~~Forecasting~~ Can model-based avalanche forecasts match the discriminatory skill of human danger ~~: human-made~~ level forecasts~~vs. fully automated model-driven predictions~~? A comparison from Switzerland

Frank Techel[1], Ross S. Purves[3], Stephanie Mayer[1], Günter Schmudlach[2], and Kurt Winkler[1]

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

Correspondence: Frank Techel (techel@slf.ch)

Abstract. In recent years, ~~the integration of physical snowpack models coupled~~ physics-based snowpack models combined with machine-learning techniques ~~has become more prevalent~~ have gained momentum in public avalanche forecasting. When ~~combined~~ integrated 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~~ raises a key question: Are such ~~detailed spatial~~ spatially detailed model predictions sufficiently accurate for ~~use in operational avalanche forecasting~~operational use? We evaluated the performance of three ~~spatially-interpolated~~spatially interpolated, model-driven forecasts of snowpack stability and avalanche danger ~~by comparing them with human-generated public avalanche forecasts~~ in Switzerland over ~~two seasons as benchmark. Specifically, we compared the predictive performance of model predictions versus human forecasts using observed avalanche events (natural or human-triggered) and non-events. To do so, we calculated~~ three winters. As a benchmark, we used the official public avalanche danger forecasts, specifically focusing on the forecast danger level including the sub-levels. We assessed the ability of both model and human forecasts to discriminate between reference conditions and avalanche events – either naturally released or triggered by humans – by calculating event ratios as proxies for ~~the probability of avalanche release due to natural causes or due to human load, given either interpolated model output or the human-generated avalanche forecast. Our findings revealed that the event ratio increased strongly with rising predicted probability of avalanche occurrence, decreasing~~ release probability. Our results show that event ratios increased clearly with higher predicted avalanche probability, lower snowpack stability, or ~~increasing avalanche danger. Notably,~~ higher forecast sub-level. Overall, both model predictions and human forecasts showed ~~similar predictive performance. In summary, our results indicate thatthe investigated models captured regional patterns of snowpack stability or avalanche danger as effectively as human forecasts, though we did not investigate forecast quality for specific events. We conclude that these model chains are ready for systematic integration in the forecasting process~~a comparable ability to discriminate between reference and event conditions, with the event ratio increasing exponentially with increasing model-predicted probabilities or forecast sub-levels. However, the human forecasts – which incorporate model output – achieved a small but statistically significant advantage in discriminatory skill. This indicates that, while the evaluated models alone do not yet reach the full discriminatory power of human forecasters, their performance is

already approaching operational usefulness in a setup as used in Switzerland. As model quality is expected to improve further in coming years, it is essential to ensure their optimal integration into the operational forecasting workflow to realize the full potential of model-based support. Further research ~~is needed to explore how this can be effectivelyachieved~~ should explore how to implement this effectively, how to integrate real-time avalanche occurrence data into model prediction pipelines, and how to ~~communicate model-generated forecaststo forecast users~~validate increasingly high-resolution avalanche forecasts.

1 Introduction

Public avalanche forecasts aim to inform and warn recreational and professional forecast users about the danger of snow avalanches at a regional scale. In many countries, the expected probability of avalanche release, given a specific triggering level, and the potential size of avalanches is described by ~~summarizing~~ generalizing this information in one of five avalanche-danger levels (lowest: 1 (low) to highest: 5 (very high), EAWS, 2023; avalanche.org, 2024). Avalanche ~~danger is~~ conditions are then communicated using a mix of formats, including tabular, graphical or text formats including a mix of symbols, classes, or words (e.g., Hutter et al., 2021). These forecasts are produced by professional forecasters making judgments based on a variety of data sources, including measurements, observations, numerical weather prediction models, and – increasingly – predictions from ~~physically-based~~ physics-based snowpack models as *Crocus* or *SNOWPACK* ~~(i.e., Morin et al., 2020)~~(e.g., Morin et al., 2020) . The latter ~~are now often being used in combination~~ models are often combined with statistical models or machine-learning approaches ~~(Pérez-Guillén et al., 2022; Fromm and Schönberger, 2022; **?**)~~(e.g., Pérez-Guillén et al., 2022; Fromm and Schönberger, 2022; H , which aim ~~at making the~~ to make complex, multi-layered snow-cover simulations more accessible to forecasters by extracting and summarizing information relevant to the forecasting task ~~(e.g., **?**Herla et al., 2022; Maissen et al., 2024)~~(e.g., Horton et al., 2020; Herla . While forecasting chains such as *SAFRAN-Crocus-MEPRA* have been used operationally for many years ~~, as for instance *SAFRAN-Crocus-MEPRA*~~ in France (Durand et al., 1999), ~~it is now possible to run~~ recent advances now allow simulations at much higher spatial and temporal resolutions ~~than those at which forecasters typically operate by coupling numerical weather predictions models with physically-based snow cover models. These high resolution predictions can therefore also~~ – down to hourly scales and specific points – compared to the broader scales typically used in regional avalanche forecasting (e.g., hundreds of km^2, Techel (2020, p. 22)). These high-resolution snow-cover predictions can serve as valuable ~~hypothesis testing~~ practical tools for forecasters ~~in exploring snow-cover conditions and evolution of~~ – helping them test assumptions and refine their mental models, for instance regarding expected snowpack conditions and the evolution of snowpack stability. Moreover, spatially interpolating point or gridded predictions allows predictions for arbitrary points in space and time as well as backcasting for avalanche events at specific locations. In addition to providing reproducible forecasts at higher resolution, model-based forecasting is likely to free up expert time for other tasks – for example, communicating to professional and recreational mountain users~~with diverse backgrounds and skills.~~ .

To date, distributed ~~snow-cover~~ snow cover simulations or interpolated model predictions have ~~been validated using forecaster's best judgments (e.g., in Canada, **??**) or so-called face validity (**?**), or actual forecasts (e.g., Maissen et al., 2024). Mismatches between scales should be considered when comparing local snow-cover simulations and regional forecasts /judgments .~~

~~Thus, when~~ primarily been validated by comparison with forecasters' expert judgment – such as the forecast presence of weak layers, avalanche problems, or danger levels (e.g., Herla et al., 2024, 2025; Maissen et al., 2024). This form of indirect validation, where model outputs at very local scales are compared to regional forecasts or judgments prepared by experts makes interpreting lack of agreement challenging. When model predictions and human judgments/forecasts differ, it often remains unclear whether forecasters or models were wrong ~~(e.g., ?). Nonetheless, given numerous~~ (e.g., Herla et al., 2025). Given recent advances in snow-cover and snow-stability ~~modelling~~modeling, driven by developments in both ~~physically-based modelling~~ physics-based modeling and machine learning, we ~~ask~~ therefore pose the question: How close is public avalanche forecasting to transitioning from human-driven analysis to fully automated, model-driven methods? This raises the question: Are high-resolution model predictions "good enough" to complement or even replace those made by professional forecasters? To answer this, we need a benchmark ~~that defines~~ defining what "good enough" means. Given the challenges in validating avalanche forecasts in general, we define this benchmark through the use of traditional, primarily human-made public avalanche forecasts. Thus, we deem model-driven forecasts to be adequate when they independently ~~make similar forecasts of avalanche danger as expert forecasters~~forecast avalanche danger with a similar skill to expert forecasters, where both the human and the model forecasts are evaluated against objective data like avalanche occurrence.

Public avalanche danger scales are based on the ~~notion~~ principle that the likelihood, number, and size of avalanches ~~increases~~ increase non-linearly with ~~increasing avalanche danger (levels) (e.g., Techel et al., 2020a, 2022; Mayer et al., 2023; ?)~~rising avalanche danger levels (e.g., Schweizer et al., 2020; Techel et al., 2022; Winkler et al., 2021). In line with this ~~fundamental~~ core concept of public avalanche forecasting, we evaluate the discriminatory skill of spatially interpolated model predictions and human forecasts ~~using events (avalanches) and non-events, or proxies for non-events focusing primarily on the likelihood of avalanche release~~by comparing avalanche events with reference distributions that represent the base-rate conditions in the Swiss Alps over three forecasting seasons. This approach ~~allows us to compare models~~ enables an objective, data-driven comparison of model-based and human forecasts~~on objective data. We therefore aim at answering two~~ . Specifically, we address the following questions: (1) ~~Is the expected increase in the number~~ Do spatially interpolated model predictions reflect the observed increase in avalanche occurrence – either in terms of natural avalanches or ~~in~~ locations susceptible to ~~human-triggering of avalanches predicted by spatially interpolated model predictions? and~~ human triggering? (2) Do fully data- and model-driven predictions ~~achieve performances comparable to human-made avalanche forecasts~~of snowpack stability or avalanche danger distinguish between avalanche-relevant conditions with similar skill to that of human danger level forecasts in the public avalanche bulletin?

2 Models in support of avalanche forecasting

2.1 Recent developments

Recent years have seen rapid growth in the use of models aiming to support avalanche forecasting. Based on ~~physical~~ physics-based snow-cover simulations using the *SNOWPACK* or *CROCUS* models ~~(??)~~(Lehning et al., 2002; Vionnet et al., 2012), numerous statistical and machine-learning models have been developed to provide predictions of potential snow-cover instability

(e.g., Mayer et al., 2022)(e.g., Monti et al., 2014; Richter et al., 2019; Mayer et al., 2022), the likelihood of natural avalanche occurrence (??Mayer et al., 2023)(Viallon-Galinier et al., 2023; Hendrick et al., 2023; Mayer et al., 2023), the presence and characterization of specific avalanche problems (e.g., ???)(e.g., Reuter et al., 2022; Perfler et al., 2023), predictions of danger levels (Fromm and Schönberger, 2022; Pérez-Guillén et al., 2022; Maissen et al., 2024) or similarity assessments of simulated snow-cover profiles (Bouchayer, 2017; Herla et al., 2021) allowing spatial clustering of distributed snow-cover simulations (e.g. ?). Often (e.g. Horton et al., 2025). Typically, these models were trained and validated using observations or judgments made by observers or professional forecasters (e.g., Pérez-Guillén et al., 2022; ?; ?)(e.g., Pérez-Guillén et al., 2022; Herla et al., 2025; Pér
. With the aim to support forecasters in their decision-making process, some of these models have been included models are now used in operational forecasting processes - for instance in Canada (Horton et al., 2023), France (Morin et al., 2020), or Switzerland (van Herwijnen et al., 2023).

2.2 Models used in Switzerland

In the following, we briefly introduce three models, which are used in this study. These models provided live predictions during two forecasting seasons (2022/2023 and 2023/2024) in Switzerland . Avalanche in Switzerland and were accessible to avalanche forecasters at the WSL Institute for Snow and Avalanche Research SLF, responsible for producing the national public avalanche forecast, had access to these model predictions during forecast production. These three models are used to assess dry-snow (slab) avalanche conditions.

2.2.1 Danger-level model

The *danger-level model*, a random-forest classifier (Breiman, 2001), was trained with a large data set of quality-checked danger levels danger levels that had undergone judgment-based quality control (Pérez-Guillén et al., 2022). The model uses features describing both meteorological conditions and snow-cover properties simulated with the SNOWPACK model. The classifier predicts the probabilities (Prprobabilities ($\Pr(D = d)$) for four of the five avalanche danger levels (1 (low) to 4 (high)). This model was live-tested by forecasters during the winter seasons 2020/2021 - 2023/2024. From these, the most likely danger level can be extracted. In addition, two further model outputs are used operationally: (1) a continuous value derived from the probability-weighted sum of the danger levels, $\sum_{d=1}^{4} \Pr(D = d) \cdot d$ (Maissen et al., 2024; Pérez-Guillén et al., 2025); and (2) probability values linked to specific danger levels, such as $\Pr(D = 4)$.

2.2.2 Instability model

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). The 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

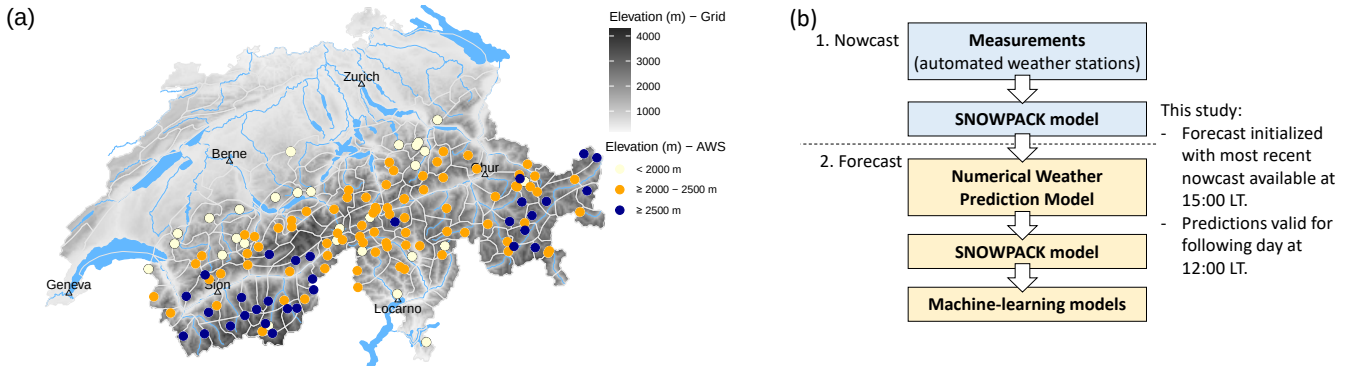

Figure 1. (a) Distribution of automated weather stations (AWS) in the Swiss Alps, at which *SNOWPACK* simulations were run. ~~The numbers show the elevation~~ DEM (source: Federal Office of ~~the station in a.s.l. divided by 100.~~ Topography swisstopo) (b) Schematic representation of the operational model pipeline for computing the *nowcast* and *forecast* predictions.

layers are assessed using this procedure. In the setup used for forecasting, the layer with the highest probability of instability (~~Pr_{instab}~~ $\mathrm{Pr}_{\mathrm{instab}}$) is determined and considered as decisive in characterizing this profile, as suggested by Mayer et al. (2022).

2.2.3 Natural-avalanche model

The *natural-avalanche model* is a simple one-parameter logistic regression model and comes in several variations: Its input either consists of the 1-day or 3-day sum of the simulated new snow or the output of the *instability model* (~~Pr_{instab}~~ $\mathrm{Pr}_{\mathrm{instab}}$) (Mayer et al., 2023). Trained ~~with~~ on a data set of natural avalanches ~~in the vicinity of an automatic weather station~~ near automated weather stations (AWS), ~~the models~~ models then predict the probability of at least one dry-snow ~~avalanches occurring in~~ avalanche of size 2 or larger occurring at the same aspect and elevation as used in the snow-cover ~~simulation~~simulations. In the operational setup, predictions from three models are combined using a weighted mean~~of the predictions using new snow amounts or Pr_{instab} as input is used by weighting the predictions from the~~ : the 1-day and 3-day new-snow models ~~with~~ each contribute a weight of 0.25~~and~~ , while the instability model ~~with~~ based on $\mathrm{Pr}_{\mathrm{instab}}$ contributes 0.5 (Trachsel et al., 2024). Including both short-term snow accumulation and snowpack instability ensures that the combined forecast reflects both the triggering potential due to new snow loading and structural weaknesses in the snowpack. We refer to the ~~predicted probability from this weighted approach as Pr_{natAval}~~ resulting probability as $\mathrm{Pr}_{\mathrm{natAval}}$.

2.2.4 Operational setup in Switzerland

In Switzerland, ~~operationally-used~~ operational snow-cover simulations are available at the locations of 147 AWS (SLF, 2024) , of which 142 are located throughout the Alps (Fig. 1a). Most of these stations are located at the elevation of potential avalanche starting zones ~~.~~ (median elevation: 2265 m a.s.l., min-max: 1258-2953 m a.s.l.). For *nowcast* predictions, SNOWPACK is driven using half-hourly or hourly measurements obtained from the network of AWS (Fig. 1b, step 1). The snow cover is simulated at 3-hour intervals at the location of the AWS for flat terrain and for four virtual slopes (North, East, South,

West) with slope angles of 38°, corresponding to typical avalanche terrain (Morin et al., 2020). In *forecast* mode (Fig. 1b, step 2), snow cover simulations are initialized using the most recent *nowcast* simulations~~(step 2)~~. Simulations are ~~then driven using the COSMO-1~~ driven by the 1 km-resolution numerical weather prediction ~~model~~ (NWP) ~~with 1 resolution as input (COSMO = Consortium for Small-scale Modeling (website)),~~ models operated by MeteoSwiss (COSMO1 until 2023/2024 (COSMO, 2025), ICON-EPS-CH1 from 2024/2025 onwards (MeteoSwiss, 2025)). All SNOWPACK input parameters obtained from the NWP models – such as wind speed, radiation, air temperature, humidity, and precipitation – are downscaled to the location of the AWS (Mott et al., 2023) using methods described in detail in Mott et al. (2023, Table 1). This provides forecast snow cover simulations up to 27 hours ahead with a temporal resolution of three hours. ML models provide predictions for flat terrain and for the virtual slopes at the location of the AWS for each of the 3-hour *forecast* and *nowcast* time steps. ~~For interpretation purposes~~Operationally, model predictions are primarily visualized on maps, sometimes as time series of predictions aggregated by region or elevation. To ease recognition of spatial patterns, predictions are interpolated in two-dimensional space.

3 Data and Methods

We ~~used data from the two avalanche forecasting seasons 2022/2023 and 2023/2024, described in detail the following Sections. Figure~~structure this section on data (Section 3.1) and methods (Section 3.1) around our two research questions (RQ): (1) Do spatially interpolated model predictions reflect the observed increase in avalanche occurrence – either in terms of natural avalanches or locations susceptible to human triggering? (2) Do fully data- and model-driven predictions of snowpack stability or avalanche danger discriminate conditions with similar skill as the human danger level forecasts published in the public bulletin?

Figure 2 provides an overview of the study design and illustrates how the data and methods ~~used for obtaining the three event-type specific data sets~~sections align with the two RQs. We first introduce the data: model predictions (Section 3.1.1), event and reference grid point datasets (Section 3.1.1), and the human avalanche forecast used as a benchmark for RQ2 (Section 3.1.1). We then describe the interpolation approach (Section 3.1.1) and the derivation of event ratios used in both analyses (Section 3.1.1). Finally, for RQ2, we explain how model outputs are transformed and analyzed to enable direct comparison with human danger-level forecasts (Section 3.1.2). Each step in the methods corresponds to elements in Figure 2 and links directly to the data processing and evaluations required to address the two research questions.

3.1 Data

We used data from three avalanche forecasting seasons (2022/2023 to 2024/2025).

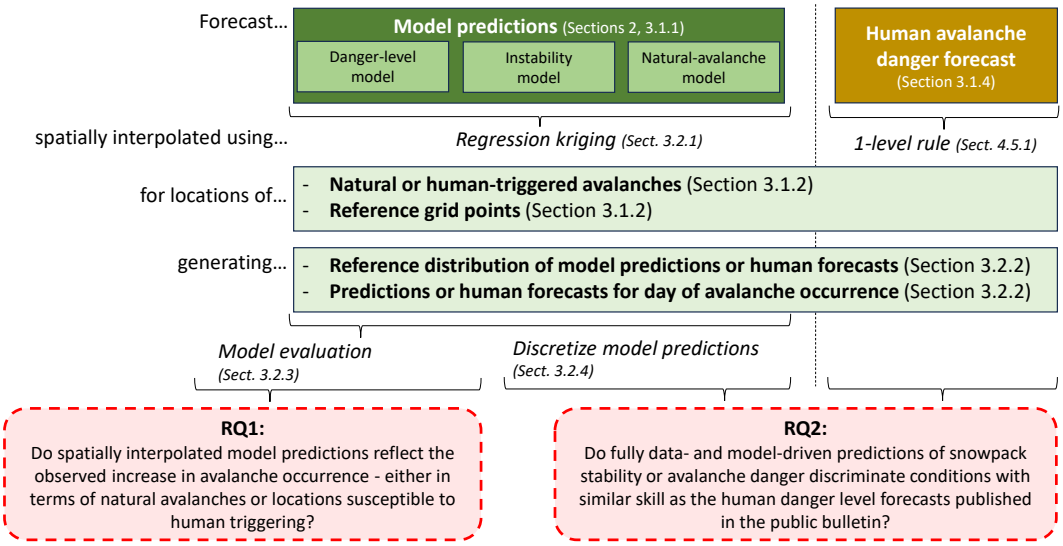

Figure 2. ~~Study layout~~**Study layout.** Forecasts – either model predictions (left) or human danger level forecasts (right) – are interpolated to locations of interest using regression kriging (for models) or the 1-level rule (for human forecasts). These locations include avalanche events and reference grid points. This results in two parallel datasets: ~~data~~ one focused on natural avalanche occurrence, and ~~data preparation~~one on locations susceptible to human triggering. From these, we derive distributions that reflect both the range of forecast conditions during the study period (base-rate distribution) and the conditions on event days. Depending on the research question (RQ, red boxes), either only models are analyzed (RQ1) or model predictions are compared to human forecasts (RQ2).

3.2 ~~Model predictions~~

3.1.1 Model predictions

We ~~analyse the model output~~ analysed the model predictions as described in Section 2.2.4. In all cases, predictions for specific aspects (e.g., N, E, S~~& W) are~~ , W) were used: the probabilities obtained with the *instability model* (~~Pr_{instab}~~$\Pr_{\mathrm{instab}}$) and the *natural-avalanche model* (~~Pr_{natAval}), and the predicted~~ $\Pr_{\mathrm{natAval}}$). For the *danger-level model*, we derived the probability for danger level ~~$D \geq 3$ (considerable) (*danger-level model*)~~$D \geq 3$(considerable), referred to as ~~$Pr_{\mathrm{D}\geq 3}$. In the latter case, we opted for $Pr(D \geq 3)$ rather than the predicted danger level D~~ $\Pr_{\mathrm{D}\geq 3}$. We chose this formulation instead of using the most likely danger level or the continuous ~~expected danger value, as this permitted analyzing the model in a similar way to the other models though at the cost of loosing some discrimination power at avalanche conditions representing 1 (low) and~~ probability-weighted sum of the danger levels described in Section 2.2.1, as it reduces the output to a probability between 0 and 1. Moreover, and in contrast to $\Pr_{\mathrm{D}=2}$ or $\Pr_{\mathrm{D}=3}$, $\Pr_{\mathrm{D}\geq 3}$ by itself differentiates well between the predicted most likely danger levels ~~(2 moderate) danger. However, $Pr(D \geq 3)$ was strongly correlated to the expected danger value (~~see also ~~Appendix Figure A1).~~ Figure A1 in the Appendix and Pérez-Guillén et al. (2025)) and is strongly correlated with the continuous probability-weighted value.

Table 1. Data overview for reference distribution (ref) and events (Ev) for the respective models and data subsets. Shown are the respective number of days or data points used in the analysis.

event type	model	days	ref	Ev
natural avalanches	natural avalanche[a]	299	8×10^5	1960
	instability	395	11×10^5	2660
	danger level[b]	386	11×10^5	2547
human-triggered avalanches	instability	395	11×10^5	1078
	danger level[b]	386	11×10^5	1046

[a] no data in 2022/2023, [b] no data in Dec 2023 and Jan 2024

This allowed us to interpolate (Sec. 3.1.1), visualize and analyze (Sec. 3.1.1) the danger-level model predictions in the same way as the other two models.

For the purpose of this analysis, we relied exclusively on model predictions calculated in real time during the forecasting season. Crucially, this means our evaluation is not based on reanalysis data, but rather forecasting of events in an operational context. ~~For the *forecast* predictions, we~~ We used simulations available at 15.00 local time (LT), the time when forecasters meet to discuss and produce the forecast for the following day (Figure 1b). From these, we extracted the ~~prediction~~ *forecast* predictions valid for the following day at 12.00 LT. ~~In addition to forecast predictions, we also used *nowcast* predictions, allowing us to estimate the effect of biases in the weather forecast input. For *nowcast* predictions, we extracted the same 12.00 LT time step as for *forecast* predictions.~~ Note that sometimes data were missing, either because the model was not available at the time (i.e., no data for natural-avalanche model in *forecast*-mode in 2022/2023 season, as it was only developed in 2023; Mayer et al., 2023), or due to a re-engineering of the data-model pipeline (no *forecast* predictions for danger-level model for parts of the 2023/2024 season).~~.~~

3.2 ~~Snow-line estimates~~

~~The AWS used for avalanche forecasting in Switzerland are primarily located at elevations at or above tree line, with few situated below 1700 . Due to the sparsity of data points at elevations below 1700 , we required an estimation of the elevation below which there was no continuous snow cover on steep slopes and where therefore no avalanche releases were possible. To obtain this threshold, we used daily estimates of the approximate elevation above which a continuous snow cover exists for steep North and South facing aspects as reported by study plot and field observers in Switzerland. This snow line is reported in 200 increments. In case the snow line cannot be seen by the observer, no estimate is being made. In total, about 19000 such estimates were available for the two seasons (North and South combined: about 100 per day~~(Table 1).

3.2 ~~Events, non-events and reference distributions~~

3.1.1 Events and reference grid points

We consider the reported occurrence of an avalanche triggered by natural causes or by human load as an event. ~~Events are described in detail in Section ??.~~

~~Defining non-events, on the other hand,~~ However, defining non-events is much more challenging ~~as~~ since non-events are ~~often~~ typically not reported, and ~~as~~ the absence of an observed avalanche does not mean there was no avalanche ~~(?Mayer et al., 2023)~~ (Hendrick et al., 2023; Mayer et al., 2023). We therefore ~~followed two paths: (1) we generated reference distributions – described in Section ??,~~ also extracted a subset of grid points from a digital elevation model (Federal Office of Topography swisstopo) to derive distributions representing the range of conditions over the study period as ~~reference, and (2) we used GPS points as proxies for non-events – described in Section ??. The latter can be considered to represent non-events – a person was at a specific point and triggered with rather high certainty no avalanche. These data have been repeatedly used for this purpose (Sykes et al., 2020; Winkler et al., 2021; Hendrikx et al., 2022; Degraeuwe et al., 2024). In contrast, using reference distributions limits comparisons to evaluating event conditions against the full spectrum of possible conditions. Therefore, reference distributions may be particularly suitable to evaluate the prediction performance for natural avalanches, as these do not require humans to be in avalanche terrain. In contrast, for human-triggered avalanches, where the presence of humans is required in avalanche terrain, the GPS tracks add another layer of informatio~~a reference. Since no assumptions are made about the occurrence of events at these locations, they may include the locations and conditions of avalanche events. However, ~~these data not only provide information on the presence of humans, they also reflect adjustments in human behaviour (i.e., choice of ski tour and or slopes skied) due to forecast or encountered avalanche conditions .~~ since avalanche events are rare, these reference distributions predominantly reflect non-event conditions and provide a baseline for comparing the conditions under which avalanches occurred.

3.1.2 ~~Events: avalanches~~

Events: avalanches

In Switzerland, approximately 80 observers or members of local avalanche commissions provide daily reports of avalanches occurring in their area of observation. Apart from avalanches documented by these observers, additional reports may come from field observers, who are also part of the observer network, or from the general public. The reported details of avalanches include their location and estimated time of occurrence, size categorized on a scale of 1 to 5 as per EAWS (2019), moisture content (classified as dry or wet), avalanche type (such as slab or loose-snow avalanche), and the triggering mechanism (such as natural release or human-triggered), following guidelines from SLF (2020). Location information generally refers to the top of the starting zone (coordinates, slope aspect, elevation). For the purpose of this analysis, we consider an **event** to have occurred at the location and date as reported.

Natural avalanches. We extracted all avalanches of *size 2 or larger*, classified as a *dry slab* avalanche with trigger type *natural release*. In total, ~~1855~~ 2977 avalanches fulfilled these criteria during ~~these two~~ the three seasons. These were located at a median elevation of ~~2505~~2509 m (IQR: ~~2280~~ 2307 - ~~2676~~2696 m).

Human-triggered avalanches. For human-triggered avalanches, we considered reported dry-snow slab avalanches with trigger type *human*, if the avalanche was either classified as *size 2 or larger* or if a person was caught in the avalanche. As a large share of these avalanches was reported by the public, we checked the location, size and moisture content for plausibility whenever possible. In total, during the ~~two seasons, 801~~ three seasons, 1223 avalanches fulfilled these criteria. Of these, ~~34% (273~~32% (386) were avalanches with at least one person being caught. Human-triggered avalanches were located at a median elevation of ~~2481~~2480 m (IQR: ~~2241~~ 2245 - ~~2713~~2708 m).

3.1.2 ~~Non-events: backcountry touring activity (GPS points)~~

~~We used a data set of GPS tracks collected on *www.skitourenguru.ch* (website), where users can upload GPS 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 GPS tracks~~ Reference grid points

In many regions and on many days, we lacked reliable information on locations where avalanches did not occur. To overcome these limitations and to enable meaningful normalization of event frequencies, we defined a reference set of locations that captures the range of conditions typically encountered across the forecast area and time period – without making assumptions about avalanche occurrence or human activity.

This artificial set serves as a neutral baseline against which observed events can be compared. It allows us to compute event ratios (described in detail in ~~Winkler et al. (2021) and Degraeuwe et al. (2024), this data set contains in total $>$ 850000 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~~Section 3.1.1) that are interpretable as relative likelihoods of avalanche occurrence under specific forecast or model-predicted conditions. By anchoring the denominator in a consistent, well-distributed sample of terrain relevant to both natural avalanche release and winter recreation (which is assumed to correlate with locations where human-triggered avalanches were recorded), we minimize biases due to missing or unevenly distributed non-event data since we normalize across potential avalanche rather than all terrain.

To generate reference points, we randomly sampled from a 1 km resolution elevation grid within an elevation range between 1600 and 3000 m a.s.l. This range was chosen to match the elevations at which most avalanche events were observed and where most of the AWS are located (Section 2.2.4), and as these are the elevations typically described in the public avalanche forecast and frequented by winter backcountry recreationists (e.g., Winkler et al., 2021). To better reflect the observed distribution of avalanche occurrences, we applied a kernel density estimate to the elevation distribution of observed avalanches when randomly

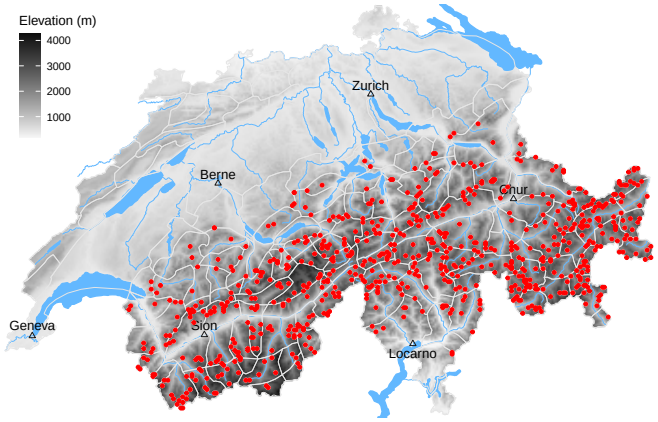

Figure 3. Map of Switzerland showing the spatial distribution of the randomly sampled subset of grid points used to obtain reference distributions. White polygon outlines delimit the spatial extent of the micro regions, the smallest spatial units used in the human avalanche forecast. (DEM and rivers/lakes - source: Federal Office of Topography swisstopo)

sampling 5% of grid points. As this approach did not fully cover all micro regions, the smallest spatial units used in human forecasts (see white polygon boundaries in Figure 3), we additionally sampled two random points above 1600 m a.s.l. from each of the remaining regions. The resulting ~~data set comprised 3998 points in avalanche terrain representing backcountry touring activities~~ reference set shown in Figure 3 comprises 709 grid points with a median elevation of ~~2298~~2402 m (~~inter-quartile range IQR: 1977~~ IQR: 2202 - ~~2569~~2603 m)~~, a slope angle of the steepest section within 70 m distance of 35° (32 - 39°). The median horizontal distance between any two points from the same track was 566~~. It is therefore both geographically well distributed and representative of the elevation range relevant to avalanche forecasting and winter backcountry recreation while permitting comparably efficient computation.

3.1.2 Snow-line estimates

The automatic weather stations (AWS) used for avalanche forecasting in Switzerland are primarily located at or above the tree line, with few stations situated below 1700 ~~(359 - 1039) and the difference in elevation 172~~ m (~~106 - 285 m). To have a corresponding data set of human-triggered avalanches,~~ Section 2.2.4). Due to sparse data coverage at lower elevations, we required an estimate of the elevation below which no continuous snow cover existed on steep slopes, and where avalanche release was therefore not possible. For this purpose, we used daily estimates of the ~~data-set of human-triggered avalanches described in the previous Section ?? were also filtered by distance to ski runs ≥~~ snow line – the approximate elevation above which a continuous snow cover was observed on steep north- and south-facing slopes – based on reports from study plot and field observers across Switzerland. The snow line is recorded in 200 m ~~. The resulting spatial distribution of GPS points (non-events) and corresponding avalanches (events) is shown in Figure3a.~~ intervals SLF (2020). If the snow line could not be visually confirmed, no estimate was reported. On average, over 100 snow-line estimates were available per day for each aspect.

Maps of Switzerland showing the spatial distribution of (a) GPS points and human-triggered avalanches (backcountry touring data set) and (b) the randomly sampled subset of grid points used to obtain reference distributions.

3.2 Avalanche forecast

3.1.1 Public avalanche forecast

To answer research question 2, we used the public avalanche forecast as benchmark forecast (Figure 2).

We extracted the forecast danger level (D) and the associated sub-level qualifier ($_{\text{s}}$, combined s, combined as 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 from the forecast published daily by the WSL Institute for Snow and Avalanche Research SLF at 17:00 local time (LT), and valid until 17.00 . This forecast is valid until 17:00 LT the following day. We extracted information related to the severity of dry-snow avalanche conditions, including the sub-level refining the danger level, and the indicated elevation threshold and aspect range. For danger level 1 (low), no sub-levelis available, elevation threshold, or aspect range is provided (SLF, 2023).

The sub-levels have been in use Sub-levels have been used internally since 2017 (internally) and since Dec 2022 they have been published (Techel et al., 2020b) and were publicly introduced in the Swiss avalanche forecast in December 2022 (Lucas et al., 2023). They provide a finer-grained and therefore more nuanced representation of the severity of avalanche conditions than the five danger levels alone. In general, higher 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 correspond to a greater number of locations prone to avalanche release and to more avalanches of larger size a higher likelihood of larger avalanches (Techel et al., 2022).

4 Methods

3.1 Methods

3.2 Spatial interpolation

We begin by describing the methods common to both research questions, and then outline additional steps specific to RQ2.

3.1.1 Spatial interpolation

We spatially interpolated point data (– specifically, model predictions and snow line estimates) to arbitrary points snowline estimates – to arbitrary locations in avalanche terrain, in our case to the locations of events, non-events and the random subset of grid points used as reference distributionincluding observed avalanche locations and the randomly sampled reference points. To do so, we utilized regression kriging [1] employed regression kriging (RK) , a geo-statistical interpolation (Hengl et al., 2007)

[1] see Hengl et al. (2007) for a detailed introduction

, a geostatistical method that combines ~~the regression analysis of the dependent variable using additional data and spatial interpolation of the residualsfrom the regression (Hengl et al., 2007). In our case, we used the location coordinates and the elevation as dependent variables to interpolate model predictions.~~ a deterministic regression model with kriging of the residuals.

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. 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.

Some ~~events or non-events~~ avalanche 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~~ compound aspects (e.g., NE, SW). For these cases, we approximated model predictions by averaging between the corresponding primary aspects. For example, for NE, we used the mean of ~~the~~ North and East predictions. We ~~proceeded in a similar way with the estimated snow line: for example, the snow line for Eastaspects was the mean of~~ applied the same logic to snowline estimates (Section 3.1.2): for East, we averaged North and South~~, the snow line for North-East was~~ ; for NE, we used a weighted mean of North (~~weight~~ 0.75) and South (~~weight~~ 0.25). For ~~locations and elevations, where observers estimated the snow line to be (Sect. 3.1.2), we set $Pr = 0$.~~ snowline elevation, any point below or at the snowline was assigned $\mathrm{Pr} = 0$.

To ~~find the best-possible kriging settings given the data~~optimize kriging performance, we tested several settings using leave-one-out ~~cross validation~~ cross-validation on a random subset of three cases ~~for each of the three models~~per model. Following best practice (e.g., Hengl et al., 2007), we ~~used~~ applied a logistic transformation ~~of Pr~~ to all Pr values prior to interpolation, setting ~~$Pr = 0$ and $Pr = 1$ to $Pr = 0.001$ and $Pr = 0.999$~~$\mathrm{Pr} = 0$ and $\mathrm{Pr} = 1$ to 0.001 and 0.999, respectively. ~~The kriging settings used for interpolation can be found in the R-script~~[1].All interpolation steps were implemented in R using the `sp` and `gstat` packages (Pebesma, 2004; Gräler et al., 2016; Pebesma, 2018; Pebesma and Bivand, 2023).[1]

3.2 ~~Benchmark forecasts: the Swiss avalanche forecast, interpreted using the 1-level rule~~

~~We used the forecasts as published in the Swiss avalanche bulletin (Section 3.1.1) as our benchmark for comparison. To do so, we checked whether a point was within the~~ Although we interpolated to specific locations – such as the known coordinates of avalanche release zones – we emphasize that interpolation yields regional patterns, not slope-specific predictions. As illustrated in Figure 4, holding 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~~ constant, the ~~forecast. The 1-level rule is a rule-of-thumb, which has proven reliable to estimate the severity of avalanche conditions~~

[1] ~~see link to script repository at end of manuscript~~

[1] Interpolation scripts are available in the code repository listed at the end of this manuscript.

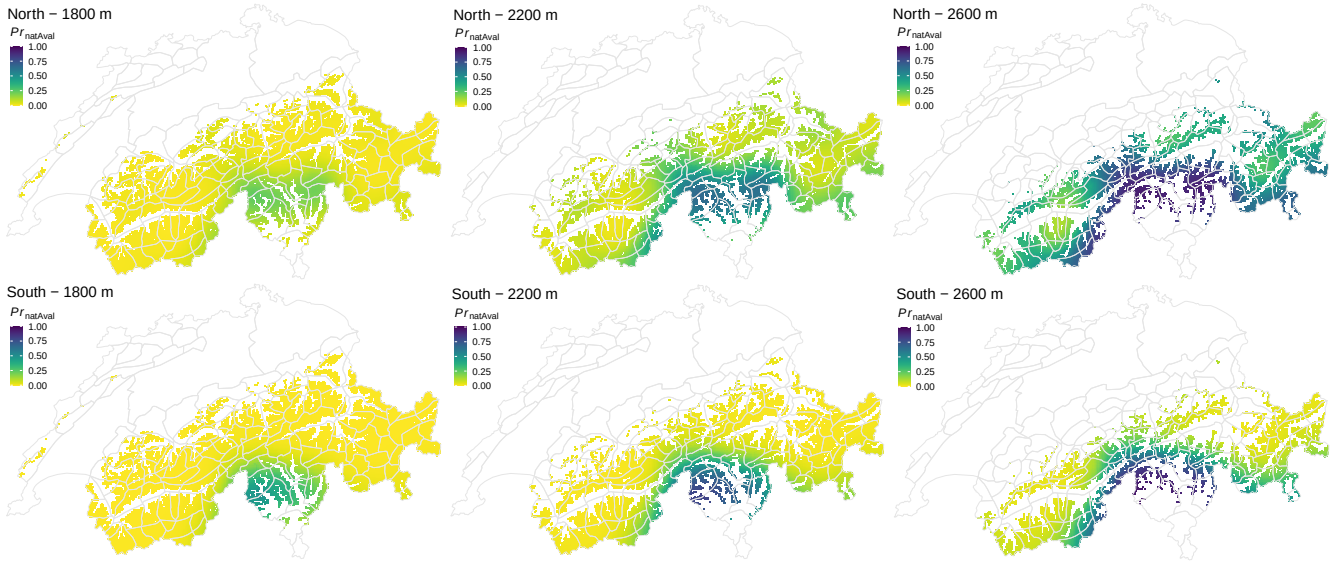

Figure 4. Interpolated predictions (natural-avalanche model, forecast-mode), valid for 11 March 2024 for three elevations (left: 1800 m, middle: 2200 m; right: 2600 m) and two aspects (upper row: North, lower row: South). For visibility, grid points 400 m below and all points above the indicated elevation are coloured. The elevation of these grid points is held fixed.

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 $D_s^* < 1$resulting spatial variation primarily reflects larger-scale gradients. Within-region differences arise mainly from variations in elevation (captured in the regression component) and aspect-specific snow cover modeling (reflected in the aspect dimension of the model inputs).

3.2 Deriving reference distributions for model predictions and avalanche forecast

As we often lacked reliable information for the non-occurrence of avalanches at specific locations (see previous Section **??**), we created representative reference distributions of D_s^* and model predictions. These reference distributions describe the range of conditions encountered throughout the investigated period without making assumptions about whether an event occurred or not.

3.1.1 Base-rate distributions of model predictions

To generate these reference distributions, we first defined an artificial data set of points, placed throughout the Swiss Alps, at elevations of interest for avalanche forecasting and winter recreation by randomly sampling 2.5% of the grid points from a 1 digital elevation model within the elevation range 1600 to 3100 . Within this elevation range, 95% of the natural and human-triggered avalanches were observed, with 1% at lower, and 4% at higher, elevations. Moreover, the critical elevation as indicated in the bulletin essentially always lies between 1800 and 2800 a. s.l. Applying this filter resulted in 666 points, with a

~~median elevation of 2203 a.s.l. (IQR: 1892 - 2507 a.s.l.). The spatial distribution of these points is shown in Figure 3b. Next, we calculated for each of these grid points and for~~ To describe the forecast conditions over the three winter seasons, we derived base-rate distributions for both the model predictions and the ~~four aspects~~ human-generated danger levels. For each reference location and for each of the four primary aspects (North, East, South, West~~, the danger level as forecast in the avalanche bulletin applying the 1-level rule (Section ??). Similarly to applying the human forecast to these locations, we computed~~), we first interpolated the model-predicted ~~probability~~ probabilities using regression kriging (Section 3.1.1). ~~This was done~~ These interpolations were performed for all days ~~, for which model predictions were available . Last, we combined model predictions with the avalanche bulletin by location and date. The resulting data sets contained only cases, when both model predictions and D_s^* were available.~~ on which model output was available (Table 1).

3.2 ~~Analysis~~

3.1.1 Evaluating model predictions with events and reference distribution

~~We first explored the agreement between the two prediction types (*forecast* and *nowcast*) for each model by calculating the Pearson correlation coefficient (r) and the difference between *forecast* and *nowcast* predictions. This allowed us to assess whether systematic bias exist in *forecast* compared to *nowcast* predictions, and to obtain an understanding of the magnitude of variation between these two.~~

~~We then addressed our first question, namely~~ To address the first research question – whether the models reflect the expected increase in avalanche occurrence probability with increasing ~~model-predicted probability . Binning the~~ predicted probabilities – we used a bin-wise event ratio approach that enables a direct and interpretable comparison between model outputs and observed avalanche activity. This evaluation framework quantifies how often avalanches occur relative to a reference baseline for each probability level. We consider this approach a reasonable approximation of the avalanche occurrence probability under specific (forecast) conditions. Furthermore, it is well suited for comparing probabilistic forecasts across models, as it highlights discriminatory power while accounting for differing event frequencies.

To account for sampling uncertainty — especially relevant given the relatively small number of events — we applied bootstrap sampling with replacement, repeating the procedure 100 times.

For each bootstrap sample, we binned the model-predicted probabilities (~~Pr) to bins of width 0.05 (for natural and human-triggered avalanches) and of width 0.1 (for backcountry), we~~ Pr) into intervals of width 0.05. For each model, we then counted the number of predictions ~~for representative grid points~~ falling into each bin from both the reference locations (ref) ~~or GPS track points (nEv) and for~~ and the avalanche events (Ev)~~falling into a bin, for each model - prediction-type combination. To investigate whether the probability of avalanche occurrenceincreases with increasing model-predicted probabilities, we calculated~~ . To quantify the relationship between predicted probabilities and avalanche occurrence, we computed the event ratio ~~R for cases when we relied on the reference distribution:~~ $R_{m,i}$ as:

$$R_{m,i} = \frac{N(Ev)_{m,i}}{N(ref)_{m,i}}, \tag{1}$$

~~and when using non-events:~~

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

where N ~~is~~ denotes the number of data points in ~~each~~ bin i ~~, and for each model and prediction type~~ for model m. ~~To assess whether the increase in~~ This ratio describes how often an avalanche event occurred relative to the reference baseline for each probability bin.

To facilitate comparison across models, we also computed a relative ratio (RR) by normalizing the event ratio in each bin using the overall base-rate event ratio R_m, defined as:

$$R_m = \frac{N(Ev)_m}{N(ref)_m}, \tag{2}$$

and

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

This normalization enables direct comparison of patterns across models, independent of absolute event frequencies.

Unless stated otherwise, we report $R_{m,i}$ or $RR_{m,i}$ as the median from the 100 bootstrap samples, and plot the corresponding 90% percentile intervals in the figures. To assess whether R ~~with increasing Pr was monotonic~~increases monotonically with Pr, we calculated the Spearman rank-order correlation coefficient (ρ) between $R_{m,i}$ and ~~$Pr_{m,i}$~~the midpoints of the corresponding probability bins $\mathrm{Pr}_{m,i}$.

~~To address our second objective, the comparison of spatially-interpolated model predictions with our benchmark forecast , the combination of~~

3.1.2 Analysis for RQ2: Evaluating the discriminatory skill of model predictions and human forecasts

To address RQ2, we assess how well interpolated model predictions and forecast sub-levels discriminate between conditions at reference locations and at avalanche events. To enable a direct comparison between continuous model outputs and the discrete sub-levels of the public avalanche forecast, we apply a consistent procedure for assigning forecast values to specific locations and for binning model probabilities. This is necessary because forecast avalanche danger is always the highest danger found in a region, usually for a specified range of terrain types, elevations and aspects.

This involves two additional steps (see also Figure 2).

Translating the regional avalanche danger forecast to local points

We used the published danger level and sub-level (D_s) ~~interpreted using the~~ from the Swiss avalanche bulletin (see Section 3.1.1) as a benchmark for comparison. For each point of interest, we checked whether it was located within the elevation and aspect range specified in the forecast. If so, we assigned the published D_s to that location. Otherwise, we applied the 1-level ~~rule (~~rule, subtracting one level to estimate local danger conditions. This rule-of-thumb has proven effective for approximating avalanche

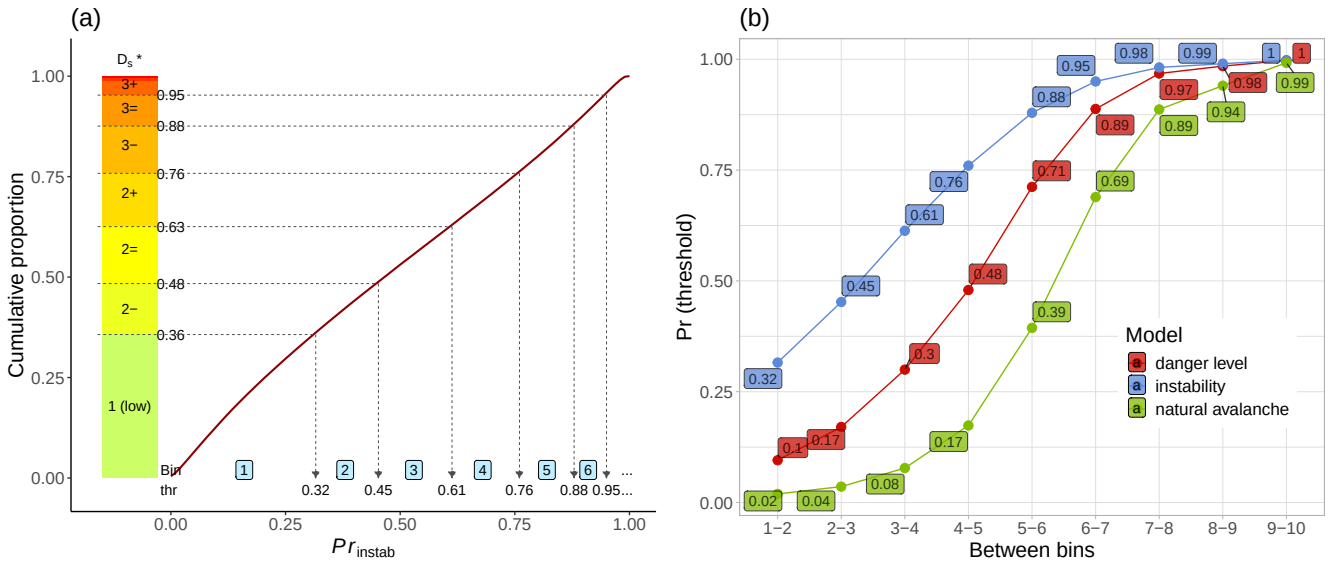

Figure 5. (a) Example using the instability model: cumulative distribution of predicted probabilities (Pr_{instab}), with bin thresholds derived from the proportion of forecast sub-levels D_s^* (y-axis). Model-predicted probabilities are assigned to bins (light-blue labels 1 to 6) such that each bin contains the same proportion of data as the corresponding sub-level class. The bold line shows the cumulative distribution of D_s^*; dotted lines indicate the thresholds (thr) separating bins. Note that thresholds for $D_s^* \geq 4-$ and bins ≥ 7 are not shown. (b) Probability thresholds corresponding to sub-level bins for the three models in forecast mode. The values shown for the instability model match those in panel (a).

danger outside the specified forecast ranges (SLF, 2023; Winkler et al., 2021). The resulting adjusted value is referred to as D_s^*, ~~Section3.1.1), a further preparatory step was necessary. As the resolution of~~ .

This approach allowed us to assign a representative danger rating to each location and aspect, consistent with how forecasts are typically interpreted in practice (SLF, 2023). It also approximates the expected reduction in danger with decreasing elevation or less exposed aspects (Winkler et al., 2021), and thus forms the basis for both the reference distribution of human forecasts – comparable to the reference distributions obtained for model predictions (Sec. 3.1.1) – and the conditions present when avalanches occurred.

Finally, we merged the model predictions and human forecast values by location and date, retaining only cases where both D_s^* ~~is limited to a discrete number of classes, we transformed model predictions to reflect these~~ and model predictions were available. The discriminatory power of both approaches is then evaluated using normalized event ratios, as described in Section 3.1.1. In addition, we calculate summary metrics to quantify discrimination strength. The following paragraphs outline the full procedure.

Binning model predictions to match forecast sub-levels

To compare spatially interpolated model predictions with the benchmark forecast (D_s^* ~~-classes by assigning them to bins of size equal to the sub-levels. Though being aware that this may potentially split model predictions in an unfavorable way, we deemed this a generally valid approach, as prior research had shown that model predictions correlated with D_s^* (Techel et al., 2022). Moreover, this step allowed to directly compare the underlying patterns in~~), we first transformed the continuous model outputs into a format compatible with the discrete forecast sub-levels. Since D_s^* consists of a limited number of ordinal classes, we binned model predictions such that the relative frequency of values in each bin matched the frequency distribution of D_s^*. This approach ensured that each bin corresponded to an equally sized sub-group, reducing distortions in the event ratio (R~~without being distorted by differences in the size of the respective groups.~~

~~To obtain bins containing an equal number of data points for human forecasts and for model predictions , we first ordered the~~) that might otherwise result from unequal sample sizes. It also enabled a fair and interpretable assessment of discriminatory performance across comparable categories.

To derive the bin thresholds, we ranked all model-predicted ~~probabilities from lowest to highest. To assign them to bins in a way that these were of equal size as the corresponding~~ probabilities in ascending order and divided them according to the cumulative proportions of D_s^* ~~-subsets, we derived the respective Pr-thresholds for each bin~~ in the corresponding reference set. As shown in Figure 5a, this procedure yields bin thresholds (thr) that align with sub-level frequencies. For example, in the subset ~~containing the predictions for the instability~~ combining the human avalanche forecasts and the corresponding instability model predictions, 36% of all D_s^* values were rated as 1 (low). Setting the upper threshold for bin 1 at $\Pr = 0.32$ ensured that 36% of the model ~~, the~~ predictions fell into this bin. Thresholds for subsequent bins (e.g., 2–6) were derived in the same way, based on the cumulative distribution. The resulting thresholds for all three models are shown in Figure 5b.

We selected this proportion-based binning strategy to allow direct and fair comparisons with the human forecast, which is issued in fixed sub-level ~~proportions 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~~ proportions. Alternative approaches such as unsupervised clustering or categorization based on internal model 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 ,~~ (e.g., the three-category classification in the instability model proposed by Mayer et al. (2022)) were considered. However, clustering lacks a direct correspondence to the forecast structure, and fixed thresholds are either unavailable or inconsistent across models. In contrast, 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~~ proportion-matching method provides a consistent and interpretable framework for comparing discriminatory performance between models and human forecasts.

After assigning bins, we proceeded as described in Section 3.1.1: we counted the number of observations from avalanche events (Ev~~falling into each bin . Similar to before, we then calculated the~~) and reference points (ref) in each bin or for each sub-level i, and computed the corresponding event ratio $R_{m,i}$ ~~.~~ (Equation 1) and relative event ratio $RR_{m,i}$ (Equation 3) for each model m.

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

$$R_m = \frac{N(Ev)_m}{N(ref)_m},$$

Evaluating the discriminatory power of model predictions and human forecasts

Lastly, to assess the discriminatory power between neighboring bins or sub-levels, we derived the factor F, which summarizes the average fold-increase in the relative event ratio RR between adjacent bins. For model m and bin i, we define:

$$F_{m,i} = \frac{RR_{m,i+1}}{RR_{m,i}} \tag{4}$$

~~when relying on the reference distribution, and~~

$$R_m = \frac{N(Ev)_m}{N(Ev)_m + N(nEv)_m}$$

~~for non-events. From these, we calculated the relative ratio as~~

$$RR_{m,i} = \frac{R_{m,i}}{R_{\mathrm{m}}}.$$

From the resulting set of values $F_{m,i}$, we computed two summary metrics: the median fold-increase $\overline{F}_m$, and the total fold-increase $F_{\mathrm{total},m}$, defined as the ratio of RR in the highest to the lowest bin:

$$F_{\mathrm{total},m} = \frac{RR_{m,\max}}{RR_{m,\min}} \tag{5}$$

~~The factor by which RR increases between two consecutive bins (i.e., from bin 1 to bin 2, or from 1 (low) to 2-) describes how well~~

Higher values of $\overline{F}_m$ and $F_{\mathrm{total},m}$ indicate a clearer separation of avalanche-relevant conditions, consistent with the expected, non-linear, increase in avalanche release probability across forecast sub-levels~~(D_s^*) and model predictions discriminate between neighbouring bins~~/, as shown by Techel et al. (2022). Conversely, values of $F_{m,i} \leq 1$ suggest that this expected monotonic increase between neighboring bins or sub-levels ~~. We therefore derived the median of the factors F over all consecutive bins summarizing how well sub-levels (D_s^*) and model predictions discriminate on average~~ is not observed. ~~In addition, we compared RR-values for human forecasts and model predictions using a *Chi-Square test*.~~

4 Results

~~Before addressing the two research questions, we compared *nowcast* and *forecast* predictions and the correlation between different model predictions (detailed results can be found in Table ?? in the Appendix). *Nowcast* and *forecast* predictions of the same model showed very high correlations ($r \geq 0.92$). No clear pattern emerged with regard to a bias. The correlation between different models was generally high ($r \geq 0.68$). The correlation between the natural-avalanche model and the instability model was lowest ($r \leq 0.70$), the correlation between danger-level model and the instability model the highest ($r \geq 0.88$).~~

4.1 ~~Model-predictions~~

~~Figure~~

4.1 Model predictions

Figure 6 summarizes model predictions ~~by model ((Pr) for each of the three models~~ (danger level, instability, natural avalanche)~~and prediction type (forecast, nowcast), separated for~~ , separated by avalanche type: natural avalanches (left column) ~~,~~ and human-triggered avalanches (~~middle column), and data related to back-country touring~~right column).

4.1.1 Reference distributions~~and non-events~~

~~Examining the reference distributions shows that low Pr-values of the natural-avalanche model were dominant, with almost 60% of the predictions at $Pr_{\mathrm{natAval}} < 0.05$, and increasingly fewer occurrences at higher Pr-values (Figure6a). Similar though less pronounced patterns can be noted~~ The reference distributions in Figure 6a and b represent the full range of model-predicted conditions across the study period. Note that the curves for the danger-level model ~~($Pr_{\mathrm{D}\geq 3} < 0.05 \approx 35\%$) and the instability model ($Pr_{\mathrm{instab}} < 0.05$: 15 – 20%), which is best seen in Figure 6b. Deriving the respective median value from the reference distributions shows that on average the probability for natural avalanches is low in the large majority of cases ($Pr_{\mathrm{natAval}} = 0.04$~~ and instability model in ~~*forecast*-mode, see Table ??). The median value for the~~ Figures 6a and b are identical, as they are based on the same underlying grid points and model predictions (see also Table 1).

For the natural-avalanche model, predictions were strongly skewed toward low probabilities: over 50% of grid points had predicted probabilities $\mathrm{Pr}_{\mathrm{natAval}} < 0.05$. The danger-level model exhibited a similar but less pronounced skew, with approximately 25% of predictions below $\mathrm{Pr}_{\mathrm{D}\geq 3} < 0.05$. In contrast, the instability model showed a more even distribution of predictions across the probability range.

The median predicted probabilities further underscore these differences: for the natural-avalanche model, the median was $\mathrm{Pr}_{\mathrm{natAval}} = 0.04$, suggesting that natural avalanche activity was rarely predicted at reference locations on an average day. For the danger-level model~~($Pr_{\mathrm{D}\geq 3} = 0.12$) corresponds to about the threshold between a predicted danger level~~ , the median prediction was $\mathrm{Pr}_{\mathrm{D}\geq 3} = 0.17$, indicating that conditions at reference points would typically have been predicted as levels 1 (low) ~~and~~ or 2 (moderate)~~(see also Figure A1b in Appendix), while for the instability model the median ($Pr_{\mathrm{instab}} = 0.31$) would be classified as *stable* according to the classification proposed by Mayer et al. (2022). Comparing the distribution of non-events (GPS tracks, Figure 6c) with the reference distribution in Figure 6b shows similar patterns and comparable median~~

values (Table ??). The instability model had a notably higher median value of $\mathrm{Pr}_{\mathrm{instab}} = 0.47$, although this still falls within the "stable" category, as defined by Mayer et al. (2022, stable if $\mathrm{Pr}_{\mathrm{instab}} < 0.5$).

Data availability and median probabilities (Pr) for events (Ev), reference distribution (ref) or non-events (nEv) for the respective models, prediction types and data subsets. event type model prediction type days ref / nEv Ev ref / nEv Ev natural avalanches natural avalanche nowcast 283 672383 1754 0.03 0.45 forecast[a] 143 351964 855 0.04 0.42 instability nowcast 283 672383 1754 0.30 0.80 forecast 236 562475 1554 0.31 0.80 danger level nowcast 298 711023 1791 0.14 0.76 forecast[b] 219 516735 1435 0.12 0.74 human-triggered avalanches instability nowcast 283 672383 737 0.30 0.74 forecast 236 562475 648 0.31 0.76 danger level nowcast 298 711023 762 0.14 0.60 forecast[b] 219 516735 621 0.13 0.60 backcountry[a] instability nowcast 129 3309 260 0.37 0.76 forecast 124 3173 244 0.36 0.77 danger level nowcast 129 3309 260 0.15 0.59 forecast[b] 73 1642 176 0.14 0.58

4.1.2 Events

The distribution of events showed different patterns (Figure 6d-f) between the models but also compared to the model-predicted distributions differed strongly between the reference distributions and non-events the event data set (Figure 6a-c).

While natural c and d). Focusing on natural avalanches first, we observe that avalanche events were approximately similarly distributed across the entire range of $Pr_{\mathrm{natAval}}$$\mathrm{Pr}_{\mathrm{natAval}}$-values (Figure 6d), the c), with a visible increase only for $\mathrm{Pr}_{\mathrm{natAval}} > 0.85$. For the other two models the number of natural avalanches increased considerably with increasing Pr and almost continuously with increasing Pr-valuesfor the other two models. The median value was $Pr_{\mathrm{natAval}} = 0.42$$\mathrm{Pr}_{\mathrm{natAval}} = 0.56$, indicating that on average the model predicted almost more than a 50% chance of a natural avalanche occurringat least one natural dry-snow avalanche occurring on event days. The median values were high for the danger-level model $Pr_{\mathrm{D}\geq 3} = 0.8$ and instability model $Pr_{\mathrm{instab}} = 0.74$ were high($\mathrm{Pr}_{\mathrm{D}\geq 3} = 0.79$) as well as for the instability model ($\mathrm{Pr}_{\mathrm{instab}} = 0.86$), which correspond to a model-predicted danger level well within danger level 3 (considerable) (Figure A1) and to about the threshold between profiles classified as potentially unstable and as potentially *unstable* according to the classification by Mayer et al. (2022) Mayer et al. (2022, unstable if $\mathrm{Pr}_{\mathrm{instab}} \geq 0.77$).

Human-triggered avalanches were more frequent when the models danger-level model and the instability model predicted higher probabilities (Figure 6e, fd). This pattern was much more pronounced for the instability model, with particularly many events when $Pr_{\mathrm{instab}} \geq 0.9$. The median values for $\mathrm{Pr}_{\mathrm{instab}} \geq 0.8$. Median Pr-values were lower for the data set of human-triggered avalanches were similar for the subsets containing all human-triggered avalanches (Fig. 6e) and the subset of events during backcountry touring activities (Fig. 6f). Pr-values were lower for human-triggered avalanches $\mathrm{Pr}_{\mathrm{D}\geq 3} = 0.57$, $\mathrm{Pr}_{\mathrm{instab}} = 0.78$) compared to natural avalanche events(Table ??).

PrPr-values differed significantly between events and non-events or reference distributions across all models and data subsets, regardless whether these were calculated in *nowcast*- or *forecast*-mode reference distributions for all models (Wilcoxon rank-sum test: $p < 0.001$).

4.1.3 Event ratio

The ratio ~~R~~ between the number of ~~events in relation to the reference distributions (Eq. 1) or to the number of non-events (Eq. ??),~~ avalanche events and the corresponding reference distribution, normalized by the overall mean event ratio, is referred to as ~~event ratio, provides an~~ the relative event ratio RR (Eq. 3). This metric provides a direct answer to our first research question~~, as it shows whether~~ : whether the model-predicted probabilities ~~capture the expected increasing frequency of (potential) triggeringlocations. As can be seen in Figure6g-h, the ratio R increased strongly, and in some cases in an exponential fashion, with increasing Pr for all models and data sub-sets~~ (reflect the expected increase in avalanche occurrence – either due to natural processes or human triggering.

As shown in Figure 6e, RR increased markedly with increasing probability value across all models for the natural avalanche dataset. The natural-avalanche model exhibited a strictly monotonic increase, with a perfect Spearman rank-order correlation ~~($\rho \geq 0.9$) .~~

~~Mayer et al. (2022) suggested thresholds to classify predictions by the instability model into predictions indicating stability ($Pr_{\text{instab}} < 0.5$) , potential instability ($Pr_{\text{instab}} \geq 0.77$),~~ ($\rho = 1$) between neighboring bins. The other two models also showed strong monotonicity ($\rho = 0.99$). Median increases between adjacent bins ranged from $\overline{F} = 1.24$ for the natural-avalanche model, to $\overline{F} = 1.35$ for the danger-level model, and $\overline{F} = 1.40$ for the instability model. The total increase in RR between the lowest and highest bins (corresponding to $\Pr < 0.05$ and ~~potential instability but with a high false-alarm rate (Pr-values in between). In the backcountry touring data set, the ratio R for~~ $\Pr > 0.95$) ranged from 72 (instability model) to 266 (natural-avalanche model), highlighting the models' ability to differentiate between stable and unstable conditions with respect to natural avalanche occurrence.

Results for human-triggered ~~avalanches was 5.1 times higher when~~ avalanches (Figure 6f) followed a similar pattern ($\rho \geq 0.98$), though the magnitude of increase was less pronounced. The corresponding median bin-to-bin increases were $\overline{F} = 1.22$ for the instability model ~~(*forecast* mode) indicated potential instability compared to the model predicting stable conditions, and 2.4 times higher compared to the in-between class. While using these three classes may help in interpreting modeloutputs, the resulting coarse classification clearly results in a loss of discriminatory power~~and $\overline{F} = 1.12$ for the danger-level model.

4.2 Comparison with benchmark forecast, the avalanche bulletin

We now compare ~~the model predictions (*forecast* mode) with the~~ model predictions with the sub-levels forecast in the public avalanche bulletin. To ~~make this comparisonpossible~~enable this comparison, we assigned the rank-ordered model-predicted probabilities to bins ~~containing equal proportions~~ such that each bin contained the same proportion of data points as the corresponding sub-level ~~distributions~~ class in the bulletin (~~described in Section3.1.1), from which we derived R. As there were (almost) no data points for backcountry touring data (human-triggered avalanches, GPS points)~~ see Section 3.1.1). Due to the very limited number of event data points at ~~$D_s^* > 3+$, R could not be calculated.~~ $D_s^* > 3+$ for the human-triggered avalanche data set, we combined the highest three sub-levels into a single class representing danger level 4 (high) or bin 8. In contrast,

for natural avalanches~~a sufficiently large number of data points~~ , a sufficient number of events was available even ~~for~~ in the highest bins. ~~As a consequence, the number of bins varies between the three event-type specific data sets, ranging between seven bins (from 1 (low) to $D_s^* \geq 3+$) for the backcountry data set (Fig. 7c) and ten bins (from 1 (low) to $D_s^* = 4+$) for natural avalanches (Fig. 7a). To ease interpretation of the event ratios R and to make them comparable across data sets~~ To allow meaningful comparisons across datasets and models, ~~these were normalized~~ we again normalized the event ratios R using the base-rate ~~event ratio~~ ratio obtaining RR (Eq. 3). Figure 7 shows the resulting relative ratios RR~~, while additional information on the respective~~ ; additional details on the distributions of events and non-events ~~is~~ are provided in the Appendix (Figures~~B1 - ??~~ B1–B2).

~~Overall, models and bulletin showed generally monotonically increasing relative ratios. Exceptions were the respective two highest bins for the human forecast for the~~ Both the models and the bulletin forecasts exhibited monotonously increasing RR values with increasing bin number or sub-level, with one exception: in the human-triggered ~~avalanches~~ avalanche dataset (Fig. 7b)~~and for the models in the backcountry data (Fig. 7c). The larger scatter between models in Figure 7c and the dropin RR in the respective highest bin (Fig. 7b, c) is likely due to the combination of few data points causing greater variability and the fact that the data is influenced by human behaviour.~~ , the final bin for the human forecast showed a drop, likely reflecting user adaptation to clearly dangerous conditions when danger level 4 (high) was published.

~~RR increased most strongly for natural avalanches~~ Overall, the results show that both model predictions and human forecasts discriminate avalanche occurrence with a comparable level of skill (Fig. 7a)~~, with the (median) factor F describing the increase from one bin to the next higher one ranging between 1.67 and 2.1, and a total increase between the highest and lowest bins by a factor F between 455 and 1520 (Tab. ??). For~~ . In both datasets – natural and human-triggered ~~avalanches and for the backcountry data (Fig. 7b, c), the median increase between bins ranged between $F = 1.39$ and $F = 1.82$. While the median increase is a robust measure describing the average RR-difference between neighbouring bins and not susceptible to individual extreme values,~~ avalanches – the relative event ratio (RR) increased steadily with higher model probabilities and forecast sub-levels, as expected. While the human forecasts showed slightly stronger increases, the ~~factor describing the increase~~ differences were modest. The average increase between adjacent bins ($\overline{F}$) ranged from 2.20 to 2.26 for 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 ~~is highly sensitive to small variations in the respective lowest and highest bins. For example, for the instability model in the backcountry data set F was merely 5, much lower than all the other values (Tab. ??). The highest bin for this model contained only two events and 20 non-events. If there would have been just one more event in this bin, this would result in $F = 7$, five events more would result in $F = 16$. Thus, the factor describing the total increase between the respective highest and lowest bins are indicative at best. Comparing the corresponding curves between model predictions and human forecasts showed no significant differences~~ (~~chi-square-test: $p > 0.05$). In summary,~~ F_{total}) ranged from 1206 to 1274 for the human forecasts, and from 286 (instability model) to 1163 (danger-level model). Statistical testing confirmed these differences were significant in most cases (Wilcoxon rank-sum test, $p < 0.001$), with the exception of the natural-avalanche model ($p = 0.08$). These findings highlight that, even though the human forecasts achieved slightly better

discrimination, the spatially interpolated model predictions – without human refinement – performed at a broadly similar level.

For human-triggered avalanches (Fig. 7b), the observed increases in relative event ratio (RR) were less pronounced than for natural avalanches, reflecting smaller effect sizes associated with human triggering. Still, both the human forecasts and model predictions ~~and human forecasts exhibited similar levels of discriminatory ability between bins~~showed a consistent increase in RR with higher sub-levels and probabilities. The human forecasts achieved $\overline{F}$ values between 1.53 and 1.67, compared to 1.46 and 1.48 for the models. Although the numerical differences were smaller than in the natural avalanche dataset, they remained statistically significant ($p < 0.01$). This again suggests that, while expert forecasts showed slightly better discriminatory power, model predictions without human input still captured meaningful differences in triggering likelihood.

~~Factor F summarizing the increase between any two neighbouring bins (Figure 7). Shown are the median increase, and the factor between the respective highest bin/sub-level and lowest bin/sub-level. event type model bulletin models bulletin models natural avalanches danger level 2.09 2.10 540 1520 instability 2.08 1.67 666 455 natural avalanche 1.79 1.86 748 1013 human-triggered avalanches danger level 1.80 1.53 28 114 instability 1.60 1.39 29 71 backcountry danger level 1.56 1.82 55 25 instability 1.75 1.58 65 5~~

5 Discussion

~~We analyzed the performance of three spatially-interpolated models predicting the probability~~ This study addressed two research questions: (1) whether spatially-interpolated model predictions for natural and human-triggered avalanches reflect observed variations in avalanche occurrence, and (2) whether these model-based predictions discriminate avalanche-relevant conditions as effectively as the human-generated sub-level forecasts published in the Swiss avalanche bulletin. We found that all three models captured patterns of avalanche occurrence ~~due to natural causes or due to human load, and showed that increasing~~ well: model-predicted probabilities ~~correlated positively and strongly~~ were strongly and positively correlated with the event ratio R (Figure 6), ~~which we consider a proxy for the probability of avalanche release . Moreover, we showed that model and human forecasts,~~ our proxy for avalanche release probability. Both the model predictions and the human forecasts – interpreted using the 1-level rule ~~, reach approximately equal performance in terms of discriminating between sub-levels~~ – showed clear, exponential increases in event ratio with rising bin or sub-level ~~equivalent model-predicted probabilities (Figure7). However, the underlying data and applied methods are prone to several limitations, which we discuss first, before reflecting on the interpretation of these findings in light of assumptions we made. Finally~~(Figure 7), with the human forecasts maintaining a small but consistent advantage.

In the following, we discuss ~~implications for avalanche forecasting more generally with respect to the adoption of model-driven forecasting processes.~~

5.1 ~~Limitations and assumptions~~

The data-sets of natural and human-triggered avalanches and GPS tracks represent only a small fraction of actual activity. For example, Degraeuwe et al. (2024) estimated that the GPS tracks in the dataset represent the activity of only about 1 in 2000 backcountry users. Moreover, eventsare likely not missing-at-random but are related to factors such as visibility, avalanche size, and the severity of incidents (e.g., Jamieson and Jones, 2015; Mayer et al., 2023). Moreover, there is uncertainty related to the exact location and timing of avalanches. For human-triggered avalanches, starting-zone coordinates and release date were checked for plausibility (and corrected if needed) during the seasons, for natural avalanches this was not possible.

We analyzed the predictions obtained from the operational model pipeline in real time. We made no attempts to improve any part of the pipeline or to remove outliers as these errors are part of the pipeline as are human-made errors in the case of the human forecasts.

We focused on the probability of avalanche occurrence, either due to natural causes or related to human-triggering. Avalanche size, which is expected to increase with increasing danger level (Schweizer et al., 2020; Techel et al., 2020a), was not analyzed in detail. Avalanche size, however, is reflected in human forecasts and is therefore also implicitly contained in the predictions by the danger-level model, as this model was trained using a historic data set of quality-checked avalanche forecasts (Pérez-Guillén et al., 2022). In contrast, both the natural-avalanche comparable discriminatory skill of model and the instability model were trained with a focus on estimating the probability of avalanche release due to natural causes or due to a human load.

For the purpose of this analysis, we assumed that human forecasts, the 1-level rule is a good approximation to apply the information provided in the human 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 (e.g., SLF, 2023), there are likely more suitable 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-orderedchallenges of verifying distributed predictions for rare and severe events, 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. key assumptions and limitations of our study, and implications for the future of avalanche forecasting – particularly regarding the integration of model-driven processes.

For part of this analysis, we relied on data reflecting human behaviour in avalanche terrain, which is known to be impacted by avalanche conditions (e.g., Winkler et al., 2021). If humans were fully ignorant of conditions and did not change their behaviour in response to forecasts or encountered conditions, the event ratio R would represent the probability of avalanche release due to human load. In contrast, if humans were perfectly able to detect all locations susceptible to avalanche release and avoid these, we could not use this data as a proxy for the probability of avalanche release. However, as reality lies somewhere in between these two extremes, we consider R to be a suitable proxy even though we do not know to what degree human behaviour impacts R-values, and whether and how this differs between model and human forecasts.

5.1 Comparable discriminatory skill – but humans still maintain a slight lead

5.2 ~~Comparison with similar studies~~

Model-driven forecasts can be considered successful when they independently achieve a level of discrimination comparable to that of expert-generated forecasts – and, crucially, do so at the spatial and temporal scales relevant to operational avalanche forecasting. Our results indicate that this aim is increasingly within reach: both model predictions and human forecasts showed consistently strong and exponential increases in the event ratio from stable to unstable conditions, suggesting that both approaches effectively reflect variations in avalanche occurrence probability.

~~Previously, several studies have shown that model predictions correlated~~ However, this similarity must be interpreted in light of an important asymmetry. While forecasters had access to model output during forecast production — likely influencing the final danger levels — the model predictions were generated without access to any human-generated information such as avalanche observations, recent activity reports, or field assessments. The comparison was therefore unbalanced: it contrasts purely model-based predictions with human forecasts ~~considering danger levels, sub-levels, but also aspect and elevation information (e.g., Techel et al., 2022; Mayer et al., 2023; **?**; **?**). In all these studies, point predictions were compared with the regional forecast or human judgments. Here, we applied the information provided in the human forecast and the model predictions to the exact location of events or non-events, similar to the studies by Winkler et al. (2021) and Degraeuwe et al. (2024) , who analyzed large data sets of backcountry touring activity using the avalanche forecast as input. They observed an increase in the chance to trigger and be caught in an avalanche with increasing danger level, calculated as in Equation **??**, resulting in avalanche risk increasing in a similar fashion as our findings . Our results are also in line with Soland (2024), who explored spatial predictions of the instability model~~ that integrated model data and benefited from broader situational awareness. Forecasters can draw on recent avalanche activity, on-the-ground snowpack observations, and knowledge of persistent weak layers – qualitative insights that are difficult to encode in current models but which meaningfully affect human forecast decisions. That model predictions nonetheless performed at a comparable level highlights the maturity of these modeling approaches and suggests they may already offer a robust and reliable foundation in situations with limited observational data – for example, in ~~*nowcast*-mode using a multi-year data set of GPS tracks and human-triggered avalanches. For instance, Soland obtained similar median values for the instability model for non-events (2 years: $Pr_{\mathrm{instab}} \approx 0.35$) and for events (4 years: $Pr_{\mathrm{instab}} \approx 0.75$; compare to Table **??**). Similarly, in a pilot study comparing the predictions of the natural-avalanche model with avalanches detected using automated detection systems for two systems in Southwest Valais (Switzerland), Trachsel et al. (2024) obtained median $Pr_{\mathrm{natAval}} = 0.50$ for events and $Pr_{\mathrm{natAval}} = 0.13$ for non-events. Again, these values are similar to what we observed.~~ remote regions.

In summary, our findings highlight that model chains are no longer merely supplemental tools – they are approaching a level of discriminatory skill that qualifies them as credible standalone components within avalanche forecasting workflows.

5.2 Verification of distributed predictions of rare and severe events

Avalanches are generally rare but potentially severe events. Exceptions are situations of widespread instability or when avalanches are very small. Public avalanche forecasts communicate the probability of these rare and severe events in a region through danger levels, or by using symbols or narrative text descriptions (e.g., EAWS, 2023; Hutter et al., 2021). They can therefore be considered a type of *rare and severe event forecast* (RSE), following the notion of Murphy (1991). Verifying RSE forecasts is ~~particularly~~ challenging due to the rarity of events, their localized nature, and the mismatch in scales between regional forecasts and local events. In practice, for a specific point in avalanche terrain within a region, the probability of avalanche occurrence is very low in most cases. We accommodated these challenges by interpolating to specific points and by evaluating the discriminatory power of human forecast and model predictions considering the increase in event ratio with increasing sub-level or model-predicted probability rather than by classifying forecasts and predictions using absolute terms as 'right' or 'wrong'. By doing so, we avoided comparing (distributed) model predictions ~~relying on~~ and treating forecasters' best judgments as ground truth as is often done due to a lack of objective data ~~(e.g., ?).~~ (e.g., Herla et al., 2025; Maissen et al., 2024). We consider this novel approach to be an important contribution of our study, since it allows us to objectively link avalanche forecast danger levels to events and reference locations representing the range of avalanche terrain – a long standing challenge (Schweizer et al., 2003).

Avalanche records are indicators of events; unfortunately, these are notoriously incomplete (e.g., Hafner et al., 2021). Automated avalanche detection systems using ground-based or airborne technologies have the potential to allow a much more systematic and continuous detection of events (e.g., Eckerstorfer et al., 2016; Fox et al., 2024; Hafner et al., 2022), particularly with regard to occurrence and absence of natural avalanches. However, the avalanche detection rate is impacted by avalanche properties including the type (wet or dry) and size of avalanches (e.g., Mayer et al., 2020; Hafner et al., 2021). Nonetheless, these systems likely provide the best means for obtaining increasingly complete avalanche records in the future, though they still do not resolve the issue of recording non-events under additional loads.

While an avalanche is a clear and objective indication that the snowpack was susceptible to triggering given a certain triggering mechanism (i.e., natural causes or additional loads from human activities) at the location and time of release, it ~~is conceptually~~ remains conceptually and practically more challenging to be certain of non-events, as these require continuous monitoring of avalanche activity at a specific location, and ~~–~~ – in case of triggering given additional loads such as a skier ~~– this~~ – also requires knowledge about whether a person skied a slope without releasing an avalanche. To our knowledge, GPS tracks are currently the most-widely used means to track actual terrain choices of ~~recreationists~~ backcountry users (e.g., Sykes et al., 2020; Winkler et al., 2021; Degraeuwe et al., 2024), and notionally also provide information on non-events. Note though that near misses or sloughs may have occurred, so these data are at best a proxy.

5.3 ~~Spatially highly-resolved predictions~~Limitations

~~As SNOWPACK simulations are driven at the location of automated weather stations (AWS)in Switzerland, we generated spatially distributed predictions by interpolating these point predictions. Although we interpolated to very specific locations~~

– such as the exact coordinates of the avalanche start zones, it is important to emphasize that interpolation can only provide regional patterns of individual parameters and not for specific slopes. As can be seen in Figure 4, keeping elevation and aspect fixed, interpolation results in primarily larger-scale patterns. Differences within a region are either related to differences in the elevation of grid points – the elevation gradient is modelled as part of regression kriging, and slope aspect – representing variations in aspect-specific snow-cover simulations. The data-sets of natural and human-triggered avalanches represent only a small fraction of actual activity. Moreover, there is uncertainty related to the exact location and timing of avalanches. For human-triggered avalanches, starting-zone coordinates and release date were checked for plausibility (and corrected if needed), for natural avalanches this was not possible.

Interpolated predictions (natural-avalanche model, forecast-mode), valid for 11 March 2024 for three elevations (left: 1800 , middle: 2200 ; right: 2600) and two aspects (upper row: North, lower row: South). For visibility, grid points 400 below and all points above the indicated elevation are coloured. The elevation of these grid points is held fixed.

In warning services in other countries, SNOWPACK simulations are directly driven on NWP grids for avalanche forecasting purposes (e.g., in Canada, **?**). Running grid-based snow-cover simulations directly at many locations offers the opportunity to analyze all simulated features at many more points, allowing for instance characterizing the type, stability and depth of simulated weak layers in a region or for specific grid-points (e.g., **?**). While this is an advantage compared to simulating the snow cover at a small number of AWS, there is a clear benefit of forcing SNOWPACK simulations with data from AWS up to the actual time and then adding the NWP forecast data to produce a forecast (setup described in Sect. 2.2.4) as the resulting snow cover simulations have greater similarity to reference profiles than simulations driven exclusively with NWP data (Herla et al., 2021; Binder et al., 2024). Similar observations were made when comparing predictions and actual occurrences of wet-snow avalanche activity. Again, the currently used forecasting-approach in Switzerland showed better correlations than a purely NWP-driven setup (e.g., **?**). We analyzed predictions made by an operational model pipeline in real time. We made no attempts to improve any part of the pipeline or to remove outliers as these errors are part of the pipeline as are human-made errors in the case of the human forecasts.

5.4 Human vs machine, or: human and machine?

We focused on the probability of avalanche occurrence, either due to natural causes or related to human-triggering. Avalanche size, which is expected to increase with increasing danger level (Schweizer et al., 2020; Techel et al., 2020a), was not analyzed in detail. Avalanche size, however, is reflected in human forecasts and is therefore also implicitly contained in the predictions by the danger-level model, as this model was trained using a historic data set of quality-checked avalanche forecasts (Pérez-Guillén et al., 2022) . In contrast, both the natural-avalanche model and the instability model were trained with a focus on estimating the probability of avalanche release due to natural causes or due to a human load.

We deem model-driven forecasts to be adequate when they independently make similar forecasts of avalanche hazard to those produced by an expert team. Keeping in mind the limitations related to data and methodology, our 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 avalanche release. Note, however, that forecasters had

access to model predictionsduring 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 For the purpose of our analysis, we assumed that the widely used 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 rule is a good approximation to apply the information provided in the human 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 (e.g., SLF, 2023), there are likely more suitable 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). Furthermore, for comparison between model predictions and 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. 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 well as a model pipeline with no access to additional verification data.

For this analysis, we generated a reference distribution that reflects the range of conditions encountered across the three forecasting seasons. This distribution served as our benchmark for comparison. Such an approach is particularly suitable for evaluating the probability of natural avalanche occurrence. However, in this study we did not investigate specific situations, which may be missed by models due to a lack of similar conditions represented in the training data. These situations may be important but infrequent, and will therefore hardly influence the global performance measures. Thus, we see a need for further studies utilizing different data sets, ideally from other forecasting areas, and methods, and focusing on rare conditions. Furthermore, future models designed to predict snow instability or avalanche danger should incorporate features that reflect the latest advances in understanding the physical processes behind avalanche formation. By integrating more physics into these models, predictive performance can be improved, even in conditions not represented in the training datahuman behavior in avalanche terrain is influenced by both forecasted and perceived conditions (e.g., Winkler et al., 2021). As a result, the reference distribution does not fully represent the true exposure of backcountry users, especially at higher danger levels: human presence in avalanche terrain tends to decrease at level 3 (considerable) and drops sharply at level 4 (high) (Winkler et al., 2021; Techel et al., 2024b). Nonetheless, the observed patterns in Figure 7b resemble those found by Techel et al. (2024a)[2]. Conceptually, if users were unaware of conditions and did not adapt their behavior, the event ratio R would approximate the true probability of avalanche release due to human load, at least at a relative scale. Conversely, if users perfectly identified and avoided all unstable slopes, R would underestimate this probability. In reality, behavior likely falls between these extremes. Therefore, while we consider the reference distribution appropriate for validating the probability of natural avalanche release, its representativeness for human-triggered avalanches remains uncertain – particularly because behavioral adaptations may differ depending on whether users see the human forecast (visible) or remain unaware of the model output (invisible).

[2] See also the respective analysis in the preprint of this study

5.4 Outlook and future directions

Our results show that ~~the performance of~~ avalanche forecasting model chains ~~has increased considerably~~ have matured substantially in recent years~~, reaching a level where these models can achieve a performance comparable to that of~~ . Their ability to discriminate avalanche-relevant conditions is now broadly comparable to human-made regional ~~avalanche forecasts ,~~ danger-level forecasts when interpreted using a simple 1-level rule.

~~It is evident from this study, as well as from several~~ This and other recent studies ~~(e.g., ??Techel et al., 2022; ?; Maissen et al., 2024; Trac~~ (e.g., Herla et al., 2024, 2025; Techel et al., 2022; Pérez-Guillén et al., 2025; Maissen et al., 2024; Trachsel et al., 2024) highlight that the time has come to integrate ~~forecasting models~~ model-based approaches more systematically into ~~the avalanche forecastingprocess. We therefore suggest that fully data- and model-driven forecasting pipelines become an integral part of avalanche forecasting. The integration of model predictions may be done either by relying on models as an additional data-source, by utilizing models to summarize relevant information (e.g., ?), or by providing independent "second opinions" valuable for the~~ operational avalanche forecasting. Such integration could take different forms: models may be used as additional data sources, as intelligent summarizers of key information (e.g., Horton et al., 2025), or as independent "second opinions" that support decision-making ~~process (e.g., Purves et al., 2003; Maissen et al., 2024; Winkler et al., 2024). In the future, as~~ (e.g., Purves et al., 2003; Maissen et al., 2024 .

To ensure constant alignment with real-world conditions, model chains must incorporate additional data streams – particularly real-time avalanche detection systems – and remain robust when confronted with unfamiliar conditions not represented in the training data. Advances in snowpack physics and hybrid approaches such as physics-informed machine learning (Raissi et al., 2019) offer promising avenues. Equally important is the spatially consistent integration of uncertainty.

As model performance continues to improve and ~~eventually surpasses~~ approaches – or surpasses – that of human forecasters, ~~the shift to increasingly automated avalanche forecasting may become a possibility. 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. Moreover, it must be ascertained that models are capable of predicting out-of-the-box situations, for which they had no training~~a shift toward more automated forecasting becomes increasingly feasible. In this context, forecasting systems must not only be accurate but also interpretable and resilient. Forecasting pipelines should include fallback strategies for data outages or infrastructure failures, with human expertise acting as a critical safeguard, especially in data-sparse or rapidly evolving situations.

~~Given the rapid growth of models to support avalanche forecasting, we believe~~ We also anticipate that avalanche forecasts ~~can~~ will be produced at ~~greater~~ increasingly high 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 . Moreover, spatial clustering of highly-resolved predictions~~ respect the limits imposed by data quality and availability. To ensure usability, clustering and aggregation techniques will be needed to ~~effectively and efficiently communicate avalanche conditions to forecast users. More generally, forecasters may spend more time explaining and communicating forecast outputs than generating themin the future.~~ communicate these detailed outputs effectively. As automation advances, the role of human

forecasters may shift – from producing forecasts to interpreting, validating, and communicating them. These tasks will remain essential for ensuring trust, credibility, and user comprehension.

~~Increasingly higher-resolved, automated predictions will eventually also become available to~~ Eventually, model-based forecasts will also become directly available to end users. In ~~theory, such information~~ principle, such products could already be ~~provided to users, and~~ offered today – ~~considering this study , with little reduction in prediction performance compared~~ as this study suggests, with minimal loss in predictive performance relative to human forecasts. However, ~~how will forecast users interpret such spatially highly-resolved information, which is still regional and~~ interpreting these high-resolution outputs remains a challenge. Despite their apparent precision, model predictions are not slope-specific~~? The platform~~ . The platform ~~skitourenguru (website)~~Skitourenguru ~~already provides one potential route to communicating~~ exemplifies one way to bridge this gap: by combining regional forecasts with terrain-based heuristics to produce location-specific ~~risk ratings of avalanche hazard through coupling the human forecast (regional scale) with highly-resolved terrain information (Winkler et al., 2021; Degraeuwe et al., 2024)~~ assessments (Winkler et al., 2021; Degraeuwe et al., 2024).

~~Our research suggests that the data~~In sum, avalanche forecasting is undergoing a transformation akin to that seen in weather forecasting over the past decades (Young and Grahame, 2024). Data, computing power~~and modelling techniques have finally~~ , and modeling capabilities have reached a point where machine-generated avalanche forecasts can ~~be produced which are broadly comparable with those created by humans. These developments mimic those in many other fields of human endeavour generally, and most specifically with respect to weather forecasting, where operational forecasting production has undergone a revolution in recent years (Young and Grahame, 2024). It will be important to discuss the limitations of using machine-learning approaches in avalanche forecasting (e.g., in forecasting rare events with very limited training data and modelling conditionsnot found in training data), developing methods to communicate higher resolution information, and defining the future~~ rival human forecasts in many respects. Key questions now lie ahead: How can ML-based models be adapted to rare or yet unseen conditions? How should high-resolution forecasts be communicated effectively? And what will be the evolving role of human ~~forecasters in avalanche forecasting. Furthermore, resilience of forecasting chains with respect to the potential unavailability of computational resources or automated measurements requires that human forecasting skill is maintained and developed.~~ expertise in an increasingly automated avalanche forecasting landscape?

6 Conclusions

We have shown that three spatially-interpolated models predicting avalanche danger, the probability of avalanche release, and snowpack instability are capable of predicting expected increasing probabilities of avalanche release due to natural causes or human load. Moreover, the ~~predictive~~ performance of these model predictions with regard to discriminating avalanche conditions with different severity are broadly comparable to the ~~performance~~ respective skill of human forecasters in an operational setting expressing this using danger levels including sub-level modifiers. Thus, fully data- and model-driven avalanche forecast pipelines – such as the ones discussed in this study, are ready to become an integral part of the avalanche forecasting process, mimicking changes to operational weather forecasting which have occurred over the last decades. Based on these find-

ings, we conclude that public avalanche forecasting may be reaching a point where a transition from primarily human-made forecasts to machine-generated forecasts is appropriate. The extensive network of real-time data and observations in Switzerland, coupled with high resolution weather forecasting model output, may provide a particularly appropriate setting for such developments. Nonetheless, more work is needed, including improving each step of forecasting pipelines, reliably predicting infrequently-occurring conditions, validating distributed or spatially-interpolated predictions, optimally integrating models in the forecasting process, and – lastly, but crucially – effectively communicating spatially and temporally highly-resolved forecasts and their uncertainties to forecast users.

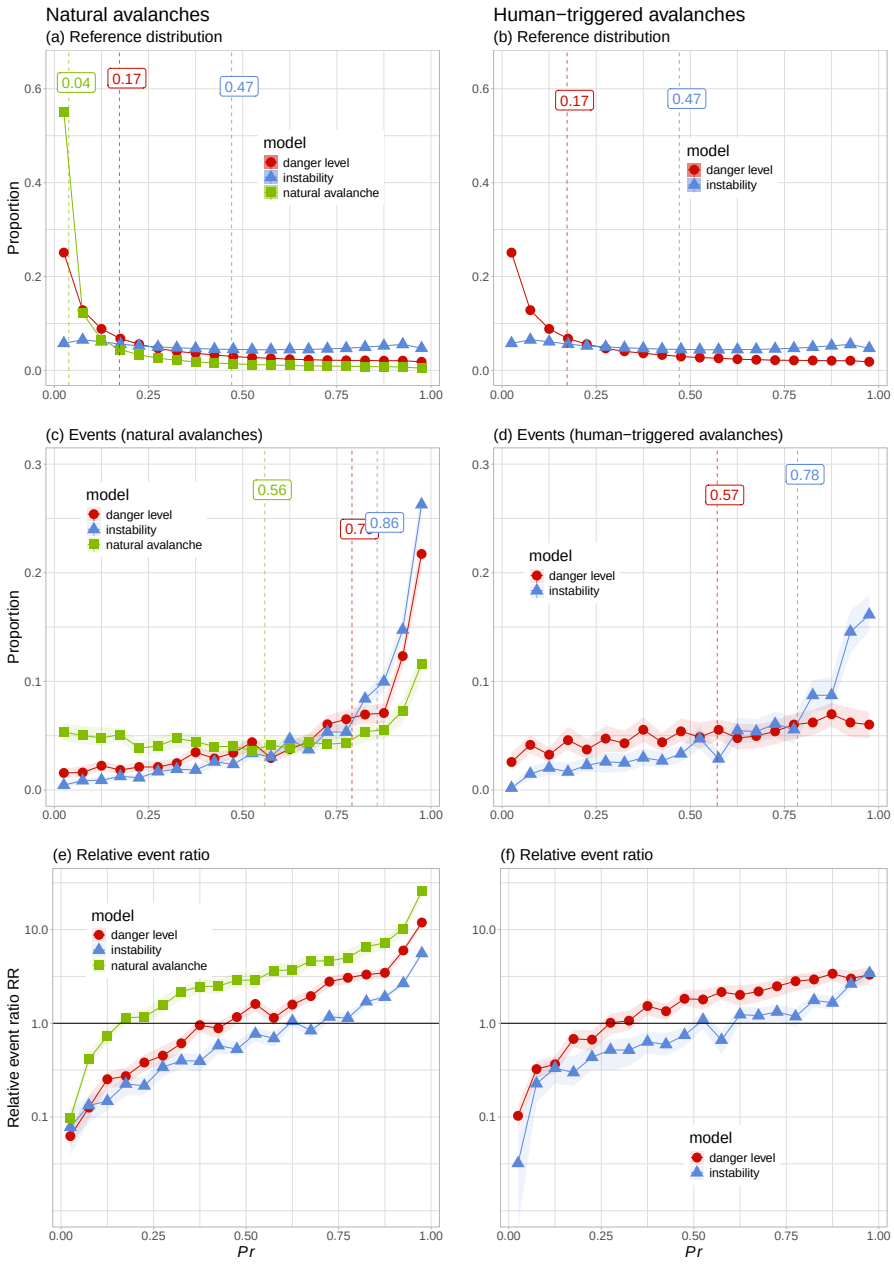

Figure 6. Model predictions (Pr). Columns show the respective results for (left) natural avalanches ~~, (middle) human-triggered avalanches,~~ and (right) ~~for data stemming from back-country touring activities~~human-triggered avalanches. ~~Upper~~ Top row: (a, b) reference distributions~~,~~ ~~simulated on the representative grid and (c) non-events(GPS points)~~; middle row: ~~events with~~ (c, d) ~~natural avalanches, (e) human-triggered avalanches and (f) human-triggered avalanches during backcountry touring~~event distributions; ~~lower~~ bottom row: (~~g-~~ie, f) event ratios. ~~Shown~~ Results are ~~the results~~ shown for each model~~(colour) and prediction type (shape)~~. To ~~allow better~~ facilitate comparison between models, proportions rather than absolute numbers are shown, where 100% ~~relates~~ corresponds to the ~~numbers shown~~ number of data points in Table~~??~~ 1. Median values are indicated. Shading in (c) to (f) indicates the 90% confidence interval. In (a) and (b), this is not visible due to minimal variation in the reference sampling.

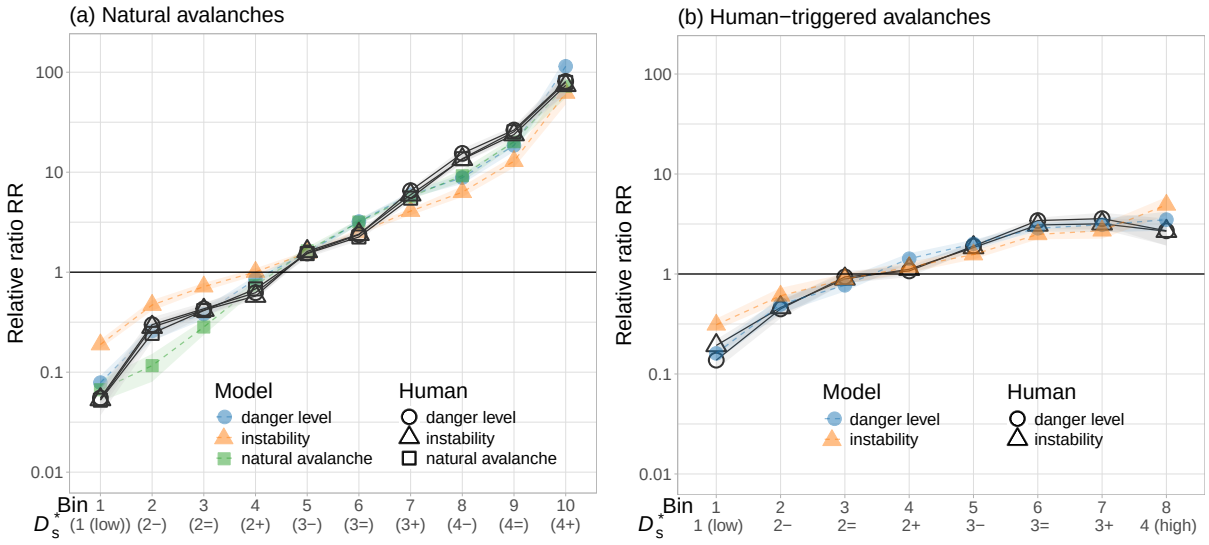

Figure 7. Ratio of avalanche events ~~(avalanches)~~ to ~~(a, b)~~ the reference distribution of model predictions ~~or (c) non-events (GPS tracks)~~and human sub-level forecasts, normalized by the overall base-rate ratio ~~of events~~ (Eq. 3) for (a) natural avalanches ~~,~~and (b) human-triggered avalanches~~, and (c) the backcountry data set~~. Note the log-scale on the y-axis. Model predictions are shown with coloured symbols; human forecasts for the same model subsets are shown with hollow black symbols.

7 ~~Appendix~~

Relationship between (a) $Pr_{\mathrm{D}\geq 3}$ and the expected danger value $E(D)$ and (b) the model-predicted danger level D and $Pr_{\mathrm{D}\geq 3}$ for the 56000 predictions of the danger-level model in *forecast*-mode during the 2022/2023 season.

Spearman correlation r between models and prediction types (*forecast*, *nowcast*). For the comparison between prediction types for the same model, the mean μ and standard deviation sd are shown. model prediction type correlation difference danger forecast-nowcast $r = 0.98$ $\mu = 0.005$, $sd = 0.06$ instab forecast-nowcast $r = 0.97$ $\mu = 0.0004$, $sd = 0.08$ natAval forecast-nowcast $r = 0.92$ $\mu = -0.001$, $sd = 0.09$ danger-instab forecast $r = 0.90$ danger-natAval forecast $r = 0.92$ instab-natAval forecast $r = 0.71$ danger-instab nowcast $r = 0.88$ danger-natAval nowcast $r = 0.85$ instab-natAval nowcast $r = 0.68$

Natural avalanches. Left column: model predictions, right column: human avalanche forecast. Upper row: reference distributions, middle row: events (natural avalanches), bottom row: relative ratio RR (Eq. 3).

Human-triggered avalanches. Left column: model predictions, right column: human avalanche forecast. Upper row: reference distributions, middle row: events (human-triggered avalanches), bottom row: relative ratio RR (Eq. 3).

Backcountry touring data. Left column: model predictions, right column: human avalanche forecast. Upper row: non-events (GPS tracks), middle row: events (human-triggered avalanches, subset backcountry touring), bottom row: relative ratio RR (Eq. 3).

Acknowledgements. We greatly appreciate the in-depth constructive reviews by Florian Herla, Christoph Mitterer, and Pascal Haegeli. We thank Marc Ruesch and Andrea Helfenstein, who implemented the operational SNOWPACK - model pipeline during 2023/2024, including the three models described in Section 2. We also benefited from Katia Soland working in parallel on her Master thesis (supervised by FT and RP), in which she explored kriging algorithms using the instability model and a larger data set of backcountry touring data (Soland, 2024). Marc Ruesch, Andrea Helfenstein, Cristina Pérez-Guillén, and Katia Soland provided feedback on an extended abstract of this manuscript, submitted to the *International Snow Science Workshop 2024* in Tromsø, Norway (Techel et al., 2024a).

Appendix A: Methods

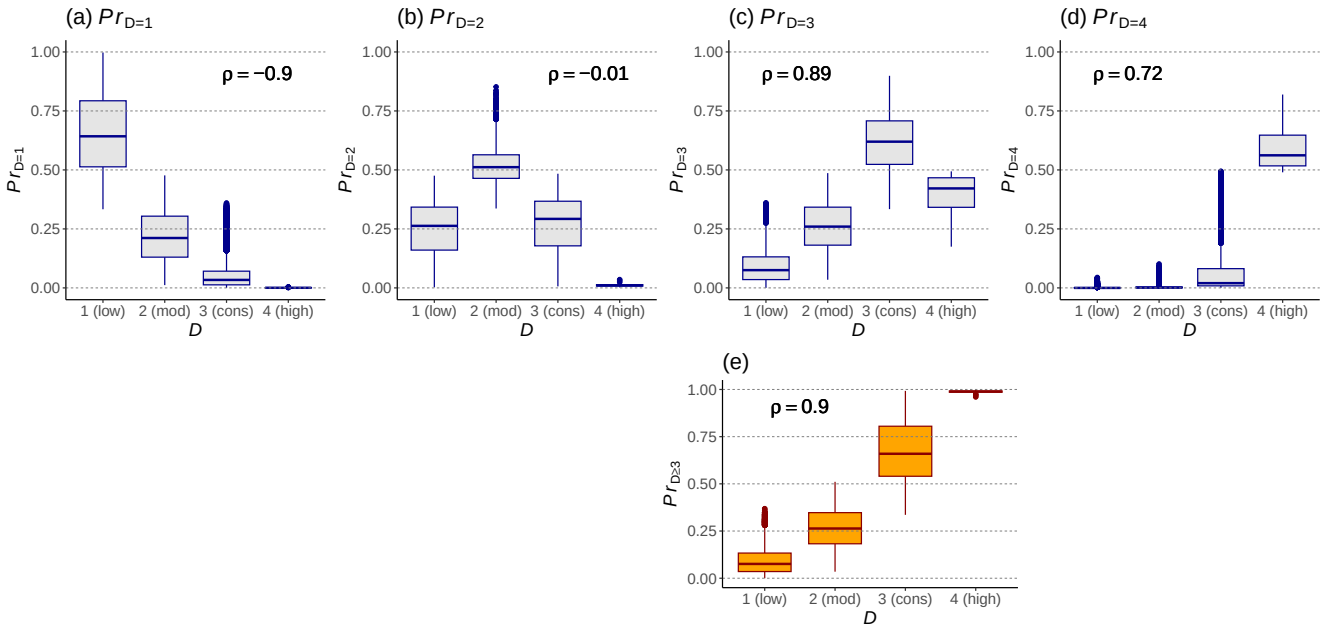

Figure A1. Relationship between the model-predicted danger level D and the predicted probabilities for individual danger levels $\Pr_{D=1}$, $\Pr_{D=2}$, $\Pr_{D=3}$, and $\Pr_{D=4}$ shown in panels (a)–(d), and for $\Pr_{D\geq 3}$ in panel (e). All predictions were generated by the danger-level model in *forecast* mode for the 2022/2023 season ($n = 56000$). Each panel also shows the corresponding Spearman rank correlation coefficient ρ between D and the predicted probability. Panel (e) shows the aggregated probability $\Pr_{D\geq 3}$, which was used in the main analysis of this study. It has, together with $\Pr_{D=1}$, the highest (absolute) correlation with D. For an in-depth analysis refer to Pérez-Guillén et al. (2025).

Appendix B: Results

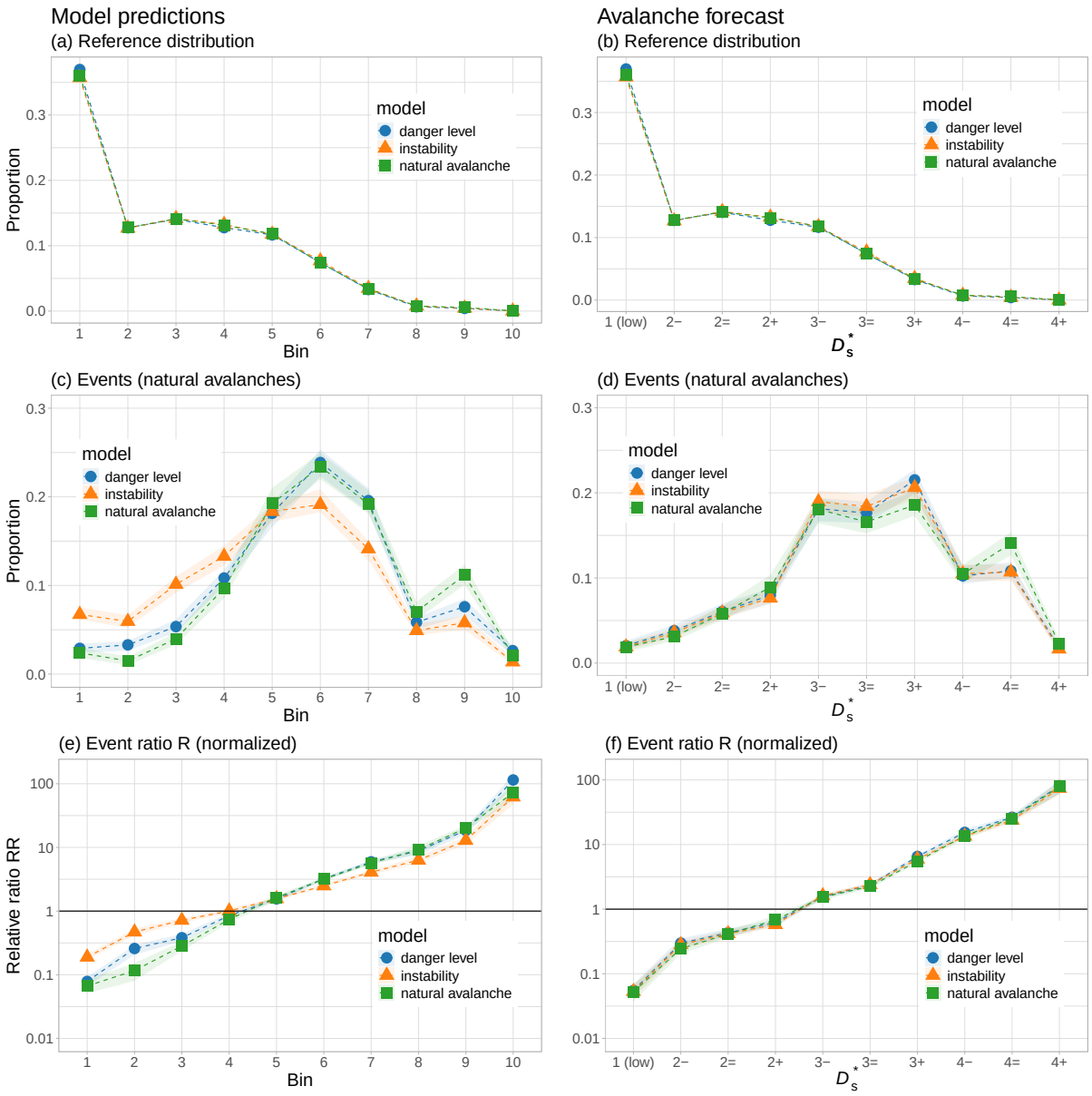

Figure B1. Natural avalanches. Left column: model predictions, right column: human avalanche forecast. Upper row: reference distributions, middle row: events (natural avalanches), bottom row: relative ratio RR (Eq. 3).

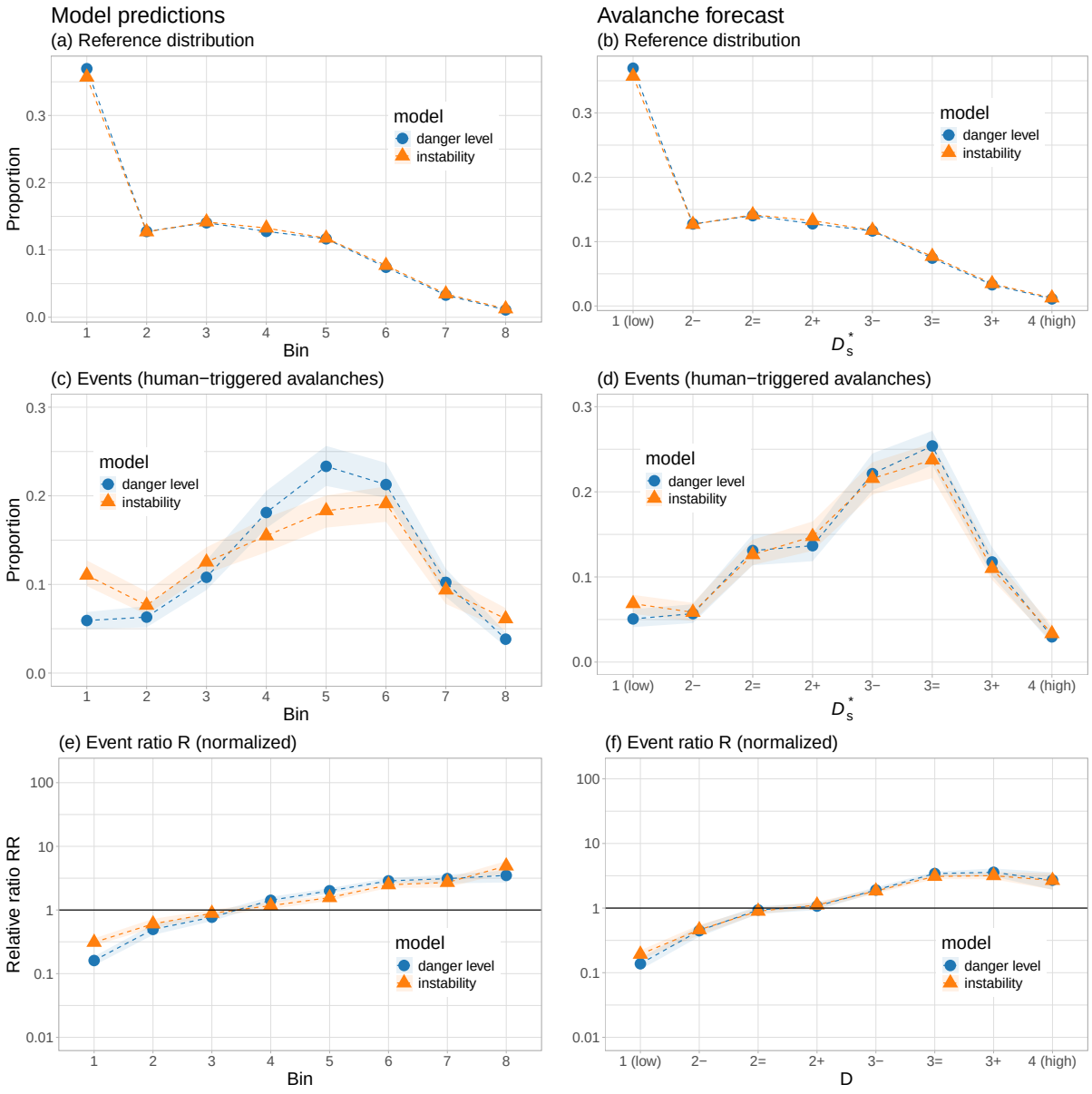

Figure B2. Human-triggered avalanches. Left column: model predictions, right column: human avalanche forecast. Upper row: reference distributions, middle row: events (human-triggered avalanches), bottom row: relative ratio RR (Eq. 3).

References

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

Binder, M., Perfler, M., Herla, F., Techel, F., and Mitterer, C.: Initializing and evaluating SNOWPACK simulations with recurring snow profile observations, in: International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, pp. 457 – 464, https://arc.lib.montana.edu/snow-science/item.php?id=3174, 2024.

Bouchayer, C.: Synthesis of distributed snowpack simulation relevant for avalanche hazard forecasting, mathesis, https://doi.org/10.13140/RG.2.2.21665.20329, 2017.

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

COSMO: COSMO - Consortium for small-scale modeling, https://www.cosmo-model.org/content/model/cosmo/overview.htm, last access: 6 May 2025, 2025.

Degraeuwe, B., Schmudlach, G., Winkler, K., and Köhler, J.: SLABS: An improved probabilistic method to assess the avalanche risk on back-country ski tours, Cold Regions Science and Technology, 221, 104 169, https://doi.org/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: avalanche size, https://www.avalanches.org/standards/avalanche-size/, last access: 13 May 2022, 2019.

EAWS: Standards: European Avalanche Danger Scale, Tech. rep., last access: 2024/07/07, 2023.

Eckerstorfer, M., Bühler, Y., Frauenfelder, R., and Malnes, E.: Remote sensing of snow avalanches: Recent advances, potential, and limitations, Cold Regions Science and Technology, 121, 126–140, https://doi.org/https://doi.org/10.1016/j.coldregions.2015.11.001, https://www.sciencedirect.com/science/article/pii/S0165232X15002591, 2016.

Fox, J., Siebenbrunner, A., Reitinger, S., Peer, D., and Rodríguez-Sánchez, A.: Automating avalanche detection in ground-based photographs with deep learning, Cold Regions Science and Technology, 223, 104 179, https://doi.org/10.1016/j.coldregions.2024.104179, 2024.

Fromm, R. and Schönberger, C.: Estimating the danger of snow avalanches with a machine learning approach using a comprehensive snow cover model, Machine Learning with Applications, 10, 100 405, https://doi.org/https://doi.org/10.1016/j.mlwa.2022.100405, https://www.sciencedirect.com/science/article/pii/S2666827022000809, 2022.

Gräler, B., Pebesma, E., and Heuvelink, G.: Spatio-Temporal Interpolation using gstat, The R Journal, 8, 204–218, https://journal.r-project.org/archive/2016/RJ-2016-014/index.html, 2016.

Hafner, E. D., Techel, F., Leinss, S., and Bühler, Y.: Mapping avalanches with satellites – evaluation of performance and completeness, The Cryosphere, 15, 983–1004, https://doi.org/10.5194/tc-15-983-2021, https://tc.copernicus.org/articles/15/983/2021/, 2021.

Hafner, E. D., Barton, P., Daudt, R. C., Wegner, J. D., Schindler, K., and Bühler, Y.: Automated avalanche mapping from SPOT 6/7 satellite imagery with deep learning: results, evaluation, potential and limitations, The Cryosphere, 16, 3517–3530, https://doi.org/10.5194/tc-16-3517-2022, 2022.

Hendrick, M., Techel, F., Volpi, M., Olevski, T., Pérez-Guillén, C., Herwijnen, A. v., and Schweizer, J.: Automated prediction of wet-snow avalanche activity in the Swiss Alps, Journal of Glaciology, 69, 1365–1378, https://doi.org/10.1017/jog.2023.24, 2023.

Hendrikx, J., Johnson, J., and Mannberg, A.: Tracking decision-making of backcountry users using GPS tracks and participant surveys, Applied Geography, 144, 102 729, https://doi.org/https://doi.org/10.1016/j.apgeog.2022.102729, https://www.sciencedirect.com/science/article/pii/S014362282200100X, 2022.

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

Herla, F., Horton, S., Mair, P., and Haegeli, P.: Snow profile alignment and similarity assessment for aggregating, clustering, and evaluating snowpack model output for avalanche forecasting, Geoscientific Model Development, 14, 239–258, https://doi.org/10.5194/gmd-14-239-2021, https://gmd.copernicus.org/articles/14/239/2021/, 2021.

Herla, F., Haegeli, P., and Mair, P.: A data exploration tool for averaging and accessing large data sets of snow stratigraphy profiles useful for avalanche forecasting, The Cryosphere, 16, 3149–3162, https://doi.org/10.5194/tc-16-3149-2022, 2022.

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, Nat. Hazards Earth Syst. Sci., 24, 2727—2756, https://doi.org/10.5194/nhess-24-2727-2024, 2024.

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, Natural Hazards and Earth System Sciences, 25, 625–646, https://doi.org/10.5194/nhess-25-625-2025, https://nhess.copernicus.org/articles/25/625/2025/, 2025.

Horton, S., Nowak, S., and Haegeli, P.: Enhancing the operational value of snowpack models with visualization design principles, Natural Hazards and Earth System Sciences, 20, 1557–1572, https://doi.org/10.5194/nhess-20-1557-2020, https://www.nat-hazards-earth-syst-sci.net/20/1557/2020/, 2020.

Horton, S., Haegeli, P., Klassen, K., Floyer, J., and Helgeson, G.: Adopting SNOWPACK models into an operational forecasting program: successes, challenges, and future outlook, in: Proceedings International Snow Science Workshop, Bend, Oregon, 2023, pp. 1544–1549, https://arc.lib.montana.edu/snow-science/item.php?id=3095, 2023.

Horton, S., Herla, F., and Haegeli, P.: Clustering simulated snow profiles to form avalanche forecast regions, Geoscientific Model Development, 18, 193–209, https://doi.org/10.5194/gmd-18-193-2025, https://gmd.copernicus.org/articles/18/193/2025/, 2025.

Hutter, V., Techel, F., and Purves, R. S.: How is avalanche danger described in textual descriptions in avalanche forecasts in Switzerland? Consistency between forecasters and avalanche danger, Natural Hazards and Earth System Sciences, 21, 3879–3897, https://doi.org/10.5194/nhess-2021-160, 2021.

Jamieson, B. and Jones, A.: The effect of under-reporting of non-fatal involvements in snow avalanches on vulnerability, in: Proceedings 12th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP12, Vancouver, Canada, 2015.

Lehning, M., Bartelt, P., Brown, R., Fierz, C., and Satyawali, P.: A physical SNOWPACK model for the Swiss avalanche warning; Part II. Snow microstructure, Cold Reg. Sci. Technol., 35, 147–167, 2002.

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.

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, Geoscientific Model Development, 17, 7569–7593, https://doi.org/10.5194/gmd-17-7569-2024, https://gmd.copernicus.org/articles/17/7569/2024/, 2024.

Mayer, S., van Herwijnen, A., Ulivieri, G., and Schweizer, J.: Evaluating the performance of an operational infrasound avalanche detection system at three locations in the Swiss Alps during two winter seasons, Cold Regions Science and Technology, 173, 102 962, https://doi.org/10.1016/j.coldregions.2019.102962, 2020.

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

MeteoSwiss: ICON forecasting system, https://www.meteoswiss.admin.ch/weather/warning-and-forecasting-systems/icon-forecasting-systems.html, last access: 6 May 2025, 2025.

Monti, F., Schweizer, J., and Fierz, C.: Hardness estimation and weak layer detection in simulated snow stratigraphy, Cold Regions Science and Technology, 103, 82 – 90, https://doi.org/10.1016/j.coldregions.2014.03.009, 2014.

Morin, S., Horton, S., Techel, F., Bavay, M., Coléou, C., Fierz, C., Gobiet, A., Hagenmuller, P., Lafaysse, M., Ližar, M., Mitterer, C., Monti, F., Müller, K., Olefs, M., Snook, J. S., van Herwijnen, A., and Vionnet, V.: Application of physical snowpack models in support of operational avalanche hazard forecasting: A status report on current implementations and prospects for the future, Cold Regions Science and Technology, 170, 102 910, https://doi.org/https://doi.org/10.1016/j.coldregions.2019.102910, https://www.sciencedirect.com/science/article/pii/S0165232X19302071, 2020.

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.

Murphy, A.: Probabilities, odds, and forecasts of rare events, 6, 302–307, 1991.

Pebesma, E.: Simple features for R: Standardized support for spatial vector data, The R Journal, 10, 439–446, https://doi.org/10.32614/RJ-2018-009, 2018.

Pebesma, E. and Bivand, R.: Spatial Data Science: With applications in R, Chapman and Hall/CRC, https://doi.org/10.1201/9780429459016, https://r-spatial.org/book/, last access: 2025-04-24, 2023.

Pebesma, E. J.: Multivariable geostatistics in S: the gstat package, Computers & Geosciences, 30, 683–691, https://doi.org/10.1016/j.cageo.2004.03.012, 2004.

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, https://doi.org/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, Natural Hazards and Earth System Sciences, 25, 1331–1351, https://doi.org/10.5194/nhess-25-1331-2025, https://nhess.copernicus.org/articles/25/1331/2025/, 2025.

Perfler, M., Binder, M., Reuter, B., Prinz, R., and Mitterer, C.: Assessing avalanche problems for operational avalanche forecasting based on different model chains, in: Proceedings, International Snow Science Workshop, Bend, Oregon, pp. 128–134, Bend, Oregon, https://arc.lib.montana.edu/snow-science/objects/ISSW2023_O4.05.pdf, 2023.

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, https://doi.org/10.1016/S0165-232X(03)00075-2, 2003.

Raissi, M., Perdikaris, P., and Karniadakis, G.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics, 378, 686–707, https://doi.org/https://doi.org/10.1016/j.jcp.2018.10.045, 2019.

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, https://doi.org/10.1016/j.coldregions.2021.103462, 2022.

Richter, B., Schweizer, J., Rotach, M. W., and van Herwijnen, A.: Validating modeled critical crack length for crack propagation in the snow cover model SNOWPACK, The Cryosphere, 13, 3353–3366, https://doi.org/10.5194/tc-13-3353-2019, 2019.

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.: A routing algorithm for backcountry ski tours, in: Proceedings International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, pp. 1489 – 1495, https://arc.lib.montana.edu/snow-science/item.php?id=3341, 2024.

Schweizer, J., Kronholm, K., and Wiesinger, T.: Verification of regional snowpack stability and avalanche danger, Cold Reg. Sci. Technol., 37, 277–288, https://doi.org/10.1016/S0165-232X(03)00070-3, 2003.

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.

SLF: SLF-Beobachterhandbuch (observational guidelines), https://www.dora.lib4ri.ch/wsl/islandora/object/wsl%3A24954/datastream/PDF/WSL-Institut_f%C3%BCr_Schnee-_und_Lawinenforschung_SLF-2020-SLF-Beobachterhandbuch-%28published_version%29.pdf, 55 p.; last access: 1 May 2022, 2020.

SLF: Avalanche bulletin interpretation guide, WSL Institute for Snow and Avalanche Research SLF, september 2023 edn., 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, https://doi.org/10.16904/envidat.406, 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, https://lean-gate.geo.uzh.ch/typo3conf/ext/qfq/Classes/Api/download.php/mastersThesis/1048, 2024.

Sykes, J., Hendrikx, J., Johnson, J., and Birkeland, K. W.: Combining GPS tracking and survey data to better understand travel behavior of out-of-bounds skiers, Applied Geography, 122, 102 261, https://doi.org/https://doi.org/10.1016/j.apgeog.2020.102261, https://www.sciencedirect.com/science/article/pii/S0143622819302115, 2020.

Techel, F.: On consistency and quality in public avalanche forecasting: a data-driven approach to forecast verification and to refining definitions of avalanche danger, Ph.D. thesis, Department of Geography, University of Zurich, Zurich Switzerland, https://doi.org/10.5167/uzh-199650, 2020.

Techel, F., Müller, K., and Schweizer, J.: On the importance of snowpack stability, the frequency distribution of snowpack stability and avalanche size in assessing the avalanche danger level, The Cryosphere, 14, 3503 – 3521, https://doi.org/10.5194/tc-2020-42, 2020a.

Techel, F., Pielmeier, C., and Winkler, K.: Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random?, Cold Regions Science and Technology, 180, 103 162, https://doi.org/10.1016/j.coldregions.2020.103162, 2020b.

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, Natural Hazards and Earth System Sciences, 22, 1911–1930, https://doi.org/10.5194/nhess-22-1911-2022, https://nhess.copernicus.org/articles/22/1911/2022/, 2022.

Techel, F., Helfenstein, A., Mayer, S., Pérez-Guillén, C., Purves, R., Ruesch, M., Schmudlach, G., Soland, K., and Winkler, K.: Human vs Machine – Comparing model predictions and human forecasts of avalanche danger and snow instability in the Swiss Alps, in: Proceedings

International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, pp. 31 – 38, https://arc.lib.montana.edu/snow-science/item.php?id=3108, 2024a.

Techel, F., Mayer, S., Purves, R. S., Schmudlach, G., and Winkler, K.: Forecasting avalanche danger: human-made forecasts vs. fully automated model-driven predictions, Natural Hazards and Earth System Sciences Discussions, 2024, 1–31, https://doi.org/10.5194/nhess-2024-158, 2024b.

Trachsel, J., Richter, B., Staehly, S., Mayer, S., Wahlen, S., van Herwijnen, A., and Techel, F.: Combining model predictions and radar avalanche detections for operational avalanche forecasting, in: Proceedings International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, pp. 1079 – 1084, https://arc.lib.montana.edu/snow-science/item.php?id=3271, 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.

Viallon-Galinier, L., Hagenmuller, P., and Eckert, N.: Combining modelled snowpack stability with machine learning to predict avalanche activity, The Cryosphere, 17, 2245–2260, https://doi.org/10.5194/tc-17-2245-2023, 2023.

Vionnet, V., Brun, E., Morin, S., Boone, A., Faroux, S., Le Moigne, P., Martin, E., and Willemet, J.-M.: The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2, Geoscientific Model Development, 5, 773–791, https://doi.org/10.5194/gmd-5-773-2012, http://www.geosci-model-dev.net/5/773/2012/, 2012.

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.

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 model predictions, in: Proceedings International Snow Science Workshop, Tromsø, Norway, 23-29 Sep 2024, pp. 124 – 131, https://arc.lib.montana.edu/snow-science/item.php?id=3123, 2024.

Young, M. V. and Grahame, N. S.: The history of UK weather forecasting: the changing role of the central guidance forecaster. Part 7: Operational forecasting in the twenty-first century: graphical guidance products, risk assessment and impact-based warnings, Weather, 79, 72–80, https://doi.org/10.1002/wea.4488, 2024.