Reply to Referee #2

Referee's Comment: The manuscript present a series of methods to extrapolate point computations of avalanche danger from Pérez-Guillén et al, 2022 over space and determine a avalanche hazard for all forecasting regions of Switzerland (the minimal units used by Swiss avalanche forecasters to produce bulletins on dynamical areas depending on the situation). The goal of the method is to produce an automatic forecast of the avalanche danger level on Switzerland from snow modelling operationally run on points (automatic weather stations). The paper is nevertheless limited to dry snow problems, while wet snow or mixed dry/wet snow avalanche problems may contribute to the overall hazard, but this is clearly acknowledged. The goal of the paper as well as the overall presentation is well suited for the readership of GMD.

The manuscript clearly present the methods, is quite well organized an easy to read and present interesting insights into scale changes for avalanche forecasting (from point scale to 1km grid and forecast regions). The evaluation seem quite complete with different scales treated (regional, global, daily or seasonal, by avalanche danger...) and give a great overview of advantages and drawbacks of the presented work. The provided code seem clear and usable. I have mainly minor comments that I detail below and that can be considered by the authors before final publication.

Author's Reply: We greatly appreciate the positive and thorough review of our manuscript and the constructive comments. We will revise the paper accordingly. Our responses to the suggestions and the intended revisions are detailed below (in blue).

General comments

Referee's Comment: Work of Pérez-Guillén do not have to be presented again, it is an input of your study and you can point to the published paper for details. However, you may give a focus on changes made from the published method. Sometimes you re-explain the model used by Pérez-Guillén which does not seem necessary to me. However, these parts are not sufficiently important to prevent general comprehension of the paper and added value of this work. I detail most useless parts in the detailed comments.

Author's Reply: We agree that there are some re-explanations related to [2] classifier and the data preparation strategy. We did so on purpose as we wanted to repeat the most important elements from [2] with the objective to facilitate the understanding of the model without necessarily requiring the reader to consult [2]. When revising the manuscript, we'll make an effort to reduce the re-explanations to a minimum.

Referee's Comment: The spatial resolution chosen for extrapolation is a 1km grid. Coarser resolutions are tested and not selected. However, nothing is said of finer resolution whereas complex topographies in mountainous regions are known to be poorly represented at coarse resolutions. Authors then use advanced methods to compute topographic variables to reduce the impact of a coarse resolution (such as Gaussian Pyramids, which is of high interest) but never discuss why they selected a 1km resolution and if their model could be used at a finer resolution, which would be of interest for the reader and for further uses of such method.

Author's Reply: We opted not to explore lower resolution grids primarily due to how the training data was compiled for the RF classifier training in [2]. Specifically, the official avalanche bulletin indicates a regional avalanche danger level on a scale significantly larger than 1km. Hence, the predictions of the RF classifier at AWS should not be interpreted as bare point-predictions but rather valid for areas close to the AWS. Consequently, considering smaller grid cells would only add computational complexity. However, the basic concept of the model pipeline could certainly be applied to finer-resolution grids if the point predictions from the initial classification stage are more spatially accurate.

Referee's Comment: The avalanche danger scale is not linear. Difference between level 3 and 4 is much higher than than difference between risk 1 and 2. Does this influence the results when computing expected danger level (equation 3) and how does this impact the different methods you used? This could also be the main reason explaining the mean method performs poorly in Fig. 5. I have seen no discussion of this important characteristic of the data you manipulate.

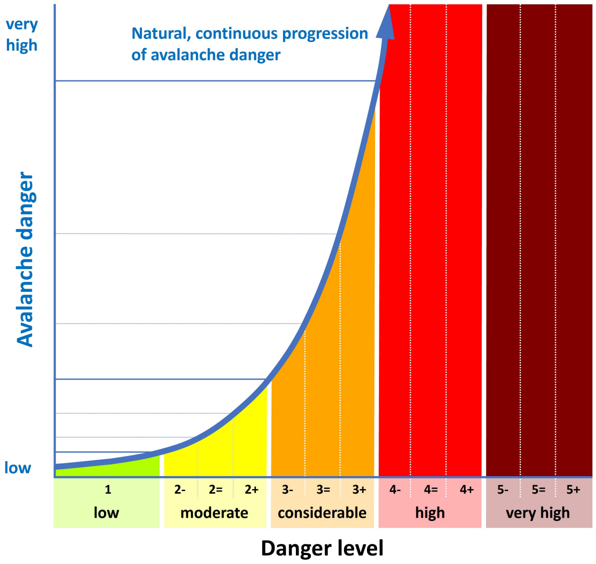

Figure 1: Sketch highlighting the relationship between avalanche danger levels (x-axis) and avalanche danger (y-axis) (figure taken from SLF website - bulletin interpretation). In addition to the danger levels, sub-levels as used in the Swiss forecast [3, 1], are indicated.

Author's Reply: Although the exact shape of the function is unknown, it is correct that avalanche danger (or the severity of avalanche conditions) increases exponentially with the avalanche danger level resulting in a relationship similar to Figure 1. However, our focus is solely on avalanche danger level (x-axis in Figure 1). In other words, the expected danger level is calculated for the levels, not for danger, thereby maintaining the non-linear relationship when calculating the expected danger level (y-axis). We agree that providing more of an explanation will be helpful, and will, therefore, add a small paragraph explaining this concept.

The mean method's poor performance is attributed to how to account topography. This is due to the fact that within a warning region, there is a higher number of low-elevation cells compared to high-elevation cells. Typically, as we mention in line 251-252, the low elevation zones are assigned to a lower danger level, leading to underestimation of the regional avalanche danger level (mean is biased towards lower danger levels than those applicable where actual dangerous conditions are). This motivated development of an elevation-based aggregation strategy, aimed at mitigating this issue by averaging grid cells within specific elevation ranges.

Referee's Comment: Some of the methods are presented in the results rather than in the material and methods section. For instance, we discover the partition in different areas in Fig. 6 in the results or the presentation of F1 score that appear only in section 6.

Author's Reply: The warning regions used in the Swiss forecast are introduced in Figure 1, the aggregation of warning regions for the purpose of model evaluation is shown in Figure 6. While the warning regions are relevant for the model chain, the latter are only used for model evaluation and not the model chain per se. We decided to introduce this particular division in Section 6 since we utilized it solely for evaluation purposes. This enabled us to present these regions and their respective results more seamlessly in the manuscript. When revising the manuscript, we will include references to the corresponding figures to help the reader locate this information.

In terms of introducing evaluation procedures (e.g., LOOCV) and metrics (e.g., accuracy, F1-scores), we find these to be quite standard practices, hence an introduction on the fly suffices in our opinion. However, we do include a definition of the scores in the appendix, and we will clearly point the reader to it.

Detailed comments

Referee's Comment: Line 109-110: "from a more recent operational SNOWPACK version": please be specific and provide clearly the identification of the code used (release number or git tag or commit) here or in the code and data availability section. You can also briefly explain if there is major changes between yours and Pérez-Guillén version.

Author's Reply: We agree, and will include this information in the data section. There are no other major differences from the work of [2].

Referee's Comment: Section 4.1 and 4.2 may be rewritten more straightforwardly. Authors introduce a lot of mathematical notations that are not used elsewhere. In particular, the mathematical description of random forest seem to be out of the scope of this paper. You can directly refer to Pérez-Guillén et al., 2022 and/or Breiman, 2001.

Author's Reply: We agree that Section 4 contains a lot of mathematical reasoning behind the methods used in the model chain. However, it gives the necessary background to understand crucial design choices, and ensures that the paper remains self-contained. For instance,

- Understanding the basic theoretical foundation of the RF classifier is essential for grasping the concept of the expected danger level, which we use as the target for interpolation. Moreover, breaking down the RF classifier into weak estimators (Equation 1) allows us to reason about adjusting the discretization thresholds in lines 278ff.
- Regarding the mathematical background on Gaussian processes, we feel that it is necessary to better justify the noise model, the constant mean function to avoid target standardization, and point out that the most crucial part of GPs are defining the kernel function.

However, Reviewer 4 made a similar recommendation, we aim to incorporate additional intuitive explanations, facilitating the understanding.

Referee's Comment: On section 4.2, several sentences present generally the interpolation method. The reader may be helped by having a presentation of exactly what you do in the paper immediately after the introduction of each notion rather that keeping general ("One of the most popular and widely used kernel function" may be transformed as "we used the most popular kernel function which is...", same for "can refer to geographical location" or "one can construct kernels").

Author's Reply: Thank you for your feedback. We will make adjustments and try to make text more streamlined.

Referee's Comment: On Figure 3b, the big red dots are not informative and prevent for viewing the background data that is the result of your method especially in the Alps area. Maybe you can keep the dots but unfilled or reduce their size.

Author's Reply: Indeed, the dots are a bit too big with this figure size. We will make the adjustments to the plot in the revised manuscript.

Referee's Comment: I am not sure I fully agree with the statement line 265 : "danger level for dry-snow avalanche increases with increasing elevation". Do you have data or references for that? For instance, situations with persistent weak-layers at mid-altitudes that are not present at higher altitudes are not so uncommon.

Author's Reply: It is correct that avalanche danger doesn't always increase with elevation as is highlighted in the example mentioned by the reviewer. However, in the Swiss forecast, only one elevation threshold is indicated. For dry-snow avalanche conditions, this is always the lower boundary of where the indicated danger level prevails. For level 1 no such information is provided in the forecast. In other words, avalanche danger describing dry-snow avalanche conditions generally increases with elevation in the Swiss forecast. Moreover, the statistical analysis by [4] also shows that, in general, avalanche danger (or avalanche risk in the case of [4]) increases with elevation. - We'll add a statement along the line that 'in general' dry-snow avalanche danger increases with elevation.

Referee's Comment: On the interpretation of Table 1: differences are very small between the different results. Do you have some clue to think that they can be significant? If yes, please provide and if no, you may underline the uncertainty in the interpretation (line 327-334).

Author's Reply: Indeed, the differences between the interpolation models are very small, but they are consistent with respect to the remaining scores/errors of Section 5. In lines 338-339, we underline this consistency between the LOOCV errors and the statistical properties of the learned coefficients of the kernel function. Similarly, for the overall performance of the model chain (see lines 356-358).

Referee's Comment: Line 377: you use only one year for evaluation. As snow coverage can largely vary between years, how does this influence your results. In particular, I suspect that this may have a larger impact on small areas with few observations and a rather tight diversity of snowpacks such as the Jura area.

Author's Reply:

We opted to use two winter seasons (i.e., winter seasons 2018/19 and 2019/20) for model selection/calibration (see Section 5.2, 5.3), to ensure a more robust model, because of possible seasonal variations you mentioned. We agree that an evaluation of the best model (Section 6) across multiple seasons would be beneficial, but at the time of the analysis we only had access to curated data until winter season 2020/21, so we had to make this particular choice of splitting the data.

We have not analyzed in detail how the varying snow coverage between years correlate with performance of the model chain. However, the snow coverage dictates the amount of usable weather stations, since we only consider weather stations at locations with snow for a given day. In the Jura, only two of the five stations are located at higher elevation (around 1500 m a.s.l.), leading to hardly any sampling points for a given day. Consequently, the interpolation (or rather extrapolation) often proves to be inaccurate in this area, as you suggested.

Referee's Comment: Figure 7 and 9: All the bars are not directly comparable as RF is evaluated on points and other on forecasting regions and the number of forecasting regions varies. It may be interesting to specify the number of regions/simulation points on these graphs.

Author's Reply: As the mentioned figures will become too cluttered when adding more information, we'll do the following:

- We'll indicate the number of warning regions in each of the climate regions in lines 379-385, potentially including additional details regarding the average number of sampling locations per day.
- We'll provide an indication on the number of forecast days contained in the validation and test data sets in lines 120-125 when introducing the data splits.

References

[1] C. Lucas, J. Trachsel, M. Eberli, S. Grüter, K. Winkler, and F. Techel. Introducing sublevels in the swiss avalanche forecast. In *International Snow Science Workshop ISSW 2023, Bend, Oregon, USA*, 2023.

[2] 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.

[3] F. Techel, S. Mayer, C. Pérez-Guillén, G. Schmudlach, and K. Winkler. On the correlation between a sub-level qualifier refining the danger level with observations and models relating to the contributing factors of avalanche danger. *Natural Hazards and Earth System Sciences*, 22(6):1911–1930, 2022.

[4] Kurt Winkler, Günter Schmudlach, Bart Degraeuwe, and Frank Techel. On the correlation between the forecast avalanche danger and avalanche risk taken by backcountry skiers in switzerland. *Cold Regions Science and Technology*, 188:103299, 2021.