Reply to further review comments

Reply to Referee #3

Referee's Comment: The authors mix three/four different term: avalanche danger - avalanche danger level - expected avalanche danger (level). The differences are subtle, but very substantial for the results they present and I think the manuscript needs a clear statement and a consistent use of the meaning for this terminology.

Author's Reply: Thank you for pointing this out. We will go through the manuscript to check for consistent use of these terms.

Reply to Referee #4

Referee's Comment: I enjoyed reading the authors' description of a novel model chain for producing regional-scale avalanche danger predictions in Switzerland. As it turns out, my review is no longer essential to the publishing process. However, since I have already read this manuscript, I decided to add my comments as they may help improve the clarity of this paper. Ref 1 and 2 seem to have already covered many topics in their general comments. I will try to minimize multiple comments on the same topic. Take it or leave it, but if nothing else, please look at my comments In Appendix I and review your equations.

Specific comments are in the attached PDF file.

Author's Reply: We appreciate the positive review of our manuscript and the constructive comments. We will make the necessary revisions to the manuscript. Below, you will find our responses to the points raised, including addressing the specific comments provided in the attached PDF file not covered by the general comments.

General Comments

Referee's Comment: As you read the manuscript, the breakdown and roles of the different models in the model chain are unclear. Clearly stating the roles of the models, like in the conclusion earlier in the manuscript, will improve the clarity of the model.

Author's Reply: In the beginning of Section 4 (lines 143-156), we list and explain each stage in a single sentence and refer to the overview figure (Figure 2 in the manuscript) for a visual representation of the model chain structure, intended to help the reader to get a rough idea before reading the individual subsections. However, we will take up this feedback When revising the manuscript by briefly introducing the three steps in the Introduction already.

Referee's Comment: The authors go into great detail to explain the mathematical reasoning behind these models. These sections may be unclear to non-data scientists. Adding a short, intuitive explanation (like in line 168: "also known as majority voting") will clarify the manuscript.

Author's Reply: We agree that it might not be straightforward to understand the equations. We will therefore add more intuitive explanations in the the revised manuscript in the respective paragraphs, as also suggested by other reviewers.

Referee's Comment: The authors mention in several places that the GP aggregation can be used to account for terrain features. However, after they explored several combinations of terrain features derived at various scales, the most successful interpolation model relied solely on the geographical location (coordinates) and elevation (Pxyz). Consider removing some of the focus from the GP step for accounting for terrain features, as it did not add much value to the selected model.

Author's Reply: We did consider to only present the 'best' results, which were relevant for the final pipeline. However, we considered the inclusion of regional-scale terrain features a possible next step to enhance the quality of regional avalanche-danger level interpolations. Moreover, we expect that this approach may be considered by others as well. Therefore, we would rather keep these results in the paper for future reference, as it would be nice if future research could pick up on this point. Moreover,

reviewers 1 and 2 were interested in this approach and our reasoning behind it, and requested that we better describe this process. We'll make an effort to accommodate the various feedback received by carefully reviewing the respective sections.

Specific Comments from the attached PDF

Referee's Comment: Line 123: What was the strategy behind choosing the Training, validating, and testing set arbitrarily or because of the seasons' characteristics? Were the 2018/19 to 2020/21 "normal" seasons? Were all the outlier seasons parts of the training set? Please elaborate in a sentence or two.

Author's Reply: One of reasons of choosing the winter seasons from 1997/1998 to 2017/2018 for training, the winter seasons of 2018/2019 and 2019/2020 for validation, the winter season of 2020/2021 for testing is to be consistent with the work conducted by [1]. Additionally, choosing two winter seasons for the validation set allows for a more robust selection of the best interpolation and aggregation methods. Ideally, we would have also used two winter seasons for the test set. However, at the time, we did not have access to the curated data for the most recent seasons.

Referee's Comment: Line 224: Why did you chose RBF? did you test other kernels?

Author's Reply: We did not explore other kernels in depth. We opted for the RBF kernel due to its popularity and the fact that it is infinitely differentiable, leading to a very smooth Gaussian process.

Referee's Comment: Line 242: How many AWS do you have in one square km? Is the GP step only applied to those grid cells with several AWS? How did you treat terrain features typically much smaller than this grid cell?

Author's Reply: In a single 1 km square grid cell, there cannot be two or more AWS. The GP is fitted with the expected danger level at the location of AWS, and in a next step used to predict the expected danger level for every grid cell. All terrain features are extracted from a 25 m resolution DSM after resampling the DSM to 1 km.

Referee's Comment: Line 389: This is interesting. Did one or more of the validation years and outlier winter?

Author's Reply: We did not delve further into examining the performance gap. Nonetheless, the validation years (i.e., winter seasons of 2018/10 and 2019/20) coincide with those in [1]. As mentioned in line, we suppose that the gap is caused by the distinct versions of SNOWPACK used to compute the input features for the RF classifier.

Referee's Comment: Line 432: Needs to italic

Author's Reply: Thank you. We'll change accordingly.

Referee's Comment: Line 553: Adding a simple none technical description to precision will improve the clarity of the manuscript for most people. Maybe something like: The proportion of correct positive identifications by the classifier.

Author's Reply: We agree. We'll make the suggested change.

Referee's Comment: Line 554: I believe it should be TP/(TP + FP). Please verify and correct.

Author's Reply: You're right, thank you for pointing that out. We'll change accordingly.

Referee's Comment: Line 555: See comment above, consider adding more inventive description like: The proportion of actual positives that were identified correctly by the classifier.

Author's Reply: We agree and will make the suggested change, similar to the description for precision.

Referee's Comment: This is somewhat unclear formula. Consider changing it to: 2*precision*recall/(precision + recall) for better clarity.

Author's Reply: We chose this version of the formula, since it relates the F1 score to the harmonic mean, usually defined as $\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$, of precision and recall. We will add this relation for a more intuitive description, and additionally state the alternative formula, as you suggested.

Referee's Comment: Line 562: This title is confusing. Consider changing it to something like model chain cost function or evaluation function/metric.

Author's Reply: Thank you for bringing this to our attention. We agree, and we will change to a more suitable title for this section.

References

[1] C. Pérez-Guillén, F. Techel, M. Hendrick, M. Volpi, A. van Herwijnen, T. Olevski, G. Obozinski, F. Pérez-Cruz, and J. Schweizer. Data-driven automated predictions of the avalanche danger level for dry-snow conditions in Switzerland. *Natural Hazards and Earth System Sciences*, 22(6):2031–2056, 2022.