Summary of major revisions

Frank Techel, and co-authors

We sincerely thank editor Pascal Haegeli (ph) for providing detailed feedback on our revised manuscript. Below, please find our point-to-point reply, highlighted in blue, with changes made in the manuscript shown in red.

- L7 – Abstract: The term "reference conditions" is not clear in the abstract. It would be best to either add an explanation of the term or avoid it and describe it in different words. - We rephrased to (L7-9): We assessed the ability of both model and human forecasts to discriminate between reference distributions of conditions – typically not associated with avalanche activity – and actual avalanche events, either naturally released or triggered by humans, by calculating event ratios as proxies for release probability.

- L118: It might be useful to explicitly list the types of ML predictions to make it clear that this refers back to the danger rating, instability, and natural-avalanche models. - We rephrased to (L120): ML models like the danger-level model, the instability model, and the natural-avalanche model provide...

- L140 – Justification of Pr(D>=3): I think that some slight tweaks in this section could strengthen your justification. How about "... described in Section 2.2.1 as it simplifies the multiple probabilities model output (one probability for each danger level) into a single and more easily interpretable probability value. ... Converting the danger rating model output into single value between 0 and 1 allows us to use a consistent analysis approach for all three ML models." - We changed to (L144-147): ...described in Section 2.2.1, as it simplifies the multiple probabilities model output (one probability for each danger level) into a single and more easily interpretable probability value. Converting the danger rating model output into single value between 0 and 1 allowed us to use a consistent analysis approach for all three ML models.

- L142: Delete "and" after moreover - Done

- L158 – Description of reference grid points: I am not sure whether the explanation provided on L158-163 is necessary since you describe this in more detail (and easier to understand) just a little bit later. - We removed these lines.

- L175, 181 and 201 – Medians and IQR: I believe that the median and IQR values do not represent the precision of the actual observations accurately. I assume that the elevation values are reported to the closest 10 or 100 meters. In my opinion, the summary stats should reflect this as well, and values rounded to the closest 10 or 100 meters seem more appropriate. - Avalanches are recorded by setting a point on a map, which represents the top of the release area of the avalanche. For this point, slope aspect – with an accuracy of $\pm 25^{\circ}$ and rounded to the eight aspects used in this study – and elevation, with an accuracy of ± 5 m a.s.l., are derived from a digital elevation model. We rounded to 10 m increments.

– L212: Consistent with other mentions of the research questions in the manuscript, use 'RQ' consistently. This first sentence might not be necessary anyway since it already refers to the analysis approach. - Done

– L225 – Arbitrary locations: It is unclear to me why you only extrapolate to arbitrary locations and not to all avalanche locations and reference points. I do not think this additional sampling is explained in more detail anywhere else. I am sorry if I missed it. It might be related to the bootstrapping (L262), but I am not sure. On L253, you state that you interpolated the predictions at all reference locations. - We interpolated to reference locations AND avalanche locations. We rephrase as follows to make this clearer: We spatially interpolated point data – specifically, model predictions and snowline estimates – to the locations of observed avalanche locations and to the randomly sampled reference points.

– L225 – Explanation of RK: I suggest that the first two paragraphs are combined since they discuss the same topic. Done. I also suggest that the text starting with "Compared to the simple ordinary kriging, . . . " and going all the way to the end of the paragraph is move up right at the end of the first paragraph. This fully explains the methods before you describe its application in the current study. - The entire paragraph now reads (L225-234): We spatially interpolated point data – specifically, model predictions and snowline estimates – to the locations of observed avalanche start zones and to the randomly sampled reference points in avalanche terrain. To do so, we employed regression kriging (RK) (**?**), a geostatistical method that combines a deterministic regression model with kriging of the residuals. Compared to simple ordinary kriging, RK enables the inclusion of environmental gradients, such as the varying magnitude of change with elevation. Compared to purely deterministic interpolation, it reduces bias introduced by unmodeled spatial autocorrelation. This hybrid method therefore offers improved interpolation accuracy and physical plausibility in mountainous terrain, where elevation-dependent and location-specific patterns dominate. This approach was well suited for our application, as it captures both spatial and elevational variation in avalanche conditions. In our implementation, elevation was used as a predictor in the regression component. The remaining spatial structure – unexplained by elevation – was interpolated using kriging, allowing us to better preserve local variability.

– L236 – Aspects: I am not sure whether "compound aspects" is the right term here, because NE is not a combination of N and E, it is rather the aspect between N and E. How about "intermediate aspects"? I would also expand the sentence to make it clearer why this is a challenge: . . . recorded on intermediate aspects (e.g., NE, SW) for which we did not simulate the snowpack and avalanche hazard indicators. - Changed to: intermediate aspects

– L264: I do not quite follow your sampling strategy and the binning. This relates to my earlier comment on L225. It might be useful to explain this in more detail. - We rephrased to make the sampling clearer (L263-264): To account for sampling uncertainty – especially relevant given the relatively small number of events – we applied bootstrap sampling with replacement before calculating the event ratios, repeating the procedure 100 times.

– Fig. 6 – Caption: Change to ". . . for natural avalanches (left column) and human-triggered avalanches (right column) . . . . Top row (a, b): reference distribution; middle row (c, d):. . . ". - Changed to Model predictions (Pr) for natural avalanches

(left column) and human-triggered avalanches (right column). Top row: (a, b) reference distributions; middle row: (c, d)...

– L369: The first sentence might not be necessary since you already described this in the methods section. - We removed this sentence now starting: As shown in Figure **??**e, the normalized event ratio, RR ((Eq. **??**)), increased markedly...

– L384: The first two sentences might not be necessary since you already described this in the methods section. Should the additional binning of the human-triggered avalanche data set already be described in the methods section? - We deleted the first two sentences. We moved the binning methodology to the Methods section 3.2.4 (L314-316).

– L398: I do not understand where the ranges are coming from. Aren't these single summary values (maybe for natural and human triggered avalanches) and not ranges? - These are single summary values for natural avalanche and the human triggered avalanche data sets for each of the three models. The "range" only refers to the range between the three models. We rephrased to make this clearer (L391-395): The average increase between adjacent bins ($\overline{F}$) ranged from 2.20 to 2.26 for the three model-specific data sets (instability, danger level, and natural avalanche model) of the human forecasts, compared to 1.63 for the instability model, 2.07 for the natural-avalanche model, and 2.0 for the danger-level model. The total increase from the lowest to highest bin (F_{total}) varied between 1206 and 1274 for the three model-specific data sets of the human forecasts, and from 286 (instability model) to 1163 (danger-level model).

– L401: While the difference is statistically significant, does it have practical relevance? I do not completely understand what the sample of these comparisons are? - No, we believe it is of comparably little practical relevance, beside the fact that human forecasts – integrating a wide range of data sources – still have a slight advantage in terms of discrimination power. We rephrase to make clear what we compare (L395-396): Statistically comparing the respective distributions of the 100 bootstrap samples for each of the model-specific data sets of human forecasts and models, confirmed that these differences were significant in most cases (Wilcoxon rank-sum test, $p < 0.001$), with the exception of the natural-avalanche model ($p = 0.08$).

– L466: Sykes et al. (2025) (https://nhess.copernicus.org/articles/25/1255/2025/) is another study using GPS tracks of actual terrain choices. - We now cite Sykes et al. (2025).

– L476: The fact that you are using the danger level as a proxy for probability of avalanche release even though the definition of the danger scale also includes avalanche size should probably be mentioned earlier (i.e., in Section 3.1.4). - On L54/55 (Introduction) we introduce that danger levels are based on the principle that the likelihood, number, and size of avalanches increases with increasing danger levels, and on L219-220 in Section 3.1.4, we say that forecast sub-levels correspond to greater number of locations prone to avalanche release and a higher likelihood of larger avalanches. In addition, we added (L220-221): In this study, we use the forecast sub-level as a proxy for the probability of avalanche release.