figure style if you prefer; I just wanted to highlight that the figures could be further improved by considering overall readability and visual clarity.

When preparing your revision, please include:

- (1) A point-by-point author response addressing each reviewer comment. You may build on the material you have already prepared and uploaded.
- (2) A version of the manuscript with tracked changes (or an equivalent diff) clearly indicating all revisions.

Please let me know if you have any questions or anticipate any delay in preparing the revision. I look forward to receiving your revised manuscript.

With best regards,

Matthias Schlögl

[NHESS Editor]

Anonymous Referee #1, 11 March 2026

This paper evaluates the efficacy of various cross-validation partitioning methods in capturing spatial dependencies during hyperparameter tuning for landslide spatial prediction (susceptibility). Additionally, it examines the impact of integrating gradient-boosted decision trees with mixed-effect modeling to account for spatial non-stationarity in landslide susceptibility predictions. They use this approach also explore the influence of forest properties on the spatial prediction of landslide occurrence during a 2023 major rainfall event in southern Norway. SHAP values were used to explore relative and rank of variable importance when comparing models with different spatial dependency structures and tuning approaches. Partial dependency plots of SHAP values were used to peak at general landslide response-predictor trends.

Overall, there are some very good methods and process-based findings in this paper that would serve the landslide susceptibility modelling community quite well:

Methods:

- SHAP values and rank can be influenced by the (spatial) partitioning methods used in nested CV for model hyperparameter tuning.
- Accounting for spatial dependence in model residuals does not result in better model performance (but still important for non-biased interpretation of response-prediction relationships in ML models)

Process:

- Tree properties can have a strong influence on landslide predictions

The authors should be cautious not to over-interpret results beyond what their methods can strictly support, in particular how the number of predictors and correlations between them may influence the variable importance results. Furthermore, the current structure of the Results and Discussion sections makes the narrative difficult to follow. Because interpretations are interspersed throughout both sections, it is challenging to identify coherent summaries of the data or clear trends in the authors' explanations.

We have completely restructured and rewritten large parts of the results and the first part of the discussion. Note that text has been shuffled around so often that virtually all text appears red in the document with tracked changes.

We still describe the tendency in the models with spatial CV to better account for spatial autocorrelation at the expense of predictive power and contrast these with the other extreme of models with random CV which show highest predictive power without accounting for spatial autocorrelation. We no longer argue that the spatial model with small block CV can be viewed as a compromise, and report that we could not recommend an optimal model. The reasons for this are due to likely bias caused by spatial

confounding, which we discuss in detail. For interpretation purposes and to address objective 2, rather, than exploring three contrasting models in detail, we focus the results on a comparison of variable importance across all 16 models presented in the main body of the article and explore subsequently variables with generally consistent importance in greater detail. The new Figures 4 and 5 now show a comparison of variable importance across all model structures, while the new Figures 6 and 8 are a combination of the previous Figures 3+4 and 5+6. Note that Figure 7 has been moved forward and is now Figure 2 which has consequences for numbering of the old Figures 2 to 6.

Major Comments

A major limitation of this paper is the authors' tendency to overinterpret the results supported by their methods. For example, variable importance ranking is highly variable depending on the collinearity between predictors/variables. The authors often provide conjectures on why variables, like elevation are important, without accounting for possible correlations (e.g. with precipitation and landuse).

Adding in a correlation analysis of the predictor variables can help strengthen the findings or further provide explanation for why variable rankings differ between different model setups. For example, I would guess the importance of elevation and precipitation are flipped between spatial and non-spatial due to them being highly correlated.

Following Reviewer's suggestion, we have included a summary of predictor correlation structure for the full and forest-only models in Appendix D. Correlations are visualised separately for signed (Spearman; Figs D1 and D3) and unsigned (η , Cramér's V; Figs D2 and D4) measures to avoid conflating measures with and without a meaningful direction of association. We refer to these results at relevant passages in the results and discussion. In short, the supplementary analysis shows only moderate correlation among a few directly related variables (planar vs. profile curvature and three pairs of forest variables) and one strong correlation (surficial deposits vs. deposit thickness) suggesting that raw predictor collinearity is unlikely to fully explain the observed differences in variable importance across modelling approaches.

Also, adding in a figure/map to illustrate what the different spatial partitioning approaches look like when applied to the landslide data would help in interpreting the results. Especially, to help differentiate if some spatial partitioning approaches are favoring a more 'locally' (vs. globally) tuned model due to their spatially coherent groups.

We have prepared two figures with four maps each showing the spatial folds in the full and forest-only models based on: (a) spatially clustered CV, (b) large block CV, (c) small block CV and (d) random CV (Figs C1 and C2 in Appendix C). We refer to these supplementary figures in relative passages of the chapter on Machine learning methods. Further, we conducted a supplementary analysis to explore the spatial structure of the included explanatory factors (Tables D1 and D2). The latter shows that relative

precipitation has the highest spatial structure of all predictors and spatial autocorrelation remains at the characteristic distances under all CV schemes. Elevation, groundwater, total 'Hans' precipitation and precipitation intensity have high remaining autocorrelation at the characteristic distance for random and small block CV only.

Additionally, the interpretation of the results relies heavily on process-based knowledge. I think the findings would benefit greatly from a discussion that considers the spatial (structure) heterogeneity of the predictors and their potential to differ across space where landslides occur, and how these are captured by the different spatial partitioning approaches, and modelling methods.

The first section of the discussion (6.1 Quantitative model evaluation versus plausibility) has been rewritten with input from the above-mentioned supplementary analysis (presented in Appendix C and D). We now build a more careful argument highlighting the likelihood of spatial confounding by referring to: (1) predictor collinearity, (2) spatial autocorrelation range of each top predictor (via variogram analysis), (3) quantification of the degree of train-test spatial proximity (leakage) under the different CV schemes and (4) clearer comparison of variable importance across the full factorial design of CV scheme (random/spatial) × model structure (with/without spatial random effect).

As a bottom line, section 6.1 acknowledges that there are biases in all of our models, and it is not possible to identify an optimal model. However, our rigorous test across different model schemes and CV structures does reveal some consistencies in importance (or unimportance), allowing a discussion of plausibility. Further, we note that the shape, direction and magnitude of predictor trends are consistent across models and we argue that this permits us to discuss them in the light of their mechanistic contribution. In particular, as Reviewer #2 is happy with the discussion around soil indicators and tree type, we choose to not make substantial changes to this part. Throughout the discussion, we are nevertheless careful to note the uncertainty in our findings.

Minor Comments

L15. Need to provide context to why forest-only dataset is used in the abstract. Currently missing.

The abstract was partly rewritten to reflect changes in the framing of the results and is close to the word limit. Context to why forest-only dataset is used is provided in the final sentence of the abstract:

In the forest-only models, tree type was consistently an important predictor, with higher landslide susceptibility in deciduous forest, underscoring the potential of forest variables and forest-specific threshold values in shallow landslide susceptibility mapping.

L46. Need to be more explicit to what is meant by data-driven. E.g. a physical model can be data-driven if observation data is used for it's tuning. Maybe empirical models could be used instead.

We introduce these terms with reference to Corominas et al 2014 and would like to stick to the established terminology. We have made an addition to the introductory sentence to clarify:

Studies of shallow landslide susceptibility are commonly differentiated into knowledge-driven, physically based and data-driven methods; all of which use landslide inventory data for validation, but only the data-driven method also applies this data as input (Corominas et al., 2014).

Later in the text we take out the same information:

Data-driven approaches ~~use landslide inventories both as input and for validation of statistical and machine learning models. They~~ build on the assumption that combinations of topographic, geological and meteorological factors that have triggered shallow landslides in the past can be used to make quantitative predictions of future susceptibility to shallow landslides (Corominas et al. 2014).

L58. “almost perfect levels of landslide prediction accuracy” – I would argue there is no perfect prediction accuracy 😊. Maybe word they are highly overoptimistic.

We would like to retain this wording as it speaks to studies that report AUC values of 1 or 0.999, and since an AUC of 1 would mathematically imply perfect prediction, the sentence is accurate.

L62. “Machine learning” to general here since later decision tree structure is mentioned in the sentence... SVM's don't have decisions trees.

“Machine learning methods” has been replaced with “Decision tree models”.

L75. Brenning 2012 in IEEE IGRSS missing from references (original spatial CV paper).

As this sentence deals with ecological ML applications, we rather included Brenning 2012 as references in the previous sentence and made some adjustments to the wording. The passage now reads:

However, to account for spatial autocorrelation in geospatial data, spatial strategies of fold assignment may be better suited (Brenning, 2012). Specifically, in some ecological machine learning applications with spatially dependent data spatial cross-validation approaches have been found to be more appropriate and less likely to lead to overly optimistic results (Linnenbrink et al., 2023; Meyer and Pebesma, 2022; Schratz et al., 2019).

L83. May want to look at Schratz et al 2019 in Ecological Modelling to add to the discussion – they similarly explore spatial partitioning schemes on hyperparameter tuning for predictions with spatial data.

Schratz et al 2019 has rather been added to the passage mentioned above (originally L75) as this speaks of ecological rather than landslide susceptibility ML applications.

L84. “The majority of recent .. studies appear to apply random or no CV”. The authors' cannot state this without a reference to a review paper or them doing the review.

We changed the wording from “the majority of” to “many”.

L105. Authors state they didn’t create a susceptibility prediction map because it was event specific data. I suggest they do add one to qualitatively assess the geomorphological plausibility of the prediction results, which can provide further evidence to support one method over another.

We argue that this work is already extensive and that the creation of shallow landslide susceptibility maps is beyond the scope of this article.

L115. Goetz et al 2015 in NHESS also assess (regionally) the influence of forest structure on landslide spatial prediction (beyond landuse).

We added Goetz et al. 2015 to the reference:

However, examples of data-driven shallow landslide susceptibility studies which consider vegetational or forest variables beyond landuse are more rare (Moos et al. 2016; Goetz et al., 2015a; Spiekermann et al., 2022; Spiekermann et al., 2023).

L144-147. I would suggest removing this section. It is vague and doesn’t help explain the methods used.

We deleted this section and added a short summary of the data preparation instead:

The data preparation consisted of editing of a published shallow landslides inventory (Rüther et al., 2024), sampling of control or non-landslide points and the collection of attribute data to be included as explanatory factors.

Under Code and data availability, we write:

The Python code used for data preparation (Haualand, 2026) and R code and dataset used for machine learning modelling (Gillespie, 2026) are available in code repositories.

L149. “The statistical and machine learning”, only ML were applied in this analysis, not statistical.

We removed “statistical and” from L149.

L160. “aim to establish causal relationships”, SHAP values used in the analysis represent contribution, not causation.

We deleted “with the aim to establish causal relationships,”.

L176. Reference Fig. 7 here to point the reader to what this sampling of control points looks like.

As forward referencing is not supported, we decided to move the old Figure 7 to the section on control point sampling and it is now called Figure 2. We added reference to Figs1b and 2 in L167.

L.333. A figure/map of what the sampling partition groups look like when applied to the landslide and control data would help the reader better understand the results.

We have prepared two figures with four maps each to show the spatial folds in the full and forest-only models based on: a) spatially clustered CV, b) large block CV, c) small block CV and d) random CV. These are added to Appendix C.

L400. Figure 3. If making model comparisons, the x-axis values should be the same between each SHAP value summary plots – e.g. -0.4 to 0.75. This helps make it easier for the reader to make comparisons.

We included revised Figures 4 and 6 to show variable importance across all combinations of model structure (spatial vs. non-spatial) and CV schemes (spatial cluster, large block, small block, random) with common x-axes. Revised Figures 5a+b and 7a+b now contain Shapley value summary plots but we chose to not adjust the range of the x-axis to not dilute the effect of the figures. The main purpose of these figures is to show the difference in rankings, but also the differences in the level of detail of the SHAP value patterns. We make the following addition to the figure text:

Figure 5 a)-b) Shapley value summary plots for model 1a (spatial model with spatially clustered CV) and model 2a (non-spatial model with spatially clustered CV). Note the dynamic SHAP value ranges on the x-axes. The numbers after variable names are mean absolute SHAP values, representing relative importance. Points are coloured by normalised feature value. c)-f) Partial dependence plots for the top four variables of model 1a. Each point represents an observation and is plotted with its feature value on the x-axis and SHAP value (contribution to model output) on the y-axis. SHAP values above 0 indicate that a feature value is likely to make a positive contribution to the probability of a landslide. Points are coloured by a second feature chosen to minimise the variance of the SHAP value given both features and represents the strongest interaction. Solid red lines are LOESS smoothed lines and shading represents 95% confidence intervals.

L.562. “Shallow landslides did not occur in areas with highest rainfall intensities” – the authors interpret this as possible unreliable data. It’s probably more likely that higher rainfall intensity does not equal more landslide frequency. Most landslides will likely be triggered at threshold below the maximum rainfall intensity.

The Reviewer is correct; models show greatest landslide probability at precipitation intensities of 10-16 mm per 3 hours. We now refer to this effect of precipitation intensity in the results with reference to the revised Figures E1a and E2a. The sentence referred to now reads:

The spatial distribution of precipitation intensity during the ‘Hans’ storm (Fig 8a), suggests that shallow landslides did not occur in the areas with highest rainfall intensities which is in line with modelling results suggesting highest landslide susceptibility for 10-16 mm precipitation per 3 hours (Figs E1a and E2a).

L575. Can bedrock really explain why landslides were triggered in southeastern areas – the SHAP value for bedrock was ~0.?

The mean SHAP value for bedrock varies with different models. The way we present the results now makes this clearer (revised Figures 4 and 6 showing variable importance across all models). Bedrock has mean absolute SHAP value of 0 in the spatial model with spatial cluster and large block CV in the full models or rather small and large block CV in

the forest-only models. Its importance is relatively high in all other models. Also the dependence plot for Bedrock in Fig. 7d does show some clear delineation between bedrock weatherability classes which we think are instructive to discuss.

L585-588. Too much conjecture. Hard to explain the spatial patterns of rainfall act as spatial structure proxies in non-spatial models, especially without seeing what the spatial partitions for model tuning look like. It could be possible that the non-spatial model looked for more global patterns vs. the spatial methods that may have relied on more local patterns for model tuning (difficult to comment without seeing what the partitions look like).

We agree that our discussion of spatial confounding (citing Urdangarin et al., 2023) was not sufficiently supported by direct evidence, and that the alternative explanation raised by the Reviewer — that non-spatial models tuned under random cross-validation (CV) may optimize for more globally-averaged patterns, while spatially cross-validated models are tuned toward locally-generalizable structure — is also plausible. However, based on the additional provided analyses in Appendix D, this alternative explanation is not supported by the calculated nugget-to-sill ratios for all CV schemes. Rather, we find that the top variables in both spatial and non-spatial models consist of a mix of variables with high and moderate to low spatial structure. The discussion in 6.1 is rewritten entirely to build this argument more carefully.

L603. “flow accumulation threshold of 5000 m²” – we cannot make this statement from the plots, we can say there is a trend around their based on the modeled data, but due to many predictors and likely high correlations, we have to be cautious to over interpret the “bumps” in the partial dependence plots.

We noticed a mistake on our side as the threshold should have been 100 000 m² (flow accumulation on logarithmic scale). We agree that partial dependence plots need to be interpreted cautiously but note that flow accumulation is not subject to strong correlation with other variables. The threshold appears to be consistent across several of the tested models, the sentence is changed to:

Our results suggest a consistent flow accumulation threshold of 5000 m² throughout the study area that flow accumulation greater than 100 000 m² contributed to increased landslide release likeliness during the ‘Hans’ storm (Fig 4e). This is a relatively high flow accumulation value which would normally correspond to permanent streams (Rüther et al., 2022) and it may indicate an inventory bias towards debris floods which were supposedly excluded from the dataset (Sect. 2 and 3.1). But could also be the result of considering the maximum flow accumulation for a 30-meter buffer surrounding landslide and control points (Sect. 3.3).

L620-625. The authors may be mis-interpreting the results. They suggest the meteorological, vegetation and soil data may be too coarse to be good. It might be true, but it could also be highly correlated with elevation, tree cover or between each other, and thus have a lower ranking, or the parameters are splitting the importance. Hard to tell without a correlation analysis.

We did not imply that meteorological, vegetation and soil data are too coarse to be good, just mentioned it as a potential explanation. To avoid any misunderstandings, we changed this text passage to the following:

Elevation indirectly impacts landslide occurrence due to variations in environmental variables like temperature, rainfall regime and vegetation (Catani et al., 2013; Quevedo et al. 2022) and may thus be a placeholder for explanatory factors which are either not included or represented by imperfect datasets. ~~While meteorological, vegetational and soil parameters have been considered as explanatory factors in this study, the included datasets may be too coarse to adequately represent important variations.~~

L716. Lower ranking of other tree properties may be because they are highly correlated with tree type.

They are not highly correlated. Tree type and tree number correlation is -0.51. Tree number vs tree type = 0.07. However, we can concede that this could also be a reason.

L784. “We have shown that ML results are highly sensitive” – good to be specific here. Only gradient boosting was assessed, so cannot generalize that all ML is sensitive to CV techniques and spatial modeling choices

We changed “ML results” to “gradient boosted models”.

Anonymous Referee #2, 01 July 2026

The manuscript makes a methodologically relevant contribution to increasing the understanding of the determinants of shallow-landslide/debris flows. Its main focus is the systematic comparison of spatial and non-spatial gradient-boosted models under different cross-validation strategies, using a relatively well-documented event inventory from the 2023 'Hans' storm. The focus on spatial autocorrelation, nested cross-validation, and the instability of predictor-importance rankings is valuable, and the finding that small-block spatial CV offers a pragmatic compromise between predictive performance and residual spatial structure is well supported. It is rare to see results presented so well in a manuscript. The forest-only analysis is also a useful addition, especially because forest characteristics are rarely treated in more detail than broad land-use classes, so this was great to see. Particularly the discussion on quantitative model evaluation vs plausibility was a meaningful component of the manuscript.

Overall, this is a strong hypothesis-driven study that tests specific questions with a useful methodological contribution, and the results deliver important insights concerning different cross-validation choices. I would recommend publication after minor revisions, since a few aspects need clarification before the manuscript is fully publishable. These concern mainly the framing and context, as follows:

1. The first paragraphs of the introduction should be tightened a little. The justification for the study - by situating it in the context of global figures - seems a little off for a Northern-European study, and the text would also benefit from clearer separation between landslide hazard, exposure, impacts/loss. It is unclear whether the authors mean that landslides themselves contribute to increasing losses globally, or whether this is primarily an exposure and climate-change argument.

We shortened this first paragraph to clarify the argument. In the opening part, we toned down the emphasis on the growing global landslide risk:

Rainfall-induced shallow landslides pose a significant and escalating threat to societies worldwide. ~~Estimates from the United Nations Office for Disaster Risk Reduction suggest over 500 000 fatalities due to climate-related disasters in the first two decades of the 21st century, almost doubling compared to the last two decades in the 20th century (Mizutori, 2018). Likewise, landslide-related economic losses increased from an estimated US\$ 50 billion per year in the 1980s to approximately US\$ 200 billion per year in the 2000s (Albano and Sole, 2018). This increase is in part driven by growing exposure of humans and assets to landslides and in part by a climate change driven increase in the landslide hazard itself (Marín Rodríguez et al., 2024). In this study, we define shallow landslides as gravitational mass movements of the slide and flow type involving soft, surficial material at rooting depth; this includes debris slides, debris avalanches, debris flows and debris floods (Hungr et al., 2014). Shallow landslides can be triggered seismically during earthquakes, anthropogenically due to deforestation and constructions, by additional loading such as rockfalls as well as due to a rise in porewater pressure by rainfall and/or snow melt. The majority of shallow landslides are, however, rainfall-triggered (Sim et al., 2022); e.g. close to 90% of fatal, recent shallow landslides in Europe and Bangladesh were rainfall-triggered (Haque et al., 2016; Sultana, 2020; Sim et al., 2022).~~

In the following part, we added the argument of the necessity to improve both dynamic and static susceptibility assessments in the light of climate change. We also removed the focus on risk and therefore do not stress the necessity of runout models as this is beyond the scope of this paper:

~~Therefore, a continued acceleration of landslide risk is expected in the future as increased~~Under future climate scenarios, an increase in frequency and magnitude of heavy precipitation over mid-latitude and tropical mountainous areas will likely lead to more rainfall-induced shallow landslide events (Crozier, 2010; Jakob and Owen, 2021; Stoffel et al., 2014; Gariano and Guzzetti, 2022). Future shallow landslide hazard assessments will thus require the inclusion of rainfall variables in dynamic models (Moreno et al., 2025; Steger et al., 2024) to predict when landslides will occur as well as improved static models to predict where future landslides will occur. This calls for an urgent need to improve shallow landslide susceptibility assessments.

2. The event inventory and landslide processes require more explanation. Initially, the authors refer to the “Hans” shallow landslide inventory. Then, in L152 the terms debris slides, debris flows, and debris floods are introduced, followed by debris avalanches in L163. I read sections 1 and 2 with shallow landslides in mind, which in my experience have relatively small release areas (100-1000 m²) with reduced run-out length compared to debris flows and depending on the water content. Therefore, the differences of the named processes should be defined more explicitly. The exclusion of debris floods is reasonable, but the authors should explain more clearly how these were identified and why they are not treated as part of the same release-process population. Example photographs or supplementary figures would help. This is important for interpreting and understanding the modelling results correctly, which is currently only possible by accessing the inventory description by Rüther et al., 2024.

Our apologies that Rüther et al. (2024) is not available as open access. Unfortunately, our publisher agreement did not cover the newly created journal type *Landslide News* at the time of publication. A read-only version can be found here: <https://rdcu.be/dzhtq>

While we are happy to provide supplementary figures to clarify the mapping procedure further, our impression is that the required information can be conveyed sufficiently in text form. The confusion arises partly from the fact that our definition of shallow landslides was not specified and differs from the Reviewer’s understanding. We have therefore specified the definition in the opening paragraph of the introduction (**section 1**):

Rainfall-induced shallow landslides pose a significant and escalating threat to societies worldwide. In this study, we define shallow landslides as gravitational mass movements of the slide and flow type involving soft, surficial material at rooting depth; this includes debris slides, debris avalanches, debris flows and debris floods (Hung et al., 2014).

In **section 2**, we provide more details on how debris floods were identified in the ‘Hans’ event inventory:

Shallow landslides triggered during this storm were systematically mapped based on the dNDVI (differential Normalised Difference Vegetation Index) calculated for composite pre- and post-event satellite images with 10 m spatial resolution (Rüther et al., 2024). The mapping was validated with all available data sources including higher resolution-satellite images with 3 m spatial resolution (29% of mapped landslides), post-event orthophotos (54% of mapped landslides), as well as drone footage and/or field observations (17% of shallow landslides). The inventory data is differentiated into debris slides (smallest events for which no flow behaviour was detectable), debris flows (well-defined release area and inferred flow behaviour, including debris avalanches) and debris floods (initiated due to continued riverbed and riverbank erosion). Debris floods were distinguished from debris flows and avalanches both sedimentologically, based on sorting of deposits, and geomorphologically due to lack of levees (Bondevik and Rüther, 2025). This approach is possible for landslides validated based on post-event orthophotos and field observations, while coarser geomorphological characteristics like channel proximity and slope angle were more decisive for landslides validated with satellite imagery. Compared to the Norwegian crowd-sourced landslide database (NSDB), the systematic mapping approach resulted in a 246% increase in the number of detected shallow landslides (from 263 to 648) and a 35% reduction of the spatial bias towards the road network for the occurrence of shallow landslides within 50 m of a road (from 61.6 to 25.9%; Rüther et al., 2024). The ‘Hans’ shallow landslide inventory allows for the investigation of explanatory factors of a rare event compared to results based on general or non-event-specific inventories.

And in **section 3**, we add a short argument why debris floods are excluded from this study:

The published version of the dataset consists of 648 mapped polygons which were differentiated into 56 debris floods, 388 debris flows (including debris avalanches) and 204 debris slides. The refinement of the dataset included a re-evaluation of the placement of starting points, and a revisited differentiation into debris flows and debris floods assisted by all available data including slope angle. Inferred debris floods were then excluded from the point dataset as they initiate by continued riverbed and riverbank erosion. because they belong to a different release process population than debris slides, flows and avalanches. While debris slides are initiated at the surface by continued riverbed and riverbank erosion (Church and Jakob, 2020), the latter have well-defined release areas in which failure occurs at shallow depth due to water infiltration, rise in porewater pressure and loss in shear strength (Bogaard and Greco, 2016).

In my view the control-point sampling strategy needs further justification. The approach is in general innovative and a good one. Limiting controls by elevation, slope, and relative precipitation is defensible, but land use is strongly imbalanced, with very few control points in agricultural land – despite this likely being an important predictor which later had to be removed due to the sample size. I suggest the authors discuss the potential bias more explicitly. While I don't believe a sensitivity test including land-use stratified controls is necessary due to the inclusion of high-resolution woody vegetation products, it would be important to discuss whether the sampling strategy of control points (within the feature space of selected variables) potentially removes otherwise meaningful predictors, or reduces their otherwise expected importance. This was noted in the discussion with respect to slope (L724), but the implication of this sampling approach could also be mentioned in the methods (adding to L175/176).

We have moved the old Figure 7 to the Data preparation section on control point sampling. Along with it, we include further justification:

Our control point sampling approach ensures parity in distribution across relative precipitation, slope and elevation (red areas in Fig. 2). This means that the contribution of these predictors is reduced to effects within the observed attribute ranges and will generally be lower than for approaches with completely randomised selection of non-landslide points (Dou et al., 2023). Controlling for elevation and slope jointly has a similar effect as masking trivial areas from the absence sampling area, e.g. exclusion of certain landuse categories such as water bodies and glaciers (Schlögl et al., 2025). Moreover, as the influence of precipitation and slope on shallow landslide susceptibility is well established (Lima et al., 2022), our approach may force machine learning models to dive deeper into the effect of less obvious predictors.

In the synthesis part of the discussion, we then expand on the control point sampling argument:

In contrast, the sampling of non-landslide points does not commonly receive much attention but has been shown to have significant influence on model outcome as well (Dou et al., 2020; Hong et al., 2019). Our control points are sampled within the observed attribute ranges of slope, elevation and relative precipitation in the landslide point inventory. While we expect reduced importance for these variables, both elevation and relative precipitation are consistently important explanatory factors; the latter is underscoring the significance of incorporating rainfall relative to the local climate, in particular when the rainfall event affects large regions. The influence of slope is moderately low in our full models, which is a clear contrast to slope angle being amongst the top predictors in While 88% of recent publications on data-driven susceptibility maps find slope angle to be amongst the top predictors (Lima et al. 2022), ~~in this study the influence of slope is moderately low for the full models as control points were sampled within observed slope ranges for landslide starting points.~~ Nevertheless, slope plays a more decisive role in the forest-only models ...

The Reviewer also mentions the imbalance in the landuse categories. We don't think that it is directly linked to our control point sampling method, but this could of course be tested in future studies.

The hydrometeorological predictors need better justification. The use of a 3-day precipitation window follows the storm period, but the authors should explain why this is sufficient without including longer antecedent windows such as 7 or 14 days. Many previous studies have shown the importance of pre-event soil moisture conditions. While the inclusion of groundwater benefits the model, antecedent rainfall/wetness is also mechanistically important to shallow landslide initiation and could have influenced the results. I would recommend testing the effect of including a predictor to represent this influence of antecedent soil moisture in the model.

We agree that weather history and antecedent soil moisture are important. As eastern Norway is generally much drier than western Norway, antecedent precipitation with 7- or 14-day windows will likely not be helpful. Rather precipitation, groundwater and soil moisture are best reported relative to normal for each pixel. Such a dataset is not readily available for a 7- or 14-day pre-event window. We therefore argue that the included explanatory factor Groundwater is the best available dataset for the pre-event soil moisture conditions. It represents the modelled groundwater level for 7 August 2023 (until 8 am) relative to normal groundwater level for the same day. As the storm reached eastern Norway midday

7 August 2023, this predictor corresponds to the pre-event groundwater level. The Norwegian authorities (NVE and MET) also produce a soil moisture map (*Degree of soil saturation*), but this did not show sufficient pre-event moisture to be a valuable input to the model.

3. The discussion around soil indicators and tree type is well done, drawing extensively on previous studies. However, the results are event-specific and several variables may act as proxies for unobserved soil, land-use, or mapping effects. While this is mentioned, e.g. in Lines 800-804, the sample size is relatively small which has been shown to influence the effect size of predictors (e.g., <https://doi.org/10.1016/j.geomorph.2023.108795>). In this context, it would add value to comment on the advantages and disadvantages of event-based (e.g., knowledge of rainfall patterns) vs. multi-temporal inventories (e.g., knowledge of possible changes through time, possible attribution of landslide response to low susceptibility rather than lack of rainfall) for the development of landslide susceptibility models, which are described in the introduction as being urgently needed.

We have added the argument that small sample size is a general disadvantage in event-based inventories in the context of the discussion of the forest effects:

This may suggest that ~~satellite-derived~~ the forest inventory SR16 datasets at 16 m resolution (Astrup et al. 2019; Breidenbach et al. 2021; Hauglin et al. 2021) ~~are~~ is too coarse to adequately represent forest structure, but could also result from these variables' moderate collinearity with canopy cover and tree type. In general, the forest-only dataset was probably too small for our analysis methods, highlighting a key weakness of event-based susceptibility studies: when sample sizes are small, variables are poorly represented, model uncertainty is higher and predictors more likely to be insignificant (Smith et al., 2023). We recommend that future landslide susceptibility studies explore the importance of trees further, incorporating larger, multi-temporal inventories and more detailed forest structure datasets.

While a broader discussion of event-based versus multi-temporal inventories beyond the scope of this paper we acknowledge existing studies on the topic in the synthesis part of the discussion:

While ~~t~~The novelty in our study lies in addressing the popularity of extreme gradient boosting ML methods and suggesting a similarly robust approach in this space. ~~However,~~ this analysis is based on ~~one dataset collected during an unusual weather event~~ a relatively small, event-based inventory with associated advantages and limitations (Oliveira et al., 2024). Therefore, and we recommend that future studies should consider exploring, adapting and testing this methodological approach further afield based on larger, multi-temporal inventories and with the aim ~~With further testing, future studies can then begin to develop a more reliable machine learning protocol for landslide susceptibility.~~

Minor comments:

L31: Additional loading such as rockfalls: Please clarify what is meant by this.

Debris flows can get triggered by rockfalls, to clarify we deleted “additional loading such as”:

Shallow landslides can be triggered seismically during earthquakes, anthropogenically due to deforestation and constructions, by ~~additional loading such as~~ rockfalls as well as ~~due to~~ a rise in porewater pressure by rainfall and/or snow melt.

L34-37: As is mentioned with reference to “acceleration of landslide risk”, an increase in the magnitude and frequency of heavy precipitation events requires not only improved susceptibility assessments, but also advances in dynamic hazard and risk assessments. This is because susceptibility is aimed at predicting *where* landslides are expected to occur in future and not how the frequency and magnitude may change. The relative susceptibility of slopes is unlikely to change significantly with increasing magnitude and frequency of precipitation events. Therefore, future assessments should include rainfall predictors to support dynamic modelling (e.g. <https://doi.org/10.1029/2024jf008219>), as well as run-out models to support decision-makers in evaluating the risk to infrastructure. I suggest the authors revise this paragraph accordingly.

As mentioned in the answer to major comment 1, we changed the wording from “risk” to “hazard” and added the argument of the necessity to improve both dynamic and static susceptibility assessments in the light of climate change. We no longer mention risk and therefore do not stress the necessity of runout models as this is beyond the scope of this paper.

L88: It’s not immediately clear what is meant by “spatial structure in response data”. Perhaps add an example to clarify, (e.g. what the issue was in the inventory used by Knevels et al., 2023).

To clarify, we changed the wording to the following:

Another source of potential over-optimism in landslide susceptibility models is ~~a lack of model fitting procedures that incorporate spatial structure in response data. A~~ apart from a recent study which incorporates spatial structure in a Generalised Additive Modelling approach (Knevels et al., 2023), the inherent spatial structure in landslide datasets is rarely modelled directly.

L115: It is true that there are limited examples. However, other examples do exist, e.g. <https://doi.org/10.1016/j.geomorph.2021.107993>, [10.1016/j.geomorph.2023.108870](https://doi.org/10.1016/j.geomorph.2023.108870).

We changed the wording slightly and added Spiekermann et al., 2023; 2024 to the references.

L127: Luck is not a valid explanation, so please remove.

Luck was removed from the argument.

L133: I was unable to access Rüther et al., 2024. It would be important to briefly mention what the results of the mapping validation were. 10m resolution imagery in my experience is very coarse for mapping shallow landslide release zones which can be as large as 1 pixel. The imagery may also have implications for mapping in forested areas by obscuring the view beneath canopy. Could this have influenced the results related to tree type?

As mentioned in response to major comment 2, we made the additions to section 2 to provide more information on the mapping validation. It is correct that amongst landslides validated based on 3m resolution satellite imagery (29% of mapped landslides), the percentage of starting points in spruce is significantly lower (11.4%) while deciduous forest is higher (37.3%) than for those landslides with orthophoto and fieldwork as validation methods (spruce: 29.1%, deciduous forest: 27.9%). However, this does not necessarily suggest a mapping bias as shallow landslides which were validated with satellite imagery are commonly situated in remote and mountainous regions where deciduous forest dominates, while the valley hillslopes where new orthophotos were produced have large shares of planted spruce.

L146: Rephrase “worked better”. Sounds like the local environment was preferable due to the advanced processing required for shapefiles. I think the final sentence in this paragraph is not needed.

Upon request of Reviewer #1, we deleted this section and added a short summary of the data preparation instead:

The data preparation consisted of editing of a published shallow landslides inventory (Rüther et al., 2024), sampling of control or non-landslide points and the collection of attribute data to be included as explanatory factors.

Under Code and data availability, we write:

The Python code used for data preparation (Haualand, 2026) and R code and dataset used for machine learning modelling (Gillespie, 2026) are available in code repositories.