A Large-scale Validation of Snowpack Simulations in Support of Avalanche Forecasting Focusing on Critical Layers

Florian Herla[1], Pascal Haegeli[1], Simon Horton[2], and Patrick Mair[3]

[1]Simon Fraser University, Burnaby, BC, Canada
[2]Avalanche Canada, Revelstoke, BC, Canada
[3]Harvard University, Cambridge, MA, USA

Correspondence: Florian Herla (fherla@sfu.ca)

Abstract. Avalanche warning services increasingly employ large-scale snow stratigraphy simulations to improve their insight into the current state of the snowpack. These simulations contain information about thin, persistent critical avalanche layers that are buried within the snowpack and are fundamental drivers of avalanche hazard. However, the data volume, data complexity, and unknown validity have so far limited the value of the simulations for operational decisions. We attribute this at least partially
5 to a lack of research that validates the simulations for their capability to represent the existence and instability of known critical layers at the regional scale. To address this knowledge gap, we present methods that enable meaningful comparisons between regional assessments of avalanche forecasters and snowpack simulations that are distributed across entire forecast regions. We applied these methods to operational data sets of ten winter seasons and three public forecast regions in western Canada and thereby quantified the performance of the Canadian weather and snowpack model chain to represent persistent critical
10 avalanche layers. We found that the overall probability of detecting a known critical layer in the simulations can be as high as 75 % when accepting a low probability of 40 % that any simulated layer is actually of operational concern in reality. Furthermore, we explored patterns that characterize which layers were represented well and which were not. Faceted layers, for example, were captured well but also caused most false alarms, whereas surface hoar layers tended to be less prevalent but in return were mostly of operational concern when modeled. Overall, our results suggest that the simulations provide a valuable
15 starting point for targeted field observations as well as a rich complementary information source that can help alert forecasters about the existence of specific critical layers or provide an independent perspective on their instability. However, we do not believe that the existing model chain is sufficiently reliable to generate assessments purely based on simulations. We conclude by presenting our vision of a real-time operational validation suite that can help forecasters develop a better understanding of the simulations' strengths and weaknesses by continuously comparing assessments and simulations in a user-friendly manner.

1 Introduction

Understanding snow avalanche hazard conditions requires information about the layered snow stratigraphy (LaChapelle, 1980; McClung, 2002b). Traditionally this information is gathered with manual snow pit observations at targeted point locations (McClung, 2002a; Campbell et al., 2016). Avalanche forecasters then combine this information with their prior mental model of the snow conditions and information about the weather as well as avalanche occurrences to evaluate the conditions at the

25   scale of the forecast application (e.g., regional public forecast) (LaChapelle, 1980). Of particular interest are the distribution and sensitivity of critical avalanche layers (Statham et al., 2018a) that were buried over the course of the winter season and persist across the region, since these layers can cause multiple recurrent avalanche cycles. We use the term critical layer to refer to persistent weak layers and crust layers throughout this manuscript. While assessing the current snowpack conditions (i.e., nowcasting) for entire regions based on limited data availability is already an exercise that demands considerable expertise,

30   forecasting for snowpack changes several days ahead requires considerable subjective judgment that can be associated with substantial uncertainty (McClung, 2002a; Statham et al., 2018b).

    Weather and snowpack model chains represent an additional, independent data source for detailed snow stratigraphy information with high spatiotemporal availability (Morin et al., 2020). Snowpack simulations can therefore help fill data sparsity gaps occurring during the early winter season, in remote regions (Storm, 2012; Storm and Helgeson, 2014), or they can com-

35   plement assumptions on the evolution of the sensitivity to triggering avalanches on known critical layers (Reuter et al., 2021; Mayer et al., 2022). However, the simulations inherently accumulate errors from the weather inputs—precipitation being the main source of uncertainty for snowpack structure and instability (Raleigh et al., 2015; Richter et al., 2020)—and from deficiencies in the snow model formulations (Lafaysse et al., 2017). Thus, avalanche forecasters are concerned about the validity of the simulations and only hesitantly integrate these novel information sources in their operational assessment processes (Morin

40   et al., 2020).

    Over the last 20 years, extensive research has been conducted to improve the capabilities of snowpack models and explore their application for avalanche forecasting. Many studies that validated simulated snow stratigraphy applied a process-based approach (e.g., Brun et al., 1992; Fierz, 1998; Durand et al., 1999; Lehning et al., 2001, 2002b; Fierz and Lehning, 2004; Bellaire and Jamieson, 2013; Richter et al., 2019; Calonne et al., 2020; Viallon-Galinier et al., 2020). These process-based

45   studies tested the capabilities of the models to represent specific physical processes, mostly using high-quality validation data sets at the point scale. On the regional scale, much effort has gone into validating bulk snow properties such as snow depth or snow–water equivalent (Vionnet et al., 2012; Lafaysse et al., 2013; Quéno et al., 2016; Vionnet et al., 2019; Morin et al., 2020; Horton and Haegeli, 2022). Several studies evaluated the model capabilities to simulate snowpack instability based on regional observations to support its application in avalanche forecasting (Schweizer et al., 2006; Schirmer et al., 2010; Vernay et al.,

50   2015; Bellaire et al., 2017; Mayer et al., 2022).

    Despite this large body of snowpack validation studies, the operational needs of (Canadian) avalanche forecasters have not been satisfied yet. While process-based validations at the point scale based on high quality data are crucial for model development and improvement, they do not provide tangible and relevant guidance for forecasters and are not necessarily representative of the real skill for operational simulations. Regional-scale validation studies of simulated snowfall and simulated

55   snow instability further contribute essential information to the valuation of snowpack simulations for avalanche forecasting, but the existing research does not paint a comprehensive picture yet: to our knowledge no large-scale study exists that created a specific link between simulated layers and known critical avalanche layers. Forecasters therefore only have a limited understanding of how to interpret individual critical layers in the simulations. To address these validity concerns and increase the operational adoption of snowpack simulations in avalanche forecasting, forecasters need an application-specific validation

approach at the regional scale that evaluates the detailed hazardous layering of the snowpack, most optimally in an operational, real-time format. The only study that followed a similar approach is Horton and Jamieson (2016), who uncovered substantial uncertainties in the prediction of hazardous surface hoar formation and its post-burial evolution during two winter seasons. All of these observations highlight the importance of an updated validation study and future real-time validation tools.

The objective of this study is to evaluate the performance of operational snowpack simulations to represent critical avalanche layers in support of operational avalanche forecasting at the regional scale. Based on forecaster assessments in ten winter seasons and three mountain ranges, we characterize the overall model performance with the likelihood that hazardous layers are captured by the simulations (i.e., probability of detection) and the likelihood that modeled critical layers are actually of concern in reality (i.e., model precision). To make our results most informative for forecasters, we explore patterns in the presence or absence of critical layers that can be explained with attributes of the layers themselves, and we present the results in tangible ways that help forecasters interpret daily model scenarios. We also quantify the degree of agreement between simulated and reported layer instability using indicators of the variation and timing of instability. By relying on operationally available data from human avalanche hazard assessments, the methods developed in this study can be used to design future operational validation suites. Overall, our study quantifies the capabilities of an operational weather and snowpack model chain to represent critical avalanche layers, which contributes to making snowpack simulations more transparent and applicable for operational applications.

We continue this paper by introducing the used data sets in Sect. 2 before explaining our methodology in detail in Sect. 3. Section 4 presents the results from their highest level to the most detail. We direct our discussion in Sect. 5 first to avalanche practitioners and share the insights gained through this study for operational application as well as share our vision for the future development of operational tools before discussing insights relevant for the snowpack modeling community.

2 Data

The data sets used in this study consist of snowpack simulations and operational avalanche hazard assessments from avalanche forecasters in western Canada over ten winter seasons (2013–2022).

2.1 Study area

The study focuses on three data-rich public avalanche forecast regions in British Columbia and Alberta, Canada, each one characterized by a different snow climate: Whistler in the Sea-to-Sky region (S2S) in the maritime Coast Mountains, Glacier National Park (GNP) in the transitional Columbia Mountains, and Banff-Yoho-Kooteney National Park (BYK) in the continental Rocky Mountains (Shandro and Haegeli, 2018).

While the human assessment data set is inherently compiled for these distinct forecast regions, we had to select model grid points to represent these regions in the simulations. We used all grid points within the region boundaries that were within 10 km of accessible roads for GNP and BYK, and within 20 km of Highway 99 for S2S. Since Canadian National Parks
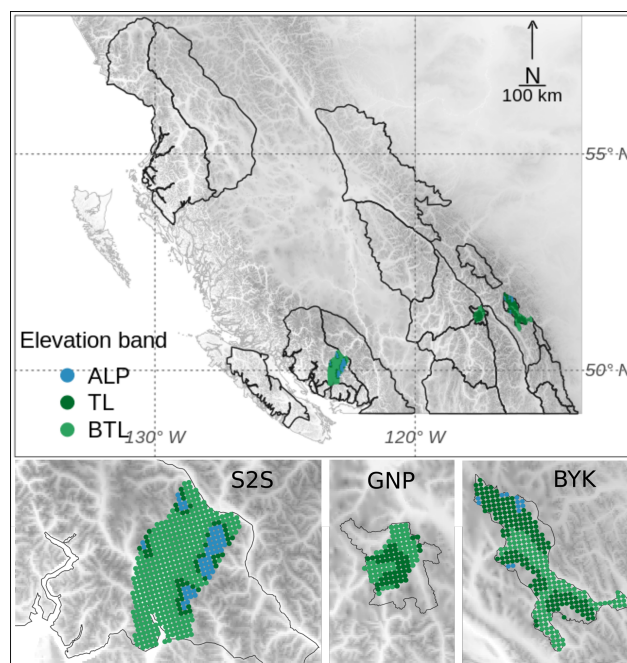
Figure 1. The public avalanche forecast regions of western Canada. The model grid points selected for this study are visualized by their elevation band (alpine, treeline, below treeline) and lie within the forecast regions Sea-to-Sky (S2S), Glacier National Park (GNP), and Banff-Yoho-Kootenay National Park (BYK).

prohibit motorized terrain access, such as via sleds, we assumed the hazard assessments to be highly biased towards these parts of the regions. Overall, we selected 1004 grid points (Fig. 1).

We classified each grid point into an elevation band class, 'alpine' (ALP), 'treeline' (TL), and 'below treeline' (BTL) to match the terrain classification in the human assessments. We used forecaster consensus of TL elevation for the classification:
95   1600–1800 m in S2S, 1800–2100 m in GNP, and 2000–2400 m in BYK.

2.2 Snowpack simulations

The snowpack simulations used in this study mirror the setup of the operational simulations that are provided to Canadian avalanche forecasters. The raw output of the numerical weather prediction model HRDPS (Milbrandt et al., 2016) is fed into the detailed snow cover model SNOWPACK (Bartelt et al., 2002; Lehning et al., 2002b, a) (v3.4). HRDPS is run operationally
100 by the Meteorological Service of Canada on a 2.5 km horizontal grid. SNOWPACK is run with weather data lead times of 6–12 h, using air temperature and relative humidity at 2 m above ground, wind speed at 10 m above ground, incoming shortwave and longwave radiation fluxes at the surface, and accumulated precipitation. We used the SNOWPACK stability scheme by Michlmayr et al. (2008), and turned off any snow redistribution by wind. To output the detailed snow stratigraphy at each grid

point, SNOWPACK's feature of merging similar layers was also turned off. All simulated profiles were valid between 4–5 PM
105 local time representing flat field conditions.

To visualize the grain types of snow layers, this manuscript uses the hazard-focused color coding suggested by Horton et al.
(2020a). We abbreviate grain types with the following acronyms: precipitation particles (PP), decomposing fragments (DF),
surface hoar (SH), depth hoar (DH), facets (FC), rounding facets (FCxr), round grains (RG), rain crust (IFrc), sun crust (IFsc),
or temperature/melt–freeze crust (MFcr), and melt forms (MF). All grains types are defined in the International Classification
110 for Seasonal Snow on the Ground (Fierz et al., 2009).

2.3 Avalanche hazard assessments

Avalanche hazard assessments compiled for this study were issued by public avalanche forecasters every day of the winter
season. The assessments represent forecasters' best knowledge of the current conditions (i.e., nowcasts) and were issued in the
afternoon for one elevation band (ALP, TL, BTL) in one forecast region. Following the Conceptual Model of Avalanche Hazard
115 (Statham et al., 2018a), forecasters partitioned the avalanche hazard into different avalanche problems and characterized each
problem by their type, location, likelihood, and size.

In addition to avalanche problems, the hazard assessments also contain records about critical avalanche layers. Forecasters
typically track the evolution of these critical layers and associate them with the relevant avalanche problems at their times of
concern. To facilitate the tracking across space and over time, avalanche forecasters name critical layers with date tags and
120 their grain type(s), e.g. "Jan 17th surface hoar layer". Reported date tags mostly represent the beginning of snowfall periods
that bury layers that were exposed to the snow surface before the snowfall and are therefore likely to contain weak grain types.
Sometimes the date tags can also represent rain events that form a crust at the snow surface. While the date tags provide general
markers to simplify communication, the actual formation or burial dates can vary within a forecast region due to differences
in local weather and its timing. When a tracked layer is associated with an avalanche problem at a given day, the forecaster
125 expects the problem to be governed by that layer.

The grain types of the tracked layers were either reported as one specific grain type or a mix of grain types, such as SH, FC,
DH, IFrc, IFsc, or more generically MFcr, where each layer is usually primarily a persistent weak layer (SH, DH, FC) or a
crust layer (IF, MFcr).

3 Methodology

130 Our validation approach consists of three main components. First, we assessed the general performance of the simulations with
regard to modeling critical avalanche layers (Sect. 3.1) by examining two practical questions: a) What is the likelihood that a
layer of operational concern is captured by the simulations? and b) How likely is it that a modeled critical layer also exists in
reality? Second, we explored whether there are distinguishable patterns between critical layers of concern that were captured
well by the model and those that were not (Sect. 3.2). Finally, we examined the agreement between simulated and reported
135 layers in more detail taking into account the timing, magnitude, and variation of layer instability (Sect. 3.3).

3.1 Assessing the general simulation performance

We used a confusion matrix approach to quantify the performance of the simulations to represent critical avalanche layers. A confusion matrix, or a 2x2 contingency table, presents the joint frequency distributions of binary forecasts and observations (Wilks, 2019, p. 374f) as the foundation for calculating various performance attributes. In our case, the binary events were

140 the presence or absence of critical layers at some point during the season in the human hazard assessments and snowpack simulations. To identify present and absent critical layers in both data sets and properly fill the four cells of the confusion matrix, we had to perform several pre-processing steps. This involved a) identifying the relevant layers in both data sets, b) matching the corresponding layers, and c) translating the spatially distributed simulation information into a binary variable representing presence or absence at the regional scale. We describe each of these steps in detail in the following sections, and

145 Figure 2 provides a concise summary of the process.

3.1.1 Pre-processing human data set

We identified critical avalanche layers in our human hazard assessment data set by filtering all layers that were tracked by avalanche forecasters to only those that were associated with either storm slab, persistent slab or deep persistent slab avalanche problems at least once during a season. Since these layers are a key ingredient to these avalanche problems, we refer to them

150 as *layers of concern* throughout the manuscript. In addition, we will refer to persistent and deep persistent avalanche problems simply as persistent problems.

To account for the various degrees of concern and level of confidence in the human assessments, we manually reviewed all forecaster comments associated with the identified layers of concern and summarized them in a qualitative data quality rating. The forecaster comments often include information about observed avalanches, associated triggers, other instability

155 observations, the absence of observed instability, or mere assumptions. Combining the qualitative comments with the number of avalanche problem days and the assessed likelihood of triggering avalanches on these layers, we assigned each layer of concern to one of four data quality classes. Class 1 (Excellent) contains all layers with consistent reporting, evidence that the layer caused avalanches, and more than 10 persistent avalanche problem days. Class 2 (Good) represents layers that were still of operational concern, but either the reporting was slightly inconsistent, or the layer was associated with a persistent problem

160 for less than 10 days, or the comments showed that the layer was rather unreactive to triggering. Class 3 (Uncertain) contains layers whose assessments hold substantial uncertainty. While the assessments suggest that these layers existed, they were only linked to very few storm or persistent avalanche problems, showed very inconsistent reporting, or were of only limited operational concern due to missing avalanche reactivity. Class 4 (Unsure) consists of layers that were barely reported upon, making it questionable whether they existed at all.

165 3.1.2 Pre-processing simulations

To identify critical avalanche layers in the snowpack simulations, we searched all simulated snow profiles for layers that were characterized by persistent grain types and poor stability. Persistent grain types included all faceted layers (FC, DH), surface
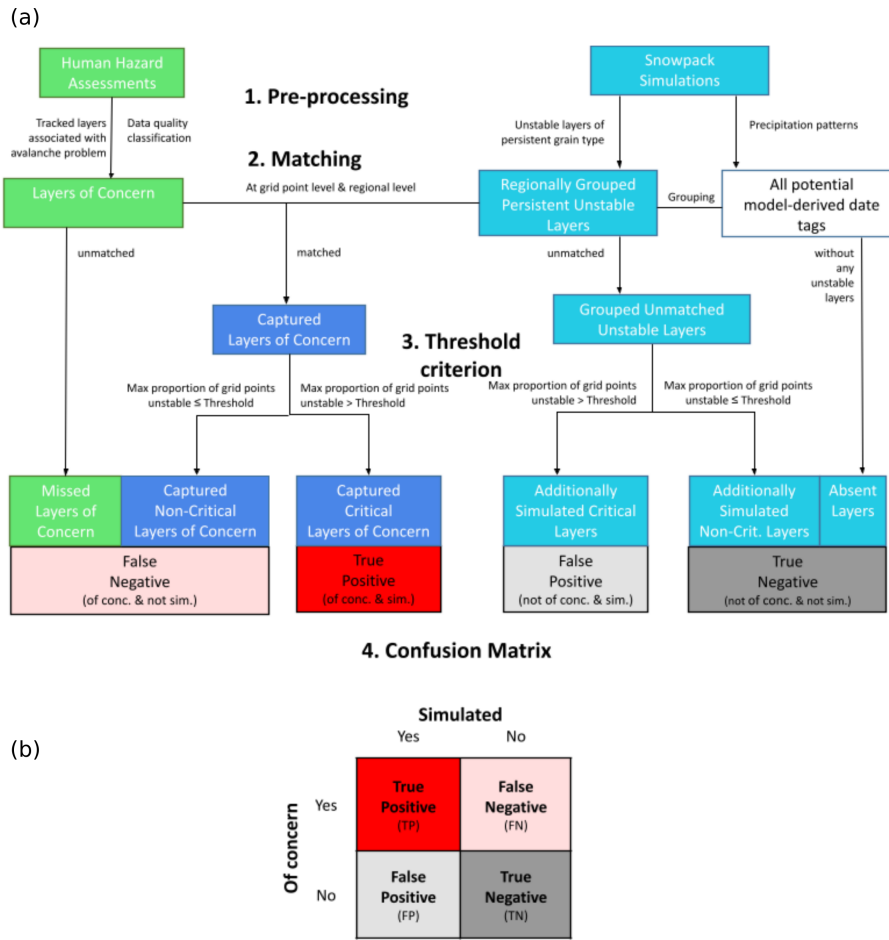
Figure 2. (a) A flowchart for assembling the four cells of the confusion matrix from the human and modeled data sets. (b) The resulting confusion matrix.

hoar layers (SH), and crust layers (MFcr, IF). We employed two different approaches to assess dry snow instability: a recently developed statistical approach and a process-based approach based on our current understanding of avalanche release. Since

170   the existing literature does not yet provide any guidance on which of these approaches is superior or more appropriate for our application, we evaluated the entire confusion matrix analysis described in Sect. 3.1 twice, once employing the statistical approach and once the process-based approach. To improve readability of our manuscript, layers or grid points with poor stability are also referred to as unstable layers or grid points.

     The statistical approach used the random forest classifier developed by Mayer et al. (2022). This model was trained with a

175   high-quality data set of observed snow profiles recorded around Davos, Switzerland. Based on the observed instability of the weakest layer in each profile, the model learned to predict the probability of layer instability ($p_{unstable}$) from a set of simulated

predictor variables. As suggested by Mayer et al. (2022), we considered layers with $p_{unstable} \geq 77\ \%$ as critical avalanche layers with poor stability.

Inspired by the work of Monti et al. (2014) and Reuter et al. (2021), our process-based approach used a combination of
180 three indices to assess the instability of a layer: a) the relative threshold sum approach RTA (Monti and Schweizer, 2013; Monti et al., 2014), b) the multi-layered skier stability index SK38 (Monti et al., 2016), and c) the critical crack length r_c (Richter et al., 2019). While potential weak layers were pre-selected based on RTA, their propensity for failure initiation and crack propagation was assessed based on SK38 and r_c, respectively, which accounts for the two main processes governing slab avalanche release (Viallon-Galinier et al., 2022). While the literature agrees on thresholds for RTA and SK38, less consensus
185 exists for r_c. Hence, we applied two thresholds to r_c and therefore identified critical avalanche layers with poor stability if RTA ≥ 0.8, SK38 ≤ 1, and $r_c \leq 0.3$ or $r_c \leq 0.4$.

While the relevant layers of human concern represent regional assessments with a single date tag (Sect. 2.3), the layers identified in the simulations consist of observations from different grid points and at different times that have not been connected yet. To make the simulated layers comparable to the regional layers of concern, we grouped the simulated layers using
190 model-derived date tags. These model-derived date tags represent regional markers for times when layers with the potential to become critical avalanche layers got buried. Analogously to human date tags, these regional markers are primarily represented by the onset dates of pronounced storms with substantial snowfall.

To establish the model-derived date tags, we first highlighted dry periods by identifying all days with less than a trace amount of snowfall or rain for each season and region. The actual model-derived date tags are defined by the start of substantial storm
195 periods following dry periods. However, since potentially critical layers can also be buried by small amounts of snow over multiple days, we inserted additional date tags in between these main storm periods if the non-storm periods were sufficiently long and characterized by substantial accumulations of small daily snowfall amounts. The exact rules and thresholds for establishing the model-derived date tags are described in detail in the Appendix A.

Once the model-derived date tags were established, we assigned each simulated unstable layer to the closest date tag older
200 than the layer's formation date. This means that all the layers associated with the same date tag can be found within a narrow band of the snow stratigraphy that got buried close to the associated date tag. This set of rules allowed us to meaningfully group simulated layers across each region and season based on precipitation patterns and their formation and burial times similarly to how forecasters label layers of concern.

In the final step, we determined the grain type class of each group of simulated layers. To label the layer groups with either
205 a single grain type (FC, SH, or MFcr) or by a mix of two classes (e.g., SH/FC), we first identified the two most prevalent grain types within the group of layers associated with a date tag at each grid point for each day. This allowed us to identify the two most prevalent grain types across all grid points for each day, and eventually for the entire lifetime of the layer. To address the well-known SNOWPACK behaviour of transforming most SH layers into DH layers after several days of burial, we judged DH layers as FC layers if only a negligible amount (less than 10%) of SH was encountered during the layer's lifetime. If the
210 fraction of SH layers was higher than 10 %, we judged all DH layers as SH layers. Finally the layer groups were labeled with the resulting one or two prevalent grain types.

3.1.3 Matching of layers of concern and simulated layers: How and when?

The matching of layers between the human data set and the simulations is arguably the most important step in the derivation of the confusion matrix, because it influences all four cells of the matrix. To ensure that all layers were classified meaningfully,

215 our matching approach consisted of two steps. We first identified layers of human concern at individual grid points before checking at the regional level to ensure we did not miss any relevant layers based on the grouping of layers by model-derived date tags. The following paragraphs explain these two steps in detail.

To identify layers of human concern in the simulated profiles at individual grid points, we constructed search windows around the human date tags of individual layers of concern. SNOWPACK creates labels that store the date and time when a particular

220 layer is formed or deposited. Following Richter et al. (2019), we also computed burial date labels based on the deposition date label of the overlying layer. To account for biases in the reported date tags and errors in the simulated precipitation patterns, a simulated layer was considered matched if it formed within the formation window or got buried within the burial window of a human layer of concern. The formation window ranged from the last day of the prior storm event to the day of the human date tag, and the burial window ranged from one day prior of the human date tag to the first day of the actually simulated storm

225 event (Fig. 3). We used thresholds for cumulative amounts of new snow (10 cm) and liquid precipitation (5 mm) to identify the end of the preceding storm event for the start of the formation window (accumulating backward) and to identify the start of the next snow fall event for the end of the burial window (accumulating forward). To avoid unreasonable matches, we limited the length of the formation window to 30 days and the length of the burial window to four days.

While the search windows identified all simulated layers that align with a specific human date tag, the layers still needed to

230 be filtered for the reported grain types to produce a meaningful match. To accomplish this, we employed up to three different grain type searches for each layer of concern: a) a strict grain type search that only accepted the specific *reported* weak grain types (e.g., SH, SH/FC, etc.) in the simulated layers; b) a relaxed grain type search that accepted *all* persistent weak grain types (SH/DH/FC) in the simulated layers; and c) a grain type search for any crust grain type (IF, MFcr) if crusts were reported to play a role in the given layer of concern.

235 The final consideration for the matching of the layers was the timing of the matching. Since persistent critical layers persist for a long time, there are several possible options for when to match the layers from the two data sets (e.g., at their time of burial, first concern, most concern). To account for temporal variability in the presence and instability of these persistent layers, the matching routine was run daily from 5 days before the reported date tag up to 90 days after. All layers of concern that were matched to simulated layers at any time within this 96-day validation window were considered *captured*.

240 As described above, the first step of the layer matching routine requires a threshold for the accumulated precipitation amounts to identify the formation and burial windows. We purposely chose a low threshold to ensure that only layers within a narrow band of the snow stratigrahy were matched. However, our initial explorations revealed that approximately 5–10 % of unstable layers were erroneously assigned to model-derived date tags that were only a few days apart from the relevant human date tag. This situation can occur at grid points that either experienced a slightly lagged arrival of the storm producing later model-

245 derived date tags or above-average precipitation amounts which results in earlier model-derived date tags. Since these layers
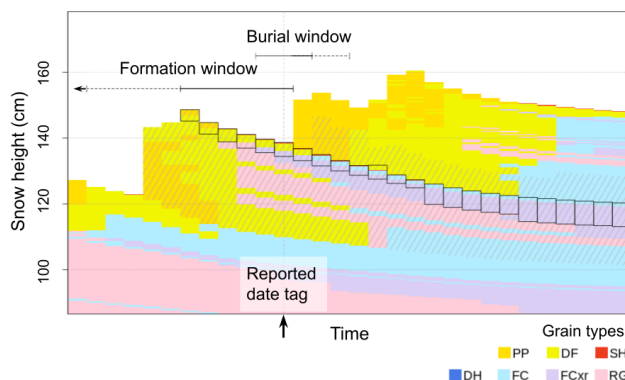
Figure 3. An illustration of the search windows for layers of human concern around the human date tag. The panel zooms in on the near-surface layers of a time series of a simulated snow profile. The extent of the formation and burial windows are highlighted by horizontal lines, where the solid lines indicate the extent of each window based on the timing of the storms. The black boxes highlight the layers that fall within at least one of the trimmed windows, while the gray hatched lines highlight the layers resulting from the fixed windows. Colors refer to snow grain types defined in Sect. 2.2.

are actually part of the same group of layers that were assigned to the correct human date tag, we changed the assignment from the model-derived date tag to the human date tag for these layers if a) the date tags were only one day apart regardless of snowfall amounts or b) the date tags were up to three days apart and no more than 10 cm of new snow was simulated in between.

3.1.4   Applying a threshold criterion to fill the confusion matrix

After the pre-processing and matching of the data sets, we can compute the counts for each cell of the confusion matrix (Fig. 2a). All layers of concern that remained unmatched by the simulations were counted as missed layers of concern and directly contributed to the False Negative cell of the matrix. Analogously, all potential model-derived date tags that were not assigned a simulated layer during the grouping were counted as absent layers and directly contributed to the True Negative cell of the matrix. While these missed layers of concern and absent layers conceptually only exist at the regional scale and do not have a direct link to the simulations at individual grid points, the captured layers of concern and the remaining grouped unmatched unstable layers are linked to the simulations at grid points, and we can calculate the daily proportion of unstable grid points, which we call the *proportion unstable*. To translate this continuous variable into a binary variable and decide whether a layer should be counted as simulated (or not), we applied a threshold criterion. If the maximum daily proportion unstable within the 96-day validation window was greater than the threshold, the layer was counted as simulated, otherwise as not simulated. Hence, the captured layers of concern either added to the False Negative cell or informed the True Positive cell. Likewise, the grouped unmatched unstable layers either added to the True Negative cell or informed the False Positive cell.

To summarize our methodology, we illustrate the concepts applied so far with the 2019 winter season in GNP at TL (Fig. 4). The time series of the average profile (Herla et al., 2022) for the region shown in the bottom left panel provides context for the
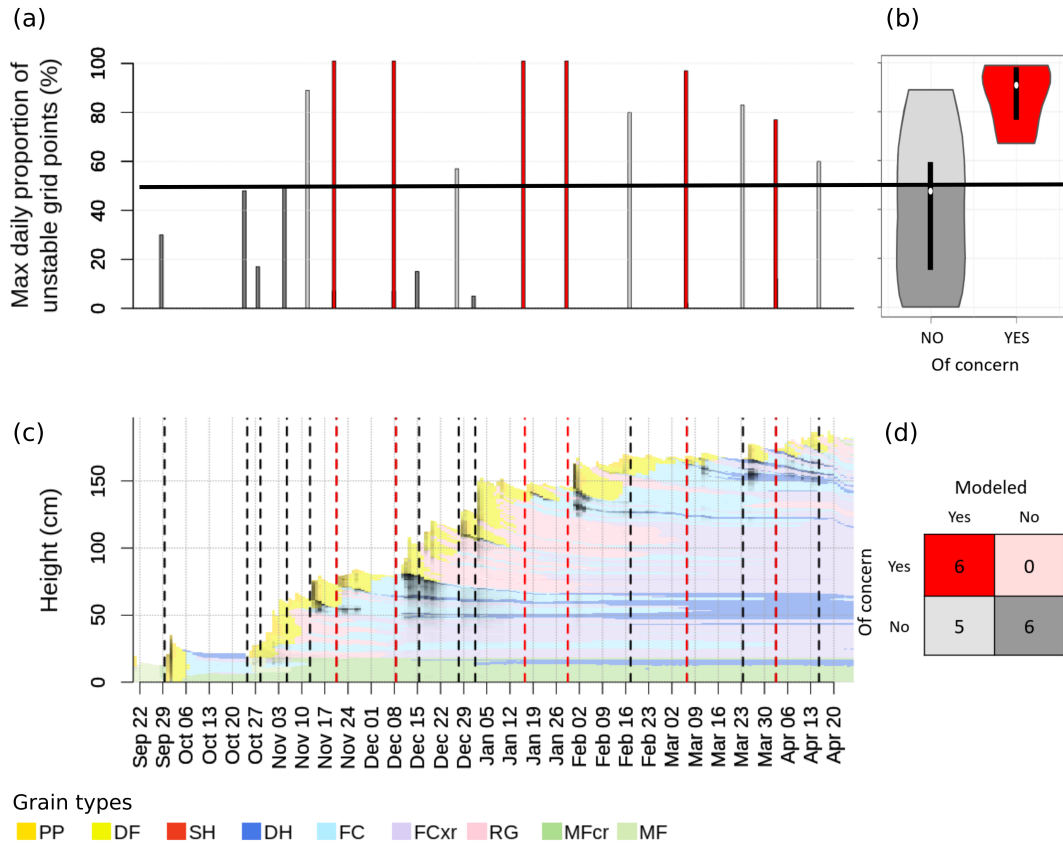
Figure 4. An illustration of the concepts applied to a case example of the 2019 season in Glacier National Park (GNP) at treeline elevation (TL). (a) Maximum daily proportion unstable for each critical layer (red: of human concern, gray: not of human concern) and (b) the resulting distributions. (c) The average profile for the region and season (only shown to improve context) with the human and model-derived date tags (red and black vertical lines, respectively). (d) The confusion matrix evaluated for an exemplary threshold criterion of 50 %. Acronyms of snow grain types are defined in Sect. 2.2.

265   human-reported date tags (red vertical lines) and the additional model-derived date tags (black vertical lines). The bar chart in in the top left panel represents the maximum daily proportion unstable for each critical layer represented by its date tag. Using 50 % as our threshold criterion for this example, each bar is colored according to its corresponding cell in the confusion matrix in the bottom right panel. The violin plots shown in the top right panel present the distributions of the maximum daily proportion unstable in more detail. In this particular case, all six layers of concern were well captured, and all the other

270   simulated layers generally had lower proportions of unstable grid points. However, there are four layers that were considered critical by the model using the 50 % threshold but not the human forecasters.

3.1.5 Statistical performance measures

According to (Wilks, 2019, p. 374f), three complementary attributes are required to comprehensively capture the model performance. We present the probability of critical layer detection, the model precision, and the false alarm rate. Probability
275 of detection, also known as recall or model sensitivity, is the ratio of True Positives / (True Positives + False Negatives), which describes the probability that a critical layer of concern is captured by the simulations. This addresses the first practical forecaster question we outlined in our research objectives. Model precision, also called positive predictive value, is the ratio of True Positives / (True Positives + False Positives). It represents the probability that a modeled critical layer is indeed of operational concern, which corresponds to the second practical forecaster question. Finally, the false alarm rate, also referred
280 to as 1 - model specificity, is the ratio of False Positives / (False Positives + True Negatives). It expresses the probability that a layer that the human forecasters are not concerned about is identified as a critical layer by the model.

To present our results, we computed the three performance attributes for all possible threshold criteria between 0–100 %. The attributes were then visualized in a precision–recall curve and a relative operating characteristics (ROC, historically also known as receiver operating characteristics) curve using the software package 'ROCR' by Sing et al. (2005) written for the R Language
285 and Environment for Statistical Computing (R Core Team, 2020). Each point on these curves describes the model performance for one specific value of the threshold criterion, which can then be compared against a no-skill baseline. The no-skill baseline represents the performance of a random model in the ROC curve, and the climatological frequency of critical layer formation in the precision–recall curve. To account for uncertainty in the assessment of model performance, we computed 95 % confidence bands around each curve by bootstrapping the confusion matrix results at each threshold following Davison and Hinkley
290 (1997) and using the R package 'boot' (Canty and Ripley, 2022). To provide insight on the effect of grain type, elevation band and forecast region on performance, we stratified the overall model performance by these variables and visualized them in additional precision–recall and ROC curves.

In addition to the three performance attributes, we computed two skill scores, the Peirce skill score (PSS) and the F_1 score, that summarize the different aspects of model performance in a single measure. Both scores express model performance relative
295 to a perfect model using different ways of combining the cells of the confusion matrix. While these scores are less tangible to interpret, they simplify comparing model performance at different threshold criteria and for different approaches of modeling instability. The PSS, also known as the true skill statistic, accounts for all cells of the confusion matrix to express the actual skill of the simulations (Eq. (1)), whereas the F_1 score disregards the True Negative cell and is therefore biased towards the prediction of events (Eq. (2)). Since the prediction rate of rare events can be high simply by chance, PSS is the preferred
300 score for our application (Ebert and Milne, 2022). Despite this limitation of F_1, it provides a complementary view on model performance that is independent of our approach to compute the True Negative cell based on model-derived date tags. For more information on skill scores in rare event forecasting please refer to Ebert and Milne (2022) who provide a comprehensive overview and discussion.

The PSS and F$_1$ scores are defined as

305 $$PSS = \frac{(TP \cdot TN) - (FP \cdot FN)}{(TP + FN) \cdot (FP + TN)} \tag{1}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}, \tag{2}$$

where TP and TF are true positive and true negative counts, and FP and FN are false positive and false negative counts, respectively (Fig. 2).

3.2 Finding more detailed patterns in all missed or captured layers of concern

310 The next components of our critical layer validation analyze all missed or captured layers of human concern in more detail. This section describes how we searched for patterns in these two groups of layers to better understand the subtleties of model performance.

While the confusion matrix analysis described in the last section only included layers with poor stability, the analysis described here also includes layers of concern that where captured structurally but remained stable. Analogous to our approach

315 of using *proportion unstable* to characterize relevant layers in the last section, we use *proportion captured* to characterize the mere structural presence of each missed or captured layer independently of its simulated stability. The computation and analysis of the proportion captured follows the same steps as the analysis of the proportion unstable as illustrated in Fig. 2 except that it skips the first pre-processing step of filtering for unstable layers. Identical to the confusion matrix analysis, we computed the maximum daily values for the proportion captured and proportion unstable for the 96-day validation window. By including

320 both the proportion captured and the proportion unstable, we acknowledge that the simulations first need to structurally capture the layers of concern before they can provide insight about instability.

To provide insight on what layer attributes and contextual factors influenced the maximum proportion captured and unstable, we used conditional inference trees (CTree) (Hothorn et al., 2006), a type of classification tree that uses a statistical criterion for finding splits. CTrees recursively partition the distribution of a response variable based on the statistically most significant

325 splits along a set of explanatory variables. For the present analysis, we used the ctree function in the R package 'partykit' (Hothorn and Zeileis, 2015). Due to their ease of use and easily interpretable results, CTrees have already been used several times in snow and avalanche research for exploratory analyses (e.g. Horton et al., 2020b). While the top node of a CTree represents the most significant split that divides the entire sample, the resulting subsamples are recursively split into smaller subsamples until the algorithm cannot find any significant splits in the the response variable anymore. The resulting terminal

330 nodes describe subsets of the data set with distinct distributions of the response variable that can be can be linked to specific combinations and thresholds of the explanatory variables.

We included the following potential explanatory variables in our analyses, which contained both direct layer attributes as well as more contextual information:

– grain type (and grain type search routine)

335 – grain size of the simulated layer at burial

- data quality class

- month of burial

- number of associated avalanche problem days

- simulated length of the dry spell before burial

340   - season

- region and elevation band.

3.3   Estimating agreement indicators between modeled and reported instability of captured layers of concern

The analysis described so far validated the existence and instability of critical layers purely based on the assumption that layers tracked by forecasters were of operational concern at some point during the seasons. However, the daily assessments of
345   the avalanche problems according to the conceptual model of avalanche hazard provide more detailed information about the associated layers. In particular, the reported likelihood of a persistent avalanche problem is closely related to the instability of the linked layer of concern.

To assess the quality of the simulated unstable layers in more detail, we compared the layers' simulated proportion unstable to the reported likelihood of the associated persistent avalanche problem. Forecasters expressed the likelihood of an avalanche
350   problem on a five level ordinal scale ranging from *unlikely*, over *possible*, *likely*, *very likely*, to *almost certain* (Statham et al., 2018a). To quantify the agreement in layer instability Ψ, we defined several agreement indicators related to the variation and timing of instability. For variation, we used the Spearman rank correlation (ρ_Ψ) (Wilks, 2019, p. 59f) between the proportion unstable and the assessed likelihood. ρ_Ψ ranges between [-1, 1] with a perfect positive correlation at 1 and no correlation at 0. For the timing we used (i) the lag in the onset of the layer starting to be a problem (Λ_{onset}), (ii) the lag between the layer ceasing
355   to be an issue in the simulations and forecasters removing the associate avalanche problem in the bulletin ($\Lambda_{turn-off}$), and (iii) the difference in the total number of days that the layer appeared to be a problem ($\Delta_{duration}$). To quantify the relevant times in the simulations we used the proportion unstable with different thresholds. After initial explorations with 5, 20, and 50 % for the threshold, we settled on using 20 % for all calculations.

To offer insights into possible reasons for why certain layers were modeled better than others, we performed a series of
360   CTree analyses on the individual indicators similar to how we used CTrees in the previous section. To focus this analysis, we reduced our data set to layers of concern at TL and layers for which all agreement indicators could be calculated. For example, we removed all layers that were characterized by a constant likelihood, which prevented calculating the correlation coefficient ρ_Ψ. We also removed all layers whose maximum proportion unstable was lower than the chosen threshold for Λ. For a threshold of 20 %, for example, the resulting data set contained 96 layers of concern.

4 Results

4.1 Inventory of all layers of concern

Our database contained a total of 167 layers of concern, of which 107 layers were identified to be primarily of weak grain type (SH, DH, FC) and 60 to be primarily crust layers (MFcr, IF). Most layers of concern were reported and tracked in GNP (on average 8.4 layers per season), followed by BYK and then S2S (on average 4.7 and 2.2 layers per season, respectively) (Fig. 5b, d, f). While GNP and BYK reported twice as many weak layers than crust layers, S2S reported an equal ratio. The data quality of the layers was assessed as excellent or good in 50–60 % of cases. The detailed annual inventory of all layers of concern revealed a high inter-seasonal variability in the number of reported layers without a long-term trend (Fig. 5a, c, e).

The number of days with persistent avalanche problems showed a slightly different pattern (solid lines in Fig. 5a, c, e). While S2S experienced persistent avalanche problems at about 30 % of all days of forecasting operations, 60 and 70 % of the days in GNP and BYK had a persistent avalanche problem. In S2S and GNP the vast majority of issued persistent problems were linked to tracked layers of concern, whereas in BYK these problems were linked to tracked layers in only 70 % of the cases (dotted lines in Fig. 5b, d, f). The number of persistent problem days fluctuated substantially from year to year in S2S and BYK (Fig. 5a, e). Although the number of persistent problem days was more steady in GNP, two seasons deviate substantially from the seasonal average (Fig. 5c).

4.2 General simulation performance

We first discuss the overall performance of the simulations for representing layers of weak grain types (SH, DH, FC) using the statistical approach of simulating layer instability for all regions, seasons, and elevation bands combined. The maximum daily proportion of unstable grid points obtains significantly different distributions for layers that were of operational concern and for those that were not (Wilcox rank sum test: P < 0.001, Fig. 6). Most layers of concern tend to be associated with more unstable grid points than layers that were not of concern. The resulting performance curve shows a linear decrease of precision with increasing probability of detection (Fig. 7a). As expected, the maximum precision of 60 % is attained with a high threshold for proportion unstable of 95 %, and precision drops slightly under 40 % for proportions unstable below 20 %. In contrast, the probability of detection increases with decreasing proportion unstable and maxes out slightly above 80 % (Fig. 7a, b). The false alarm rate increases with decreasing proportion unstable, first slowly for high proportions and then stronger for low proportions unstable. The precision–recall and ROC curves both indicate that the simulations have skill in capturing layers of concern, but their performance is far from perfect. A perfect model would produce a horizontal line at the top of the precision–recall curve, and the line in the ROC curve would follow the y-axis to the upper left corner and then across at the top of the chart. On the other side, the no-skill base line is represented by a horizontal line in the lower part of the precision–recall curve and by the diagonal in the ROC curve (dashed lines in Fig. 7).

Analyzing the overall model performance with the process-based approach to assessing layer instability reveals several findings. Overall, the process-based approach yields better results with a threshold for the critical crack length of $r_c \leq 0.3$ than $r_c \leq 0.4$ (Fig. 7c, d). While the performance of the process-based approach with the higher threshold is significantly lower than
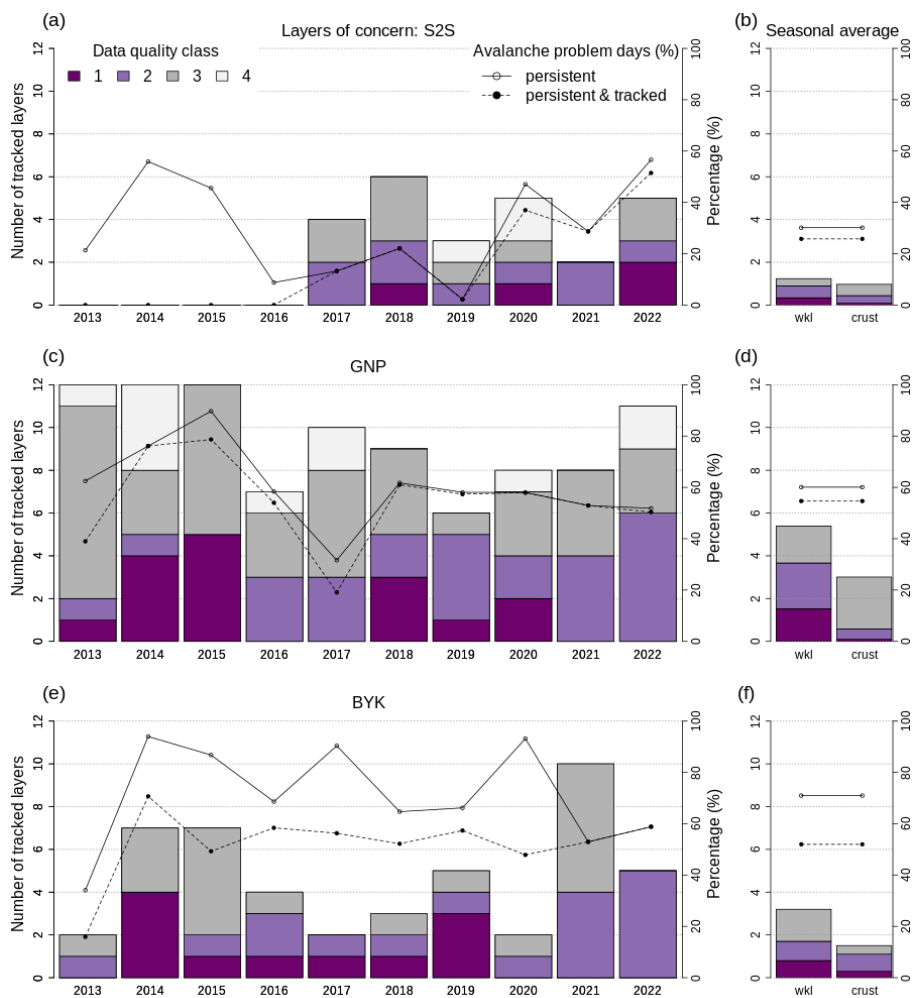
Figure 5. An inventory of tracked layers of concern that were extracted from the human hazard assessments for the regions Sea-to-Sky (S2S), Glacier National Park (GNP) and Banff-Yoho-Kooteney (BYK). Detailed inter-seasonal patterns are visualized for each region in (a, c, e), and seasonal averages are shown in (b, d, f) for weak layers (wkl) and crusts. The colored bars sum up the number of layers in each data quality class, and the lines show the percentage of days per season where a persistent avalanche problem was issued (solid line and open circles), and linked to a tracked layer of concern (dashed line and filled circles).
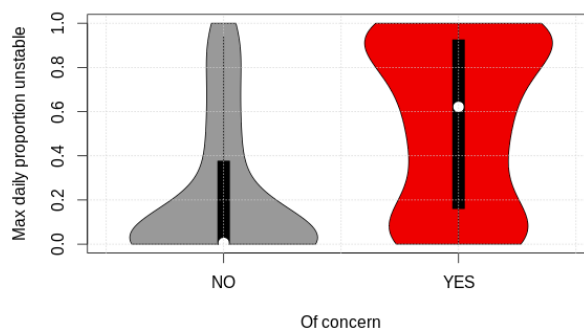
Figure 6. The overall distributions of the maximum daily proportion unstable for all weak layers (surface hoar/depth hoar/facets) that were and were not of operational concern.

the statistical approach on the 95 % confidence level, the process-based approach with the lower threshold yields comparable results and lies within the mutual confidence bands (Fig. 7a, c). Furthermore, while the process-based precision–recall curve

400 with the higher threshold remains constant at approximately 40 % for all proportions unstable, the process-based approach with the lower threshold follows a similar slope as the statistical approach and reaches a comparable maximum precision. At equal proportions unstable both process-based curves show a 10-15 percentage points lower probability of detection than the statistical approach, which can be seen in all panels of Fig. 7. Lastly, the false alarm rate of the process-based approach with the lower threshold follows the same pattern of the statistical approach, but at identical proportions unstable the process-based

405 approach yields an approximately 8 percentage points lower false alarm rate.

Condensing the three performance attributes into the Peirce skill score (PSS) reveals a very flat peak of the statistical approach at a threshold of approximately 20 % for the proportion of unstable grid points and a PSS of 0.41 (Fig. 8a). The process-based approach with the lower threshold obtains a maximum PSS that is five percentage points higher, albeit only for very low thresholds of the maximum proportion unstable of around 3 %. For higher proportions the PSS of the process-based

410 approach decreases linearly below the PSS of the statistical approach. The F_1 curves show a similar pattern, but with an even flatter peak and a slower decrease with increasing proportions unstable. For our subsequent analysis, we focus on the statistical approach and neglect the process-based approach.

Calculating separate performance curves for grain types revealed that SH layers are modeled with a precision between 80-100 %, whereas FC layers only between 25-40 %. False alarm rates are also significantly lower for SH layers. The probability

415 of detection, however, is slightly better for FC (Fig. 9a, b). Crust layers are virtually absent from the simulations with probabilities of detection between 0-10 % (Fig. 9a, b). The performance curves for the different elevation bands highlight that the simulations at TL score substantially better than in the ALP with higher precision at higher proportions unstable and a generally slightly better probability of detection (Fig. 9c, d). False alarm rates remain comparable. The performance curves for BTL is between the TL and ALP, but more closely aligned with TL. The curves for the different forecast region show that precision
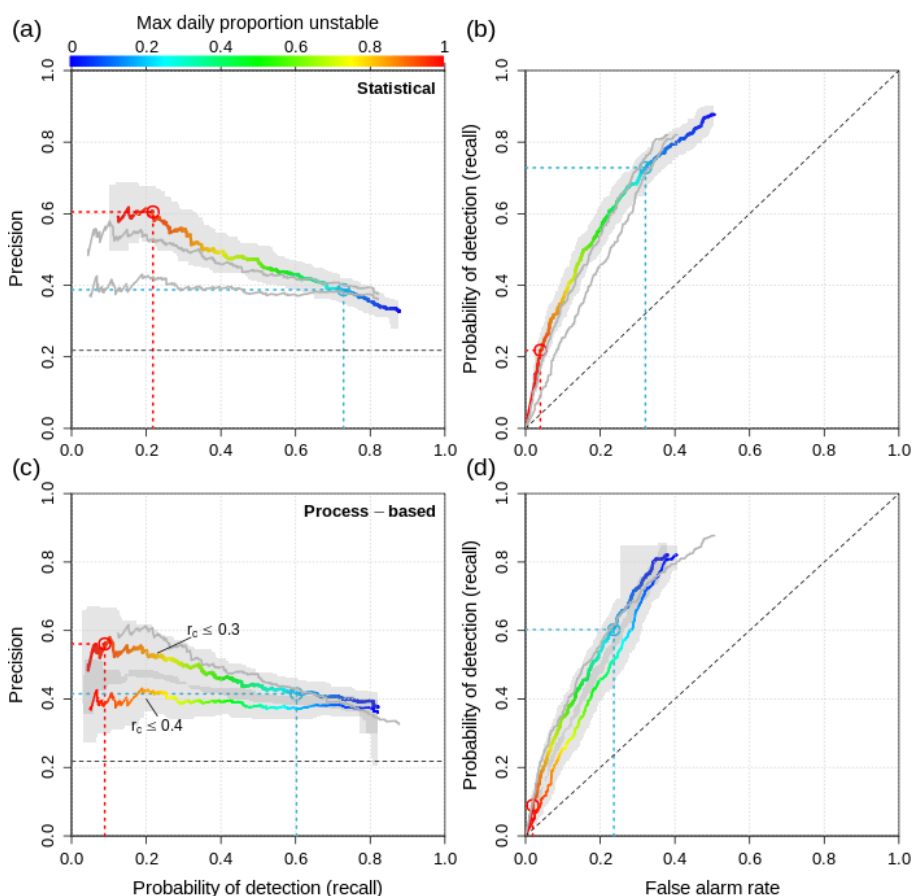
Figure 7. (a), (c) Precision–recall and (b), (d) ROC curves characterizing overall model performance for weak grain types in all forecast regions and at all elevation bands given different thresholds of the maximum daily proportion unstable (color). (a), (b) for the statistical approach of simulating layer instability, and (c), (d) for the process-based approach employing two different thresholds for the critical crack length r_c. Gray bands indicate 95 % confidence bands, gray lines shadow the curves of the other panel.
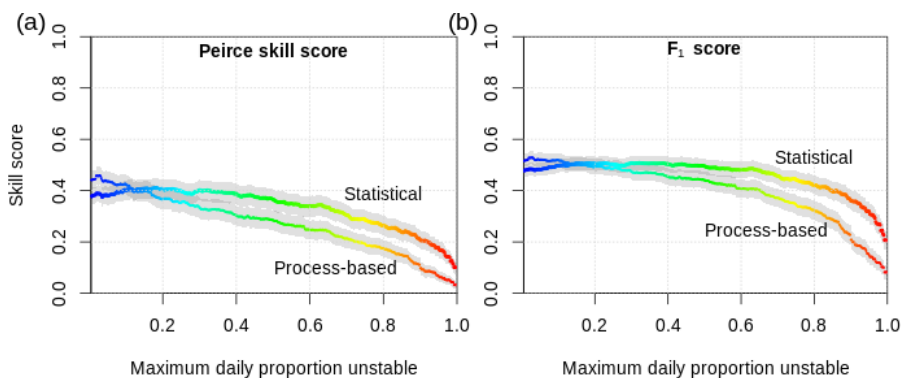
Figure 8. Skill scores for the statistical and process-based approaches to modeling layer instability as a function of the threshold for maximum daily proportion of unstable grid points. The process-based approach in this curve uses a threshold of the critical crack length $r_c \leq 0.3$. Gray confidence bands represent the 95 % level. Color refers to the same value as the x-axis and is shown only to facilitate cross-referencing to Fig. 7.

420   is substantially higher in GNP (50-90 %) than in S2S and BYK (20-50 %), while the probability of detection is identical or slightly lower in GNP (Fig. 9e, f). False alarm rates are also lower in GNP than in S2S and BYK, although the effect is not as pronounced as for precision. The confidence bands highlight a substantially higher uncertainty in S2S than in BYK and than in GNP, which is due to differences in sample sizes.

4.3   Patterns in representing layers of concern

425   ### 4.3.1   Missed or structurally captured layers of concern

The CTree analysis examining the differences between layers that were structurally captured and those that were not revealed that the proportion captured most significantly depends on the grain type (Node 1–first split of the CTree, Fig. 10). While sun crusts (IFsc) are virtually absent from our simulations (Node 13), rain crusts and temperature crusts (IFrc, MFcr) are more likely to be captured, particularly when they formed in the early season (Nodes 11 and 12). The proportion captured of SH
430   layers is substantially influenced by the length of the dry spell before the layer got buried. While the majority of SH layers that got buried after a dry spell of at least seven days are structurally captured by our simulations (Node 7), the proportion captured varies widely within the entire spectrum if the dry spell is shorter than seven days (Node 8). In contrast to crusts and SH layers, all FC layers of concern are structurally present in the large majority of grid points (Nodes 4 and 5), although some FC layers of lower data quality class are represented by lower proportions captured (Node 5).

435   While the CTree algorithm found additional significant splits in the terminal nodes shown in Figure 10, many of them became difficult to interpret and did not provide additional insight. Hence, we are not showing the full tree. However, some splits revealed patterns worth mentioning. For example, FC were almost always present next to SH in GNP and BYK even if they were not reported. Early season IFrc and MFcr (Oct, Nov, Dec) were well captured at BTL and moderately well captured
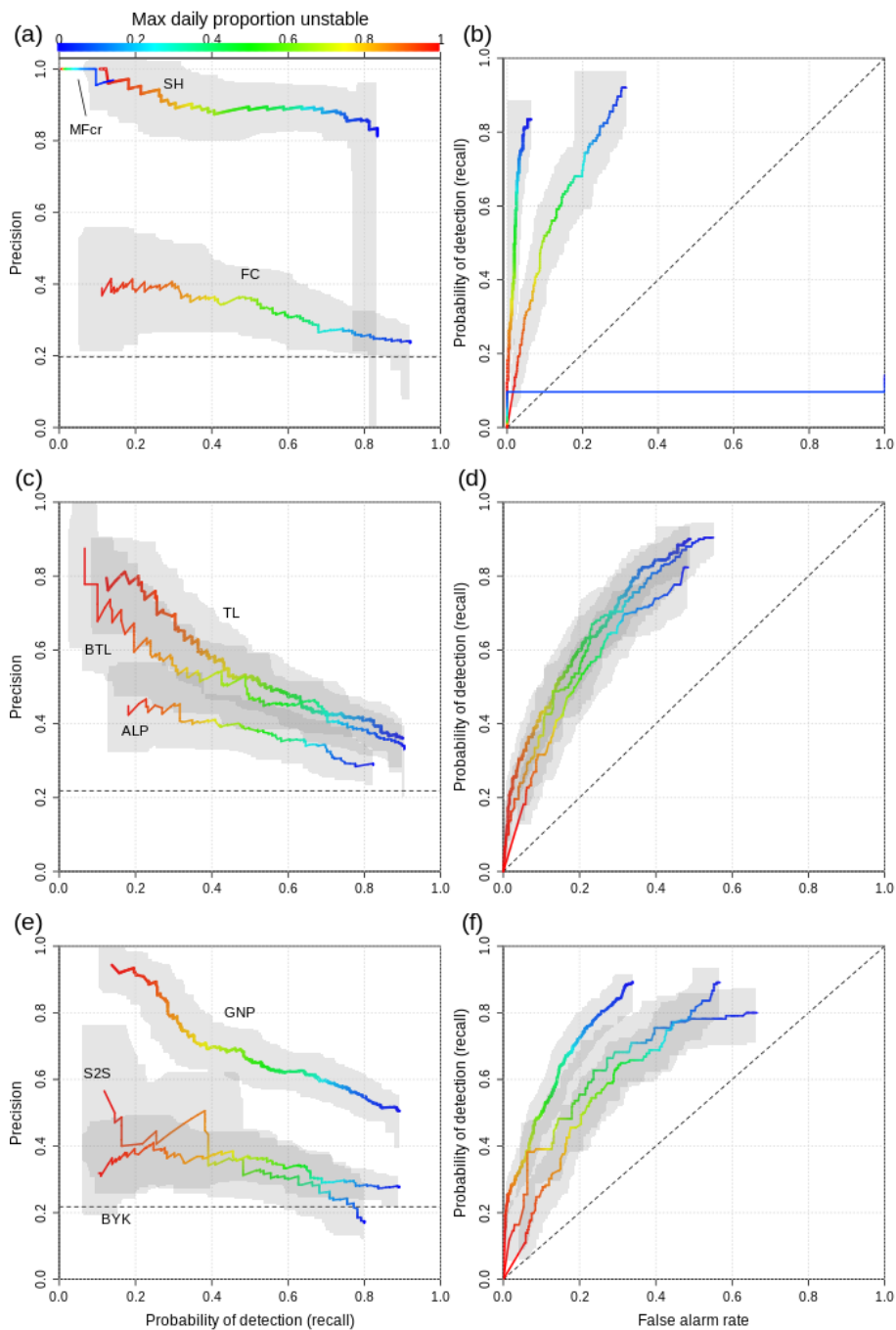
Figure 9. (a), (c), (e) Precision–recall and (b), (d), (f) ROC curves characterizing model performance for different grain types, forecast regions, and elevation bands given different thresholds of the maximum daily proportion unstable (color). All panels were computed with the statistical approach of modeling layer instability. Gray bands indicate 95 % confidence bands.
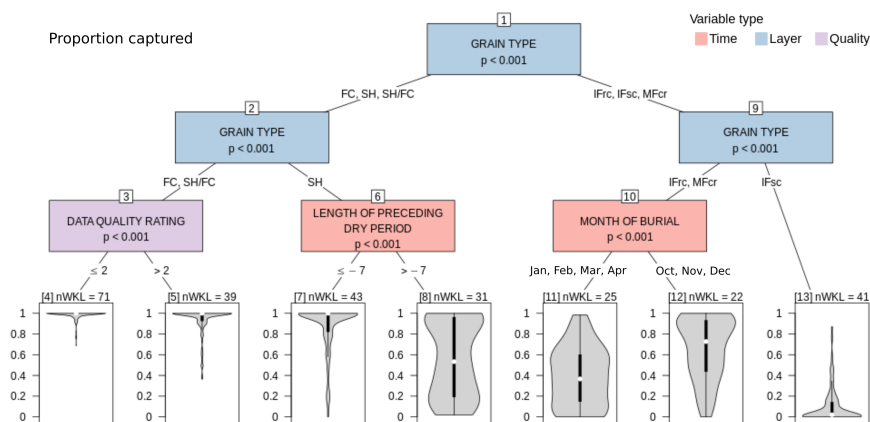
Figure 10. CTree for proportion captured examining the differences between layers of concern that were structurally captured by many grid points and those that were captured by few grid points or none. Terminal nodes are labeled with the number of distinct layers they contain (nWKL).

at ALP and TL (albeit with large variability). Mid- and late season IFrc and MFcr were less well captured in all elevation
440   bands.

4.3.2   Missed or structurally captured and unstable layers of concern

Due to the poor model performance in capturing crust layers, we limited the CTree analysis of all missed or structurally captured and unstable layers of concern to weak grain types (SH, DH, FC). Upon fitting a CTree with the full list of potential explanatory variables, we found that 'season' continuously emerged as an important significant variable, highlighting that
445   model performance varies substantially between seasons. Seasons with good performance include 2014, 2016, and 2019, whereas 2017 and 2020 were characterized by particularly poor performances (not shown). Since the influence of 'season' made the resulting CTree challenging to interpret, we removed it from the list of potential explanatory variables for the analysis presented in Fig. 11.

After removing season effect, the governing variables were the same as in the CTree for just structurally captured layers:
450   grain type, length of the dry spell, and data quality. Hardly any SH layers became unstable in any grid points if the length of the preceding dry spell was shorter than 7 days (Node 9). If the dry spell was longer, the proportion unstable was substantially larger for most SH layers (Node 8). Layers that consisted of FC or mixtures of SH/FC became unstable at most grid points and for most layers of concern, if the length of the dry spell exceeded eight days or was shorter but associated with a high data quality class (Nodes 3 and 5). Layers that were characterized with a shorter dry spell and lower data quality class showed
455   substantially more spread (Node 6).
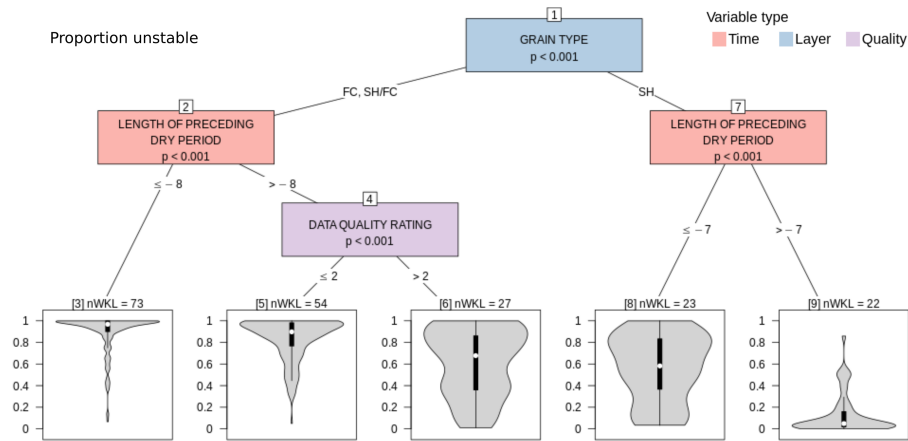
Figure 11. CTree for proportion unstable examining the differences between layers of concern that were modeled unstable by many grid points and those that were modeled unstable by few grid points or none. Terminal nodes are labeled with the number of distinct layers they contain (nWKL).

Although the simulated grain size of SH layers at their time of burial did not show up as a significant variable in the CTrees that were fitted with multiple variables, a univariate analysis confirmed a significant relation with the proportion unstable: larger grain sizes tended to be associated with higher proportions unstable (not shown).

4.3.3 Agreement indicators between modeled and reported instability of captured layers of concern

460 The agreement in modeled and reported variation of instability ρ_Ψ was primarily influenced by the number of days each layer was associated with a persistent avalanche problem (Fig. 12a). A strong negative correlation tended to be attributed to layers whose instability was very short-lived (Node 2) and can likely be attributed to phase lags of the peak instability. The best agreements were found with layers that were of concern for more than 14 days (Node 5), and among those for layers whose modeled onset of instability was close to their reported onset (not shown). The remaining layers were characterized by a highly 465 variable agreement, many appeared to be even uncorrelated (Node 4). The agreement was not influenced by grain type or data quality class. However, a qualitative inspection of all layers that were of concern for more than 14 days but characterized by a $\rho_\Psi < 0.4$ suggested the following possible reasons for the poor agreement:

– The instability was primarily reported for crust layers, none of which were properly captured by the simulations (five layers, all in BYK)

470 – The reported instability tapered off while the modeled instability remained high (four layers of good data quality class)

– The modeled instability at the time of concern was dominated by a different weak layer (two layers of poor data quality class)

- The proportion unstable followed the variation of the danger rating, while the reported likelihood remained constant and therefore yielded worse correlations (two layers of good data quality class)

475 - A modeled precipitation event was underestimated, which led to underestimated instability in buried weak layers (two layers of good data quality class)

- The modeled instability tapered off earlier than the reported instability (one layer of good data quality class).

The agreement indicator Λ_{onset} for the onset of the layer starting to be of concern was found to be independent of the threshold used for the analysis. Furthermore, for almost all layers the modeled onset of the instability was earlier than or at the time of
480 the reported onset (Fig. 12b). For FC layers or SH and SH/FC layers with pre-burial dry spells longer than 14 days the onset date tended to be modeled substantially earlier than reported, often up to three weeks earlier (Nodes 2 and 4). The smallest deviation in the onset date and therefore best agreement was found for SH and SH/FC layers, whose dry spell was shorter than 14 days (Node 5).

Similarly to Λ_{onset}, the agreement indicator $\Lambda_{turn-off}$ for the timing of when layers ceased to be of concern was not affected
485 by the threshold used. $\Lambda_{turn-off}$, however, was evenly distributed around lag zero, which means that the modeled instability sometimes healed sooner and sometimes later than the reported instability. The number of problem days was the only driver influencing the distribution (Fig. 12c): While layers that were of concern for more than 17 days tended to be modeled stable sooner (Node 3), layers that were of concern less long tended to be modeled stable later than what the human forecasts had suggested (Node 2).

490 Consistent with the results for the timing of when layers started and ceased to be of concern, the agreement indicator $\Lambda_{duration}$ was influenced more heavily by a potential lag of the final phase of the layers' concern than during its onset. The combined effects led to most layers being modeled unstable for more days than they were associated with avalanche problems (Fig. 12d). Only layers that were of concern for more than 16 days and were characterized by a dry spell shorter than 12 days tended to be modeled unstable for fewer days (Node 9). However, any layer associated with a long dry spell tended to be modeled unstable
495 for more days (Nodes 3 and 8). Layers that were of concern for less than 16 days without long dry spell tended to be modeled unstable for only few more days (Nodes 5 and 6).

5 Discussion

5.1 Insights for the use of snowpack modeling in operational avalanche forecasting in Canada

Our results have shown that the Canadian weather and snowpack model chain is able to skillfully represent weak layers that
500 are of operational concern. However, the performance is far from perfect and caution is advised. The results of our analysis offer an opportunity to better understand a wide range of model scenarios that cover different regions, elevation bands, grain types, and a varying degree of simulated instability, which we express as the proportion of unstable grid points within a forecast region.
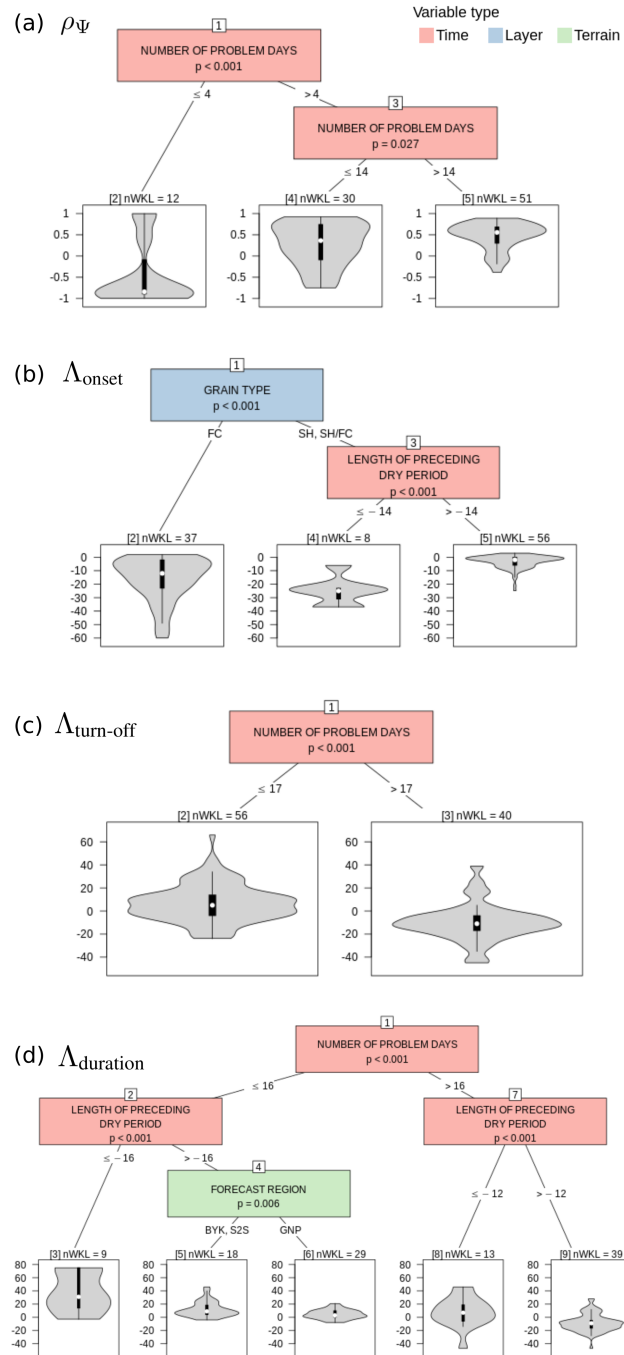
Figure 12. CTrees for different agreement indicators between modeled and reported instability. (a) CTree for the indicator of variation ρ_Ψ examining the differences between layers of concern whose reported instability agreed well and poorly. (b), (c), (d) similar CTrees but for the indicators of timing Λ_{onset}, $\Lambda_{\text{turn-off}}$, and $\Lambda_{\text{duration}}$, respectively. Terminal nodes are labeled with the number of distinct layers they contain (nWKL).

When using model simulations for assessing avalanche conditions, avalanche forecasters need to decide at what proportion of unstable grid points they should be concerned about a particular simulated layer. Naturally, the simulations capture more layers of operational concern when employing a lower threshold, but they also issue more false alarms. For a higher threshold, the number of false alarms decreases substantially, but also the number of captured layers decreases and therefore more layers of concern are missed. To illustrate this trade off, let us examine an average season, which typically includes 20 date tags of which four are associated with a layer of concern. Using a threshold of 20 % for the proportion unstable, eight date tags are identified as unstable layers in the simulation, but only three are actually of human concern whereas five are not of human concern. With this threshold, only one layer of human concern (out of four) is missed (13a). Using a higher threshold of 95 % unstable grid points, only two out of 20 date tags are associated with simulated unstable layers. Of these two simulated layers, one is identical to the human layers of concern, while one is not reported to be of concern. Three layers of human concern (out of four) are missed (13b). These examples are informed by the more general performance results presented in Fig. 7a, b. Forecasters can use these curves to understand the interaction between the proportion of unstable grid points and the resulting model performance, which can help them make informed decisions about how to integrate the simulated information into daily decisions. The precision–recall curve is particularly important for forecasters who find unstable layers in the simulations and ask themselves whether or not to act upon them. If the encountered layer is unstable in 20 % of grid points and the forecasters decide to take the layer seriously, the likelihood of capturing all layers of concern is 75 % (i.e., probability of detection), while the likelihood that the encountered layer is actually of concern is only about 40 % (i.e., precision) (blue circle and blue dotted lines in Fig. 7a). If the forecasters decide to take only layers with as many as 95 % unstable grid points seriously, the likelihood of capturing all layers of concern decreases to almost 20 %, while the likelihood that each of those layers will actually be of concern rises to 60 % (red circle and red dotted lines in Fig. 7a). Hence, only taking layers seriously that are unstable in the majority of grid points and ignoring layers that are represented by a small proportion of grid points means missing many layers of concern. By contrast, when acknowledging any simulated layer with a date tag, even if the layer is only present in a small proportion of grid points, forecasters can be confident that they have narrowed in on a subset of layers that captures the majority of layers of concern. They can then focus on examining observational data (e.g., field observations, avalanche observations) to discern which layers are truly layers to be concerned about and which ones are false alarms. In this way the simulations provide a valuable starting point for targeted observations.

The performance of the simulation strongly depends on grain type. Simulated layers whose prevalent grain type is SH—or at least a mix of SH/FC—were very often of concern. If forecasters encounter a regionally grouped, unstable layer in the simulations that is labeled as SH or SH/FC, they can therefore assume that the layer is very likely of concern if it is unstable in more than 10 % of grid points. In contrast, pure FC layers are modeled more often than they are of concern. This causes regular false alarms from FC layers on the one hand, but on the other it causes the simulations to capture slightly more FC layers of concern than SH layers. These findings are consistent with the results of the CTree analysis, which showed that the vast majority of FC layers was structurally captured at most grid points, even if FC were not reported, and particularly if a long dry period without precipitation preceded the burial of the layer. We therefore hypothesize that our simulations tend to overpredict the structural existence and instability of FC layers.
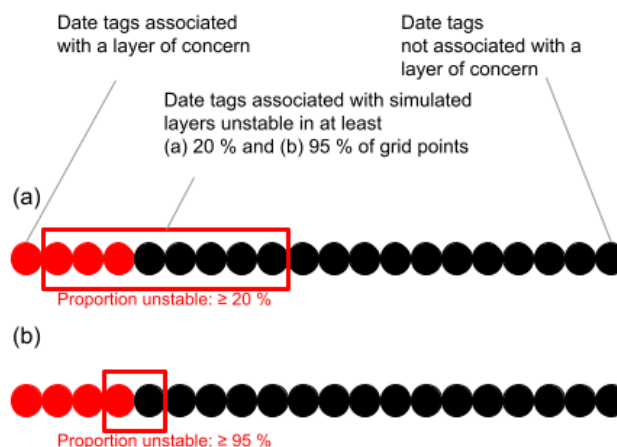
Figure 13. A visual representation of the confusion matrix results for different thresholds of unstable grid points. Filled circles indicate the average number of date tags per season that are associated (or not) with a layer of human concern and with a simulated layer unstable in (a) 20 % of grid points (corresponds to red circle and red dotted lines in Fig. 7a, b), and (b) 95 % of grid points (corresponds to blue circle and blue dotted lines in Fig. 7a, b).

Crust layers of any kind were heavily underrepresented and underestimated in our simulations. Rain and temperature crusts

540   were sometimes modeled by a limited number of grid points, most often in the early season and at lower elevations. Although sun crusts were regularly discussed by forecasters, they were never captured by the simulations. Since the formation of sun crusts is substantially influenced by aspect-dependent exposure to solar radiation, we believe that the lack of simulated sun crusts in our study is primarily caused by the the fact that we only used flat field simulations. Given the poor model performance in capturing any type of crust, we advise to use the Canadian operational weather and snowpack model chain only for assessing

545   weak layers (such as SH, DH, FC) and not crusts.

The performance of the simulations also varies with elevation band and forecast region. In the ALP, the model tends to produce many false alarms at any proportion of unstable grid points, while at BTL and TL layers are considerably more likely to be of concern when they are unstable in more grid points. These effects could be influenced by grain type, since there are likely less SH layers and more FC layers in the ALP than at lower elevations. The forecast region had a strong effect on model

550   performance, which made simulations in GNP substantially more trustworthy than in S2S and BYK. This strong effect is likely influenced by the regions' snow climates. Due to its transitional snow climate GNP is characterized by substantial snowfall amounts interspersed with frequent periods of critical layer formation (Haegeli and McClung, 2007; Shandro and Haegeli, 2018). Therefore many persistent avalanche problems exist each season, most of which can usually be linked to specific thin critical layers (Fig. 5c, d). In the maritime snow climate of S2S, critical layers form much less often and cause less persistent

555   problems (Fig. 5a, b). Although continental BYK experiences the most days per season with persistent avalanche problems, less persistent problems can actually be linked to specific critical layers (Fig. 5e, f). Instead of thin critical layers, the continental snowpack is often characterized by thick bulk layers of low cohesion. Since these thick layers often get deposited by different

snowfall events and facet over the course of many dry spells, it can be challenging to name these layers, let alone distinguish
them in the field. For these reasons our data set of tracked layers of concern is skewed towards GNP and leaves BYK and
560 particularly S2S underrepresented. Our results for S2S and BYK therefore have to be interpreted with more caution and in
light of their regional peculiarities. For example, Horton and Haegeli (2022) found that BYK (and other forecast regions in the
Canadian Rockies) consistently receive underestimated modeled snowfall amounts, which increases temperature gradients in
the snowpack and helps explain the overestimated faceting we see in our results of BYK, which in turn leads to a low model
precision in BYK. In contrast, the low model precision in S2S could be due to an underrepresentation of critical layers of
565 concern. Since instabilities are often short-lived when critical layers get buried by big storms, many critical layers potentially
get never associated with a persistent avalanche problem.

Although our analysis of agreement indicators between modeled and reported instabilities likely pushed the limits of our data
sets (see Limitations section, 5.2), the findings still suggest that the simulations could be used as a complementary information
source for the critical assessment of buried weak layers. Overall, most indicators that describe the timing of instability agreed
570 on average, but we observed large variations. Furthermore, our analyses showed that forecasters tended to be concerned about
layers that were associated with a persistent avalanche problem for longer than 17 days consistently longer than suggested
by the simulations. This is in line with interviews with Canadian avalanche forecasters who acknowledge the challenges of
taking persistent avalanche problems off the bulletin (Hordowick and Haegeli, 2022). The simulations could therefore provide
forecasters with valuable insight for when to remove persistent avalanche problems. Lastly, we found encouraging results in the
575 indicator that describes the correlation between modeled and reported instability. The longer the layers tended to be assessed,
the better the agreement turned out on average. Our more detailed qualitative follow-up analysis of layers of long-lived concern
found that poor correlations were usually caused by an instability that was modeled more severe than it was assessed. In other
words, either the simulations tend to give a conservative recommendation, which is desirable, or forecasters underestimated the
instability and would therefore benefit from model guidance. In other instances poor correlations could mostly be explained by
580 other apparent reasons, such as an underestimated snowfall event, which highlights the need for continuous real-time model
validation (see Sect. 5.3 for a discussion on potential future validation and monitoring).

5.2 Limitations

While the present results provide interesting insights for avalanche forecasters in Canada, our data sets have several limitations
that need to be considered when interpreting the results. Two of these limitations have already been discussed in the previous
585 section—the large influence of the GNP study area on the results due to the large number of persistent weak layers in a
transitional snow climate, and the limitations imposed by the flat field simulations—but the primary limitation of our analysis
is the fact that the validation data set consists of human assessments that do not necessarily represent an objective truth.

Our validation data set consists of human assessments that represent filtered and synthesized data from a comprehensive
range of relevant observations such as weather, snowpack, and avalanche observations from different sources, locations and
590 times. The resulting data set is available daily during the winter season and describes regional-scale conditions. There are
no other data sets that describe the avalanche conditions in such a structured and consistent manner for so many seasons

and over such large areas. However, since human assessments are subjective judgments that can be influenced by operational requirements and practices, they can contain biases and inconsistencies (Statham et al., 2018b; Techel et al., 2018; Horton et al., 2020b). Particularly since the Conceptual Model of Avalanche Hazard (Statham et al., 2018a) is purely qualitative and does

595   not suggest any quantitative links between observations and hazard components (e.g., likelihood of an avalanche problem), the assessment requires substantial human judgment and expertise. While the binary assessments of layers of concern are likely one of the more straight forward and therefore reliable aspects of the used assessments, more intricate attributes such as onset or turn-off dates of avalanche problems and particularly the reported likelihood of a layer to cause avalanches are more susceptible to biases and inconsistencies.

600   In addition to potential human error, sparsity of observations in general can impact the quality of the human assessments. Although we chose three relatively data-rich study regions and assessed each layer of concern for its data quality to alleviate the limitations, it remains unclear whether the simulations or the human assessments were closer to reality when the two data sets disagreed. For example, while several modeled critical layers that were not reported in the assessments may be attributed to the model being overly sensitive, some could also have been missed by human observers. Similarly, while most layers of concern

605   that were not captured by the simulations can likely be attributed to model deficiencies, a few might actually not have been as reactive as initially expected by the forecasters, or it might have been another layer that was the main cause of instability at a given day. Hence, our results can be interpreted as the lower limit of expected model performance.

5.3   A vision for the future use of snowpack simulations in operational avalanche forecasting

Based on the insights from the present study and our practical experience with the tools developed for this research, we

610   have developed a vision of how snowpack modeling can be embedded into operational avalanche forecasting dashboards more informatively. Despite the encouraging results found by our critical layer validation, we do not believe that the existing model chain is sufficiently reliable to generate assessments purely based on simulations. Instead, we view the simulations as a rich complement to other information sources that can help alert forecasters about the existence of specific critical layers or provide an additional, independent perspective on their instability. To enable their full potential as a complementary information

615   source, snowpack simulations need to be embedded in a validation suite that allows forecasters to continuously compare past assessments to past and future simulations in real time. Continuously judging the model performance will allow forecasters to develop a better understanding for the strength and weaknesses of this data source over time.

To be useful for forecasters and fit into their already busy forecasting days, such a validation suite must present the information in an intuitive way that integrates seamlessly with their existing practices (Horton et al., 2020a). Hence, we suggest to

620   present simulated layers in a grouped format similar to human layers of concern. The methods described in this paper can be used to group and compare layers in such a format. Figures 14 and 15 illustrate how we envision the dashboards of a potential validation suite to look like using the 2019 winter season at TL in GNP as an example.

At any given day, a seasonal overview as shown in Figure 14 could show the entire history of the season and potentially the future forecast simulations for the next three days. The information presented in this view combines human assessment

625   data with simulated information. Relevant human assessment data includes the danger rating, the tracked layers of concern,

and the likelihood of their associated avalanche problems (Fig. 14a–c). Relevant simulated information includes the proportion captured, the proportion unstable, and the time series of a representative snow profile (Herla et al., 2022) (Fig. 14d, e). While the profile gives a familiar and comprehensive overview of the snow stratigraphy including snow height, new snow amounts, prevalent layers and their times of instability, the proportion captured and proportion unstable offer insight on how many grid

630 points contain the current layer of human concern or suggest instability in that layer (colored circles and bars, Fig. 14d). To acknowledge other layers that are modeled unstable at the same day, we also show the proportion unstable of *all* persistent layers at each day (gray bars, Fig. 14d). A dashboard view like this effectively summarizes the large amount of simulated observations and allows forecasters to easily put their assessments of specific critical layers in relation to the simulated data. Any serious discrepancies between assessments and simulations can be spotted easily. For example, situations where the

635 assessed likelihood or the proportion unstable changes but not the other will jump out immediately. Missing layers will not show up in the representative profile and simultaneously be characterized with a low proportion captured. This information will help forecasters to decide whether the model is wrong because they know the layer exists, or whether they need more targeted pieces of evidence to confirm the existence of the simulated layer. If a layer different from the current layer of concern is modeled to be the main culprit of instability at a given day, this view will alert forecasters about this potential.

640 The second dashboard we envision would focus on a single layer of concern (Fig. 15). Analogously to before, the view combines human assessment information with relevant simulated information. Relevant assessment information could include the times of concern and the likelihood of the associated avalanche problems. To provide meaningful context for the information specific to the layer, the view still includes the proportion captured and the proportion unstable of the specific layer as well as the overall proportion unstable of all persistent layers analogously to the previous view (Fig. 15c, black line with circles, black

645 bars, gray bars, respectively). In addition, the view shows daily summaries of the simulated characteristics of the specific layer of concern, including the layer depth, the statistical stability measure $p_{unstable}$ (Mayer et al., 2022), the process-based stability measures RTA, SK38, and r_c, and other select variables of interest, such as the cohesion of the overlying slab expressed by the average density over grain size of all slab layers $< \frac{\text{density}}{\text{grain size}} >_{\text{slab}}$. These characteristics are visualized as daily distributions in the form of violin plots. To prevent loss of context in this highly zoomed in view and allow forecasters to relate to information

650 of other potential layers of concern, we also show persistent unstable layers that are *not* associated with the specific layer of concern. These other layers are visualized as separate point clouds color-coded according to their native grain types. The novelty of this view is that it selects conceptually equivalent layers based on the methods described in this paper and then visualizes all individual data points from multiple grid points and dates in an accessible way.

The dashboard views shown in Figures 14 and 15 focus on the information contained in the data sets of our present study.

655 In an operational context, additional information will be available that is highly relevant for the proposed validation suite. In particular, adding specific observations like snowfall, snow depth, or avalanche observations will be critical for providing forecasters with the full range of available evidence. For example, since the proportion unstable will have to be interpreted in light of new loading due to added snowfall as currently indicated only in the representative profile, the view could be complemented with observed and modeled snowfall amounts to alert forecasters to situations when modeled snowfall might

660 be substantially under-/overestimated or delayed.
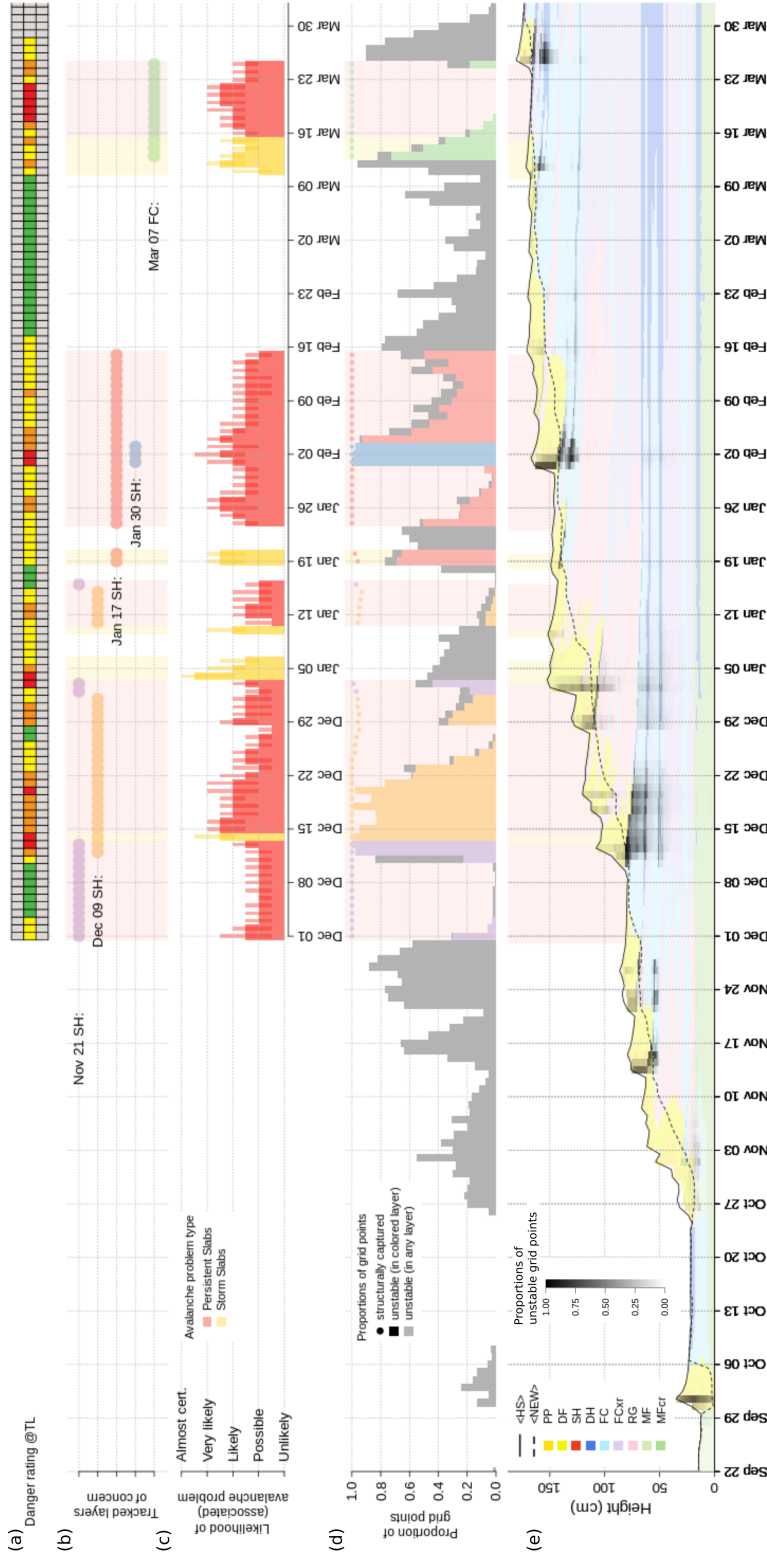
Figure 14. A dashboard view that provides a seasonal overview of (a), (b), (c) human assessments and (d), (e) model simulations. (a) The danger rating at treeline elevation (TL), (b) times when tracked layers of concern were associated with an avalanche problem, (c) the likelihood of associated avalanche problems, (d) the proportion of unstable model grid points in the forecast region, and (e) the time series of a representative simulated snow profile, where <HS> and <NEW> refer to the median height of snow and the median amount of new snow, respectively. Acronyms of snow grain types are defined in Sect. 2.2.
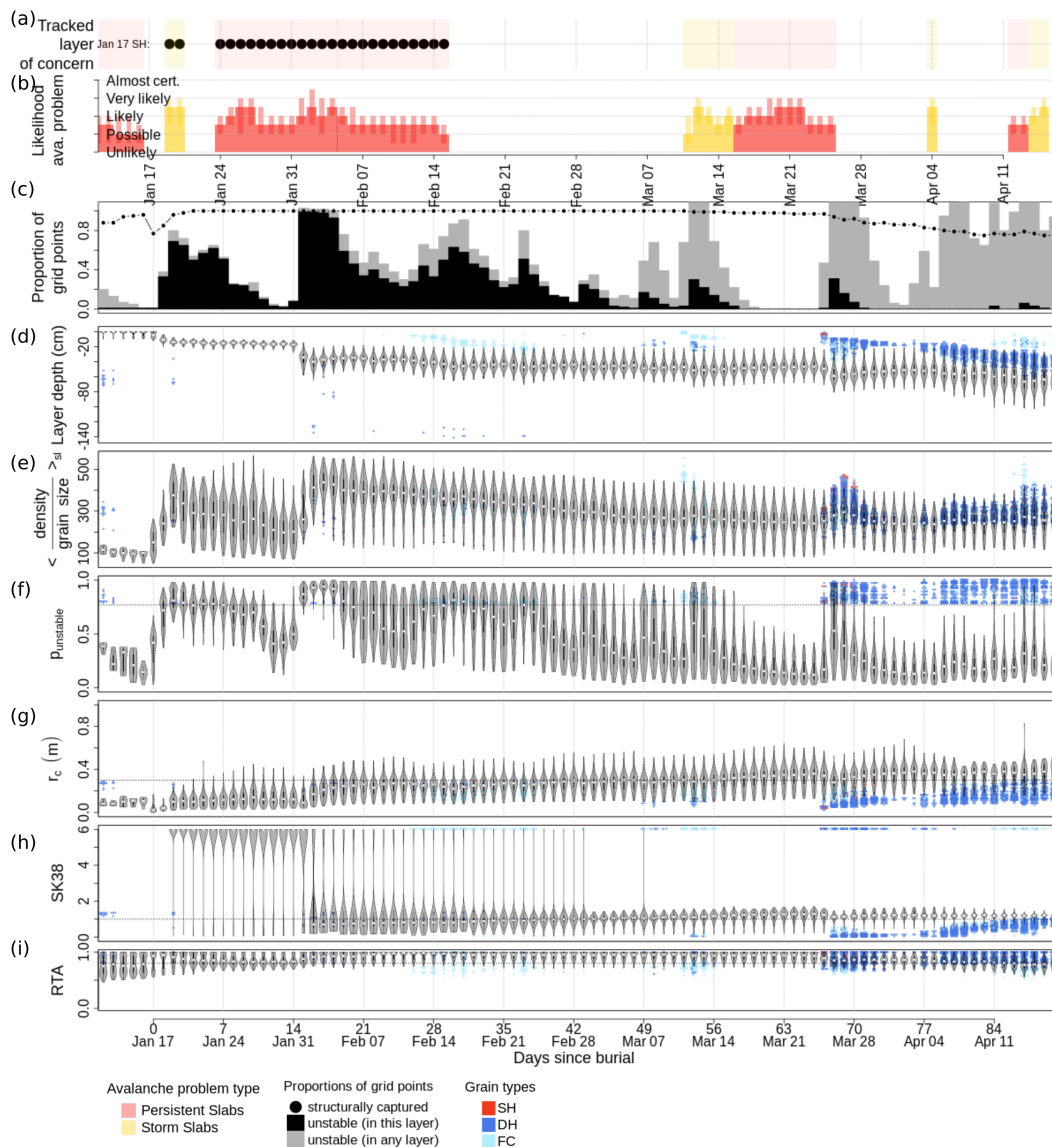
Figure 15. A dashboard view that zooms into the details of an individual layer of concern. (a) Times when the layer was associated with an avalanche problem, (b) the the likelihood of associated avalanche problems, (c) the proportion of unstable model grid points in the forecast region, and daily summaries of simulated characteristics of the target layer (violin distributions) and other layers (point clouds) like (d) layer depth (m), (e) cohesion of the slab $< \frac{\text{density}}{\text{grain size}} >_{\text{slab}}$ (kg m^{-4}), (f) statistical approach to layer instability p$_{\text{unstable}}$, (g) critical crack length r$_c$ (m), (h) stability index SK38, (i) relative threshold sum approach RTA.

5.4 Implications for snowpack modeling

While improving insight for snowpack modeling was not the main objective of this paper, our explorations offer a few interesting observations for this community.

Our large scale critical layer validation offered an opportunity to compare the overall performance of two recently promoted approaches to estimate dry snow layer instability: The process-based approach promoted by Monti and Schweizer (2013); Monti et al. (2014); Reuter et al. (2015); Monti et al. (2016); Reuter and Schweizer (2018); Reuter et al. (2021) and a statistical approach developed by Mayer et al. (2022). Our results from using the process-based approach are in line with recent research that suggested decreasing the threshold for the critical crack length r_c to 0.3 m (Reuter and Schweizer, 2018; Reuter et al., 2021; Mayer et al., 2022). Furthermore, although the statistical approach was fitted with Swiss data, our results demonstrate that it can be employed in western Canada, where it performs at a similar level as the process-based approach. The proportion of unstable grid points in a forecast region based on the statistical approach turned out to be a valuable predictor of instability. Future research might want to investigate whether a combination of the proportion of unstable grid points with the individual magnitude of instability at each grid point yields improvements.

The length of the dry period before the burial of the potential weak layer emerged as a variable with strong explanatory power in many of our CTree analyses (Sect. 3.2 and 3.3). Since longer dry periods conceptually increase the chance of weak layer formation and extended faceting on the snow surface, these findings seem plausible at first. However, the strong influence of this variable in several different CTree analyses (particularly Fig. 12d, 11, and 10) also potentially suggest a model bias towards overestimating the prevalence and instability of weak layers that were buried at the end of very long dry periods. Interestingly, the median simulated grain size at the time of burial was substantially less impactful. In fact, despite many SH layers with reported grain sizes larger than 10 mm, the median simulated grain sizes at the time of burial rarely exceeded 2.5 mm. Hence, the observed model sensitivity to the length of the dry spell must be caused by a different layer property, and the only other plausible options are density and sphericity (Mayer et al., 2022). However, even though grain sizes were simulated substantially smaller than reported in many cases, our findings still suggest that the simulated grain sizes were consistent within the simulations. Our CTree analysis has shown that median simulated grain sizes at the time of burial that exceeded 0.9 mm were associated with a higher proportion unstable than smaller grain sizes. This is in line with Mayer et al. (2022), who report an increased influence of the simulated grain size on layer instability for grain sizes above 1–1.5 mm (Fig. B2d in Mayer et al., 2022).

6 Conclusion

We evaluated the performance of an operational weather and snowpack model chain to represent critical avalanche layers based on human hazard assessments from three avalanche forecast regions in western Canada and ten winter seasons. By developing methods to identify human layers of concern in the simulations and to group simulated critical layers from different times and grid points by their time of burial, we could quantify (i) the likelihood that a layer of human concern was captured by the simulations (probability of detection), (ii) the likelihood that a simulated critical layer was indeed of concern in reality

(precision), and (3) the likelihood that the model simulated a critical layer that is not of concern (false alarm rate). Furthermore, we employed conditional inference trees (CTrees) to (a) identify patterns between layers of concern that were well captured from those that were not and (b) examine the agreement between simulated and reported layers in more detail taking into account the variation and timing of instability.

While we presented model performance curves that will allow forecasters to interpret any model scenario at their hand, the overall model performance can be summarized with a sensitivity of 75 % at a precision of 40 % and a false alarm rate of 30 % when 20 % of model grid points suggest instability in a specific weak layer. While the performance was substantially better for surface hoar (SH) layers than facets (FC), the model had no skill in representing any type of crust layer (IFrc, IFsc, MFcr). The CTree analyses confirmed the strong influence of grain type. While the analysis of agreement indicators between reported and simulated instability approached the limits of our data sets, it suggested that the simulations and human assessments tended to agree for the majority of captured weak layers.

The presented research contributed to address a knowledge gap in snowpack validation, which has primarily focused on process-based model validation at the point scale and validation of bulk properties at the regional scale. By evaluating the detailed hazardous layering at the regional scale in a way that is informative for forecasters, this contribution will help make snowpack simulations more transparent and applicable for operational avalanche forecasting.

We acknowledge that our human validation data set does not represent an objective truth, and the results therefore represent a lower limit of expected model performance. Nevertheless, we believe that the models offer a valuable complementary information source for avalanche forecasting, but we discourage the standalone use of the simulations without any field observations. The methods presented in this paper aim to provide a starting point for designing informative simulation dashboards that enable forecasters to better understand this novel information source by comparing assessments and simulations in real time.

Code and data availability. The data and code to reproduce the analysis in this paper are available from a DOI repository at https://www. doi.org/10.17605/OSF.IO/W7PJY (Herla et al., 2023).

Appendix A: Rules and thresholds to identify model-derived date tags

We used modeled precipitation amounts to identify model-derived date tags (i.e., potential layers of concern). To describe the precipitation in a study region, we aggregated all solid and liquid precipitation from all model grid points accumulated over 24 and 72 hours by calculating their 25th and 75th percentiles ($<HN24>_{75}$, $<HN72>_{25}$, $<RAIN24>_{75}$, and $<RAIN72>_{25}$). Days with less than a trace amount of precipitation were identified by 24-hour amounts of the 75th percentiles

$<HN24>_{75} < 2$ cm and

$<RAIN24>_{75} < 2$ mm.

We refer to these days as 'Days 0'. To determine which of these Days 0 were followed by a major storm, we scanned the five subsequent days 'Days 1–5' for substantial 72-hour precipitation amounts using the the 25th percentiles and following

725 thresholds

$\langle HN72 \rangle_{25}$(@Days 1–5) > 10 cm or

$\langle RAIN72 \rangle_{25}$(@Days 1–5) > 5 mm.

Whenever any day of Days 1–5 satisfied the above rule, we counted a major storm event. While the above rules recognize the onset of major storms, they can also accidentally identify situations when a major storm tapered off. To exclude these times of

730 tapering-off, we added another set of rules to ensure that the 72-hour precipitation amounts were increasing (i.e., higher than on Day 0) and not decreasing:

$\langle HN72 \rangle_{25}$(@Days 1–5) $>$ $\langle HN72 \rangle_{25}$(@Day 0) $+ 5$ cm or

$\langle RAIN72 \rangle_{25}$(@Days 1–5) $>$ $\langle RAIN72 \rangle_{25}$(@Day 0) $+ 2$ mm.

All Days 0 that satisfied the above rules were recorded as model-derived date tags ahead of major storms. To account for long

735 periods in between these main storm events that could have buried critical layers by substantial cumulative accumulations from small daily snowfall amounts, we inserted additional date tags when a period longer than 10 days accumulated more than 25 cm of $\langle HN24 \rangle_{75}$.

Author contributions. All authors conceptualized the research; FH ran the snowpack simulations and implemented the methods and analysis; all authors contributed to writing the paper; PH acquired the funding.

740 *Competing interests.* FH, SH, and PM declare they have no competing interests; PH is a member of the editorial board of the journal.

Acknowledgements. We thank Karl Klassen for valuable feedback and insight throughout the project development.

References

Bartelt, P., Lehning, M., Bartelt, P., Brown, B., Fierz, C., and Satyawali, P.: A physical SNOWPACK model for the Swiss avalanche warning: Part I: Numerical model, Cold Reg Sci Technol, 35, 123–145, https://doi.org/10.1016/s0165-232x(02)00074-5, 2002.

Bellaire, S. and Jamieson, J. B.: Forecasting the formation of critical snow layers using a coupled snow cover and weather model, Cold Reg Sci Technol, 94, 37–44, https://doi.org/10.1016/j.coldregions.2013.06.007, 2013.

Bellaire, S., van Herwijnen, A., Mitterer, C., and Schweizer, J.: On forecasting wet-snow avalanche activity using simulated snow cover data, Cold Reg Sci Technol, 144, 28–38, https://doi.org/10.1016/j.coldregions.2017.09.013, 2017.

Brun, E., David, P., Sudul, M., and Brunot, G.: A numerical model to simulate snow-cover stratigraphy for operational avalanche forecasting, J Glaciol, 38, 13–22, https://doi.org/10.3189/s0022143000009552, 1992.

Calonne, N., Richter, B., Löwe, H., Cetti, C., Ter Schure, J., Van Herwijnen, A., Fierz, C., Jaggi, M., and Schneebeli, M.: The RHOSSA campaign: Multi-resolution monitoring of the seasonal evolution of the structure and mechanical stability of an alpine snowpack, The Cryosphere, 14, 1829–1848, https://doi.org/10.5194/tc-14-1829-2020, 2020.

Campbell, C., Conger, S., Gould, B., Haegeli, P., Jamieson, J. B., and Statham, G.: Technical Aspects of Snow Avalanche Risk Management– Resources and Guidelines for Avalanche Practitioners in Canada, Revelstoke, BC, Canada, 2016.

Canty, A. and Ripley, B.: boot: Bootstrap R (S-Plus) Functions, 2022.

Davison, A. C. and Hinkley, D. V.: Bootstrap Methods and Their Applications, Cambridge University Press, Cambridge, 1997.

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, J Glaciol, 45, 469–484, https://doi.org/10.3189/s0022143000001337, 1999.

Ebert, P. A. and Milne, P.: Methodological and conceptual challenges in rare and severe event forecast verification, Nat Hazard Earth Sys, 22, 539–557, https://doi.org/10.5194/nhess-22-539-2022, 2022.

Fierz, C.: Field observation and modelling of weak-layer evolution, Ann Glaciol, 26, 7–13, https://doi.org/10.3189/1998AoG26-1-7-13, 1998.

Fierz, C. and Lehning, M.: Verification of Snow Cover Models on Slopes, in: Proceedings of the 2004 International Snow Science Workshop, Jackson Hole, WY, USA, pp. 94–98, https://arc.lib.montana.edu/snow-science/item/1059, 2004.

Fierz, C., Armstrong, R. L., Durand, Y., Etchevers, P., Greene, E., McClung, D. M., Nishimura, K., Satyawali, P., and Sokratov, S. A.: The International Classification for Seasonal Snow on the Ground, vol. 5, UNESCO/IHP, https://unesdoc.unesco.org/ark:/48223/pf0000186462, 2009.

Haegeli, P. and McClung, D. M.: Expanding the snow-climate classification with avalanche-relevant information: Initial description of avalanche winter regimes for southwestern Canada, J Glaciol, 53, 266–276, https://doi.org/10.3189/172756507782202801, 2007.

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.: Critical Layer Validation—Data and Code, https://doi.org/10.17605/OSF.IO/W7PJY, 2023.

Hordowick, H. and Haegeli, P.: Understanding avalanche problem assessments: A concept mapping study with public avalanche forecasters, mathesis, http://www.avalancheresearch.ca/pubs/2022_hordowick_mrm/, 2022.

Horton, S. and Haegeli, P.: Using snow depth observations to provide insight into the quality of snowpack simulations for regional-scale avalanche forecasting, The Cryosphere, 16, 3393–3411, https://doi.org/10.5194/tc-16-3393-2022, 2022.

Horton, S. and Jamieson, J. B.: Modelling hazardous surface hoar layers across western Canada with a coupled weather and snow cover model, Cold Reg Sci Technol, 128, 22–31, https://doi.org/10.1016/j.coldregions.2016.05.002, 2016.

780 Horton, S., Nowak, S., and Haegeli, P.: Enhancing the operational value of snowpack models with visualization design principles, Nat Hazard Earth Sys, 20, 1557–1572, https://doi.org/10.5194/nhess-20-1557-2020, 2020a.

Horton, S., Towell, M., and Haegeli, P.: Examining the operational use of avalanche problems with decision trees and model-generated weather and snowpack variables, Nat Hazard Earth Sys, 20, 3551–3576, https://doi.org/10.5194/nhess-20-3551-2020, 2020b.

Hothorn, T. and Zeileis, A.: Partykit: A modular toolkit for recursive partytioning in R, J Mach Learn Res, 16, 3905–3909, http://cran.
785 r-project.org/package=, 2015.

Hothorn, T., Hornik, K., and Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework, J Comput Graph Stat, 15, 651–674, https://doi.org/10.1198/106186006X133933, 2006.

LaChapelle, E. R.: The fundamental processes in conventional avalanche forecasting., J Glaciol, 26, 75–84, https://doi.org/10.3189/s0022143000010601, 1980.

790 Lafaysse, M., Morin, S., Coléou, C., Vernay, M., Serça, D., Besson, F., Willemet, J.-M., Giraud, G., and Durand, Y.: Toward a new chain of models for avalanche hazard forecasting in French mountain ranges, including low altitude mountains, Proceedings of the 2013 International Snow Science Workshop, Grenoble, France, pp. 162–166, https://arc.lib.montana.edu/snow-science/item.php?id=1741, 2013.

Lafaysse, M., Cluzet, B., Dumont, M., Lejeune, Y., Vionnet, V., and Morin, S.: A multiphysical ensemble system of numerical snow modelling, Cryosphere, 11, 1173–1198, https://doi.org/10.5194/tc-11-1173-2017, 2017.

795 Lehning, M., Fierz, C., and Lundy, C.: An objective snow profile comparison method and its application to SNOWPACK, Cold Reg Sci Technol, 33, 253–261, https://doi.org/10.1016/s0165-232x(01)00044-1, 2001.

Lehning, M., Bartelt, P., Brown, B., and Fierz, C.: A physical SNOWPACK model for the Swiss avalanche warning Part III: Meteorological forcing, thin layer formation and evaluation, Cold Reg Sci Technol, 35, 169–184, https://doi.org/10.1016/S0165-232X(02)00072-1, 2002a.

Lehning, M., Bartelt, P., Brown, B., Fierz, C., and Satyawali, P.: A physical SNOWPACK model for the Swiss avalanche warning Part II.
800 Snow microstructure, Cold Reg Sci Technol, 35, 147–167, https://doi.org/10.1016/S0165-232X(02)00073-3, 2002b.

Mayer, S., van 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.

McClung, D. M.: The Elements of Applied Avalanche Forecasting, Part I: The Human Issues, Nat Hazards, 26, 111–129, https://doi.org/10.1023/a:1015665432221, 2002a.

805 McClung, D. M.: The Elements of Applied Avalanche Forecasting, Part II: The Physical Issues and the Rules of Applied Avalanche Forecasting, Nat Hazards, 26, 131–146, https://doi.org/10.1023/a:1015604600361, 2002b.

Michlmayr, G., Lehning, M., Koboltschnig, G., Holzmann, H., Zappa, M., Mott, R., and Schöner, W.: Application of the Alpine 3D model for glacier mass balance and glacier runoff studies at Goldbergkees, Austria, in: Hydrological Processes, vol. 22, pp. 3941–3949, John Wiley and Sons, Ltd, https://doi.org/10.1002/hyp.7102, 2008.

810 Milbrandt, J. A., Bélair, S., Faucher, M., Vallée, M., Carrera, M. L., and Glazer, A.: The pan-canadian high resolution (2.5 km) deterministic prediction system, Weather Forecast, 31, 1791–1816, https://doi.org/10.1175/WAF-D-16-0035.1, 2016.

Monti, F. and Schweizer, J.: A relative difference approach to detect potential weak layers within a snow profile, in: Proceedings of the 2013 International Snow Science Workshop, Grenoble, France, pp. 339–343, https://arc.lib.montana.edu/snow-science/item.php?id=1861, 2013.

815   Monti, F., Schweizer, J., Gaume, J., and Fierz, C.: Deriving snow stability information from simulated snow cover stratigraphy, in: Proceedings of the 2014 International Snow Science Workshop, Banff, AB, Canada, pp. 465–469, https://arc.lib.montana.edu/snow-science/item.php?id=2096, 2014.

Monti, F., Gaume, J., van Herwijnen, A., and Schweizer, J.: Snow instability evaluation: calculating the skier-induced stress in a multi-layered snowpack, Nat Hazard Earth Sys, 16, 775–788, https://doi.org/10.5194/nhess-16-775-2016, 2016.

820   Morin, S., Fierz, C., Horton, S., Bavay, M., Dumont, M., Hagenmuller, P., Lafaysse, M., Mitterer, C., Monti, F., Olefs, M., Snook, J. S., Techel, F., 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 Reg Sci Technol, 170, 1098–1107, https://doi.org/10.1016/J.COLDREGIONS.2019.102910, 2020.

Quéno, L., Vionnet, V., Dombrowski-Etchevers, I., Lafaysse, M., Dumont, M., and Karbou, F.: Snowpack modelling in the Pyrenees driven
825   by kilometric-resolution meteorological forecasts, The Cryosphere, 10, 1571–1589, https://doi.org/10.5194/tc-10-1571-2016, 2016.

R Core Team: R: A Language and Environment for Statistical Computing, https://www.r-project.org/, 2020.

Raleigh, M. S., Lundquist, J., and Clark, M. P.: Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework, Hydrol Earth Syst Sci, 19, 3153–3179, https://doi.org/10.5194/hess-19-3153-2015, 2015.

Reuter, B. and Schweizer, J.: Describing Snow Instability by Failure Initiation, Crack Propagation, and Slab Tensile Support, Geophys Res
830   Lett, 45, 7019–7027, https://doi.org/10.1029/2018GL078069, 2018.

Reuter, B., Schweizer, J., and van Herwijnen, A.: A process-based approach to estimate point snow instability, The Cryosphere, 9, 837–847, 2015.

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 Reg Sci Technol, 194, 103 462,
835   https://doi.org/10.1016/j.coldregions.2021.103462, 2021.

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.

Richter, B., Van Herwijnen, A., Rotach, M. W., and Schweizer, J.: Sensitivity of modeled snow stability data to meteorological input uncertainty, Nat Hazard Earth Sys, 20, 2873–2888, https://doi.org/10.5194/nhess-20-2873-2020, 2020.

840   Schirmer, M., Schweizer, J., and Lehning, M.: Statistical evaluation of local to regional snowpack stability using simulated snow-cover data, Cold Reg Sci Technol, 64, 110–118, https://doi.org/10.1016/j.coldregions.2010.04.012, 2010.

Schweizer, J., Bellaire, S., Fierz, C., Lehning, M., and Pielmeier, C.: Evaluating and improving the stability predictions of the snow cover model SNOWPACK, Cold Reg Sci Technol, 46, 52–59, https://doi.org/10.1016/j.coldregions.2006.05.007, 2006.

Shandro, B. and Haegeli, P.: Characterizing the nature and variability of avalanche hazard in western Canada, Nat Hazard Earth Sys, 18,
845   1141–1158, https://doi.org/10.5194/nhess-18-1141-2018, 2018.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T.: ROCR: visualizing classifier performance in R, Bioinformatics, 21, 3940—-3941, https://doi.org/10.1093/bioinformatics/bti623, 2005.

Statham, G., Haegeli, P., Greene, E., Birkeland, K. W., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J.: A conceptual model of avalanche hazard, Nat Hazards, 90, 663–691, https://doi.org/10.1007/s11069-017-3070-5, 2018a.

850   Statham, G., Holeczi, S., and Shandro, B.: Consistency and Accuracy of Public Avalanche Forecasts in Western Canada, in: Proceedings of the 2018 International Snow Science Workshop, Innsbruck, Austria, https://arc.lib.montana.edu/snow-science/item.php?id=2806, 2018b.

Storm, I.: Public Avalanche Forecast Challenges: Canada's Large Data-Sparse Regions, in: Proceedings, 2012 International Snow Science Workshop, Anchorage, AK, USA, pp. 908–912, https://arc.lib.montana.edu/snow-science/item.php?id=1668, 2012.

Storm, I. and Helgeson, R. G.: Hot-spots and hot times: exploring alternatives to public avalanche forecasts in Canada's data sparse Northern Roeckies region, in: Proceedings of the 2014 International Snow Science Workshop, Banff, AB, Canada, International Snow Science Workshop, Banff, AB, Canada, https://arc.lib.montana.edu/snow-science/item/2034, 2014.

Techel, F., Mitterer, C., Ceaglio, E., Coléou, C., Morin, S., Rastelli, F., and Purves, R. S.: Spatial consistency and bias in avalanche forecasts -a case study in the European Alps, Nat Hazard Earth Sys, 18, 2697–2716, https://doi.org/10.5194/nhess-18-2697-2018, 2018.

Vernay, M., Lafaysse, M., Mérindol, L., Giraud, G., and Morin, S.: Ensemble forecasting of snowpack conditions and avalanche hazard, Cold Reg Sci Technol, 120, 251–262, https://doi.org/10.1016/j.coldregions.2015.04.010, 2015.

Viallon-Galinier, L., Hagenmuller, P., and Lafaysse, M.: Forcing and evaluating detailed snow cover models with stratigraphy observations, Cold Reg Sci Technol, 180, 103 163, https://doi.org/10.1016/j.coldregions.2020.103163, 2020.

Viallon-Galinier, L., Hagenmuller, P., Reuter, B., and Eckert, N.: Modelling snowpack stability from simulated snow stratigraphy: Summary and implementation examples, Cold Reg Sci Technol, 201, 103 596, https://doi.org/10.1016/j.coldregions.2022.103596, 2022.

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, Geosci Model Dev, 5, 773–791, https://doi.org/10.5194/gmd-5-773-2012, 2012.

Vionnet, V., Six, D., Auger, L., Dumont, M., Lafaysse, M., Quéno, L., Réveillet, M., Dombrowski-Etchevers, I., Thibert, E., and Vincent, C.: Sub-kilometer Precipitation Datasets for Snowpack and Glacier Modeling in Alpine Terrain, Front Earth Sci, 7, 182, https://doi.org/10.3389/feart.2019.00182, 2019.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Elsevier, https://doi.org/10.1016/c2017-0-03921-6, 2019.