Response to review by Christoph Mitterer

Frank Techel, and co-authors

Dear Dr. Christoph Mitterer,

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

1 Summary

The study explores the effectiveness and consistency of human-generated compared to fully automated, model-driven avalanche danger forecasts by addressing two main questions:

- Does the spatial interpolation of model predictions indicate an anticipated rise in natural avalanche occurrences or an increase in areas prone to human-triggered avalanches?
- Can fully automated, data- and model-driven avalanche forecasts deliver performance levels comparable to those created by human forecasters?

In order to answer these two questions, the authors compare data sets of natural and human-triggered avalanches and GPS tracks of backcountry activity to human-made and fully automated model-driven forecasts. The human-made forecasts rely on the daily published bulletin data of Switzerland for two consecutive winter seasons (2022/2023 and 2023/2024). The model-based forecasts are explored in different modes (forecast and nowcast) for three different models (danger-level model, instability model, natural-avalanche model).

Using event ratios as proxy, the authors examine the relative accuracies and consistency of human- and machine-made forecasts by comparing the spatial interpolation of the various forecasts to (1) a reference distribution containing events/non-events, and (2) recorded events of natural and human-triggered avalanches and non-events.

The results reveal that human-made and machine-made forecasts show similar relative predictive behaviour,i.e. that increase in all model probabilities are correlated to an increase in avalanche release probability. The authors did not investigate specific or absolute behaviour. This leads the authors to the final conclusion that it is timely to introduce model-based forecasting into the operational settings of avalanche warning services.

2 Evaluation

The presented manuscript has a clear story line and applies in large parts transparently and comprehensible a sound set of methods to obtain innovative results in the field of model-based forecasts assessing avalanche conditions in a regional scale. The data set is innovative, methods have been already in place by the authors with other contributions (Degraeuwe et al., 2024;

Techel et al., 2022; Winkler et al., 2021). Approach and results are scientifically relevant and represent a major impact on that specific topic for the community.

Most parts of the manuscript are concise, well-structured and nicely written. Some parts though of the Data (3.1 Model predictions) and the Methods Section (4.4 Analysis) were at least for me not easy to follow. Also within the Results Sections there are part – especially the ones pointing to Table 1 that are hard to follow and grasp. In addition, the Discussion Section is very broad and remains too vague in some parts for my taste. I would advise the authors to have the courage to draw more direct conclusions so that the manuscript can have even more impact on the community for their operational approaches and future directions.

Also Figures and Tables and especially their captions will need a bit of re-touch to make them even more clear (see comments directly in the manuscript). I am convinced that this excellent work should be published on NHESS having addressed my general and specific suggestions for improvement.

3 General comments

My general comments touch two different aspects of the manuscript: one is of a more technical nature, the other concerns the comprehensibility of the manuscript and thus sometimes drifts into questions of taste. In this respect, I am fully aware that parts of my second comment cannot, of course, be decided objectively.

4 Technical comments

- The reasoning of why the authors adjusted the danger-level model of by rather opting for $Pr(D \geq 3)$ than D (Lines 110-117) including the Figure A1 is unclear. The authors refer to an “expected danger value” which might be or not connected to Pérez-Guillén et al. (2024) or to the concept of an expected danger level presented in Maissen et al. (2024). It remains, however, unclear since no citation is given. Since this altering of the danger-level model might affect central results, the reader needs more details on why the authors decided to do so. In addition to that, I would love to see what the results would look like, if the authors do not introduce this classification criteria, but demonstrate the outcome of predicting the danger level D instead.

  The main reason for using $Pr(D \geq 3)$ rather than the expected danger level ($E(D)$) was that this allowed us to use exactly the same interpolation approach and subsequent assignment of rank-ordered predictions to bins as for the other models. If we had used $E(D)$, which is bound to values between 1 and 4, compared to Pr, bound to the range 0 to 1, the interpolation approach would have had to be adjusted considerably. - We will provide this explanation in the revised manuscript. We will make sure that terminology is consistent with either (Pérez-Guillén et al., 2024) or (Maissen et al., 2024).

- My main comment here: The authors analysed nicely now the median relative performance; is now possible to get to the more complex cases, i.e. extremes or misses of humans. The authors did a tremendous work on creating an objective test data set. I highly acknowledge that, so why not use that approach for tackling this further question.

In this study, we focused on purpose only on the overall performance as the study is already complex and lang. However, we fully agree that shedding light on the performance of models and humans in situations, which are either rare, or where either approach fails, is an obvious next step. See also our comment below.

- With this study we have learned that the machine relatively seen thinks almost identical as a human – which in turn is only to a certain extent impressive, since at least for the danger-level model analysis are in a somewhat closed circuit: The machine learned very well to pick up the forecasting culture of a well-trained and consistent forecasting team (Switzerland) and now mimics their behaviour well in a relative way.

  In this study, we moved from (a comparably small number of) point predictions, trained to predict regional-scale patterns (danger-level model, natural-avalanche model), to interpolating predictions to arbitrary locations in potential avalanche terrain (with specific slopes, aspects, and elevations). We show that this approach works, at least when compared to human forecasts for the same locations as a benchmark. Showing that human forecasts and model predictions can be interpolated to specific locations capturing the relative increase in avalanche release probability is a key first step towards higher-spatially and temporally resolved avalanche forecasts. While we have previously shown that human forecasters can predict the increasing probability of events occurring using sub-levels, at least in a relative way (Techel et al., 2022), as a side-result, we show that there is also a correlation between human forecasts and the probability of event occurrence at very specific locations.

- The most exciting question though is: when do they differ? Therefore, I would like to see at least one or two examples where either a specific region for the entire period (e.g. warning region of Davos) or a specific period (e.g. December 2023) for the entire forecasting domain.

  As we also explain below, a statistical analysis of when models and forecasters differ (and fail) is not possible within this manuscript. However, as reviewer 1 suggested such an example, we have provided 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 1). 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.

Question of taste comments:

- Please help the reader to make the two very important Sections 3.1 Model predictions and 4.4 Analysis more comprehensible. Now, sentences are very long and full of different terms referring to different "probabilities". In addition, the explanation of using $Pr(D \geq 3)$ needs more text support.

  To make it easier for the reader to see that we use three different probabilities, we'll use bullet points in L128-130. In addition, we'll provide more explanation regarding the use of $Pr(D \geq 3)$ along the lines of our reply before.

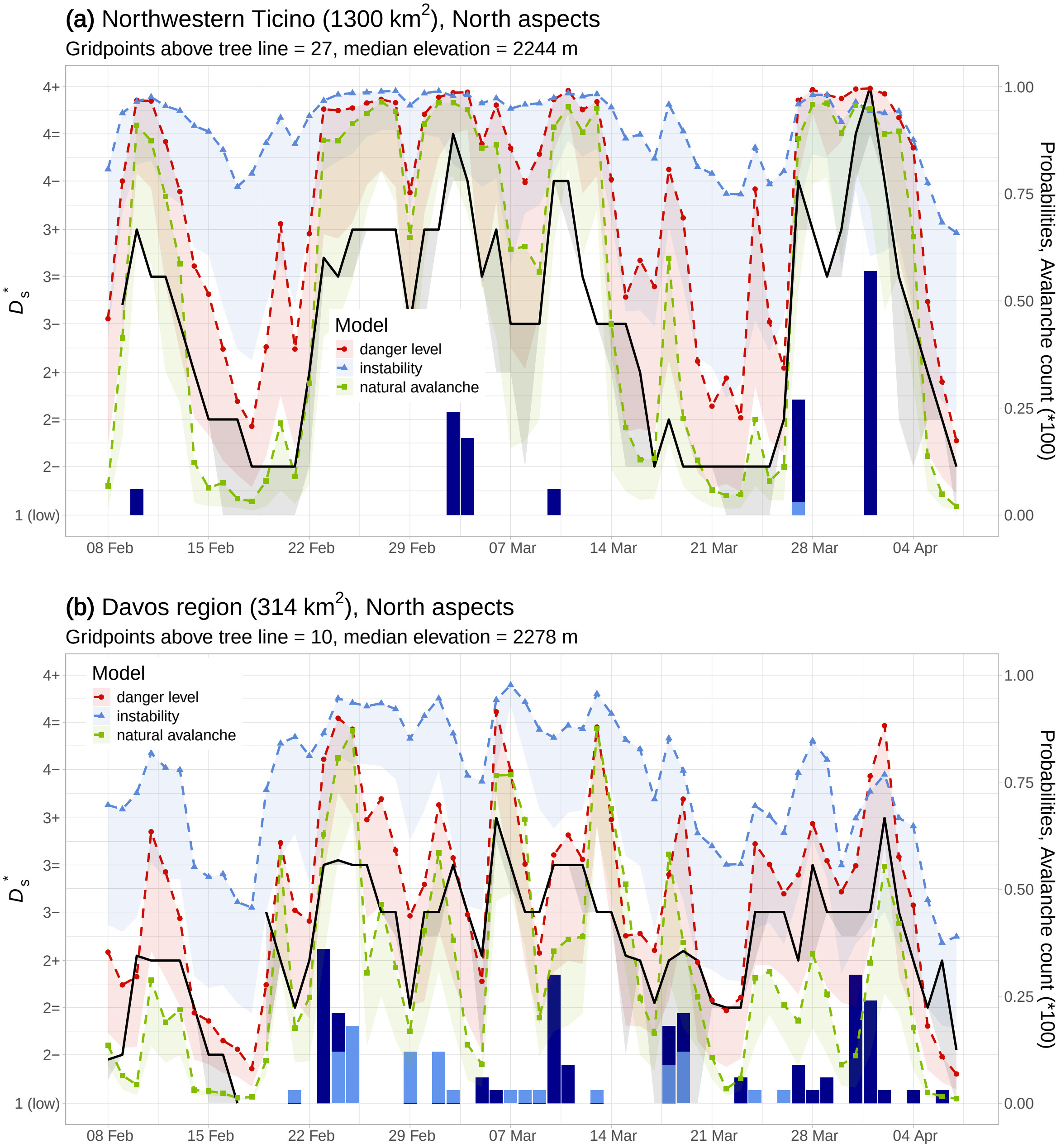

Figure 1. Time series showing 8 weeks between 8 Feb and 8 Apr 2024 for two regions in the Swiss Alps. For the randomly selected grid points in these two regions, the respective 90%-quantile of model predictions (coloured, dashed lines and points, right y-axis) and the avalanche forecast (solid, black 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).

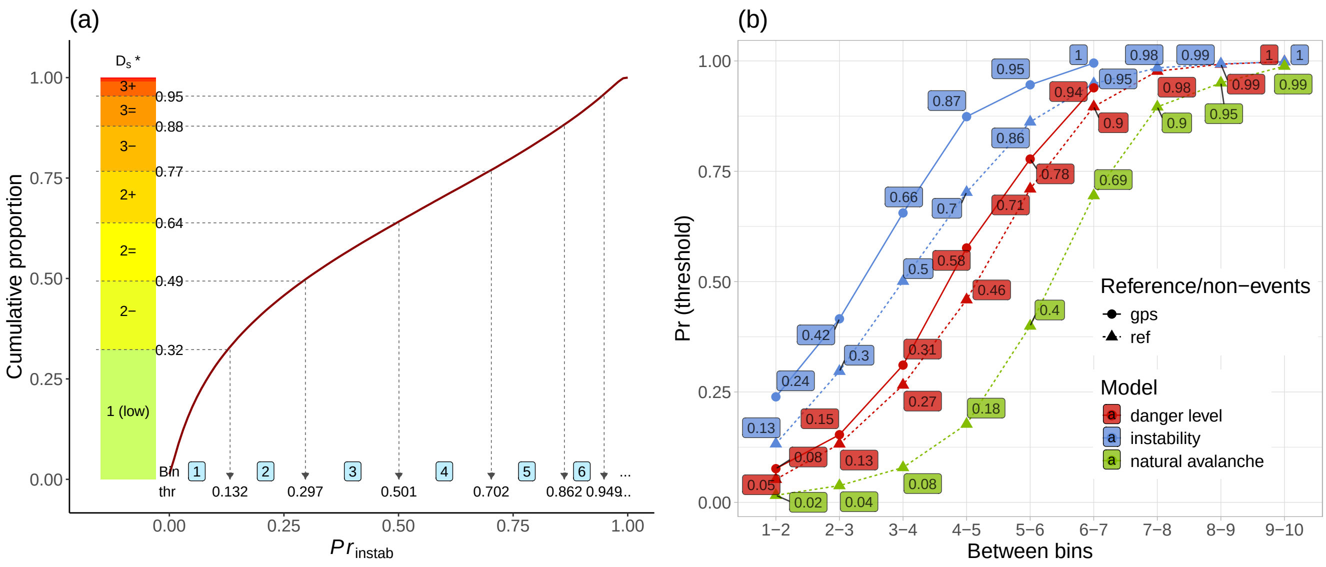

Figure 2. (a) Assigning model-predicted probabilities (here Pr_{instab}) to bins (light-blue labels 1 to 6) equal in size to the proportion of sub-levels D_s^*. Shown are the cumulative proportion of Pr_{instab} (line). Using the cumulative proportion of sub-levels, the resulting probability threshold (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).

– Maybe you add some graphs to explain the reversed binning approaching a bit better.

We'll add a figure (Figure 2) aimed at supporting the introduction of the binning approach described in Section 4.4. (Figure 2a), and showing the shape of the functions describing the probability thresholds between bins for mapping (Figure 2b), as proposed by reviewer 1.

– During the entire reading it was not clear to me that were working on two different human-triggered data sets (cf. Fig. 4e,f).

We introduced the different data sets in Section 3, providing Figure 2 as the overview linking the data and methods. We'll make this clearer.

– Table 1 is hard to understand. E.g. what is referring to ref, what is referring to nEv?

We will revise the caption of Table 1 to explain more clearly what refers to ref and what to nEv.

– I enjoyed large parts of the Discussion but have the feeling that it is too lengthy and not to the point of results that were shown. Not sure whether Section 6.4 could be shortened and incorporated into the limitation's section representing the output.

We interpolate to specific points in the start zone of avalanches. These points are defined by coordinates with 1-meter resolution. They are located in real terrain, on specific slopes. However, and on purpose, we dedicated a specific section

(Section 6.4) on the seemingly highly-resolved predictions to clearly emphasize that the interpolation approach leads to predictions which are still regional and not slope-specific. We don't consider this a limitation as such, but it is important to clearly explain the difference between resolution and scale to avoid misinterpretation. We will shorten lines L454 to 463

- I would however, love to see more work on combining Section 6.3., 6.5 by addressing the questions I posed before: When do machine and humans think differently and could this think differently help us in improving the quality of our product.

  We agree that understanding when and why models and humans differ and fail is a crucial next step. However, focusing on specific situations requires analyzing very small subsets of data, which makes it difficult to conduct robust statistical analyses. Our dataset already contains a limited number of events, and even in low-frequency bins, a few occurrences are expected. These should not necessarily be considered incorrect predictions. - We plan to add a section (likely in the Discussion) where we present time series for two regions (Figure 1), briefly comparing model predictions with human forecasts. More importantly, we will address the challenges associated with verifying predictions for specific cases. The questions you pose are natural, given our results. But this paper's focus is on the performance of models across Switzerland, and we would rather see this as a logical next step rather than overfill an already complex paper.

- While reading, the feeling arises here and there that the team of authors is trying, subtly, to polarize (e.g. choice of title). Since they have done a wonderful job either way, they have no need to do so.

  We don't see where we polarize but we certainly want to make the point that the performance of model pipelines have reached a state where they are approximately equal to the predictive performance of human forecasts - at least in the setup used in Switzerland.

5 Specific comments

See my mark-ups and comments within the attached supplement. Thank you. We will address them when revising the manuscript.

References

Degraeuwe, B., Schmudlach, G., Winkler, K., and Köhler, J.: SLABS: An improved probabilistic method to assess the avalanche risk on backcountry ski tours, Cold Regions Science and Technology, 221, 104 169, https://doi.org/https://doi.org/10.1016/j.coldregions.2024.104169, 2024.

Maissen, A., Techel, F., and Volpi, M.: A three-stage model pipeline predicting regional avalanche danger in Switzerland (RAvaFcast v1.0.0): a decision-support tool for operational avalanche forecasting, EGUsphere [preprint], 2024, 1–34, https://doi.org/10.5194/egusphere-2023-2948, 2024.

Pérez-Guillén, C., Techel, F., Volpi, M., and van Herwijnen, A.: Assessing the performance and explainability of an avalanche danger forecast model, https://doi.org/10.5194/egusphere-2024-2374, 2024.

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.

Winkler, K., Schmudlach, G., Degraeuwe, B., and Techel, F.: On the correlation between the forecast avalanche danger and avalanche risk taken by backcountry skiers in Switzerland, Cold Regions Science and Technology, 188, 103 299, https://doi.org/10.1016/j.coldregions.2021.103299, 2021.