We thank both reviewers for their thorough and constructive assessments, which will help us refine both the framing and the methodological aspects of the manuscript. In response to their suggestions, we will tighten the opening of the Introduction by shifting from a predominantly global and climate-focused framing to a more process-based argument, and we will clarify the definition of shallow landslides in the specific context of this study. We will also clarify key concepts, add missing references, and remove ambiguous or potentially misleading wording. To address concerns about overinterpretation, we will soften interpretative language in the Results and Discussion, add supplementary correlation and factorial analyses, as well as maps of the spatial cross-validation (CV) folds for both the full and forest-only models. This supplementary material will enable us to explicitly discuss the spatial structure of key predictors and its interaction with model specification and cross-validation strategy. Guided by Reviewer #2's positive assessment of the hypothesis-driven structure and clear presentation of the results, we will keep the overall organization of the Results and Discussion while implementing these targeted clarifications and additions, which will be balanced by condensing parts of the Introduction and Discussion. We believe that these changes will substantially strengthen the transparency, interpretative robustness, and accessibility of the study.

In the following, we have copied all reviewer comments in black font, while our responses are highlighted in red colour.

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.

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.

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.

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 differential across space where landslides occur, and how these are captured by the different spatial portioning approaches, and modelling methods.

We think that all major comments raised are related and we therefore formulate our answer jointly. We agree that these are important points, and we will soften out interpretative language in the results and discussion. Further, we will provide a supplementary analysis including a correlation matrix (Appendix C) and we will prepare 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 (Figs D1 and D2 in Appendix D). We suggest that all this supplementary information is best presented in the Appendix, but single figures can of course be moved to the main text if Reviewer or Editor prefer.

While we concede the point that variable importance is sensitive to correlation, we also note that there are only two moderate correlations to speak of: between Deposit class and Deposit thickness (-0.48 – Spearman's correlation as deposit class is an ordinal variable), and the two curvature variables (-0.46). All other correlations are between -0.4 and 0.4. Elevation in particular is not involved in strong correlations because of the manner in which we controlled for elevation in our control point selection. This suggests that raw predictor collinearity is unlikely to fully explain the differences in variable importance we observe between modelling approaches. To further address the concern, we propose a more thorough supplementary analysis (presented in Appendix C) examining the spatial structure of key predictors and its interaction with model specification and cross-validation strategy. Specifically, we will: (1) quantify the spatial autocorrelation range of each top predictor (e.g., via variogram analysis) and relate this to our spatial CV block/buffer scale; (2) directly quantify the degree of train-test spatial proximity (leakage) under random versus spatial CV; and (3) compute variable importance across a full factorial design of CV scheme (random/spatial) × model structure (with/without spatial random effect). This approach should allow us to isolate whether importance shifts are driven by CV-induced leakage, by the spatial random effect absorbing predictor signal, or by an interaction between the two. We expect this analysis will clarify, for example, whether predictors such as Bedrock, Precipitation, and Deposit thickness are favoured under Non-spatial + random CV models in part because their smooth spatial structure allows leakage between nearby train/test points, while predictors with finer-scale spatial variation (e.g., Northerness, Flow accumulation) become more prominent once genuine spatial independence is enforced.

Supported by this additional analysis we will revise the Discussion (primarily section 6.1) to explicitly address how spatial heterogeneity in predictors interacts with our spatial partitioning and modelling choices, rather than relying primarily on process-based reasoning to interpret importance rankings. We further plan to shorten some of the process-based arguments (primarily in section 6.2).

With support from Reviewer #2, we choose to not make major changes to the structure of the Result and Discussion sections: "It is rare to see results presented so well in a manuscript." and "a strong hypothesis-driven study that tests specific questions", the latter referring to how the structure of the Discussion mirrors the objectives presented in the Introduction.

Minor Comments

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

We argue that context is provided in L.19-21. The abstract currently contains 96 words, so adding additional information is not possible without taking out information in other places.

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 will, however, include a clarification to show that both knowledge-driven and physically based models use (landslide inventory) data for validation, while data-driven approaches use such data both as input and for validation.

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 or 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.

Will change to a less general word, maybe Decision Tree Models.

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

We will include Brenning 2012 as references here.

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.

We will add Schratz et al 2019 as references here.

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.

Will change 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).

Thanks for pointing out a highly relevant paper, we will add it to the references.

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

We would still like to keep the reference to Github code repository, but in line with Reviewer #2's suggestion we will remove the last sentence (L146-147) and change the end of the first sentence (L145-146) to: "which was run on a local computer to allow for advanced processing of shapefiles (Haualand, 2025)."

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

We will remove "statistical and" from L149.

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

Will delete "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.

Will add reference to Fig.1b in L167. I assume that referencing forward should be avoided, otherwise reference to Fig. 7 could be added as well. We have experimented with the colour for the control points in figure 7 and conclude that orange provides best overall contrast.

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

Like mentioned in the answer to the Major Comments, we will prepare 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. We suggest that these are best shown in Appendix D.

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 makes it easier for the reader to make comparisons.

Our feeling is that this would dilute the effect of the figure – the main purpose of the figure is to show the difference in rankings, but also that the random CV non-spatial model gives almost equal importance to all predictors, rather than distinguishing between them. However, we will note the differences in axes in the caption.

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 exact triggering time is only known for a few of the ‘Hans’ shallow landslides which affected buildings and infrastructure. Therefore, total rainfall magnitude and intensity are given for the 3-day-long storm period, and the data is thus not suitable to derive intensity-duration-thresholds or draw similar conclusions. We will mention this data restriction in this part of the discussion.

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

It is the mean SHAP value that is close to 0 for Bedrock weatherability, and as this is a categorical/ordinal variable the balance of SHAP values are perhaps cancelling each other out. Inspection of the SHAP contribution plot demonstrates a clear distinction between categories. We can add this plot to the supplementary material (or take this part out).

L585-588. Too much conjecture. Hard to explain the spatial patterns of rainfall act as spatial structure proxies in non-spatial models, especially without out 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 and not currently distinguished from spatial confounding in our analysis. We will add two supplementary figures (Figs D1 and D2) showing the spatial distribution of fold assignments under both random and the three spatial CV schemes to allow readers to directly assess the partition structure referenced in this comment. Further, our planned factorial analysis (Appendix C; see response to the Major Comments above) will enable us to evaluate whether variable importance shifts are better explained by spatial confounding (absorption of spatially-structured predictor signal by the random effect), by global-versus-local pattern capture during CV-based tuning, or by both, and will revise our interpretation of the precipitation/bedrock weatherability pattern accordingly rather than attributing it to spatial confounding alone.

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.

Thanks to this comment, we noticed a mistake on our side as the threshold should have been 10 000 m2 (flow accumulation on logarithmic scale). We agree that partial dependence plots need to be interpreted cautiously but note that flow accumulation is not object to strong correlation with other variables (a correlation matrix will be included in Appendix C). The threshold appears to be consistent across several of the tested models, so we think it is worthwhile pointing out in the context of Norwegian susceptibility mapping. The sentence can be changed to: “Our results suggest ~~a consistent flow accumulation threshold of 5000 m^2 throughout the study area~~ that flow accumulation greater than 10 000 m^2 contributed to increased landslide release likeliness during the ‘Hans’ storm (upper right panel in Fig. 4 and in Fig B3).”

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.

Based on the correlation analysis (to be included in Appendix C; see response to the Major Comments above), we will be able to evaluate whether the low ranking of meteorological, vegetation and soil data can be explained with high correlations amongst these or other variables.

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 will change “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 will tighten this first paragraph and make sure to clarify the argument. This links to the minor comment to L34-37, where the reviewer correctly points out that the relative susceptibility of slopes is unlikely to change significantly with increasing magnitude and frequency of precipitation events. In the opening paragraph, we will tone down the emphasis on the global landslide hazard and risk situation and on climate change and develop a stronger process-based argument.

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. We are happy to provide supplementary figures to clarify the mapping procedure and how processes were differentiated. However, our impression is that the confusion arises from the fact that our definition of shallow landslides was not specified and differs from the Reviewer's understanding. We would define shallow landslides as umbrella term of soft, surficial material including mud, sand, earth and debris moving either as a slide (translational and rotational) or flow. Thus, following the Varnes classification of landslide types (Hungr et al. 2014) debris slides, debris avalanches, debris flows and debris floods all fall within this shallow landslide definition. Early in section 1, we will specify this definition. In section 2, we will provide more detail on how debris floods were identified in the 'Hans' event inventory (see response to minor comment L133 for details) and why debris floods are excluded from this study.

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 agree and will make such an addition in the method chapter (L175/176). The effect of the control point selection is further mentioned in section 6.1 (L544-551). The additional correlation and factorial analysis in response to Reviewer #1's comments and subsequent changes that will be made to the arguments in the discussion (particularly in section 6.1), may also help to clarify the effect of the chosen control point selection with regard to landuse and other predictors.

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. This is reflected in the included explanatory factor *Groundwater*. It is 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. We will clarify this in L225-243 in a concise way.

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 agree, and in the summary part of the Discussion (Section 6.4) we will add a brief discussion of the advantages and disadvantages of event-based versus multi-temporal landslide inventories. Smith et al. (2023) is a highly relevant reference in this context and will be included.

Minor comments:

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

Debris flows can get triggered by rockfalls. We will clarify in the text.

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.

Agreed, this paragraph will be revised to clarify the climate change effect on susceptibility and hazard. The focus of the opening paragraph will be redirected from the global landslide hazard and risk situation towards a process-based argument. This will also allow us to clarify how shallow landslides are defined in the context of this paper early on.

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).

By the phrase "spatial structure in response data" we mean that the landslide occurrences and control points are likely to be spatially autocorrelated. We will make this explicit throughout the manuscript.

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.

We appreciate being pointed to this highly relevant work and will add Spiekermann et al., 2023; 2024 to the references.

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

We will remove luck 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 a 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?

Some more details on the mapping and validation will be provided at this point. 10m resolution applies to the dNDVI (differential Normalized Difference Vegetation Index) map which formed the basis of the systematic search. Post-event orthophotos were available for 53.8% of all landslides, 17.2% of all landslides were ground-truthed based on field visits by the authors and others, while 29.1% were ground-truthed based on 3m resolution satellite imagery (Planet). For the first two categories, debris floods (including the transitions from debris avalanche/debris flow to debris flood) can be distinguished sedimentologically based on sorting of deposits and geomorphologically due to lack of levees. This distinction cannot be made based on satellite imagery, thus coarser geomorphological characteristics like proximity to channels and slope angle become more decisive.

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.

Will be rephrased to “… which was run on a local computer to allow for advanced processing of shapefiles”. The final sentence in the paragraph will be removed.