Reply to Referee #1

Referee's Comment: This manuscript presents a model chain for producing regional-scale avalanche danger predictions in Switzerland. The key contribution extends a point-scale danger model (Pérez-Guillén et al., 2022) to a regional scale by interpolating across a continuous grid and aggregating within predefined regions. The interpolation and aggregation methods aim to capture relevant processes that influence avalanche danger while aligning with approaches used by human forecasters so that the model chain can be applied as a decision-support tool. The presentation of this model chain is an interesting topic that fits well within the scope of GMD.

The manuscript does an excellent job of communicating a complex topic with clarity and logical progression. It establishes clear objectives, employs sound methodological choices, and draws fair and relevant conclusions applicable to operational avalanche forecasting. I think a few details could be further clarified (explained below), but otherwise recommend the publication of this manuscript.

Author's Reply: We appreciate the positive review of our manuscript and the constructive comments. We will revise the paper accordingly. Our responses to the suggestions are detailed below (in blue). For the main comments, we point to the changes in the revised manuscript with corresponding line numbers. Adjustments in the revised manuscript are displayed in italics. For remaining minor corrections, refer to the marked-up manuscript version showing all modifications.

Specific comments

Referee's Comment: Representativeness of the stations. Providing additional information about the stations and snowpack simulations would help readers understand how effectively the training data represents the variability of avalanche conditions within a region. While Pérez-Guillén et al. (2022) likely address some of these details, including more information would offer valuable insights. For example, the number of stations in the dataset, the nature of the simulations (flat field and/or virtual slopes), and whether wind transport was simulated. Without such details, it remains unclear how well the stations capture the full range of expected conditions within each region and how this might impact the resulting predictions. Can we expect this method to predict the most unstable slopes in a region and if not does this create a bias? How well can the interpolation routine capture snowpack conditions not represented in the input data?

Author's Reply: We acknowledge that some information is currently absent. Specifically, we will include the number of data points and stations for the various splits of the dataset (training, validation and test set), as well as the precise configuration and version of SNOWPACK employed. Yes, wind transport was indeed simulated and is one of the input features of the RF classifier. For these details, including a comprehensive list of extracted features that describe the snowpack conditions, we will refer to [1] where appropriate.

Regarding the prediction of the unstable slopes, one has to consider the way the training data was gathered for the RF classifier in [1]. Specifically, it uses the official avalanche bulletin, which indicates a regional avalanche danger level on the scale of warning regions, and reflects that forecasters consider the most unstable slopes in this area when deciding on a danger level. Warning regions are of greater scale than a single slope. 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. This principle extends to the interpolation of these predictions, explaining the decision not to employ a grid with a resolution finer than 1km. Consequently, our model is unable to predict the most unstable slopes. We have included a justification for selecting the 1 km resolution grid in the revised manuscript which can be found in lines 135 f.:

"However, since the avalanche danger level in the training data was typically assessed on a scale of warning regions, adopting finer resolution interpolation grids would unnecessarily increase computational complexity. Therefore, the DSM is downsampled to 1 km × 1 km raster cells by simple averaging."

Furthermore, our proposed methodology interpolates avalanche danger level directly across a 1km resolution grid, using features from the digital surface model, particularly the geographical location and elevation. There is no explicit notion of snowpack conditions in the interpolation, which assumes that this information is intrinsically contained in the danger level values used for model training and predicted at each station location. As such, if conditions differ greatly between slopes and what is captured by the weather stations, model performance will inevitably be lower. We've added the following explanation on this relevant point in the discussion section (lines 499 f.):

"Moreover, if the conditions captured by the AWS and SNOWPACK differ significantly from those on the nearby slopes, it is expected that the RF classifier may generate more inaccurate predictions."

Referee's Comment: Terrain features. The selection of terrain features for the interpolation routine should be explained in more detail. It is not entirely clear which features are derived from the DSM, nor is the meaning of directional derivatives, difference of Gaussians, and Gaussian pyramids (lines 137 to 141). Some plain-language explanations of what these derived variables are and how they potentially relate to avalanche danger would help. Also, some clarification is needed regarding the interpretation of slope angle, curvature, and aspect at the coarse scale of 32 km^2, and why these are expected to be relevant. Further explaining the terrain variables would provide readers with important context for interpreting the results.

Author's Reply: Thanks for pointing this out. Indeed, these features represent the feature space on which the interpolation algorithm models the danger level. However, the best model only make use of the location and elevation, and not of a more complex representation of the terrain. We will add further detail and high-level interpretation of these features, but to balance out a request from Reviewer 4, we will keep this short and to the point. The paragraph in the revised manuscript (lines 145-157) reads as follows:

"Specifically, we extract elevation, slope angle, profile curvature, and the aspect from the resampled 1 km resolution DSM. Then, the technique of Gaussian pyramids (Adelson et al., 1984) is applied for the features elevation, slope angle, and profile curvature to capture patterns at lower resolution (2 km^2 - 32 km^2) in the scale of long mountain ridges, mountain groups, plateaus and valleys. Gaussian pyramids are build by constructing a sequence of images in which the resolution of the next image is half of the resolution of the previous image in the sequence, while a Gaussian filter is applied before the down-sampling operation.

Finally, these features are complemented by extracting directional derivatives and differences of Gaussian's (DoG) (Gonzalez and Woods, 2006). Both techniques are commonly used for detecting and enhancing edges and corners in image processing, thus with regard on topology, aiding capturing valleys and ridges effectively. Directional derivatives are extracted by applying a Sobel operator (Sobel and Feldman, 1973) on a blurred DSM (i.e., Gaussian filter), focusing on the north-south and east-west directions. On the other hand, DoGs are computed by subtracting two blurred versions of the DSM. Different degrees of blurring are taken into account for both of these features."

Referee's Comment: The methods section (Sect. 4) has extensive use of mathematical symbols, some of which may be excessive and cause confusion rather than clarity. This is simply a personal preference, but I think it would be clearer to use more plain language and then use symbols strategically where it helps communicate mathematical relationships. Also please check all symbols are unique and defined (e.g., alpha is used differently in line 251 vs alpha in line 281, Ne in line 263 is not defined).

Author's Reply: We agree that Section 4 contains a lot of mathematical reasoning behind the methods used in the model chain. We will add some plain language and high level explanations making this section more accessible to the reader. However, it gives the necessary background to understand crucial design choices, and ensures that the paper remains self-contained. For instance:

- Understanding the theoretical foundation of the RF classifier is essential for grasping the concept of the expected danger level, which we use as a 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.

Furthermore, we acknowledge that there was an oversight in defining certain symbols, particularly N_e

and N_g. We have addressed this issue in the revised manuscript, including resolving the ambiguity around the use of α. The revised sections read as follows:

- (lines 265 f.): *"Let $\mathcal{D}_{grid} = \{(\mathbf{s}_i, d_i)\}_{i=1}^{N_g}$ be this grid compromising N_g danger level assessments, [...]"*
- (line 282): *"Consider an ordered set of elevations $\{e_j\}_{j=1}^{N_e}$, containing N_e elements where $e_i \leq e_j$ for $i \leq j$."*
- (lines 299 f.): *"For instance, predicting a class probability of $p = 1$ requires that all base estimators predict the same class. Consequently, as the RF classifier predicts danger levels 1-low to 4-high, the expected danger level typically falls within the range of $[1 + \epsilon_1, 4 - \epsilon_2]$, for some $\epsilon_1, \epsilon_2 > 0$."*

Referee's Comment: I agree with the approach to evaluating performance with mean/median accuracy, however, am curious if there were any directional biases in the model in terms of over or under-predicting danger (e.g., for specific regions or danger levels). While this doesn't need to be fully presented, it would be interesting to comment if this was investigated.

Author's Reply: We conducted a brief analysis of the over- and under-prediction of danger levels using confusion matrices, though these findings were not included in the manuscript, but available in the code repository. Specifically, we noted an under-prediction for danger level 3 and an over-prediction for danger levels 1 and 2 during the winter seasons of 2018/19 and 2019/20 (validation set). Similar trends are depicted by [1] in Figure 6a) for the same winter seasons.

Technical comments

Referee's Comment: Line 66: "built" not "build".

Author's Reply: Thanks for pointing this out.

Referee's Comment: 1: The IMIS and ZERO-DL station networks are not defined/described anywhere in the manuscript.

Author's Reply: Thanks for spotting this. We have made sure that IMIS and ZERO-DL are introduced appropriately.

Referee's Comment: Line 109-11: Data extraction times are unclear. Public forecasts are valid until 17 LT, snow cover data is extracted at 12 LT, but then why is resampled meteorological data centered around 18 LT? Wouldn't it make sense for all data to be extracted at a single time?

Author's Reply: The meteorological time series has a 3-hour resolution, making 18:00 LT the closest match to the forecast publication time of 17:00 LT. Extracting features precisely at 17:00 LT would require interpolation and could potentially introduce bias, instead of simple averaging over a moving 24-hour window. Additionally, we decided to adopt the exact same data pre-processing strategy as [1], which include snow profile data extraction at 12:00 LT, to make sure that the interpolated product matches exactly the prediction model (i.e, the RF classfier).

Referee's Comment: Line 121: Perhaps state the total dataset size (e.g., number of station-day-danger points).

Author's Reply: Good point, we have added the total size for each set (training, validation, and test set).

Referee's Comment: Line 130-140: It is not clear how extracting terrain features at a scale of 1 to 32 km2 is capturing the smaller scale topographic properties you say influence avalanches at scales to tens to hundreds of metres. Did you derive slope angle, profile curvature, and aspect from the 25 m DSM and then upscale to coarser grids? Perhaps more details would clarify how terrain characteristics are being captured in the model.

Author's Reply: We derive slope angle and profile curvature based on the 1km DSM, and subsequently upscale them to coarser grids of scale 2km to 32km via Gaussian pyramids. Feature extraction is briefly outlined in lines 125-140, albeit in a rather generalized manner. We have made this section clearer, but following the input from Reviewer 4, we kept this concise.

Regarding feature importance, we mention in line 134, that, "It is less clear whether such properties, derived for larger scales, correlate with regional avalanche conditions." However, in Section 5.2, we assess the significance of the extracted terrain features using Leave-One-Out Cross-Validation (Table 1) and by analyzing the learned kernel combination coefficients (Table 2). Ultimately, we determined that, apart from location and elevation, only the slope angle and profile curvature shows some significance (see lines 341-342). Despite this, we exclude the these additional features from the final model due to its negligible impact on performance.

Referee's Comment: Line 140: A brief plain language description of the Gaussian pyramid technique would help.

Author's Reply: We have elaborated more on the terrain features by providing a more intuitive description. However, in response to the feedback from Reviewer 4, we kept this concise.

Referee's Comment: Fig 2. In the interpolation section, the "etc." in terrain features is confusing as the methods only list location, elevation, slope angle, curvature, and aspect. Does "etc." mean to capture the directional derivatives, DOG, and Gaussian pyramids?

Author's Reply: We acknowledge the potential for confusion and have modified the figure to include the list of all terrain features.

Referee's Comment: Sect 4.3: It is not clear that three distinct methods were tested (mean, top-alpha, bands). When reading it can be interpreted that top-alpha and band averaging are done in conjunction, rather than two distinct methods.

Author's Reply: Thanks for the feedback, we have ensured greater clarity in the revised manuscript.

Referee's Comment: Line 366-359: Perhaps I misunderstood the method, but I don't see how the elevation bands overlap. I would have assumed when you increase the bandwidth you decrease the number of bands accordingly to avoid overlap. What is the motivation for allowing overlap?

Author's Reply: The general definition of the elevation-based aggregation strategy (lines 263ff.) allows to choose an ordered set of elevations $\{e_j\}_{j=1}^{N_e}$, and a bandwidth b defining the elevation bands $[e_j - b/2, e_j + b/2]$. The bandwidth parameter b is independent of chosen elevations, making it possible to define overlapping bands. For instance, consider the strategy *elev-full* (defined in line 353) with a bandwidth of 300 m, leads to elevation bands: $1200 \pm 150, 1400 \pm 150$, etc.

Overlapping bands might be advantageous when estimating the danger level every 100 m (or even 50 m) instead of every 200 m. By overlapping the bands, a more accurate estimate can be obtained, as wider bands typically includes more grid points, Nevertheless, we have not conducted empirical testing on this aspect. Ultimately, our aim was to keep the definition of the elevation-based aggregation strategy as general as possible.

Referee's Comment: Line 356: Is the "mean method" defined or labelled anywhere? I think the meaning of this method is intuitive but slightly confusing if it is not explicitly defined/labelled anywhere.

Author's Reply: Thanks for pointing this out, it was not explicitly labelled as such, we only introduced the simple averaging method in line 250. We have explicitly labeled the method in the revised manuscript accordingly.

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.

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). For the main comments, we point to the changes in the revised manuscript with corresponding line numbers. Adjustments in the revised manuscript are displayed in italics. For remaining minor corrections, refer to the marked-up manuscript version showing all modifications.

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've made 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. We have included a justification for selecting the 1 km resolution grid in the revised manuscript which can be found in lines 135 f.:

"However, since the avalanche danger level in the training data was typically assessed on a scale

of warning regions, adopting finer resolution interpolation grids would unnecessarily increase computational complexity. Therefore, the DSM is downsampled to 1 km × 1 km raster cells by simple averaging."

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.

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 an explanation will be helpful, and have therefore added a small paragraph explaining this concept in lines 190f. of the revised manuscript:

It is important to emphasize that avalanche danger (or the severity of avalanche conditions) increases exponentially with the avalanche danger level. However, the expected danger level (see Eq. 3) is determined based on the levels rather than the danger, thereby maintaining the non-linear relationship.

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.

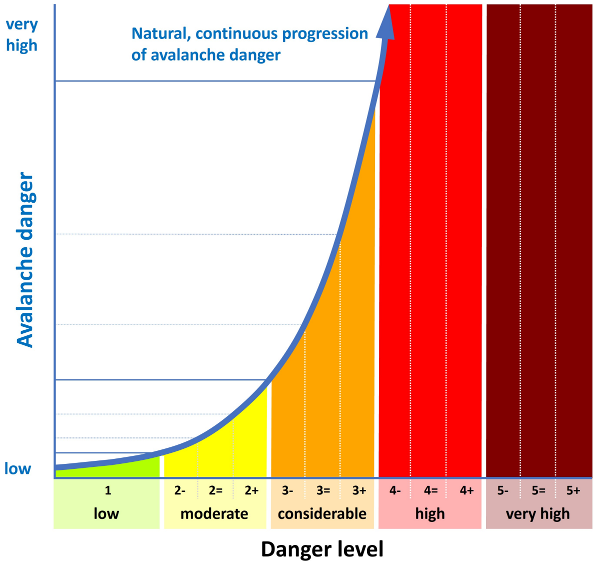

Figure 1: Sketch highlighting the relationship between avalanche danger levels (x-axis) and avalanche danger (y-axis) (figure taken from SLF website - bulletin interpretation). In addition to the danger levels, sub-levels as used in the Swiss forecast [3, 1], are indicated.

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.

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 have included 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 have made suggested adjustments in the revised manuscript.

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 have made 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 now say the following (lines 284 f.)

"The danger level for dry-snow avalanches (as opposed to wet-snow avalanches) ***typically*** *increases with increasing elevation, determining the maximum danger level iteratively from the bottom to top provides an elevation threshold for particularly affected altitude range as in human-made forecasts."*

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've indicated the number of warning regions in each of the climate regions in Figure 6.
- We've provided an indication on the number of forecast days contained in the validation and test data sets in lines 125 f. 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.

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've checked the revised manuscript for a 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. For the main comments, we point to the changes in the revised manuscript with corresponding line numbers. Adjustments in the revised manuscript are displayed in italics. For remaining minor corrections, refer to the marked-up manuscript version showing all modifications.

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 took up this feedback and briefly introduce the three steps in the introduction as follows (lines 34 f.):

"Inspired by Brabec and Meister (2001)'s ideas for regional avalanche forecasting, we develop and validate a three-stage model pipeline for regional avalanche danger forecasting (RAvaFcast v1.0.0), comprising the stages Classification, Interpolation and Aggregation. Concretely, we propose an interpolation algorithm allowing the prediction of high-resolution danger level maps for the Swiss Alps based on point-predictions at the locations of the automated weather stations, where the RF classifier from Pérez-Guillén et al. (2022a) infers danger levels. Then, a novel elevation-based *aggregation strategy infers an avalanche danger level for predefined warning regions, to ultimately produce a regional avalanche forecast that mimics human forecasts. Lastly, we compare the model's predictive performance to the point-based approach used by Pérez-Guillén et al. (2022a), and importantly to the published avalanche forecast bulletins."*

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 have therefore made efforts to add more intuitive explanations in 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 have made an effort to accommodate the various feedback received by carefully reviewing the respective sections. In particular we revised the description of extracted terrain features as follows (line 145 f.):

"Specifically, we extract elevation, slope angle, profile curvature, and the aspect from the resampled 1 km resolution DSM. Then, the technique of Gaussian pyramids (Adelson et al., 1984) is applied for the features elevation, slope angle, and profile curvature to capture patterns at lower resolution (2 km^2 - 32 km^2) in the scale of long mountain ridges, mountain groups, plateaus and valleys. Gaussian pyramids are build by constructing a sequence of images in which the resolution of the next image is half of the resolution of the previous image in the sequence, while a Gaussian filter is applied before the down-sampling operation.

Finally, these features are complemented by extracting directional derivatives and differences of Gaussian's (DoG) (Gonzalez and Woods, 2006). Both techniques are commonly used for detecting and enhancing edges and corners in image processing, thus with regard on topology, aiding capturing valleys and ridges effectively. Directional derivatives are extracted by applying a Sobel operator (Sobel and Feldman, 1973) on a blurred DSM (i.e., Gaussian filter), focusing on the north-south and east-west directions. On the other hand, DoGs are computed by subtracting two blurred versions of the DSM. Different degrees of blurring are taken into account for both of these features."

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've changed 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've implemented 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've corrected the equation.

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 implemented 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 have added this relation for a more intuitive description, and additionally stated the alternative formula.

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 have selected 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.