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).

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.

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 will add an explanation on this relevant point in the discussion section.

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.

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 will address this issue in the revised manuscript, including resolving the ambiguity around the use of α.

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 will make 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 will add 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 will make this section more clearer, but following the input from Reviewer 4, we will keep 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 will elaborate more on the terrain features by providing a more intuitive description. However, in response to the feedback from Reviewer 4, we will keep 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 will modify 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 will ensure 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 will explicitly label 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.