HUMAN VS. MACHINE - COMPARING MODEL PREDICTIONS AND HUMAN FORECASTS OF AVALANCHE DANGER AND SNOW INSTABILITY IN THE SWISS ALPS

Frank Techel[1,*], Andrea Helfenstein[1], Stephanie Mayer[1], Cristina Pérez-Guillén[1], Ross Purves[3], Marc Ruesch[1], Günter Schmudlach[2], Katia Soland[3], Kurt Winkler[1]

[1] *WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland*
[2] *skitourenguru GmbH*
[3] *University of Zurich, Zurich, Switzerland*

ABSTRACT: In recent years, the integration of physical snowpack models coupled with machine-learning techniques has become more prevalent in public avalanche forecasting. When combined with spatial interpolation methods, these approaches enable fully data- and model-driven predictions of snowpack stability or avalanche danger at any given location. This prompts the question: Are such highly detailed spatial model predictions sufficiently accurate for operational use? To explore this, we assess the performance of interpolated, model-based predictions of snowpack stability and avalanche danger, comparing them to human-generated public avalanche forecasts during the 2023/2024 winter season in Switzerland. To do so, we compare human forecasts and model predictions for locations in avalanche terrain (considering coordinates, aspect, elevation) where skiers triggered avalanches (244 events) or which were skied but where no avalanche was triggered (non-events, 3173 data points from GPX tracks). While this data reflects human behavior to some extent, we consider the event ratio as a proxy for the probability of avalanche release due to human load. We observed that with increasing model-predicted danger level or decreasing model-predicted snowpack stability, the event ratio increased. Comparing model predictions with human-made forecasts showed that the predictive performance of two operationally used models was similar to the performance of human avalanche forecasts: both predicted a strong increase in the probability of human-triggered avalanches. In summary, our results indicate that models capture regional patterns of snowpack (in)stability or avalanche danger well, and that these model chains should therefore be systematically integrated in the forecasting process.

Keywords: public avalanche forecasting, forecast verification, machine learning, model-driven predictions

1. INTRODUCTION

In recent years, the use of physical snowpack models combined with machine-learning techniques has increased in public avalanche forecasting. These model combinations predict snowpack properties, snow instability, avalanche problems, or avalanche danger (e.g., Mayer et al., 2022, 2023; Reuter et al., 2022; Pérez-Guillén et al., 2022). While forecasting chains have been used for many years, as for instance *SAFRAN-Crocus-MEPRA* in France (Durand et al., 1999), through the coupling of numerical weather predictions models with physically-based snowpack models, it is now possible to run simulations at much higher spatial and temporal resolutions than those at which forecasters typically operate. Moreover, by using geo-statistical interpolation methods, it is possible to obtain predictions for arbitrary points in space and time. Given the currently rapidly evolving suite of models and related applications, and the promising feedback following live-testing in forecast settings (e.g., van Herwijnen et al., 2023; Horton et al., 2024; Winkler et al., 2024), the question is warranted: are (high-resolution) model predictions ”good enough” to complement or even replace those produced by professional forecasters? Before we can answer this question, however, we must first define a benchmark which such model-driven forecasts must reach to be considered ”good enough”? We define this benchmark through the use of traditional, primarily human-made public avalanche forecasts. We deem model-driven forecasts to be adequate when they independently reach a similar quality in predicting avalanche danger or snowpack stability as those produced by an expert team.

Public avalanche danger scales are based on the notion that both the probability of avalanche release - described by triggering level and number of potential triggering locations - and the size of avalanches increase with increasing avalanche danger (levels) (e.g., EAWS, 2021; avalanche.org, 2024). Given the challenges in validating avalanche forecasts in general, we focus on evaluating the expected increase in the likelihood of avalanches with increas-

*Corresponding author address:
Frank Techel, WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland, email: techel@slf.ch

ing danger level or decreasing snowpack stability. To compare models and human forecasts on objective data, we utilize data representing events (human-triggered avalanches) and a proxy for non-events (GPX tracks obtained while backcountry touring) from the 2023/2024 avalanche forecasting season in Switzerland. We therefore aim at answering two questions: (1) Is the expected increase in the number of locations susceptible to human-triggering of avalanches predicted by spatially interpolated model predictions? and (2) Do fully data- and model-driven predictions achieve performances comparable to human-made avalanche forecasts?

2. DATA

2.1. Model predictions

In Switzerland, a network of automated weather stations (AWS) provides half-hourly or hourly measurements of meteorological conditions (i.e., temperature, wind speed and direction, humidity, radiation) and snow depth (SLF, 2024). Most of these stations are located at the elevation of potential avalanche release areas. Using this data as input, the physics-based snow-cover model *SNOWPACK* (e.g., Lehning et al., 1999) simulates the snow cover evolution at 3-hour intervals for flat terrain and for four virtual slopes (North, East, South, West) with a slope angle of 38° at the locations of some 147 automated weather stations, providing *nowcast* simulations. In *forecast*-mode, snow cover simulations are initialized using the most recent *nowcast* simulations. Forecast simulations are driven using the COSMO-1 numerical weather prediction model (NWP) with 1 km resolution as input (COSMO = Consortium for Small-scale Modeling, website) providing snow cover simulations up to 27 hours ahead with a temporal resolution of three hours.

Recently, several machine-learning models have been developed, which use SNOWPACK simulations and meteorological data as input for differing target outputs. Here, we introduce the *danger-level* and *instability models* used in this study. These models provided real-time predictions during the forecasting season 2023/2024 in Switzerland.

The **danger-level model** was trained with a large data set of quality-checked danger levels spanning more than 20 years (Pérez-Guillén et al., 2022). A random-forest classifier (Breiman, 2001) uses 30 features, describing both meteorological conditions (24-hour averaged values) and snow-cover properties simulated with the SNOWPACK model. The classifier predicts the probabilities (Pr) for four of the five avalanche danger levels (1 (low) to 4 (high); danger level 5 (very high) is too rare to predict) .

The **instability model** assesses snow-cover simulations provided by the SNOWPACK model with regard to potential instability related to human-triggering of avalanches (Mayer et al., 2022). A random-forest model uses six variables describing the potential weak layer and the overlying slab to predict the probability that a snow layer is potentially unstable. The output probability ranges from 0 (a layer was classified as stable by all the trees) to 1 (classified as unstable by all trees). All simulated layers are assessed using this procedure. In the setup used for forecasting, the layer with the highest probability of instability (Pr_{instab}) is determined for each simulated snow profile and considered as decisive to characterize this profile, as suggested by Mayer et al. (2022).

For the purpose of this analysis, we relied exclusively on model predictions in *forecast*-mode as calculated in real-time during the forecasting season and available at 15.00 local time (LT), the time when forecasters meet to discuss and produce the forecast for the following day. From the *forecast*-predictions, we extracted the prediction valid for the following day at 12.00 LT. We used the predictions of the instability model and the danger-level model for the four slope aspects. For the instability model, we used Pr_{instab} as described before, for the danger-level model, we used the probability that the danger level was 3 (considerable) or higher (the sum of $Pr_{\text{D=3}}$ and $Pr_{\text{D=4}}$), referred to as $Pr_{\text{D}\geq 3}$. We opted for $Pr_{\text{D}\geq 3}$ rather than the predicted danger level, as this permitted analyzing the model in a similar way to the instability model, at the cost of a slight loss in discrimination power for avalanche conditions representing danger levels 1 (low) and 2 (moderate).

As we relied on real-time model predictions, in some cases data were missing. Moreover, due to a re-engineering of the data-model pipeline, predictions for the danger-level model in forecast-mode were only available from February 2024 onwards.

2.2. Avalanche forecast

We extracted the forecast danger level (D) and associated sub-level qualifier ($_{\text{s}}$, combined D_{s}) summarizing the severity of avalanche conditions related to dry-snow avalanches together with the indicated elevation threshold and aspect range from the avalanche forecast published by *WSL Institute for Snow and Avalanches SLF* (SLF) at 17.00 local time (LT), and valid until 17.00 LT the following day. For danger level 1 (low), no sub-level is available.

The sub-levels have been in use since 2017 (internally) and since Dec 2022 they have been published in the Swiss forecast (Lucas et al., 2023). Using sub-levels allows closer tracking of expected conditions compared to danger levels. On average, a higher forecast sub-level is generally related to more locations susceptible to avalanche release and to more avalanches of larger size (Techel et al.,

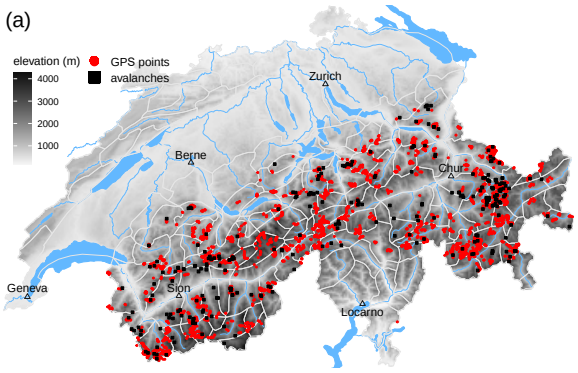

Figure 1: GPX points and human-triggered avalanches (backcountry touring).

2022).

2.3. Verification data: events and non-events

For this analysis, we used human-triggered avalanches – representing events, and points of GPX tracks in avalanche terrain – considered as non-events from the forecasting season 2023/2024 in Switzerland.

The GPX tracks were collected on *www.skitourenguru.ch* (website), where users can upload GPX tracks and have them rated with regard to avalanche risk (Schmudlach and Eisenhut, 2024). 928 different tracks, including time stamps, were uploaded during the winter of 2023/2024. Since we consider it unlikely that tracks were uploaded if people were involved in avalanches, we treat these tracks as proxies for non-events. Following post-processing of the GPX tracks – described in detail in Winkler et al. (2021) and Degraeuwe et al. (2024), this data set contains in total about 850'000 points. Following largely the criteria used by Degraeuwe et al. (2024), we extracted points if they were at a distance from controlled ski runs of $\geq$ 200 m, if they were at an elevation $\geq$ 1600 m and in potential avalanche terrain, defined by the maximum slope angle within 70 m distance (for details: Schmudlach, 2022, p. 10) being $\geq$ 30°. Lastly, in order to avoid auto-correlation, consecutive points from the same track had to be $\geq$ 200 m apart. The resulting data set comprised 3173 points in avalanche terrain representing backcountry touring activities.

From SLF's operational avalanche data-base, we extracted human-triggered dry avalanches, which were size 2 or larger, or in which at least one person was caught. As for the GPX tracks, we removed avalanches close to controlled ski runs. This resulted in 244 human-triggered avalanches.

3. METHODS

3.1. Spatial interpolation

We used *regression kriging* (Hengl et al., 2007) to spatially interpolate the point predictions from the location of the AWS to the locations of events and non-events. In addition, we set $Pr = 0$ for locations, for which observers provided an aspect-specific threshold of the snow line as this is the elevation below which avalanche release is not possible.

Some events or non-events were recorded on North-East, South-East, South-West or North-West aspects. To obtain interpolations for these points, we calculated the respective mean of the Pr-values, i.e., for North-East we calculated the mean of the North and East predictions.

3.2. Benchmark: the Swiss avalanche forecast

We used the forecasts as published in the Swiss avalanche bulletin (Section 2.2) as our benchmark for comparison. To do so, we checked whether a point (event or non-event) was within the elevation and aspect range as indicated in the bulletin. If this was the case, we assigned the forecast D_s to this point. If this was not the case, we applied the 1-level rule, subtracting one level from D_s published in the forecast. The 1-level rule is a rule-of-thumb, which has proven reliable to estimate the severity of avalanche conditions outside the indicated aspect and elevation range (SLF, 2023; Winkler et al., 2021). This adjusted danger rating is referred to as D_s^*. For the purpose of this analysis, we set $D_s^* = 1$ for cases, when the adjusted $D_s^* < 1$.

3.3. Analysis

We first determined whether the spatially-interpolated model output showed an increase in the probability of avalanche occurrence with increasing model-predicted probability. To do so, we binned the model-predicted probabilities (Pr) in bins of width 0.1. For each bin, we counted the number of non-events (nEv), in our case GPX track points, and events (Ev). Using these, we calculated the event ratio R

$$R_{m,i} = \frac{N(Ev)_{m,i}}{N(Ev)_{m,i} + N(nEv)_{m,i}}, \quad (1)$$

where $N(Ev)_{m,i}$ ($N(nEv)_{m,i}$) is the number of events (non-events) in each bin i, and for each model m.

To allow a comparison with our benchmark forecast, the combination of danger level and sub-level (D_s) interpreted using the 1-level rule (D_s^*, Section 2.2), we created bins of equal size. We thus ensured

that the comparisons between human-made forecasts and model predictions reflected the underlying patterns without being distorted by differences in the size of the respective groups. To obtain bins containing an equal number of data points for human-made forecasts and for model predictions, we first ordered the model-predicted probabilities from lowest to highest. To assign them to bins that were of equal size as the corresponding D_s^*-subsets, we derived the respective Pr-thresholds for each bin. For example, in the subset containing the predictions for the instability model, the sub-level proportions for the three lowest sub-levels were $D_s^* = 1$ (low): 40.9%, $D_s^* = 2-$: 18.0%, and $D_s^* = 2=$: 17.8%. Applying these percentiles to the ordered probabilities of the instability model resulted in thresholds for these three classes of $Pr_{\text{instab}} = [0, 0.266] \rightarrow$ bin 1, $Pr_{\text{instab}} = (0.266, 0.459] \rightarrow$ bin 2, and $Pr_{\text{instab}} = (0.459, 0.710] \rightarrow$ bin 3. Consequently, after splitting the model predictions using these thresholds, the bins contained the same proportion of data points as $D_s^* = 1$ (low) , $D_s^* = 2-$, and $D_s^* = 2=$. For higher sub-levels, we proceeded in the same way. In a second step, applying the same thresholds, we calculated the number N of nEv and Ev falling into each bin. Similar to before, we then calculated the event ratio $R_{m,i}$.

For visualisation purposes, we derived a relative ratio RR, by normalizing individual $R_{m,i}$-values using the overall base rate event ratio R_{m}, defined as

$$R_{\text{m}} = \frac{N(Ev)_{\text{m}}}{N(Ev)_{\text{m}} + N(nEv)_{\text{m}}} \tag{2}$$

This allows to calculate the normalized relative ratio as

$$RR_{m,i} = \frac{R_{m,i}}{R_{\text{m}}}. \tag{3}$$

Finally, we compared R-values for human forecasts and model predictions using a *Chi-Square test*. In addition, we derived the median of the factors F describing the increase in R for two consecutive bins (i.e., from bin 1 to bin 2, or from 1 (low) to 2-). In other words, we evaluated how well sub-levels (D_s^*) and model predictions discriminate on average between neighbouring bins/sub-levels.

4. RESULTS

Before comparing model predictions and human forecasts, we first analyzed whether models predict the expected increase in potential triggering locations.

Backcountry touring activity (non-events), as observed using GPX tracks, was highest when the danger-level model predicted low probabilities for $D \geq 3$ (considerable) ($Pr_{\text{D}\geq 3}$) and was lowest when $Pr_{\text{D}\geq 3} \approx 1$ (Figure 2a). Patterns for the instability model were similar, though much less pronounced. The distribution of events (human-triggered avalanches) showed opposite patterns with fewer events at low Pr-values, and increasingly more events with increasing Pr (Figure 2b). This increase was much stronger for the instability model compared to the danger-level model. For both models, probabilities were significantly higher when events occurred compared to non-events ($p < 0.001$, *Wilcoxon rank-sum test*). For example, the median $Pr_{\text{D}\geq 3}$ was 0.14 for non-events indicating that the danger-level model would have predicted either a danger level 1 (low) or 2 (moderate), while for events, the median $Pr_{\text{D}\geq 3}$ was 0.58, suggesting a tendency to danger level 3 (considerable) or even 4 (high). Large differences in the distributions were also observed for the instability model, with $Pr_{\text{instab}} = 0.77$ for events and $Pr_{\text{instab}} = 0.36$ for non-events.

The opposing trends seen for non-events and events indicate that the model-predicted probabilities captured the expected increasing frequency of (potential) triggering locations. This was confirmed when calculating the event ratio R (Eq. 1). As can be seen in Figure 2c, R increased strongly. Comparing the ratios between the respective highest ($Pr = 1$) and lowest ($Pr = 0$) bins shows that the ratio was 34 (danger-level model) and 24 (instability model) times higher. Note, however, that this increase is calculated by dividing with the R-value in the lowest bin, and is thus highly sensitive to small variations in the number of events. As the lowest bins are characterized by low numbers of events, a single event more or less will impact the calculated increase in R. Therefore, these numbers are, at best, indicative.

Mayer et al. (2022) suggested thresholds to classify predictions by the instability model into predictions indicating stability ($Pr_{\text{instab}} < 0.5$), potential instability ($Pr_{\text{instab}} \geq 0.77$), and potential instability but with a high false-alarm rate (Pr-values in between). Applying these thresholds, the event ratio R for human-triggered avalanches was 5.1 times higher when the model indicated potential instability compared to the model predicting stable conditions, and 2.4 times higher compared to the in-between class.

To compare model predictions with the avalanche bulletin, we assigned rank-ordered model-predicted probabilities to bins containing equal proportions of data points as recorded for the respective sub-level distributions (see Section 3.3). As can be seen in Figure 3, backcountry touring activity was highest when conditions were favorable and decreased with increasing forecast or model-predicted avalanche danger or instability (Figure 3a, b). For example, when conditions were forecast or predicted to be the most favorable – corresponding to $D_s^* = 1$ (low) or to bin 1 for the models – about 40% of the GPX points were recorded. In contrast, when avalanche danger

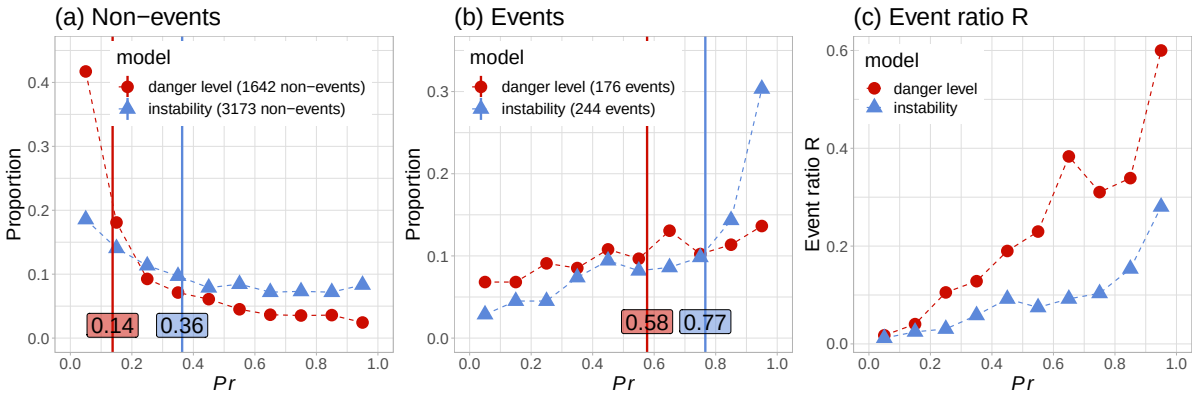

Figure 2: Distribution of model-predicted probabilities Pr, for ten bins of width 0.1 for (a) non-events (GPX track points) and (b) events (human-triggered avalanches). Shown are the proportion of data points in each bin. For each model, median Pr-values are shown for non-events and events (value and vertical line). In (c), the event ratio R is shown.

Table 1: Median factor F describing the increase in the event ratio R between consecutive bins in Figure 3e and f.

model/subset	prediction	F (median)
danger level	model	1.82
	bulletin	1.56
instability	model	1.58
	bulletin	1.75

was high – $D_s^* \geq$ 3+ corresponding to bin 7 – activity was strongly reduced ($\leq 1\%$). The distribution of non-events was similar for models and human forecasts.

Turning to the distribution of human-triggered avalanches during backcountry touring activities shows that the number of events increased strongly for the human forecast from $D_s^* = 1$ (low) to $D_s^* = 3=$, but dropped drastically at $D_s^* \geq$ 3+ (Figure 3d). In contrast, the models showed the largest number of events in bins 3 to 6, which would correspond to $D_s^* = 2=$ to $D_s^* = 3=$ (Figure 3c). Similar to the human forecast, the number of events is low(est) in the highest bin.

Having information on locations, where avalanches were triggered, and locations which were skied but where (most likely) no avalanche was triggered, allowed us to compare the relative increase in the likelihood of human-triggering of avalanches given a human forecast or a model prediction (Figure 3e, f). Overall, models and bulletin showed increasing relative ratios RR. The median increase in R between consecutive bins was approximately similar for models and human forecast (Table 1), as also confirmed by a *Chi-Square test* ($p > 0.05$).

5. DISCUSSION

We analyzed the performance of two spatially-interpolated models predicting the probability of avalanche occurrence for human-triggering and showed that increasing model-predicted probabilities correlated positively and strongly with the event ratio R (events to events plus non-events), which we consider a proxy for the likelihood of avalanche triggering by humans (Figure 2c). These results are in line with Soland (in prep.), who explored spatial predictions of the instability model in *nowcast*-mode using a multi-year data set of GPX tracks and human-triggered avalanches. For instance, Soland obtained similar median values for the instability model (non-events [2 years]: $Pr_{\text{instab}} \approx 0.35$, events [4 years]: $Pr_{\text{instab}} \approx 0.75$).

Using a smaller number of classes as the three stability classes proposed by Mayer et al. (2022) for the instability model may ease interpretation of model output, but resulted in a loss of discriminatory power between conditions predicted as stable and potentially unstable.

5.1. Limitations

For the purpose of this analysis, we assumed that the 1-level rule is a good approximation to apply the human-made avalanche forecast to locations outside the aspects and elevations indicated in the public avalanche forecast. Even though this rule-of-thumb has been used for many years to apply the bulletin to avalanche terrain during the planning phase of ski tours, there are likely better approaches, which reflect the more gradual – rather than step-wise – increase of avalanche danger with elevation and aspect (Winkler et al., 2021; Degraeuwe et al., 2024). At the same time, for the comparison of model predictions with human forecasts, we assigned rank-ordered, model-predicted probabilities to bins equal in size to the proportion of sub-levels. While this facilitated the comparison, it possibly split model predictions in an unfavorable way, potentially reducing discrimination capabilities of model predictions.

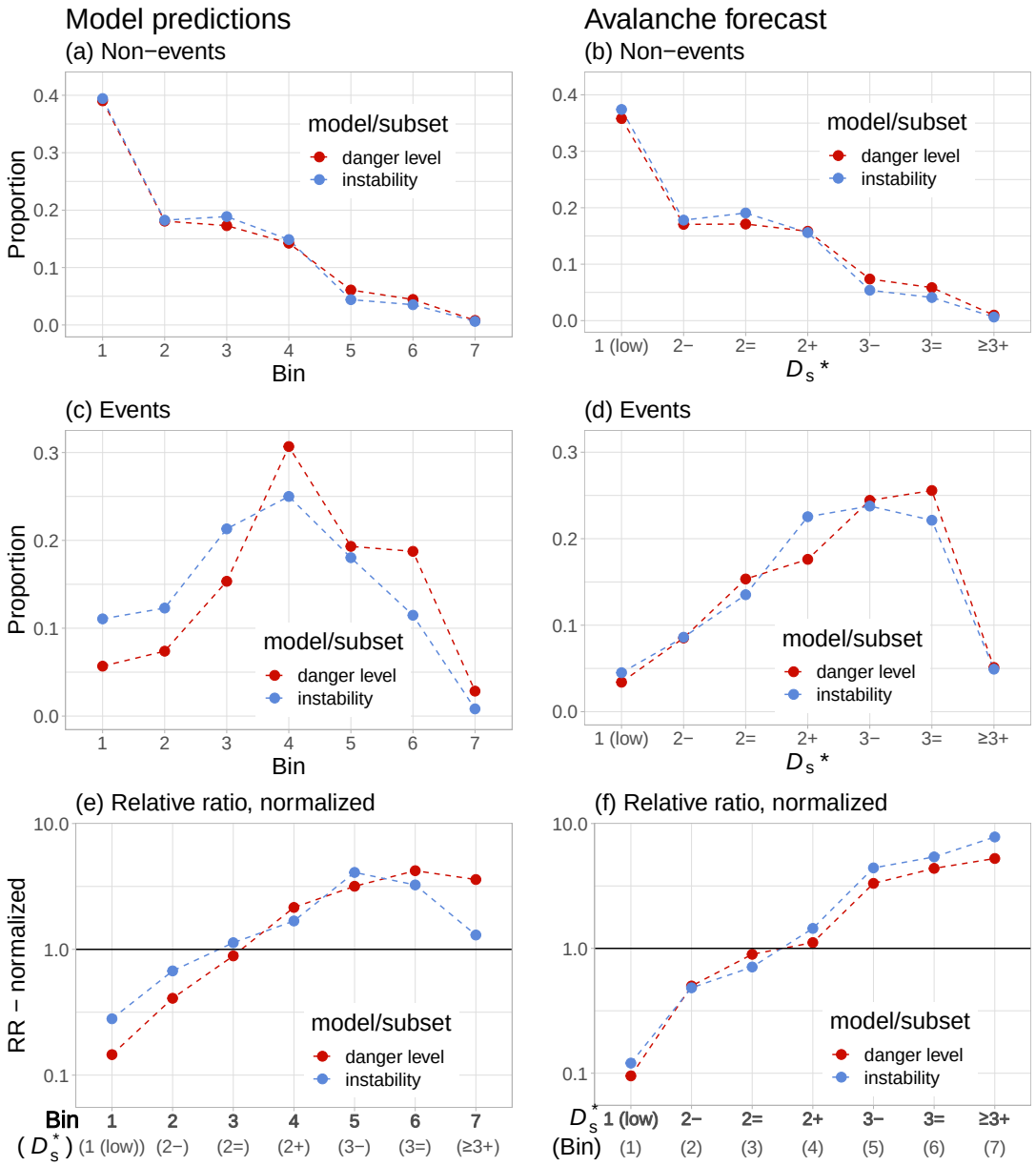

Figure 3: Comparing model predictions (left column) and avalanche forecast (right column). The distribution of non-events is shown at the top, the distribution of events in the middle row. The lowest row shows the relative ratio, normalized using the base-rate proportion of events. Note that the y-axis is log-transformed in (c). The analysis shows results for two models. To allow a meaningful comparison, the avalanche forecast contains the same data subset as available for the respective models. Danger-level model: 176 events, 1642 non-events, instability model: 244 events, 3173 non-events.

Backcountry touring activity decreases with increasing avalanche danger (level) (e.g., Techel et al., 2015). At the same time, less avalanche terrain is accessed on a tour at higher danger levels (Winkler et al., 2021). Thus, the data on events and non-events reflect adjustments in human behaviour due to forecast or encountered avalanche conditions. In contrast to human forecasts, model predictions remained unknown to forecast users.

5.2. Human vs machine

Keeping in mind the limitations related to the data and methodology, the results suggest that human-made forecasts and model predictions discriminate similarly (well) between conditions considered to be generally stable (i.e., D = 1 (low) or bin 1) and those considered the most susceptible to human-triggering of avalanches (i.e., $D \geq$ 3= or bin 6). Note, however, that forecasters had access to model predictions during forecast production and we assume that some of the information provided by the models already impacted the avalanche forecast. In contrast, no such information leakage existed the

other way round. This means that we compared purely data-driven, spatially-interpolated *model predictions* to *human-made forecasts including model predictions* interpreted using the 1-level rule. Moreover, while models only used meteorological measurements to correct for potential *forecast* errors, forecasters integrated avalanche observations and other field observations to assess current avalanche conditions. In summary, currently, the team of two or three human forecasters utilizing all available data and jointly producing the avalanche bulletin at SLF seems to perform about as good as predictions obtained from a model pipeline with no access to additional verification data.

5.3. Integrating model predictions in forecast process?

It is evident – from this study, but also from several other recent studies (e.g., Herla et al., 2023, 2024; Techel et al., 2022; Pérez-Guillén et al., 2024; Trachsel et al., 2024), that now is the time to integrate forecasting models closer and in a more systematic way into the avalanche forecasting process. We suggest that fully data- and model-driven forecasting pipelines become an integral part of avalanche forecasting, as they provide relevant input for decision-making or valuable "second opinions" (e.g., Purves et al., 2003; Maissen et al., 2024; Winkler et al., 2024).

In the future, as model performance continues to improve and eventually surpasses that of human forecasters, the shift to increasingly automated avalanche forecasting may become a possibility. However, to ensure that predictions are closely aligned with actual conditions, additional data sources must be integrated into model prediction pipelines - as for example, information from real-time avalanche detection systems (Trachsel et al., 2024).

While models performed well on average, there is a need to ensure they can handle unexpected situations, for which they had no training data. Therefore, mechanisms must be developed to detect and mitigate gross model errors during out-of-the-box events missed by a model.

Given recent developments, we believe that avalanche forecasts will be produced at greater spatial and temporal resolutions in the coming years. However, the resolution of such forecasts must correspond to the resolution that can be reasonably achieved given the available data and models. For example, in this study, we interpolated to very specific locations. However, we emphasize that spatially-interpolated model predictions only provide regional patterns (Techel et al., submitted).

Lastly, there is an ongoing discussion about the limited transparency of complex machine-learning models, such as random forest models, which are often considered 'black-box' models. Obviously, it is necessary to have at least a rough understanding regarding the relevant features in these models, and their impact on predictions. However, recently Pérez-Guillén et al. (2024) successfully applied an algorithmic approach called *SHapley Additive exPlanations (SHAP)* (Lundberg and Lee, 2017) making the danger-level models' decision-making process more transparent, not just globally but also for individual predictions.

6. CONCLUSIONS

We have shown that two spatially-interpolated models predicting avalanche danger and snowpack instability are capable to predict the expected increasing likelihood of human triggering of avalanches. Moreover, these model predictions increasingly reach the performance of human forecasts. Thus, model pipelines – as the ones discussed in this study, should become an integral data source in the avalanche forecasting process.

This conference proceedings paper is a summary of a more comprehensive analysis of human-triggered and natural avalanches, using three models, and comparing model performance in *nowcast*- and *forecast*-mode (Techel et al., submitted). In this more extensive analysis, we show that purely model-driven predictions discriminate almost as well between generally stable and rather unstable conditions as do human forecasts.

References

avalanche.org: North American Public Avalanche Danger Scale, URL https://avalanche.org/avalanche-encyclopedia/human/resources/north-american-public-avalanche-danger-scale/, last access: 12 Aug 2024, 2024.

Breiman, L.: Random forests, Machine Learning, 45, 5–32, doi:10.1023/A:1010933404324, 2001.

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, doi:https://doi.org/10.1016/j.coldregions.2024.104169, 2024.

Durand, Y., Giraud, G., Brun, E., Mérindol, L., and Martin, E.: A computer-based system simulating snowpack structures as a tool for regional avalanche forecasting, Journal of Glaciology, 45, 469–484, 1999.

EAWS: Standards: European Avalanche Danger Scale (2018/19), last access: 2021/11/17, 2021.

Hengl, T., Heuvelink, G. B., and Rossiter, D. G.: About regression-kriging: From equations to case studies, Computers & Geosciences, 33, 1301–1315, doi:10.1016/j.cageo.2007.05.001, 2007.

Herla, F., Haegeli, P., Horton, S., and Mair, P.: A large-scale validation of snowpack simulations in support of avalanche forecasting focusing on critical layers, EGUsphere [preprint], 2023, 1–38, doi:10.5194/egusphere-2023-420, 2023.

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 infor-

mation derived from distributed snowpack simulations, EGUsphere, 2024, 1–30, doi:10.5194/egusphere-2024-871, 2024.

Horton, S., Herla, F., and Haegeli, P.: Clustering simulated snow profiles to form avalanche forecast regions, EGUsphere, 2024, 1–24, doi:10.5194/egusphere-2024-1609, 2024.

Lehning, M., Bartelt, P., Brown, B., Russi, T., Stöckli, U., and Zimmerli, M.: Snowpack model calculations for avalanche warning based upon a new network of weather and snow stations, Cold Reg. Sci. Technol., 30, 145 – 157, doi:10.1016/S0165-232X(99)00022-1, 1999.

Lucas, C., Trachsel, J., Eberli, M., Grüter, S., Winkler, K., and Techel, F.: Introducing sublevels in the Swiss avalanche forecast, in: Proceedings International Snow Science Workshop ISSW 2023, Bend, Oregon, USA, pp. 240–247, 2023.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, edited by Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., vol. 30, p. 4768–4777, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf, 2017.

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, doi:10.5194/egusphere-2023-2948, 2024.

Mayer, S., Herwijnen, A., Techel, F., and Schweizer, J.: A random forest model to assess snow instability from simulated snow stratigraphy, The Cryosphere, 16, 4593–4615, doi:10.5194/tc-16-4593-2022, 2022.

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, doi:10.5194/nhess-23-3445-2023, 2023.

Pérez-Guillén, C., Techel, F., Hendrick, M., Volpi, M., van Herwijnen, A., Olevski, T., Obozinski, G., Pérez-Cruz, F., and Schweizer, J.: Data-driven automated predictions of the avalanche danger level for dry-snow conditions in Switzerland, Natural Hazards Earth System Sciences, 22, 2031–2056, doi:10.5194/nhess-22-2031-2022, 2022.

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

Purves, R., Morrison, K., Moss, G., and Wright, D.: Nearest neighbours for avalanche forecasting in Scotland: development, verification and optimisation of a model, Cold Regions Science and Technology, 37, 343–355, doi:10.1016/S0165-232X(03)00075-2, 2003.

Reuter, B., Viallon-Galinier, L., Horton, S., van Herwijnen, A., Mayer, S., Hagenmuller, P., and Morin, S.: Characterizing snow instability with avalanche problem types derived from snow cover simulations, Cold Regions Science and Technology, 194, 103 462, doi:10.1016/j.coldregions.2021.103462, 2022.

Schmudlach, G.: Avalanche Risk Property Dataset (ARPD), https://info.skitourenguru.ch/download/data/ARPD_Manual_3.0.13.pdf, last access: 2024/07/23, 2022.

Schmudlach, G. and Eisenhut, A.: Avalanche risk rating of user defined backcountry ski tours, in: Proceedings International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, 2024.

SLF: Avalanche bulletin interpretation guide, WSL Institute for Snow and Avalanche Research SLF, september 2023 edn., URL https://www.slf.ch/fileadmin/user_upload/SLF/Lawinenbulletin_Schneesituation/Wissen_zum_Lawinenbulletin/Interpretationshilfe/Interpretationshilfe_EN.pdf, edition September 2023; last access: 12 Aug 2024, 2023.

SLF: IMIS measuring network, doi:10.16904/envidat.406, URL https://www.envidat.ch/#/metadata/imis-measuring-network, last access: 12 Aug 2024, 2024.

Soland, K.: Towards automating avalanche forecasts: A kriging model to interpolate modeled snow instability in the Swiss Alps, in prep.

Techel, F., Zweifel, B., and Winkler, K.: Analysis of avalanche risk factors in backcountry terrain based on usage frequency and accident data in Switzerland, Nat. Hazards Earth Syst. Sci., 15, 1985–1997, doi:10.5194/nhess-15-1985-2015, 2015.

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, doi:10.5194/nhess-22-1911-2022, 2022.

Techel, F., Purves, R., Schmudlach, G., Mayer, S., and Winkler, K.: Forecasting avalanche danger: human-made forecasts vs. fully automated model-driven predictions, Natural Hazards Earth System Sciences, submitted.

Trachsel, J., Staehly, S., Richter, B., Mayer, S., Wahlen, S., van Herwijnen, A., and Techel, F.: On the value of radar-based avalanche detection data for the operational validation of model predictions and the regional avalanche forecast in Switzerland, in: Proceedings International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, 2024.

van Herwijnen, A., Mayer, S., Pérez-Guillén, C., Techel, F., Hendrick, M., and Schweizer, J.: Data-driven models used in operational avalanche forecasting in Switzerland, in: International Snow Science Workshop ISSW 2023, Bend, Oregon, USA, 2023.

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, doi:10.1016/j.coldregions.2021.103299, 2021.

Winkler, K., Trachsel, J., Knerr, J., Niederer, U., Weiss, G., Ruesch, M., and Techel, F.: SAFE - a layer-based avalanche forecast editor for better integration of machine learning models, in: Proceedings International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, 2024.