Response to review by Florian Herla

Frank Techel, and co-authors

Dear Dr. Florian Herla,

we greatly appreciate your detailed and constructive feedback to our manuscript. Please find our responses below (in blue).

1 Summary

The manuscript titled "Forecasting avalanche danger: human-made forecasts vs. fully automated model-driven predictions" presents a novel approach for evaluating the performance of avalanche hazard assessment models, a highly relevant topic within the scope of NHESS. By leveraging data sets of natural and human-triggered avalanches as well as GPS tracks of backcountry users, the authors pursue a statistical exercise to address two main question. (1) Can spatially interpolated model predictions of avalanche danger and snowpack stability reflect observed avalanche activity, and (2) How well do the automated predictions perform relative to human-made avalanche bulletins. The authors conclude that the model-predicted probabilities correlated strongly with their proxy variable for the probability of avalanche release, and that the model predictions discriminate between different avalanche hazard situations as well as the human-made bulletins. These are substantial findings that the authors introduce and discuss well in the context of underlying assumptions, existing literature, and the future of avalanche forecasting.

I have one main comment that could help make the manuscript even stronger. The comparison between human and model performance in discriminating between different conditions (Fig. 5) is not completely independent. In L254ff the authors explain how the model predictions are tied to the human predictions, and in L381–383 they discuss that this could reduce the estimated model performance. To make the present approach more transferable to other countries and the results more illustrative in general, I would greatly appreciate two numerical experiments that simulate (a) less-quality bulletin data and (b) worse model predictions. In the first step, all data could be held equal except for the reported danger rating, which could be perturbed with a given standard deviation. In the second step, only the model predictions would be perturbed. This experiment would add another figure similar to Fig 5, which tells us (a) whether this approach will always cap model performance at the level of human performance, and (b) how a significantly worse model prediction would line up on this rather abstract visualization. This new figure could help the reader appreciate the strong results even more, and help other warning agencies to assess whether this evaluation strategy is suited for their contexts (e.g., less consistent and accurate danger rating data).

We understand this suggestion. However, this will add another layer of complexity to an already rather complex manuscript. Moreover, perturbing the human forecast isn't all that straightforward as - on a given day, the error will be the same in an aggregate of warning regions, but it may be different (or not), in other warning region aggregates. However, we will add these suggestions as possibilities in the discussion for future work. Nonetheless, rather we did think about a (somewhat simpler)

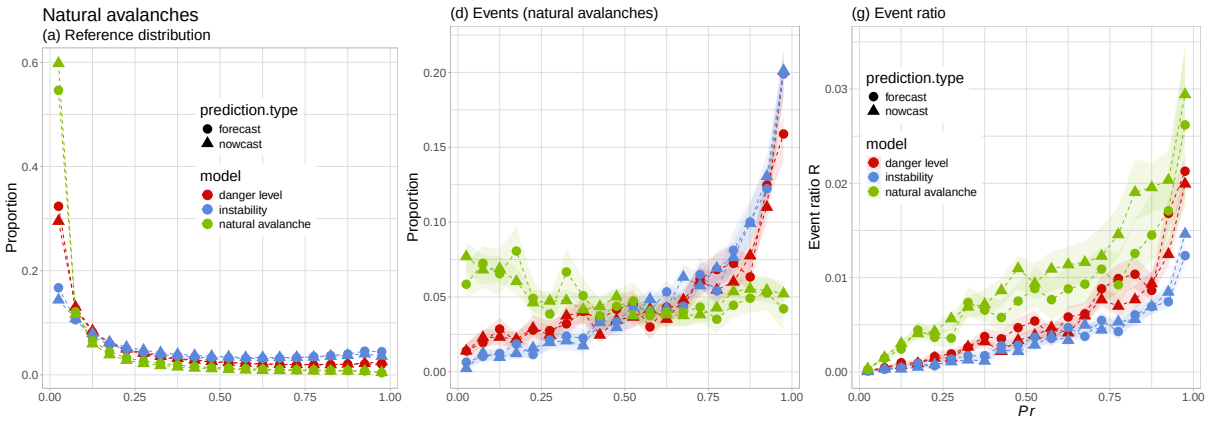

Figure 1. Model predictions. - Here an example of the natural-avalanche data displayed in Figure 4a, d, g in the manuscript, but including the bootstrap 90-% confidence interval (shading) for events and the event ratio R.

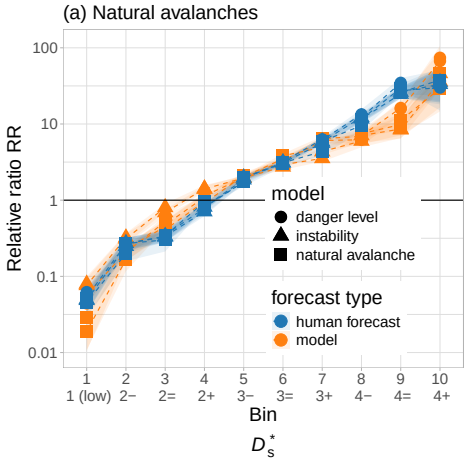

Figure 2. Model and bulletin predictions. - Here an example of the natural-avalanche data displayed in Figure 5a in the manuscript, but including the bootstrap 90-% confidence interval (shading).

way of exploring these issues by adding uncertainty related to the events, which are comparably rare. As we mentioned in the manuscript (L342-348), the distribution of events contained the respective lowest and highest sub-levels or bins impacts the relative ratio (RR). We will therefore show bootstrap-sampled confidence intervals (CI) for the event data and for RR. Examples of the updated Figures for the natural-avalanche data are shown in Figures 1 and 2. We'll provide tables summarizing the median increase between the respective lowest and highest bins including the CI in the Appendix or as a Supplement, or appended to Table 2 in the manuscript.

Another comment along similar lines, but outside the scope of this manuscript unless the authors actually investigated the following thoughts already. To not run the risk of capping model performance, one could make the bins entirely independent of

the human distribution of danger ratings. The results may be less suited for comparing the model and human predictions, but we may learn better in which interval ranges the probabilities are most capable of discriminating conditions. And lastly, I fully buy into the point made by the authors that there is a limit to the value that comparing model to human data sets has, when it is unclear which data set is closer to reality when they disagree. However, I personally would be more than curious how an actual day-to-day comparison over an entire season in a prominent region looks like in the Swiss data set (for example, similar to Fig 5 in Herla et al., 2024).

With respect to the first suggestion, although potentially interesting, that's outwith the scope of our study, which focuses on a comparison. Reviewer 2 also suggested exploring a region, and as such, we will explore providing a figure for two regions covering an 8-week period during which all three models were available in forecast mode, and during which the human-made forecast sub-level varied between 1 (low) and 4+ (Figure 3). Nonetheless, an in-depth analysis is beyond the scope of this manuscript. Moreover, as event data is rare, such a region-specific analysis will at most provide an indication of trends rather than statistically robust findings. We'll use this figure to briefly highlight differences between human forecasts and model predictions, to address the challenges with verifying specific events, and to discuss the disadvantages of making a comparison using a regional context.

The storyline is sound, focused on the research objectives, and communicated well. Congratulations to the authors for this contribution!

Thank you.

2 Detailed comments

2.1 Abstract

- L14: "We conclude that these model chains are ready for systematic integration in the forecasting process." Consider adding a statement like "in Switzerland" or giving other warning agencies in other snow climates a heads-up that other modeling pipelines might not be on par.

  We will add "in Switzerland".

2.2 Introduction

- L47: Consider using a clearer wording, for example, *...when they independently forecast avalanche danger with a similar skill as expert forecasters*.

  We will rephrase as suggested.

- L53: I find the statement of the first objective, "(1) Is the expected increase in the number of natural avalanches or in locations susceptible to human-triggering of avalanches predicted by spatially interpolated model predictions?", more complicated than it needs to be. Consider rephrasing it, e.g., "Can spatially interpolated models predict the observed

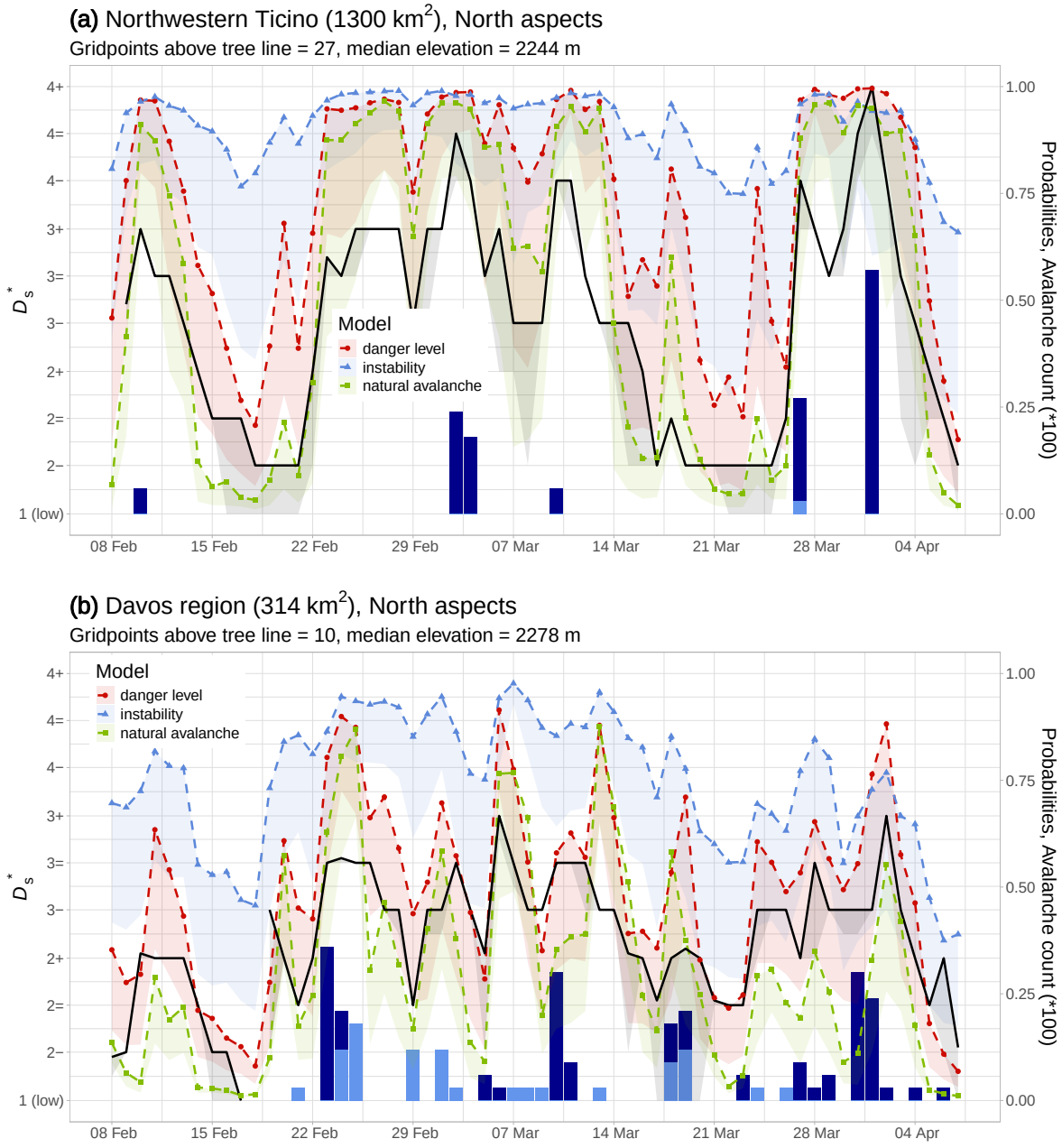

Figure 3. Time series showing 8 weeks between 8 Feb and 7 Apr 2024 for two regions in the Swiss Alps. For the randomly selected grid points in these two regions, the respective 90%-quantile of model predictions (coloured, dashed lines and points, right y-axis) and the avalanche forecast (solid, black line, left y-axis) are shown. The shaded areas expand between the respective median value and the 90%-quantile. The bars display the number of avalanches (dark blue: natural avalanches, light blue: human-triggered avalanches).

increase in the number of natural avalanches or ...".

We will rephrase as suggested.

2.3 Models

- L80: Please add that the instability model is suited for dry slab avalanches only.

  We will add this information.

- L86: Please add that the natural-avalanche model is suited for dry slab avalanches only.

  We will add this information.

- L102: Downscaling weather model output to point scales is a complex endeavour. Can you please describe the key modifications to the raw NWP output before you refer the reader to (Mott et al., 2023)?

  The downscaling of COSMO1-input data is indeed complex and a detailed description of this process is out-of-the scope of our manuscript as different methods are used to downscale wind, radiation, air temperature and relative humidity, but also snow and precipitation. However, to address this we will add a statement that all of the SNOWPACK input parameters are downscaled using a variety of approaches according to Mott et al. (2023) (Table 1).

2.4 Data

- L111: "We analysed ..." (past tense)

  We will change as suggested.

- Paragraph 3.3: In principle, the analysis would be complete with comparisons of natural and human-triggered avalanches to a reference distribution. By including Non-events (approximated by GPS tracks), you offer another perspective on evaluating performance for human triggering, a bonus so to say. Given that readers likely have a strong opinion about using GPS tracks to approximate Non-events (see next comment), I suggest you make this point ("it's a bonus") more clear to the reader. For me, it was helpful to understand that in Figure 2 the box "Events/Human-triggered avalanches" caused two arrows, one to the data set that links human-triggered avalanches to the reference distribution and one to the data set that links human-triggered avalanches to the GPS tracks. This nicely visualizes that you examined human-triggering of avalanches from two complementary perspectives: one more theoretical, and the other purely data-driven, though relying on assumptions that are not easily quantified.

- L142: GPS tracks as non-events: This approach assumes that an avalanche would have occurred if a skier loaded the snowpack and it was unstable. That assumption holds more true for surface problems than for persistent problems buried more deeply. In the latter case we know from avalanche accidents that it's not always the first skier who triggers the avalanche, particularly since the characteristics of the slab and depth of the weak layer vary within a slope. Moreover, I assume that the snowpack at popular ski tours or just outside of ski resorts is heavily modified by skier traffic throughout the season. Within the typical skier corridors, weak layers will likely be destroyed and the primary avalanche problems

will likely be new snow and wind slab problems. Can you please discuss these thoughts and their potential effect on the results in the Discussion and refer to that discussion from Sect. 3.3? It would nicely add to the paragraph in L384ff.

There are considerable uncertainties related to the event and non-event data. This is not specific to the GPS tracks. GPS data does not contain any information about whether a slope was skied for the first time or already skied before. We only assume that no avalanche was triggered during the recorded ascent, regardless of the prevailing avalanche problem, since we think it unlikely that skiers involved in avalanches then uploaded GPS tracks. If a skier loaded the snowpack at a specific point, we can be reasonably confident - but obviously not certain - that the snowpack was sufficiently stable at this location. - We'll make a remark in the respective paragraph.Off-piste terrain was excluded by only using ski touring data and also discarding data near ski lifts. We don't have robust data allowing us to systematically distinguish between popular tours and rarely used tours and will include this point in the discussion.

- L149: Consider adding "... due to forecast, encountered avalanche conditions, or previous terrain use" (or similar).
  We will change as suggested.

- L175: I suggest changing "avoid" to "minimize".
  We will change as suggested.

2.5 Methods

- L195: "the random subset of grid points used as reference distribution". This is the first location that mentions that the reference distribution is based on a randomly selected subset of grid points. Please add a statement that tells the reader that this concept will be explained in detail below in Sect. 4.3.
  We'll add this information as suggested.

- Footnote 1: I think you can simply omit the footnote, particularly since you cite the same publication at the end of the sentence anyways.
  We will delete the footnote.

- L203: Consider rephrasing the sentence to e.g., "For locations and elevations with dis-continuous or non-existent snow cover, we set Pr = 0."
  We will change as proposed.

- L207: Thanks for providing the code to this analysis. Could you still summarize the high-level tuning (i.e., the hyperparameter settings) in the text of the manuscript please?
  We are not sure which hyper-parameters these are. But we will add some more details on the settings used for the kriging method.

- Footnote 2: Please mention the software package used to to implement the regression kriging.
  We will add the libraries used for kriging interpolation.

– L221: How sensitive are the results to the choice of 2.5% of all grid points, and how did you decide on that number? I assume the analysis is computationally fairly inexpensive. In that case, could you easily re-run the analysis for other values and report on the main differences? Also, I think the elevation filter should ideally be applied before the random sampling.

We applied the elevation filter first. To decide on the number of grid points, we looked at several subsets. In the end, we decided on 5% (2.5% is an error in the manuscript) of the 13323 grid points in the elevation range. See also Figure 4. In all four variants, the median elevation was about 2200 m. Despite the 5% variant not covering 10% of the warning regions in the Alps, the resulting proportion of sub-levels was essentially identical as for larger subsets. We therefore decided on 5% of grid points to keep computational costs low and to easily rerun experiments if necessary (interpolation for all model variants took about 5 hours on a normal laptop). - We will state clearly that we first selected grid points within the elevation range before randomly sampling. We will briefly comment on the insignificant changes in elevation and sub-level distributions between smaller and larger subsets, and that the 5% variant covers 90% of the warning regions with a median number of points per region of 5.

– L234: Consider rephrasing to e.g., "systematic biases exist between the *forecast* and *nowcast* predictions"

We will change as suggested.

– L236: "whether the models reflect the expected increase in avalanche occurrence probability with increasing model-predicted probability." I found this sentence somewhat confusing and suggest to either delete 'with increasing...' or to rewrite that last part like 'by predicting higher probabilities themselves'.

We will change as suggested.

– L241 & L243: Instead of "for cases when we relied on the reference distribution:" and "when using non-events:", please call it the same way as in Fig. 2 and 4, i.e., 'for natural and human-triggered avalanches' and 'for backcountry data'. The equations tell the reader already when the reference distributions and non-events are used.

We will change as suggested.

– L237f: Can you add a brief statement why you chose different bin widths? L254: "To obtain bins containing an equal number of data points for human forecasts and for model predictions, ...". I am not sure whether this is the correct justification. I do buy into that binning approach in order to compare human and model data, but I assume this is rather necessary because the danger rating reflects a non-linear increase in hazard (Schweizer et al., 2020; Techel et al., 2022), whereas the model predictions reflect non-linear increases of other functional shapes Mayer et al. (maybe sigmoidal?; e.g., Figure 8 in 2023). In other words, there needs to be a mapping of some sort, which you implement through the binning. Do I understand that correctly?

You understand correctly. As it is unclear how to assign probabilities describing rather different phenomena (probabilities of potential instability, of natural avalanche occurrence, or of danger levels) to sub-levels, mapping the probabilities according to frequency distributions seemed a sensible choice. As shown in Figure 5b, the derived mapping functions

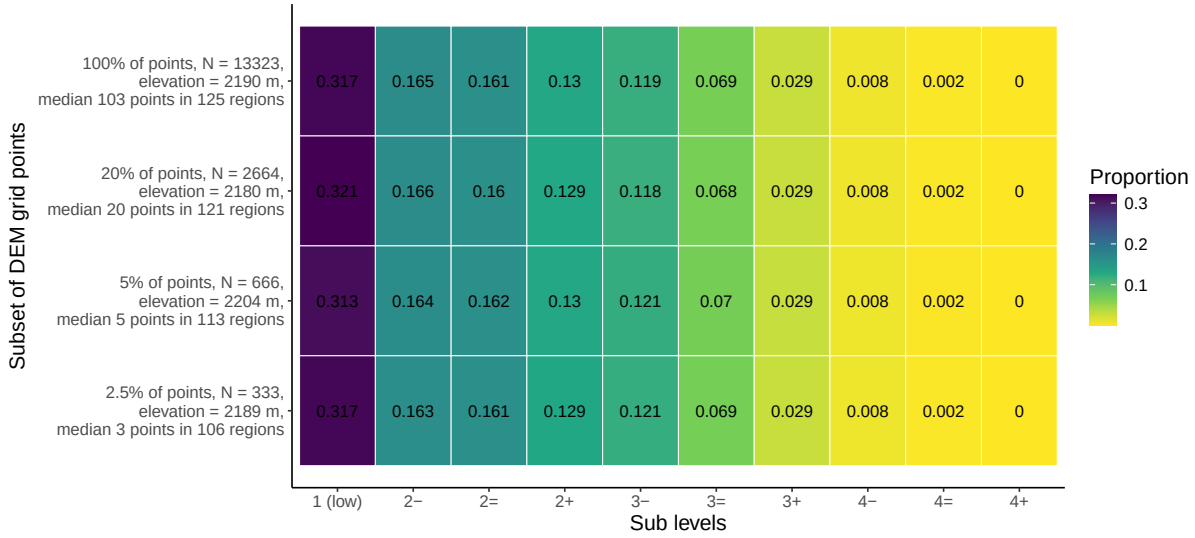

Figure 4. Number of grid points used to define the representative data set. - We will not include this figure in the manuscript.

differ in shape between the models and data set. - When revising, we will include this new figure and add an explanation along these lines.

– L257: Please add ", etc." after 17.8%. That is, only if I interpret the statement "For higher sub-levels, we proceeded in the same way" correctly.
We will change as suggested.

– L266 & L268: Same comment as for L241 & L243 above.
We will change as suggested.

2.6 Results

– Figure 4: "... middle row: events with (d) natural avalanches, (e) human-triggered avalanches and (f) human-triggered avalanches during backcountry touring;". I don't fully understand the difference between the data used for panel (e) and (f). Can you please make that more clear.
We will make that more clear.

– Sections 5.1: Very explainable and encouraging results! Great to see it all come together after an intense workout of data acrobatics beforehand ;-)
Thank you.

– Figures 5, B1–3: I am confused why the human forecast is further stratified into the models. More specifically, why are there three lines in Fig. B1 b, d, f that are colored according to different models? As far as I understand, each panel

corresponds to one specific data set, e.g. Fig. B1d contains all natural avalanches and there should be one curve that displays the proportion of issued danger ratings. Please make sure that a correct explanation is in the text and that the reader will find that explanation from the figure captions.

As shown in Table 1, the number of cases varies between models and between *forecast* and *nowcast*, as predictions were not always available. For instance, for event type *natural avalanches*, *forecast* predictions were available on 143 days for the *natural-avalanche model*, on 236 days for the *instability model*, and on 216 days for the *danger-level model*. As the analysis is always performed for subsets of the data when model predictions and bulletin were both available, each analysis includes a different number of cases. This hardly impacts the distributions of the reference distributions or the avalanche forecast, but causes variations in the event data, resulting also in variations in the event ratio. Therefore, curves are shown for human forecasts for each model. - We will mention this more clearly in the Results section.

- Additional table: In Figures 5 and B1–3, the x-axis allows for translating between the Bin and D_s*. For example, Danger level 3- corresponds to bin 5. Please add a table to the Results section or Appendix, that shows the thresholds for each of the 10 bins and for each of the 3 model types, similarly to L259.

  We will take up this recommendation and provide this information when introducing the approach of defining the bins (see Figure 5b). In addition, we will also provide a figure explaining the mapping approach described in Section 4.4 (Figure 5a).

2.7 Discussion

- L429: I suggest changing recreationists to backcountry users.

  We will change as suggested.

- Paragraph 6.6: I suggest you re-iterate somewhere in the paragraph (e.g., L461) that the conclusions are valid for danger level, probability of avalanche release, and instability, but not for avalanche problems or other specific characteristics.

  We will mention this as suggested.

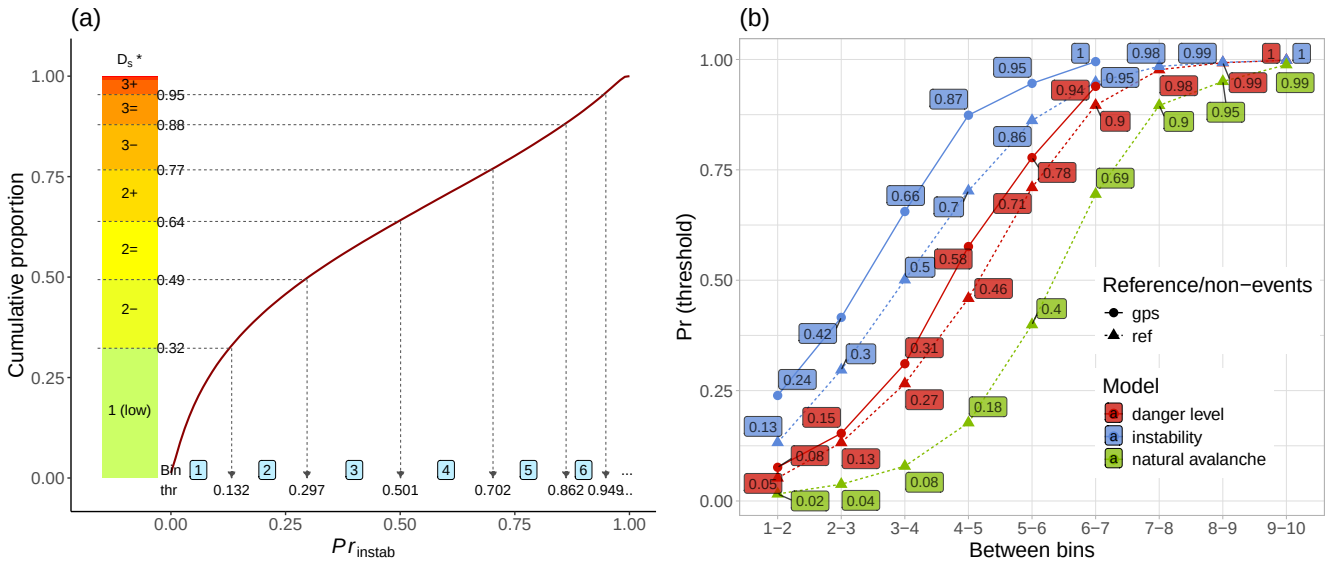

Figure 5. (a) Assigning model-predicted probabilities (here Pr_{instab}) to bins (light-blue labels 1 to 6) equal in size to the proportion of sub-levels D_s^*. Shown are the cumulative proportion of Pr_{instab} (line). Using the cumulative proportion of sub-levels, the resulting probability threshold (thr) can be derived. Note that proportions and thresholds for $D_s^* \geq 4-$ and bin ≥ 7 are not shown. See explanations in the text. (b) Probability thresholds obtained for the forecast predictions for the three models using the reference distribution (ref) or GPS tracks (gps).

References

Herla, F., Haegeli, P., Horton, S., and Mair, P.: A quantitative module of avalanche hazard—comparing forecaster assessments of storm and persistent slab avalanche problems with information derived from distributed snowpack simulations, EGUsphere, 2024, 1–30, https://doi.org/10.5194/egusphere-2024-871, 2024.

Mayer, S., Techel, F., Schweizer, J., and van Herwijnen, A.: Prediction of natural dry-snow avalanche activity using physics-based snowpack simulations, Nat. Hazards Earth Syst. Sci., 23, 3445—-3465, https://doi.org/10.5194/nhess-23-3445-2023, 2023.

Mott, R., Winstral, A., Cluzet, B., Helbig, N., Magnusson, J., Mazzotti, G., Quéno, L., Schirmer, M., Webster, C., and Jonas, T.: Operational snow-hydrological modeling for Switzerland, Frontiers in Earth Science, 11, https://doi.org/10.3389/feart.2023.1228158, 2023.

Schweizer, J., Mitterer, C., Techel, F., Stoffel, A., and Reuter, B.: On the relation between avalanche occurrence and avalanche danger level, The Cryosphere, https://doi.org/10.5194/tc-2019-218, 2020.

Techel, F., Mayer, S., Pérez-Guillén, C., Schmudlach, G., and Winkler, K.: On the correlation between a sub-level qualifier refining the danger level with observations and models relating to the contributing factors of avalanche danger, pp. 1911–1930, https://doi.org/10.5194/nhess-22-1911-2022, 2022.