RC1: Comment on egusphere-2024-2374 (Simon Horton, 14 Aug 2024)

General comments

This paper evaluates a machine learning model that predicts avalanche danger levels in Switzerland. The authors developed the model authors in a previous study, so this study focused on comparing the model's predictions with expert forecasts during the 2020-21 winter across space, time, and under different avalanche conditions. The authors also analyze which model inputs influenced predicted danger in different scenarios, adding transparency and explanations for an otherwise black-box model. The study is well-designed, clearly organized, and effectively communicated, making it both interesting and relevant. Numerical avalanche forecasting is a rapidly developing field with significant implications for public safety and natural hazard management. I recommend publishing this paper in NHESS after addressing the relatively minor comments and clarifications below.

Thank you for your detailed review of our paper. We greatly appreciate your positive feedback and are glad you consider the study well-designed and relevant to the field of numerical avalanche forecasting. Thank you also for highlighting the importance of this research for public safety and natural hazard management. We will address the comments and suggestions you provided in the new version of the manuscript. Thank you for your thoughtful review and for recommending the publication of our work. Please find below the reply to your comments.

Kind regards,

Cristina Pérez-Guillén

Specific comments

1. One limitation that could be more clearly emphasized is that the model only accounts for dry snow avalanches. It's important to clarify whether the analysis excluded situations when wet avalanches may have influenced the danger rating, especially since the study period extended into May. According to the EAWS workflow, the danger level is determined based on the highest level indicated by the EAWS matrix for each avalanche problem. If wet snow problems significantly contributed to the danger rating on certain days, those days should be excluded from the evaluation, as the ML model was designed exclusively for dry snow avalanches.

Thank you very much for your suggestion. We will emphasize clearly in the revised manuscript that the model was developed to predict danger levels for dry-snow conditions. You are correct that on some days, the danger level for wet snow conditions was higher than that for dry snow conditions. In Switzerland, two separate danger levels are issued in the public bulletin on those days, corresponding to dry and wet snow conditions. Typically, on such days, the danger levels for dry snow conditions are low (mostly danger level 1 or 2). Thus, we filtered the dataset to include only dry-snow conditions. Since we aim to evaluate the time series of the predictions comprehensively, we chose to use the complete dataset in our evaluation.

2. Within/outside categories. The within/outside categories could be described more clearly. It seems that the "within" group requires both the station elevation to be within the critical elevation range of the bulletin and the virtual slope to be a critical aspect. How are flat slopes within the critical elevation treated? The "outside" predictions are defined as not being in the critical elevation range. But then where would a simulation that falls within the critical elevation range but on a virtual slope that isn't a critical aspect belong? Also, is a subcategory for predictions on critical aspects outside the critical elevation range relevant? It might also be clearer to consistently use "critical" elevations and slope aspects, as in Fig. 1 and Sect. 3, instead of switching to other terms that appear to be synonyms such as "core zone" and "active slope", which may contribute to confusion.

We will clarify and be consistent with the terms used in the revised version of the manuscript. For flat slopes, we grouped the predictions as *Within* the core zone when the station was above the critical elevation and *Outside* otherwise. For virtual aspects, we grouped predictions *Within* when the station's elevation was above the critical elevation and the aspect was considered critical in the bulletin. *Outside* the core zone predictions refer to those below the elevation range specified in the bulletin. Predictions for virtual aspects are further categorized into those where only one criterion was not fulfilled (i.e., aspect aligns with the forecast but falls below the indicated elevation), and those where neither condition was met.

3. The deviations between the expected danger and sub-levels might stem from the fact that expected values calculated with Eq. 3 would gravitate towards average values and away from extremes. Fig. B1 suggests the sub-level assessments have a wider spread compared to the expected values. This characteristic of expected values could be worth discussing in more detail. (Section 5.2.4 / Fig. B1)

Yes, this is correct. The expected values are constrained between 1 and 4. Additionally, the model rarely predicted danger level 4 with a high probability (> 0.7) (Fig. 5), so the maximum expected values are rarely above 3.5 or within the bin of 4- (Fig. B1). The same applies to the lower limit, danger level 1. However, since no sub-level danger is assigned for this level, the agreement rate is the highest (Fig. B1). We will emphasize this aspect in the revised version.

Technical comments

I appreciate several interesting results from this study, including how the model often predicted lower danger than human forecasters, responded to increases/decreases in danger faster than humans, and showed overall poorer performance for persistent weak layer problems. It was encouraging to see the recommendations for improving performance on persistent weak layers, as this seems to be the biggest limitation for operational adoption.

Line 20: Several hundred million Swiss francs “per year”?

It can reach several hundred million Swiss francs, as seen in the catastrophic winter of 1999 [1].

Line 48: Consider a narrative in-text citation to be very clear that “a model” is precisely Pérez-Guillén et al. (2022), which may not be clear with a parenthetical citation.

Thank you very much. We will modify this.

Fig. 1 and 7: These figures label wind-drifted snow problems as “snowdrift/SD” instead of “wind slab/WS” as defined in line 68.

Thank you very much. We will change Figures 1 and 7 according to your suggestion.

Eq. 1: The lowercase probability from each tree (pt) is not defined.

Thank you very much. We defined t in line 98. To clarify further, we will modify this in the new version.

Fig. 2: This is an excellent and clear illustration that helps the reader understand a complex model system.

Thank you for your positive feedback.

Line 157-163: I found these lines difficult to understand until reading Appendix B. Consider moving some details from the appendix to the main text for clarity. This also warrants its own paragraph. I assume that the nowcast versus forecast comparison involved comparing a nowcast with the forecast issued 24 hours earlier — if so, this could be stated explicitly. Additionally, please clarify that the rounding strategy was applied to the expected danger values, as it is initially unclear which variable is being rounded for the comparison.

Thank you for your comment. We will clarify these points in the revised version. Specifically, we will move some definitions from Appendix B to the main text. Additionally, we will explicitly state that the comparison involved a nowcast and a forecast issued 24 hours earlier, ensuring the comparison of the predictions within the same time window. We will also clarify that the rounding strategy was applied to the expected danger values.

Line 183: Were the forecast predictions often higher than nowcast predictions due to a systematic bias in the COSMO forecasts compared to what was measured at stations? This could be worth discussing later, and perhaps linking with the case where COSMO underestimated precipitation from March 15 to 17.

Based on our analysis, we cannot confirm the existence of a systematic bias in the COSMO forecasts compared to the data measured at the stations. Figure 7 shows that, for predictions at the VDS and WFJ stations, the forecast predictions are not always higher than the nowcast predictions throughout the entire time series. This would be a very interesting topic to investigate in a future study.

Line 192: “Frequently” or “often” are better choices than “essentially always”.

Thank you for your comment. We will change it to "mostly".

Fig. 4: Perhaps clarify in the caption that the numbers below the percentages represent counts.

We will modify it.

Sect 5.2.2: It would be interesting to discuss possible reasons for trends in model performance on flat/south/north aspects in the discussion section, perhaps linking with which meteorological/snowpack features may be causing the differences.

Thank you for your suggestion. While this would be very interesting, linking the differences in performance to the variations in snowpack features across different aspects is beyond the scope of this paper.

Line 216: Why is Table A2 with nowcast predictions cited when sections 5.2.2 onward are supposed to focus on forecast predictions? Table 1 seems like a better citation showing many predictions were within one level.

Thank you for your comment. This is an error, and Table 1 should be cited instead. We will correct this in the new version.

Line 227: The phrase "model bias was towards the forecast in the bulletin" is unclear. Bias typically suggests a consistent directional trend, but Fig. 5b shows that when the model assigns the highest probability to a rating different from the bulletin, its second-highest probability often aligns with the bulletin rating. This doesn’t necessarily suggest a positive/negative bias.

We will rephrase the sentence to improve clarity.

Fig. 5: The second-row plot for level 2 has some random characters in the middle (7BA7F5).

Thank you very much; we will delete them and update the Figure.

Line 237: The phrase “showing a decrease in the number of samples with a larger difference” is unclear.

We will rephrase this sentence in the new version.

Line 261: The model’s response to precipitation several hours earlier might be attributed to its 3-hour temporal resolution, compared to the 24-hour resolution of the bulletin. Similarly, the patterns in Fig. 7 could reflect both the different temporal resolutions and the inherent differences between the two methods. A fairer comparison might involve using only the 1800 LT model predictions. unless the goal was to emphasize the advantages of higher temporal resolution.

To compare the model predictions with the public bulletin forecast and compute the statistics (Tables 1 and A1, Figures 4 and 5), only the model predictions at 1800 LT were used (Lines 142-145). However, in Figure 7, we display the full time series of the model predictions with 3-hour resolution to qualitatively demonstrate the advantage of higher temporal resolution.

Line 288: I agree that the SHAP distribution of MS_Snow for levels 1 and 2 are inverted compared to level 4. however, the distribution for level 3 appears to be scattered.

Yes, you are correct. We will rephrase this sentence in the new version.

Fig. 9: The plot titles for Moderate and Considerable are incorrect. Additionally, could the top panel legend for the black and blue lines use consistent terms from the rest of the paper? I assume the black line is the sub-levels forecast in the bulletin and the blue line is the expected danger from the model.

Thank you very much. We will modify this Figure.

Line 310-311: It would be more intuitive to explain the transition from level 1 to level 2 before discussing the transition from level 2 to level 3. This order would make it easier for readers to follow and avoid confusion (I initially looked at the wrong table when trying to visualize the thresholds). Additionally, it might be helpful to guide readers on how to interpret approximate thresholds from Fig. 9. You could clarify this by stating: "Approximate thresholds for a given danger level can be estimated by identifying feature values when the SHAP values switch from negative to positive" (assuming this was the method).

Thank you very much for your comment. We will improve this in the reviewed paper.

Line 318: MS_Snow is not shown for level 1.

Thank you very much for your comment. This is an error, and Figure 8 should be cited.

Line 324-326: Why would an unstable snowpack with regards to natural avalanches favour level 2? This would make sense for higher danger levels, but I would expect natural avalanches to be unlikely for level 2. The fact low Sn values favour levels 2 and 4, while high Sn values favour levels 1 and 3 suggests the impact of this variable may not be that simple.

We agree with you. As highlighted in previous studies [2, 3, 4], the Sn computed by SNOWPACK shows limited discriminatory ability, and its direct impact by SHAP values is difficult to interpret. We will rephrase this sentence to clarify this point.

Line 333: Why is Fig. 6 cited here?

Thank you very much for your comment. This is an error, and Figure 7 should be cited.

Line 371: Wrong citation style.

Thank you very much for your comment. We will correct this in the revised version.

Appendices: The grouping of tables and figures into 2 appendices seems illogical. Why is Fig. B1 included in an appendix titled “Evaluation Metrics”. Consider splitting these into separate appendices.

We will correct this.

Line 490: Dbu,a should be defined here.

We will change it in the new version.

Line 577: Is there a reason both the discussion paper and final paper are listed?

No, we will correct this.

References

1. Bründl, M. *et al.* Ifkis-a basis for managing avalanche risk in settlements and on roads in switzerland, DOI: 10.5194/nhess-4-257-2004 (2004).

2. Jamieson, B., Zeidler, A. & Brown, C. Explanation and limitations of study plot stability indices for forecasting dry snow slab avalanches in surrounding terrain. *Cold Reg. Sci. Technol.* **50**, 23–34, DOI: 10.1016/j.coldregions.2007.02.010 (2007).

3. Reuter, B. *et al.* Characterizing snow instability with avalanche problem types derived from snow cover simulations. *Cold Reg. Sci. Technol.* **194**, DOI: 10.1016/j.coldregions.2021.103462 (2022).

4. Mayer, S., Techel, F., Schweizer, J. & van Herwijnen, A. Prediction of natural dry-snow avalanche activity using physics-based snowpack simulations. *Nat. Hazards Earth Syst. Sci.* **23**, 3445–3465, DOI: 10.5194/nhess-23-3445-2023 (2023).