https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

Comparing and validating modelled snow stability metrics

Jürg Schweizer[1]*, Benjamin Reuter[2,3], Christoph Mitterer[4], Fabiano Monti[5], Stephanie Mayer[1]

[1] WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland
[2] Météo-France, Direction des opérations pour la prévision, Toulouse, France
[3] Univ. Grenoble Alpes, Univ. de Toulouse, Météo-France, CNRS, CNRM, Centre d'Etudes de la Neige, Grenoble, France
[4] Avalanche Warning Service Tyrol, Innsbruck, Austria
[5] ALPSolut, Livigno, Italy

Correspondence to: Jürg Schweizer (schweizer@slf.ch)

Abstract. Recent developments in snow stability modelling demonstrated the great potential of numerical snow cover modelling for avalanche forecasting. These recently developed metrics provided promising results, possibly even superseding traditional stability indices. To further validate these results, we compared the temporal evolution of various stability metrics to a unique dataset of measurements, including snow stratigraphy, snow microstructure, and stability obtained at the high Alpine study site Steintälli above Davos (Eastern Swiss Alps) during the winter 2015/16. There, we measured the shear strength of a prominent weak layer of depth hoar crystals using the shear frame and assessed the propagation propensity of this weak layer with propagation saw tests. Concurrently, we characterized snow microstructure with the snow micro-penetrometer (SMP). At the study site, an automated weather station is located, providing the data to run the numerical snow cover model SNOWPACK so that modelled stability metrics can be derived. Field measurements showed that weak layer strength and toughness were initially low but increased over time, in parallel with the increasing slab load; all three parameters were correlated. While field observations focused on the most prominent weak layer, modelled stability metrics were evaluated for the persistent weak layers identified by the respective modelling approaches. These generally corresponded well to the layers visually identified in simulated snow stratigraphy. Data-driven and process-based stability metrics exhibited similar temporal evolution and comparable skill in relation to local avalanche activity, achieving overall accuracies of 60–70%. This site-specific validation demonstrates that both types of numerical stability metrics provide valuable support for avalanche forecasting, although differences in their behaviour highlight the need for metric-specific interpretation. Hence, a multi-model approach may be advantageous for operational forecasting.

1 Introduction

Snow stability is a local property of the sloping, layered snowpack describing the likelihood of failure initiation and crack propagation in view of dry-snow slab avalanche release. The prerequisite for a non-negligible avalanche release probability is a snow stratigraphy that includes at least one weak layer overlain by sufficiently cohesive slab layers, provided this layering exists at the scale of a slope. Slab and weak layer properties are equally important for the processes of failure initiation and crack propagation (Reuter and Schweizer, 2018). In avalanche forecasting, the regional-scale frequency distribution of snow

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

stability, together with expected avalanche sizes, determines the danger level used to warn the public about the severity of the avalanche situation (Müller et al., 2025; Techel et al., 2020; Techel et al., 2026). Therefore, estimating, measuring, and modelling snow stability is crucial for public forecasting as well as risk assessment by backcountry recreationists.

Given the relevance of snow stratigraphy for stability evaluation, snow layer properties are either measured in the field or simulated by numerical snow cover models such as Crocus or SNOWPACK (Morin et al., 2020). Snow cover models generally provide a realistic representation of snow stratigraphy and frequently reproduce avalanche-relevant weak layers (e.g., Bellaire and Jamieson, 2013; Calonne et al., 2020; Herla et al., 2024; Wever et al., 2015). As the snowpack is a complex sequence of snow layers of different properties, its interpretation with respect to snow instability is not straightforward.

In the field, specific measurements that include pulling shear frames, performing stability tests, or crack propagation tests provide useful information (e.g., Birkeland et al., 2023; Schweizer and Jamieson, 2010). From measurements of shear strength and slab density, stability indices such as the skier stability index SK38 can be derived, which proved to be a good indicator of failure initiation in view of skier-triggered avalanches (Jamieson, 1995). The critical cut length (r_c) measured with the propagation saw test is a proxy for crack propagation propensity in the tested weak layer and reflects the slab–weak layer system response. However, the quantitative relationship between r_c and snow instability remains poorly constrained, although $r_c < 30$ cm has been proposed as a critical threshold (Reuter and Schweizer, 2018). As an alternative to tests, the snow micro-penetrometer (SMP; Schneebeli and Johnson, 1998) provides information on penetration resistance and snow microstructure that allows for the derivation of corresponding metrics for initiation and propagation (Reuter et al., 2015).

For simulated stratigraphy, the two metrics, the skier stability index (sk38) and the critical crack length (r_c), can also easily be determined (Gaume et al., 2017; Lehning et al., 2004; Schweizer et al., 2006; Viallon-Galinier et al., 2022). While these metrics can be calculated for each layer in simulated snow stratigraphy, it is not trivial to identify the potentially most critical weak layer. The layer where the value is minimal is the obvious candidate. However, this is often not the case. For example, low values of the skier stability index typically occur in weak layers just below the ski's penetration depth, where the overlying slab layer is still too thin to promote avalanche formation. Hence, in analogy to the idea of distinct layer properties (McCammon and Schweizer, 2002; Schweizer and Jamieson, 2007), some attempts aimed at finding the potentially critical weaknesses in simulated profiles based on structural differences such as grain size and hardness across layer boundaries (e.g., Bellaire et al., 2006). Monti et al. (2012) adjusted the threshold sum approach (TSA) to simulated profiles. Furthermore, Monti and Schweizer (2013) suggested considering relative rather than absolute thresholds for the structural instability parameters, creating the RTA (relative threshold sum approach). Combining the RTA (to find potential weaknesses) with the sk38 calculated for the depth of these potential weaknesses provides a measure of instability. Based on a small dataset of observed stability, they preliminarily concluded that the RTA-sk38 combination was useful to discriminate between poor/fair and good stability. However, the approach has not been validated so far.

For the originally suggested formulation of the critical cut length (Gaume et al., 2017), a validation study pointed toward some deficiencies in the parameterization, since weak layer thickness was related to the setup of the discrete element model and not

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

suited for numerically simulated stratigraphy. Therefore, Richter et al. (2019) suggested an alternative parameterization including weak layer grain size.

With these developments, two metrics describing failure initiation and crack propagation became available, reflecting the improved understanding of dry-snow avalanche release. These metrics allowed for the interpretation of snow stratigraphy in terms of stability and raised the prospect that high-resolution numerical prediction of snow instability, both in space and time, could become feasible. One of the first attempts to follow this process-based avenue was reported by Reuter and Bellaire (2018), who showed for a particular weak layer that both modelled instability metrics followed the temporal evolution of avalanche activity, which was not the case in a similar validation study reported by Schweizer et al. (2016). While these studies followed a known weak layer identified in the field, weak layers should ideally be identified automatically, tracked throughout the season, and evaluated in terms of how they contribute to the avalanche situation. To this end, Reuter et al. (2022) provided a process-based approach to detect, track, and assess instability as well as the related avalanche problem types for simulated weak layers; their approach is based on, among other metrics, the skier stability index and the critical crack length. Instead of considering these two independent metrics, Gaume and Reuter (2017) suggested a single stability criterion combining the two processes of failure initiation and crack propagation. For a given weak layer, they calculated the ratio S_p between the critical crack length and the size of the area where the skier-induced stress exceeds the shear strength of the weak layer, which in 2D was termed skier crack length l_{sk}. The onset of crack propagation occurs if the skier crack length exceeds the critical crack length ($l_{sk} > r_c$).

Concurrently, numerical avalanche prediction advanced by employing machine learning (ML) methods to link weather, snow, and snowpack variables to different types of target variables or labels. Those included, among others, the avalanche danger level (Pérez-Guillén et al., 2022), avalanche activity (Hendrick et al., 2023; Viallon-Galinier et al., 2023), and snow instability ($P_{unstable}$; Mayer et al., 2022). Although variable importance analyses identified influential predictors, their physical relevance to snow stability was not always apparent, limiting the interpretability of the resulting models. Moreover, for standard metrics of instability, such as sk38 or r_c, the variable importance varied between the studies. It was reported to be high by Viallon-Galinier et al. (2023) for both metrics. On the other hand, the importance of sk38 was low in the random forest models developed by Pérez-Guillén et al. (2022) and Mayer et al. (2022), but rather high for r_c. The reason for the discrepancy seems unclear but may be related to the input data, model setup, and how the instability metrics were implemented in the respective numerical snow cover model. In contrast, Herla et al. (2024) found similar performance results for the process-based (sk38, r_c) and the ML-based approach ($P_{unstable}$) for simulating layer instability in a large-scale validation study. Preliminary results by Schweizer et al. (2023) also suggest a similar degree of performance.

All of the above-mentioned recent ML modelling approaches included some form of validation, with results typically reported as summary performance metrics. As a result, direct comparisons among the various stability metrics are difficult, particularly given the absence of a common validation dataset. While average performance metrics may be conclusive, further improvements often become obvious only when looking at specific situations. Analysing spatial patterns or the evolution over

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

time may provide additional insights. It is presently unclear how the different approaches for assessing stability compare, given contrasting results, for instance, for the skier stability index.

Therefore, we aim to compare different stability metrics, classical (process-based) and data-driven ML-based metrics. We analyse how these various approaches perform during the winter season 2015/16 at our study site Steintälli, where we collected a comprehensive set of field measurements, including shear strength, to which we can compare the model results.

2 Methods

2.1 Field measurements

We performed manual snowpack observations on 9 days from early January to mid-March 2016 at the study site Steintälli (2442 m a.s.l.) above Davos (46.8075° N, 9.7880° E; eastern Swiss Alps; Tab. 1), where an automated weather station (AWS) is located. On each observation day, we recorded a detailed snow profile according to Fierz et al. (2009), performed stability tests (CT: compression test; ECT: extended column test) and propagation saw tests (PST; 3-5 per day) (Birkeland et al., 2023). The principal weak layer we monitored was a layer of depth hoar (2-4 mm in size). It had formed at the snow surface during December 2015, got buried on 31 December 2015, and was a persistent weak layer (PWL) of concern throughout January, February, and March 2016. By mid-March, the PWL was deeply buried (1.2 m).

PST column length was initially 1.2 m when the slab thickness was 0.25 m, and was increased to 2.4 m in mid-March when the weak layer we tested was deeply buried. PSTs were filmed, and relevant mechanical properties were derived from the video images (van Herwijnen et al., 2016).

Table 1: Key field data of monitoring PWL 1

Date	Days since burial	Slab thickness (m)	Slab density (kg m^{-3})	Shear strength ± SE (Pa)	Stability test results: Compression test	Stability test results: Extended column test
6 Jan 2016	6	0.23	159	730±55	CT17 SC	ECTN 18
15 Jan 2016	15	0.56	130	890±65	n/a	ECTP 15
22 Jan 2016	22	0.75	237	1170±75	CT11 SC	ECTP 11
29 Jan 2016	29	0.62	281	1300±75	CT11 SC	ECTP 11
4 Feb 2016	35	0.92	259	1320±55	CT13 SC	ECTP 21
9 Feb 2016	40	0.88	287	1500±50	CT21 SC	ECTP 21
18 Feb 2016	49	0.83	311	1650±90	CT15 SC	ECTP 27
26 Feb 2016	57	1.15	311	1560±65	CT14 SC	ECTP 21
10 Mar 2016	70	1.31	308	1770±110	CT23 SC	ECTP 29

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

In addition, we pulled shear frames (A = 250 cm^2; 8-16 per day) to measure the shear strength τ_s of the depth hoar layer. We did not adjust shear strength for normal load as suggested for persistent weak layers by Jamieson and Johnston (1998). Hence, $\tau_s = 0.65\,(F/A)$ with F the pulling force, A the shear frame area, and 0.65 the size adjustment factor for a frame size of A = 250 cm^2 (Föhn, 1987).

The manual snow stratigraphy observations were completed with snow micro-penetrometer measurements (SMP; 7-24 per day; Schneebeli and Johnson, 1998). On the last measurement day (10 March 2016), the SMP data quality, however, was insufficient, so the data could not be analysed. SMP data analysis was performed as described in detail in Schweizer et al. (2016). From the SMP signal, we derived the following mechanical properties: weak layer thickness h_{WL}, density ρ_{WL}, modulus E_{WL}, strength σ_c, and fracture energy w_{fSMP} (Reuter et al., 2019). The PSTs were filmed, and with PTV analysis, the following properties were extracted from the videos: critical cut length r_c, slab effective modulus E*, and specific fracture energy of the weak layer w_{fPST} (van Herwijnen et al., 2016). An effective value of toughness K_c^* was approximated from weak layer fracture energy w_{fPST}, and modulus E_{WL}:

$$K_c^* = \sqrt{\frac{E_{WL}\, w_{fPST}}{1-\nu^2}} \tag{1}$$

with $\nu = 0.25$.

For this study, we opted for these bulk mixed-mode fracture properties and did not explicitly consider the mode interactions (Adam et al., 2024).

Apart from the persistent weak layer monitored in detail, four additional persistent weak layers were identified in the simulated stratigraphy solely through visual inspection. These PWLs were easily recognizable as continuous layers of faceted crystals or depth hoar (Fig.1; Tab. 2). As SNOWPACK merged PWL 2 with the layer below on 10 February, we no longer monitored this persistent weak layer.

Table 2: Persistent weak layers present at the Steintälli field site in Winter 2015/16. PWL 1 is the primary weak layer we observed in detail. PWLs 2-5 are characterized as simulated by SNOWPACK; for PWL 1, observed (2-4 mm) and simulated (1.9 mm) grain sizes are given. Slab depth *H* (measured vertically from the snow surface) is given for 10 March 2016, when the simulated snow height was 1.54 m.

PWL number	Burial date	Grain shape and size (mm)	H (m)
1	31 Dec	DH, 2-4 (1.9)	1.31
2	7 Jan	DH, 1.6	n/a
3	29 Jan	FC, 1.4	0.89
4	10 Feb	DH, 1.6	0.58
5	1 Mar	FC, 1.3	0.38

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

2.2 Snow cover simulation

At the AWS Steintälli, measurements include incoming and outgoing short- and longwave radiation, in addition to snow depth, air and snow surface temperature, relative humidity, wind speed, and direction. Therefore, we ran the numerical snow cover model SNOWPACK (Version 3.0.0) using all these measurements. The resulting snow cover evolution is shown in Figure 1.

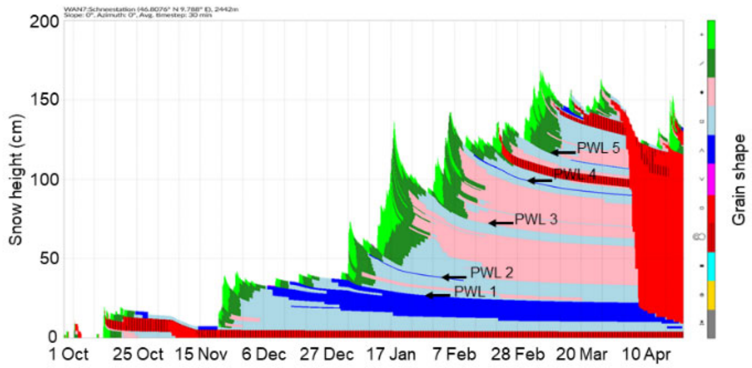


Figure 1: Snow cover evolution at AWS Steintälli during winter 2015/16 as simulated with SNOWPACK. Colours refer to grain shape. Arrows point to the five manually identified persistent weak layers (PWL). PWL 1 is the layer we specifically tracked in the field from January to March 2016.

2.3 Stability modelling

Based on the manual measurements, we calculated the skier stability index SK38 based on shear strength, slab density, and the additional skier stress $\Delta\tau = 155/(H\text{-PS})$, accounting for skier penetration PS but neglecting slab layering (Jamieson and Johnston, 1998; Monti et al., 2016).

From the SNOWPACK simulations, we extracted the RTA, the skier stability index sk38, the critical crack length r_c, and the stability metric $P_{unstable}$ for each layer once a day at 11:00 LT (local time). As for the skier stability index SK38 calculated from the field measurements, we did not consider slab layering in the modelled sk38. For the critical cut length, we considered the parameterization suggested by Richter et al. (2019). For the probability of instability, $P_{unstable}$, we extracted the maximum value and the depth at which it occurred. In addition, potential weak layers were identified using an algorithm described by Reuter et al. (2022), in which the layers are tracked from their burial date onward. From the potential weak layers identified by the algorithm, we only considered those associated with persistent avalanche problems (pAp) and used the weak layer depth of these problems. We then analysed sk38 and r_c for the depth of the most relevant critical weak layer. The RTA approach can identify several weak layers as critical, i.e., RTA > 0.8. We selected the one with the lowest value for sk38 as the most critical

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

PWL. With the pAp approach, occasionally, it occurs that the slab above the PWL is still rather shallow, so that sk38 is not defined (i.e., $PS > H$). In that case, we also selected the PWL with the lowest value for sk38 as the most critical instead. To rate stability (stable vs. unstable), we considered the following individual thresholds: sk38 < 1, $r_c < 0.3$ m, and $P_{unstable} > 0.77$ (Mayer et al., 2022; Monti and Schweizer, 2013; Reuter et al., 2022).

The 3-day sum of new snow height (HN3d) is a key indicator in avalanche forecasting and has been shown to be a highly significant feature in many data-driven forecasting models (e.g., Pérez-Guillén et al., 2022; Schweizer and Föhn, 1996; Viallon-Galinier et al., 2023). Therefore, we considered HN3d as a benchmark model. We used the daily measurements from the Weissfluhjoch study plot (2.9 km to the northeast of Steintälli) and classified days with HN3d ≥ 23 cm as unstable (Mayer et al., 2023).

For reference, we compared the stability metrics to the avalanche activity index for the region of Davos. The avalanche activity index (AAI) is a weighted sum of all avalanches observed, including human-triggered avalanches and natural avalanches (dry-snow and wet-snow avalanches) (Schweizer et al., 2003). We compared the AAI (only considering the dry-snow avalanches: AAI_{dry}) for days when an instability metric indicated unstable conditions to the AAI of the days classified as stable, using the non-parametric Mann-Whitney *U*-Test. Furthermore, we considered a value of AAI ≥ 1 as indicating unstable conditions (resulting in 26 unstable days), whereas values below 1 were considered indicative of stable conditions (48 days). This threshold corresponds to the median value for the danger level 3–Considerable as reported by Schweizer et al. (2020). Based on this classification, we rated the prediction performance of the various metrics, expressed as the accuracy or proportion of correct (PC), probability of detection (POD), and the Pearson skill score (PSS) (Wilks, 2019).

Finally, we also referred to the avalanche danger level, which was forecast in the public bulletin for the region of Davos. The forecasts were corrected for obvious misclassifications. For instance, on 9 February 2016, the avalanche activity was very high (AAI = 61.9), so that the forecast danger level 3–Considerable was, in hindsight, corrected to 4–High (Schweizer et al., 2020). This type of verification was done for a total of 5 days. Overall, significant avalanche danger prevailed during most of the time: on 51 out of 74 days the forecast danger level was 3–Considerable or higher. The reason for the general heightened avalanche danger was the poor snowpack stability. At the end of December 2015, the shallow snowpack consisted almost entirely of faceted crystals and depth hoar (Fig, 1). Consequently, on all except one day (1 March), the avalanche problem "Persistent weak layer" (EAWS, 2026) was mentioned in the forecast. During the following two and a half months, this poor snowpack was then repeatedly loaded with new snow (in total about 400 mm SWE) or wind-drifted snow. Each major loading event resulted in significant avalanche activity of mainly dry-snow natural avalanches.

3 Results

3.1 Field measurements

The temporal evolution of the manual measurements and some corresponding modelled SNOWPACK parameters are shown in Figure 2. In general, weak layer density, shear strength, and resulting skier stability index increased with time. The modelled

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

parameters showed the same increasing trend (Fig. 2b). Modelled strength, which is parameterized from density, increased more prominently than the manually measured strength. Compared to the shear frame measurements, the modelled strength was lower initially and higher at the end of the observation period. Accordingly, the modelled skier stability index was also lower initially and higher after mid-February than the stability index calculated from the measurements.

The critical cut length increased, in general; the values modelled, however, increased too strongly compared to the measurements, which varied between 20 and 35 cm. The stability tests (CT/ECT) confirmed the increasing trend of stability; failure initiation became increasingly harder, while crack propagation remained possible (Fig. 2a). On the last field day, 70 days after burial, the critical cut length was still only 32 cm ± 9.8 cm, and all tests showed cracks propagating to the very end. On the other hand, weak layer specific fracture energy w_{fPST} and toughness K_c^* had increased to approximately twice their initial values since the first measurement 6 days after burial (Fig. 2c). The increase was correlated with load: shear strength (r_{P} =0.94, $p < 0.001$), specific fracture energy ($r_{\mathrm{P}} = 0.87$, $p = 0.002$), and fracture toughness ($r_{\mathrm{P}} = 0.92$, $p = 0.001$).

Overall, the agreement between measurements and model results was reasonable. For the load, for instance, the mean differences between measurements and model results were 0.16 kPa (MAE) and 10% (MAPE) (Tab. 3). Similarly, for weak layer density, the differences were 17.5 kg m^{-3} and 10%. For the mechanical and therewith derived instability metrics, the differences were larger, namely, for weak layer shear strength, the mean absolute percentage error was 15%, for the skier stability index 21%, and for the critical cut length 45%. Still, the general trends were similar, so that in the following, we compare the various modelled instability metrics.

Table 3: Comparison between measured and modelled slab and weak layer (WL) properties: Mean absolute error (MAE), mean absolute percentage error (MAPE) ($N = 9$)

Property	MAE	MAPE
Load on WL	0.16 kPa	0.10
WL density	17.5 kg m^{-3}	0.10
WL shear strength	0.19 kPa	0.15
Skier stability index	0.16	0.21
Critical cut length	0.13 m	0.45

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

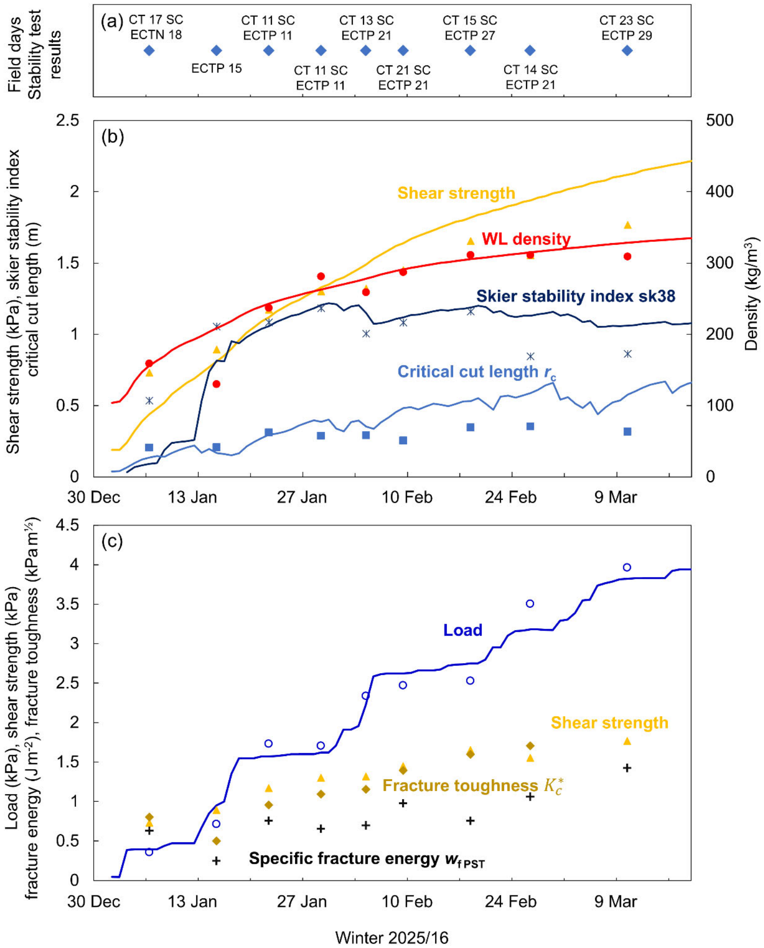


Figure 2: Field measurements and model results for the persistent weak layer buried on 31 December 2015 (PWL 1). (a) Results of stability tests (CT, ECT) on the 9 measurement days. (b) Comparison of measurements and model results for: weak layer density, shear strength, skier stability index, and critical cut length. Continuous lines show SNOWPACK parameters, symbols (of the same color) show the field measurements. (c) Load (measured and modelled), weak layer strength (same as in (b)), weak layer specific fracture energy $w_{f\,PST}$, and effective fracture toughness K_c^*.

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

3.2 Stability modelling

3.2.1 Weak layer identification

Assessing snow instability requires determining the most relevant critical weak layer, i.e., we only considered one weak layer per profile (or date). While performing stability tests in the field, this weak layer shows up automatically. In snow cover simulations, in contrast, weak layer detection is the first step in stability assessment. Figure 3 shows weak layer identification by $P_{unstable}$, RTA, and pAp. For $P_{unstable}$, the most critical weak layer was identified by the maximum value of the metric. For RTA, we considered all layers with RTA > 0.8 and from these selected the layer with the lowest value of sk38 (which we termed the most relevant critical weak layer). Similarly, for pAp, we also considered the persistent weak layer where sk38 was minimal.

With the $P_{unstable}$ and RTA methods, the automatically identified weak layer depth coincided with the depth of one of the five visually identified PWLs in about half of the days, i.e., the difference in the weak layer position was < 1 cm (Tab. 4). Until 1 February, $P_{unstable}$ suggested the depth of the most relevant critical weak layer to be a few centimetres below PWL 1, then it jumped between PWL 1 and PWL 3, and subsequently PWL 4. As of 20 February, it consistently stayed at PWL 4, and finally at PWL 5 (after 7 March). On the other hand, RTA initially suggested the depth of the most critical weak layer to be below PWL 1, i.e., above the crust that had formed very early in the winter (Fig. 1). By 8 January, the depth coincided with PWL 1 for 6 days, and then until the end of January, fluctuated between PWL 1 and 2. Subsequently, until mid-February, the suggested depth was either below or above PWL 3, and as of 21 February, it was near PWL 4, until it finally (by 5 March) jumped to PWL 5.

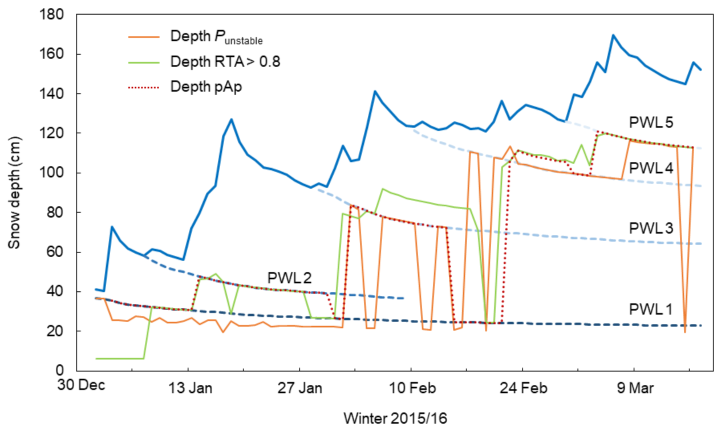


Figure 3: Snow depth (in blue) and depth of the five manually identified persistent weak layers (PWL 1-5) in winter 2015/16 at Steintälli, and depth of the most relevant critical weak layer as identified by the maximum value of $P_{unstable}$, RTA > 0.8 and min(sk38), and the weak layer tracking algorithm pAp and min(sk38).

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

Table 4: Agreement between the most relevant critical weak layer identified with one of the three algorithms and the manually identified PWLs

Algorithm	Minimum absolute difference (mean)	Number of exact matches (out of 74)	Number of matches < 1 cm (out of 74)	Number of jumps between observed PWLs
P_{unstable}	2.9 cm	26	35	12
RTA	5.0 cm	39	40	9
pAp	0.53 cm	67	67	6

Finally, with the weak layer tracking approach to identify persistent avalanche problems pAp, the suggested weak layer depth almost always (91 %) coincided with one of the five visually identified PWL's, and there were fewer jumps between layers. The depth initially followed PWL 1, and moved up to PWL 2 when a suitable slab was present. By 31 January, the next weak layer, PWL 3, had been identified, but only two days later, the overlying layers had formed a slab. During these two days, the most relevant critical weak layer became PWL 1 again, as sk38 at the depth of PWL 1 was lower than that of PWL 2. Subsequently, for the next weeks, the most relevant critical weak layer became PWL 3, and then, the above-described pattern repeated until a slab had formed above PWL 4. With pAp, there was not only a match on most days between the weak layers, but the match was also mostly exact, i.e., the difference was also small on average (0.53 cm; Tab. 4). The difference was largest for RTA (5.0 cm) followed by P_{unstable} (2.9 cm).

All three methods indicated an additional PWL between layers 4 and 5, which was buried on 20 February. It was, however, not very prominent and consisted of small facets below a crust. Still, stability tests in the field identified the weakness on 26 February, but not any more from 10 March 2015 onwards.

3.2.2 Stability metrics

We considered the numerically modelled stability metrics for the most relevant critical weak layers indicated by the different models, as shown in Figure 3. For P_{unstable}, the stability is simply the value of P_{unstable} at that particular depth; for RTA > 0.8 and pAp, it is the minimum value of sk38 as provided by SNOWPACK. For pAp, the skier stability index and the critical cut length are shown as suggested by Reuter et al. (2022).

Figure 4 shows the temporal evolution of the various stability metrics, indicates the periods when they were in the unstable range ($P_{\text{unstable}} > 0.77$, sk38 < 1, $r_c < 0.3$), and, for comparison, provides the avalanche activity index and the forecast regional avalanche danger level (partly corrected in hindsight).

As of early January, the value of P_{unstable} was high (Fig. 4a), indicating low stability, for a weak layer below PWL 1. This period of instability lasted until the end of January, when on 27 January, P_{unstable} dropped below the threshold of 0.77. With

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

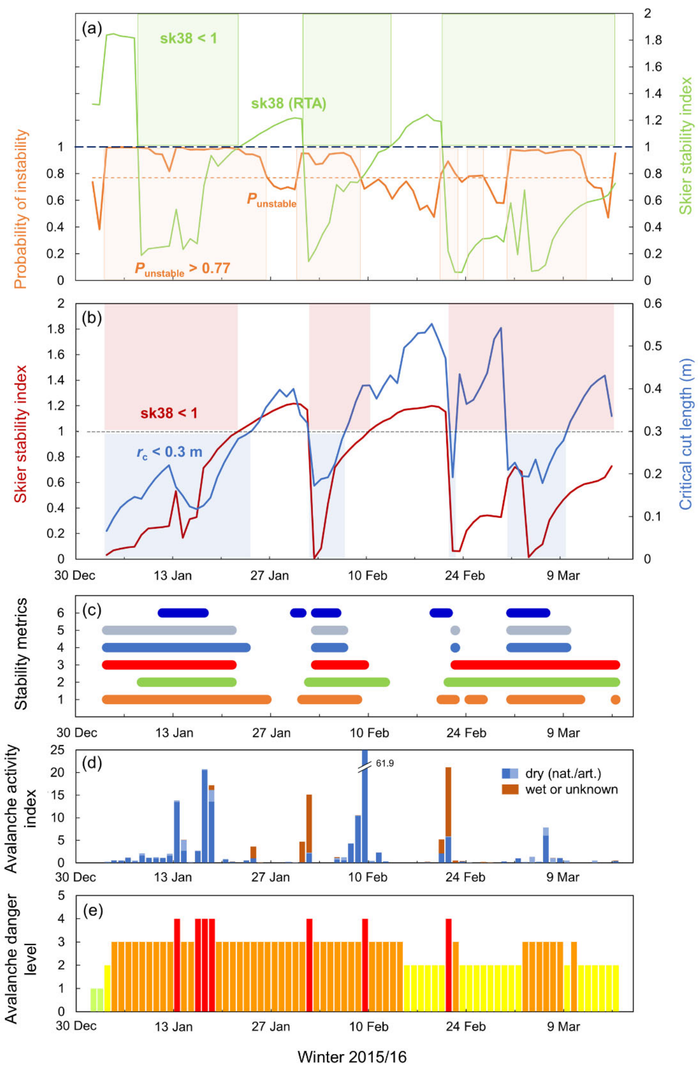
the snowfall at the end of January, P_{unstable} increased again and indicated instability on 31 January, for a layer still below PWL 1, and then on 2 February jumped to PWL 3, staying in the unstable range for another week. Then a period of higher stability (9-19 February) followed, before the value of P_{unstable} was high again (low stability) for most of the remainder of the observation period.

Overall, P_{unstable} indicated that most of the time conditions were rather unstable (on 51 out of 74 days; Tab. 5). While the metric missed the day with the highest avalanche activity (9 February; Fig. 4d), the median was $\text{AAI}_{\text{dry}} = 0$ for the stable days, and for the days rated unstable, the median was $\text{AAI}_{\text{dry}} = 0.53$. Still, overall, the correlation with AAI_{dry} was poor ($r_P = 0.01$). However, considering the two classes stable vs. unstable, the non-parametric Mann-Whitney U-test indicated that the differences between the AAI distributions were statistically significant ($p = 0.002$). The forecast performance (proportion correct), in comparison with the AAI_{dry} (threshold ≥ 1.0), was 61% (PSS = 0.36).

Identifying the most relevant weak layer with the RTA method and considering sk38, the stability was initially high (Fig. 4a), since the PWL 1 was not identified with the RTA method before 8 January 2016. For the next two weeks, it first stayed low and then steadily increased to sk38 = 1 on 22 January 2016; at the same time, P_{unstable} started to decrease. The index further increased until 31 January and then switched to low values on 1 February, when it jumped from PWL 1 to near PWL 3. Subsequently, it increased again and stayed below 1 until 12 February. As of 13 February, the index indicated rather stable conditions for 8 days. Then, PWL 4 and later PWL 5 became the relevant critical weak layers, and sk38 stayed low (< 1) until the end of the observation period in mid-March.

Most of the time, the sk38 for the layer identified with RTA > 0.8 and min(sk38), indicated rather unstable conditions (on 51 out of 74 days; Tab. 5). For the cases with sk38 ≥ 1, the median avalanche activity index was AAI = 0, and for the days rated unstable (sk38 < 1), the median was AAI = 0.5. Again, the correlation with the AAI was very poor ($r_P = -0.03$). However, considering the two classes stable vs. unstable, the non-parametric Mann-Whitney U-test indicated that the distributions of AAI in the two groups were different ($p = 0.008$). The forecast performance (proportion correct), in comparison with the AAI (threshold ≥1.0), was 58% (PSS = 0.30).

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026



300 **Figure 4: Modelled stability metrics as provided by (a) the probability of instability $P_{unstable}$ and the skier stability index sk38 as identified with RTA, and (b) skier stability index and critical cut length as identified with pAp layer tracking. The light shaded areas indicate the periods of instability; for pAp, those can be further restricted to the times when both metrics indicated instability. (c) Summary of times of instability according to the stability metrics (1) $P_{unstable}$, (2) RTA-sk38, (3) pAp-sk38, (4) pAp-r_c, and (5) pAp-sk38∩r_c, and for comparison (6) HN3d. (d) Avalanche activity index (AAI) as observed in the region of Davos; the index is a weighted sum including all types of avalanches recorded (Schweizer et al., 2003). (e) Avalanche danger level for the region of Davos.**

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

Figure 4b shows the evaluation for the third approach, i.e., weak layer detection with pAp, and for stability, considering the skier stability index sk38 and the critical cut length r_c. The sk38 and r_c were low, indicating instability, from the beginning in early January. They reached the thresholds indicating stable conditions on 22 and 24 January, respectively. On 2 February, PWL 3 became the most relevant weak layer, and sk38 and r_c both dropped to very low values, suggesting a switch from stable to unstable conditions. The metrics then increased rather sharply and after 4 days (r_c) and 8 days (sk38) indicated stable conditions again, until 21 February 2016. On that day, avalanche activity was high, but only the day after, on 22 February, both metrics indicated instability. However, the value of critical cut length was low for only one day and then increased again to values of 40 cm or larger. While the critical cut length switched again from stable to unstable by the end of February, the skier stability index remained low from 22 February for more than 3 weeks until the end of the observation period in mid-March.

Concerning prediction performance (Tab. 5), we first assess the two metrics (sk38 and r_c) separately and then jointly. The metrics suggested unstable conditions in 69% and 49% of the days, respectively (Tab. 5). Combining the metrics, assuming that for avalanche release both have to be in the unstable range, the number of days interpreted as unstable reduced to 34 out of 74 (46%). For the days with sk38 ≥ 1, the median avalanche activity was 0.1; for the days rated unstable (sk38 < 1), the median was 0.51; the respective medians of AAI for r_c and the combination of the two metrics were 0.08 and 0.94, and 0.08 and 1.0. The Pearson correlation coefficients for sk38 and r_c were 0.08 and -0.05, respectively. Comparing the distributions of AAI for days rated stable and unstable using the *U*-test yielded for sk38 a level of significance of $p = 0.034$, for r_c a value of $p = 0.002$, and for the combined metrics $p = 0.001$. The percentage of correct forecasts were 53% and 65%, and 68% for the combined metrics. However, the measure for discrimination power, the Pearson skill score (PSS), was rather low for sk38, and slightly higher for r_c, and the combined metrics, 0.18, 0.32, and 0.36, respectively.

Table 5: Characteristics and prediction performance for the various stability metrics and the danger level forecast

Approach	# days stable vs. unstable	Median AAI (dry snow) for stable/unstable days	U-test: *p*-value	Pearson correlation coefficient r_p	Performance: skill scores PC	POD	PSS
$P_{unstable}$	23 vs. 51	0.0 / 0.53	0.002	0.01	0.61	0.92	0.36
RTA>0.8 min sk38	23 vs. 51	0.0 / 0.5	0.008	-0.03	0.58	0.88	0.30
pAp min sk38	23 vs. 51	0.1 / 0.51	0.034	0.08	0.53	0.81	0.18
pAp r_c	38 vs. 36	0.08 / 0.94	0.002	-0.05	0.65	0.69	0.32
pAp sk38 ∩ r_c	40 vs. 34	0.08 / 1.0	0.001	n/a	0.68	0.69	0.36
pAp $P_{unstable}$	30 vs. 44	0.03 / 0.7	0.002	0.04	0.62	0.81	0.33
HN3d	52 vs. 22	0.18 /1.2	<0.001	0.13	0.73	0.54	0.37
Danger level forecast	23 vs. 51	0 / 0.87	0.001	n/a	0.61	0.92	0.36

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

As shown above, the weak layer detection worked best with the pAp method. Therefore, we explored the combination of weak layer detection by pAp with P_{unstable} for the stability assessment. This approach (pAp-P_{unstable}) predicted some less unstable days (59%; 44 out of 74; Tab. 5). For the days rated as stable, the median AAI_{dry} was 0.03, and for the unstable days 0.7. The correlation with the AAI_{dry} was still very poor (r_{P} = 0.04). The distributions of the AAI for the days rated stable or unstable were judged to be significantly different (p = 0.002) using the U-test. The prediction performance was similar to that of the original algorithm P_{unstable} (PC = 0.62; PSS = 0.33).

The benchmark model, HN3d, performed surprisingly well and showed similar or better skill than the stability metrics (Table 5). However, the POD was low, and the good performance resulted from the high value of PON = 0.83.

For each metric, Figure 4c summarizes the days when instability was predicted and shows fair agreement between the approaches. There were three periods when all metrics suggested unstable conditions: 8-21 January, 2-6 February, and 1-8 March 2016, and a single day: 22 February 2016. On the other hand, there were two short periods at the end of January (27-30 Jan) and in mid-February (13-19 Feb) when all metrics indicated stable conditions; in fact, almost no avalanches were reported during these periods (Fig. 4d). For comparison, also unstable periods based on HN3d, the benchmark model, are shown in Figure 4c. Obviously, these periods were shorter, but in general, they coincided with the unstable periods indicated by the stability metrics. On two occasions, HN3d suggested instability earlier than the stability metrics: on 1 and 21 February.

Finally, for comparison, we considered the regional danger level (Fig. 4e), which indicated a significant avalanche situation (danger level 3–Considerable or 4–High) for a long period (4 Jan – 14 Feb), i.e., for 42 days. This is not unusual given the existence of persistent weak layers in the snowpack. Compared to the avalanche activity index, the performance values of the public forecast were in a similar range as those of the various stability metrics (Tab. 5). As we only considered two classes of stability (stable vs. unstable), the hindsight corrections did not change the performance of the forecast.

Overall, P_{unstable} and the combination of sk38 and r_{c} (pAp sk38 ∩ r_{c}) performed slightly better than the other metrics, in terms of PSS. While P_{unstable} detected most of the unstable situations (POD = 0.92), the combination of the two stability metrics showed the most balanced performance in terms of correctly indicating stable as well as unstable situations (POD = 0.69; PON = 0.67). The skier stability RTA-sk38 performed similarly well as P_{unstable} in detecting unstable situations (POD = 0.88).

4 Discussion

We followed the temporal evolution of snow stratigraphy and stability at the Steintälli field site during winter 2015/16, and analysed numerically modelled stability metrics during the observation period (January to March). The period was characterized by a prominent persistent weak layer problem, which resulted in repeated days with significant natural dry-snow avalanche activity and heightened regional avalanche danger (levels 3 and 4). We will first discuss how measured and modelled mechanical properties and stability metrics compare. Second, we put the stability metrics into perspective with the overall avalanche situation.

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

4.1 Field measurements

For the field measurements, we focused on the persistent weak layer of faceted crystals and depth hoar that was buried on 31 December 2015 (PWL 1). Weak layer shear strength steadily increased from about 0.7 to 1.8 kPa. These values fit well within the range of data presented by Jamieson and Johnston (2001). The average increase in strength (about 20 Pa/day) was rather low. For instance, Zeidler and Jamieson (2006) reported an average rate of change of 60 Pa/day for the shear strength of layers of faceted crystals. Their higher value of the rate of change might be related to the typically thicker snowpack and lower temperature gradients found in the Columbia Mountains of western Canada. The increase in strength was highly correlated with the increase in load due to the repeated accumulations by snowfall, in line with previous studies (e.g., Chalmers and Jamieson, 2001; Schweizer et al., 2016).

We estimated fracture properties (weak layer specific fracture energy and effective fracture toughness) from analysing PST experiments and SMP measurements. They also increased with load, w_{fPST} from about 0.3 to 1.4 J m^{-2}, and K_c^* from 0.5 to 1.7 kPa m$^{1/2}$. The increase of w_{fPST} was in a similar range to that reported by Schweizer et al. (2016). The order of magnitude of the fracture properties agreed with previously reported values (e.g., Adam et al., 2024; Bergfeld et al., 2023; van Herwijnen et al., 2016). Moreover, weak layer specific fracture energy and effective fracture toughness showed a very similar temporal evolution, suggesting that similar microstructural changes may contribute to the observed increasing trends.

Overall, PWL 1 can be considered a weak layer typical of the regional snow climate (Schweizer et al., 2024) and well-suited to be monitored and used for a validation study.

The observations in the field generally agreed reasonably well with the corresponding simulated properties. However, as previously reported (e.g., Richter et al., 2019; Schweizer et al., 2016), there were also some deviations. For instance, modelled shear strength and critical cut length were overestimated from about the end of January onwards. The largest deviation was observed for the critical cut length. As modelled weak layer shear strength also increased more prominently than the observed strength, adjusting the modelled shear strength for normal load would further increase the difference. Overall, given the magnitude of the mean differences (MAPE: 10-20%; Tab. 3) in the mechanical properties and the observed temporal trends, we expect that the modelled stability parameters can be evaluated and may be useful. Nevertheless, the observed discrepancies for strength and critical cut length suggest that the corresponding parameterizations may need to be revisited at some point.

4.2 Stability modelling

4.2.1 Weak layer detection

In manual stability tests in the field, relevant weak layers are automatically identified by the tests. Without a stability test, finding the relevant weak layers is equally difficult in manually observed snow profiles (e.g., Schweizer et al., 2005) and in simulated stratigraphy. For instance, Bellaire et al. (2006) reported that the skier stability index performed poorly in terms of identifying potential weak layers in a simulated snow cover. Applying the RTA method, developed for that very purpose, Monti and Schweizer (2013) reported a detection score (POD) between 50 and 80% depending on the failure type. For P_{unstable},

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

the probability of detecting the manually picked weak layers in the training dataset with the local maxima of $P_{unstable}$ was 60%. In our case study, for the simulated snow stratigraphy in the winter of 2016, the various approaches performed fairly well. The weak layer detection algorithm pAp (Reuter et al., 2022) showed by far the best results (Tab. 4). This finding is not surprising since it is a prescriptive approach based on our understanding of dry snow instability (Herla et al., 2025). For pAp, the predicted and visually identified weak layers matched on 91% of the days. While our study only refers to one winter season, this finding is encouraging, since identifying the potentially critical weak layers in simulated snow stratigraphy has always been considered particularly challenging. Similarly positive results were recently reported by Herla et al. (2025), who also applied the approach by Reuter et al. (2022). However, we have only considered persistent weak layers, and other types of instabilities, e.g., in new snow, may be more difficult to detect.

4.2.1 Stability metrics

Given the various PWL's in winter 2015/16, the numerically modelled stability metrics suggested long periods of instability, also reflected in the forecast avalanche danger level (Fig. 4e). Overall, the trends were similar for the three approaches. In mid-January, early February, and early March, the periods of instability coincided. In between, all metrics indicated periods with somewhat higher stability. However, the timing of the transition from stable to unstable conditions often differed, typically by 1-2 days, with $P_{unstable}$ usually being the first to indicate a shift to unstable conditions. The other metrics lagged behind, often since the slab was not yet well developed, and, for instance, sk38 was not defined since penetration depth was larger than the slab depth. All modelled metrics indicated increasing stability when the highest avalanche activity was observed on 9 February 2016. Still, the skier stability indices, as identified with RTA and pAp, were slightly below 1 in the unstable range. However, this was a particular event (which was also not sufficiently anticipated in the public forecast) since the instability was due to loading by wind-blown snow following heavy snowfall, which had accumulated about 50-60 cm of new snow a few days earlier; our SNOWPACK model setup did not account for wind-drifted snow. The high value of the AAI was mainly due to natural dry-snow avalanches. On the other hand, the high avalanche activity on 1 February was well anticipated by $P_{unstable}$ and RTA-sk38. While pAp correctly predicted PWL 3 as the relevant weak layer on 1 February, the skier stability index was not defined (i.e., PS > H), so PWL 1 was selected as the critical weak layer, where sk38 was 1.2; by 2 February, PS had decreased (PS < H), meaning that PWL 3 became the critical weak layer, and sk38 (= 0.01) indicated instability.

Concerning stability, differences between the approaches $P_{unstable}$, RTA-sk38, and pAp-sk38 were not particularly prominent. On 48 out of 74 days (65%), the stability class (stable/unstable) agreed, and the trends were similar; for instance, when $P_{unstable}$ decreased, sk38 increased. In fact, $P_{unstable}$ was negatively correlated with the stability metrics pAp-sk38 (-0.38; $p = 0.001$) and pAp-r_c (-0.84; $p < 0.001$). The high correlation with the critical cut length was to be expected since r_c is one of the 6 features considered to estimate $P_{unstable}$ (Mayer et al., 2022). A quantitative comparison with the avalanche activity index showed that the metric $P_{unstable}$ and the combined metric pAp sk38 ∩ r_c performed best in terms of PSS (Tab. 5).

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

Considering the three days with the highest dry-snow avalanche activity (17 January, 18 January, and 9 February), on 17 and 18 January, all five metrics predicted unstable conditions (Fig. 4c,d), on 9 February, only RTA-sk38 and pAp-sk38 were in the unstable range. Hence, only RTA-sk38 and pAp-sk38 correctly predicted all three high activity days.

Interestingly, on 21 February, when the avalanche activity was also high (AAI = 21), but avalanche type was mostly recorded as wet, the dry-snow metrics P_{unstable}, RTA-sk38, and pAp-r_c all indicated unstable conditions. It snowed on 20-21 February with rising air temperature (rain at low elevations); the snowfall ended in the very early morning; air temperature continued to rise and was positive during the day. We do not have any information on the failure layers of the natural avalanches that were recorded as 'wet' or 'unknown', but it is clear that it briefly rained up to 2500 m a.s.l. (Fig. 1); however, apart from the top few centimetres, the snowpack remained dry, so the avalanches very likely failed in the dry snow. This would explain why the dry-snow metrics successfully predicted the high (wet-snow) avalanche activity. Similar conditions (snowfall with elevated air temperature) existed on 1 February (AAI = 15) when mostly natural avalanches with snow type 'unknown' were recorded. For this day, P_{unstable} and RTA-sk38 indicated unstable conditions.

Considering the days with $\mathrm{AAI}_{\mathrm{dry}} \geq 5$ ($N = 8$), RTA-sk38 indicated unstable conditions for all these days; P_{unstable} and pAp-sk38 each predicted 7 out of 8 of these days correctly. On all these days, the danger level in the public forecast was either 3–Considerable or 4–High. Furthermore, the performance of the regional forecast was comparable to that of the stability metrics (Tab. 5), in line with previous findings (Techel et al., 2024).

These findings suggest that the stability metrics that are supposed to assess dry-snow conditions in view of human triggering often also correctly predicted conditions when the avalanche activity was dominated by natural avalanches. However, this may only apply for moderate amounts of new snow (30-70 cm) covering a critical, persistent weak layer, in which case human triggering is still likely. Also, they consistently (and likely correctly) indicated dry-snow instabilities when avalanche observations may have been biased towards wet-snow releases since deposits at lower elevations appeared wet, whilst the snowpack in the starting zones at about 2500 m a.s.l. was still dry.

The machine-learning-based metric P_{unstable} performed only slightly better than the traditional skier stability index RTA-sk38, considering POD (92% vs. 88%). The combined metrics pAp sk38 ∩ r_c showed the most balanced performance with the lowest false alarm rate and the same overall discrimination power as P_{unstable} (PSS = 0.36). Hence, the partly contrasting results on the significance of process-based metrics that were reported (e.g., Mayer et al., 2022; Pérez-Guillén et al., 2022; Viallon-Galinier et al., 2023) may be related to differences in input data, model setups, validation datasets, and methods used across these studies. Our findings are in line with Herla et al. (2024) who, in their large-scale validation study in Canada, reported a similar level of performance between the process-based approach and the data-driven approach. While the machine-learning-based metric performed slightly better, the process-based metrics have the advantage of being intuitive and easy to understand. The finding that they often lag behind when the slab is still fairly thin, could partly be prevented by adapting or omitting the penetration depth in the calculation of the skier stability index.

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

The data-driven approach, the random forest classifier $P_{unstable}$, was trained with profiles from the region of Davos. For the sake of transparency, it should be noted that among the 146 profiles in the training dataset, there were also 11 profiles from the winter of 2015/16.

The differences and similarities among the various modelling approaches suggest that a multi-model-based approach could be advantageous for operational forecasting. Differences in timing as well as consistent results, but differing forecaster assessment may call for a more detailed review. For instance, towards the end of January, all models indicated rather stable conditions. In fact, no avalanches were recorded 26-28 January 2016, but the avalanche forecast did not change. However, this apparent discrepancy can be explained by the continued presence of signs of instability such as whumpf sounds. This sort of information is presently not included in any of the numerical prediction models and points to one of their limitations.

Finally, by comparing the stability metrics to the avalanche activity index (or the regional danger level), we related a point-scale to a regional-scale variable. This is obviously a scale mismatch. However, in avalanche forecasting, it is standard practice to extrapolate from often-called representative point observations or measurements into the surroundings, simply because data are scarce. Several studies have shown that there is often a fair to good relation between point observations from study plots and the avalanche conditions in the surroundings (e.g., Jamieson et al., 2007). Given this limitation, the relatively low overall performance of the metrics, for instance, PC = 60-70%, is nonetheless remarkable. Also, the performance of the public forecast (no scale mismatch) is not better than that of the simple metrics (Techel et al., 2025), confirming that situations with persistent weak layers are challenging to assess. Moreover, the evaluation using AAI with only two classes (stable vs. unstable) can obviously not reflect the true variety of avalanche conditions. What we call a stable situation (AAI < 1) can still be a situation when conditions are favourable for skier triggering, and a danger level 3–Considerable may be justified. We also considered a lower threshold of 0.2 for the AAI (not shown), which did not change the relative performance among the various metrics, but overall, slightly improved the accuracy (PC) of all metrics, except for HN3d; for instance, the proportion correct for the danger level forecast increased from 61% to 69%.

Beyond the findings presented here, the Steintälli validation dataset from winter 2015/16 offers a valuable benchmark for the evaluation of new stability metrics. Although the dataset only covers part of one winter season, the wealth of data may make it an attractive case study.

5 Conclusions

We followed the evolution of snow stability at the study plot Steintälli from January to March 2016. By combining field measurements and numerical modelling, we gained insight into the performance of three numerical stability approaches based on SNOWPACK output data.

The field measurements for a particular persistent weak layer, typical for the snow conditions in regions of the Swiss Alps with a predominantly transitional snow climate, revealed that the numerical modelling of snow stratigraphy and mechanical properties agreed satisfactorily overall. Weak layer shear strength as measured with the shear frame, as well as specific fracture

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

energy estimated from PST experiments, increased with time, in line with the increase in load. While cracks still propagated to the very end in PST experiments, failure initiation in the deeply buried PWL became increasingly unlikely. While the specific fracture energy of the weak layer increased over time, the concurrent increase in slab load maintained a high crack propagation propensity, as reflected by the relatively small critical crack lengths.

The modelled stability metrics often indicated similar behaviour, in terms of weak layer depth as well as stability. All three approaches most of the time followed the five visually identified persistent weak layers, with the weak layer tracking algorithm by Reuter et al. (2022) showing the closest agreement.

For the winter season analysed in this study, during which the snowpack exhibited several persistent weak layers, we found that, overall, the process-based indices and the recently developed data-driven instability model performed similarly well. The different approaches often showed similar trends and times of instability. In comparison with the avalanche activity index for dry-snow avalanches (AAI_{dry}), the ML approach, $P_{unstable}$, performed slightly better than the traditional skier stability index RTA-sk38, and showed about the same performance as the approach combining the skier stability index and the critical cut length (pAp sk38 ∩ r_c). The various metrics, originally developed to assess the probability of human-triggered avalanches, also performed well when the avalanche activity was dominated by natural avalanches, at least in the case of moderate amounts of new snow.

Some recent studies employing the random forest method with many different snowpack variables resulted in somewhat contrasting findings on the value of traditional stability parameters (Mayer et al., 2022; Pérez-Guillén et al., 2022; Viallon-Galinier et al., 2023). Whereas these numerical models included data from many winters and locations, we only followed part of one winter season at a specific location, since this allows us to get more nuanced insights into the model behaviour. Our findings suggest that it might be instructive to compare different model outputs and that a multi-model approach may be advantageous for operational forecasting. Data-driven as well as process-based approaches are useful to estimate stability and the depth of the most relevant critical weak layer, equivalent to fracture depth, and thus relevant for avalanche size. None of the days with high avalanche activity was missed by all metrics, so it seems best to consider several metrics, and in case of disagreement, have a closer look when assessing snow stability. While our findings confirm that it is feasible to assess stability from numerically simulated snow stratigraphy, it would be beneficial to revisit some of the parametrizations of mechanical properties, which often prominently depend on density only.

Data availability

The underlying research data will be made available at the data repository https://envidat.ch/.

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

Author contributions

Conceptualization and methodology: all authors; Formal analysis: J.S., F.M., S.M., R.B.; Writing - original draft: J.S.; Writing review & editing: all authors.

Competing interests

The authors have no competing interests to declare.

Acknowledgements

We thank our colleagues who assisted with fieldwork, Alec van Herwijnen for providing the PST data obtained with the PTV analyses, and Martin Perfler for help with the weak layer tracking algorithm.

References

Adam, V., Bergfeld, B., Weißgraeber, P., van Herwijnen, A., and Rosendahl, P. L.: Fracture toughness of mixed-mode anticracks in highly porous materials, Nat. Commun., 15, 7379, https://doi.org/10.1038/s41467-024-51491-7, 2024.

Bellaire, S., Schweizer, J., Fierz, C., Lehning, M., and Pielmeier, C.: Predicting snow cover stability with the snow cover model SNOWPACK, Proceedings ISSW 2006. International Snow Science Workshop, Telluride CO, U.S.A., 1-6 October 2006, 38-43, 2006.

Bellaire, S., and Jamieson, 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.

Bergfeld, B., van Herwijnen, A., Bobillier, G., Rosendahl, P. L., Weißgraeber, P., Adam, V., Dual, J., and Schweizer, J.: Temporal evolution of crack propagation characteristics in a weak snowpack layer: conditions of crack arrest and sustained propagation, Nat. Hazards Earth Syst. Sci., 23, 293-315, https://doi.org/10.5194/nhess-23-293-2023, 2023.

Birkeland, K. W., van Herwijnen, A., Techel, F., Bair, E. H., Reuter, B., Simenhois, R., Jamieson, B., Marienthal, A., Chabot, D., and Schweizer, J.: Comparing stability tests and understanding their limitations, Proceedings ISSW 2023. International Snow Science Workshop, Bend, Oregon, U.S.A., 8-13 October 2023, 64-71, 2023.

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, Cryosphere, 14, 1829-1848, https://doi.org/10.5194/tc-14-1829-2020, 2020.

Chalmers, T. S., and Jamieson, J. B.: Extrapolating the skier stability of buried surface hoar layers from study plot measurements, Cold Reg. Sci. Technol., 33, 163-177, https://doi.org/10.1016/S0165-232X(01)00043-X, 2001.

EAWS: Avalanche Problems: https://www.avalanches.org/standards/avalanche-problems/, last accessed: 1 June 2026, 2026.

Fierz, C., Armstrong, R. L., Durand, Y., Etchevers, P., Greene, E., McClung, D. M., Nishimura, K., Satyawali, P. K., and Sokratov, S. A.: The International Classification for Seasonal Snow on the Ground, HP-VII Technical Documents in Hydrology, IACS Contribution No 1, UNESCO-IHP, Paris, France, 90 pp., 2009.

Föhn, P. M. B.: The stability index and various triggering mechanisms, Symposium at Davos 1986 - Avalanche Formation, Movement and Effects, IAHS Publ., 162, 195-214, 1987.

Gaume, J., and Reuter, B.: Assessing snow instability in skier-triggered snow slab avalanches by combining failure initiation and crack propagation, Cold Reg. Sci. Technol., 144, 6-15, https://doi.org/10.1016/j.coldregions.2017.05.011, 2017.

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

Gaume, J., van Herwijnen, A., Chambon, G., Wever, N., and Schweizer, J.: Snow fracture in relation to slab avalanche release: critical state for the onset of crack propagation, Cryosphere, 11, 217-228, https://doi.org/10.5194/tc-11-217-2017, 2017.

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

Herla, F., Haegeli, P., Horton, S., and Mair, P.: A large-scale validation of snowpack simulations in support of avalanche forecasting focusing on critical layers, Nat. Hazards Earth Syst. Sci., 24, 2727-2756, https://doi.org/10.5194/nhess-24-2727-2024, 2024.

Herla, F., Haegeli, P., Horton, S., and Mair, P.: A quantitative module of avalanche hazard – comparing forecaster assessments of storm and persistent slab avalanche problems with information derived from distributed snowpack simulations, Nat. Hazards Earth Syst. Sci., 25, 625-646, https://doi.org/10.5194/nhess-25-625-2025, 2025.

Jamieson, J. B.: Avalanche prediction for persistent snow slabs, Department of Civil Engineering, University of Calgary, Calgary AB, Canada, 258 pp., 1995.

Jamieson, J. B., and Johnston, C. D.: Refinements to the stability index for skier-triggered dry slab avalanches, Ann. Glaciol., 26, 296-302, https://doi.org/10.3189/1998AoG26-1-296-302, 1998.

Jamieson, J. B., and Johnston, C. D.: Evaluation of the shear frame test for weak snowpack layers, Ann. Glaciol., 32, 59-68, https://doi.org/10.3189/172756401781819472, 2001.

Jamieson, J. B., Zeidler, A., and Brown, C.: Explanation and limitations of study plot stability indices for forecasting dry snow slab avalanches in surrounding terrain, Cold Reg. Sci. Technol., 50, 23-34, https://doi.org/10.1016/j.coldregions.2007.02.010, 2007.

Lehning, M., Fierz, C., Brown, R. L., and Jamieson, J. B.: Modeling instability for the snow cover model SNOWPACK, Ann. Glaciol., 38, 331-338, https://doi.org/10.3189/172756404781815220, 2004.

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

Mayer, S., Techel, F., Schweizer, J., and van Herwijnen, A.: Prediction of natural dry-snow avalanche activity using physics-based snowpack simulations, Nat. Hazards Earth Syst. Sci., 23, 3445–3465, https://doi.org/10.5194/nhess-23-3445-2023, 2023.

McCammon, I., and Schweizer, J.: A field method for identifying structural weaknesses in the snowpack, Proceedings ISSW 2002, International Snow Science Workshop, Penticton BC, Canada, 29 September-4 October 2002, 477-481, 2002.

Monti, F., Cagnati, A., Valt, M., and Schweizer, J.: A new method for visualizing snow stability profiles, Cold Reg. Sci. Technol., 78, 64-72, https://doi.org/10.1016/j.coldregions.2012.02.005, 2012.

Monti, F., and Schweizer, J.: A relative difference approach to detect potential weak layers within a snow profile, Proceedings ISSW 2013. International Snow Science Workshop, Grenoble, France, 7-11 October 2013, 339-343, 2013.

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

Morin, S., Horton, S., Techel, F., Bavay, M., Coléou, C., Fierz, C., Gobiet, A., Hagenmuller, P., Lafaysse, M., Ližar, M., Mitterer, C., Monti, F., Müller, K., Olefs, M., Snook, J. S., van Herwijnen, A., and Vionnet, V.: Application of physical snowpack models in support of operational avalanche hazard forecasting: A status report on current implementations and prospects for the future, Cold Reg. Sci. Technol., 170, 102910, https://doi.org/10.1016/j.coldregions.2019.102910, 2020.

Müller, K., Techel, F., and Mitterer, C.: The EAWS matrix, a decision support tool to determine the regional avalanche danger level (Part A): conceptual development, Nat. Hazards Earth Syst. Sci., 25, 4503-4525, https://doi.org/10.5194/nhess-25-4503-2025, 2025.

https://doi.org/10.5194/egusphere-2026-3913
Preprint. Discussion started: 28 July 2026

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

Reuter, B., Schweizer, J., and van Herwijnen, A.: A process-based approach to estimate point snow instability, Cryosphere, 9, 837-847, https://doi.org/10.5194/tc-9-837-2015, 2015.

Reuter, B., and Bellaire, S.: On combining snow cover and snow instability modelling, Proceedings ISSW 2018. International Snow Science Workshop, Innsbruck, Austria, 7-12 October 2018, 949-953, 2018.

Reuter, B., and Schweizer, J.: Describing snow instability by failure initiation, crack propagation, and slab tensile support, Geophys. Res. Lett., 45, 7019-7027, https://doi.org/10.1029/2018GL078069, 2018.

Reuter, B., Proksch, M., Löwe, H., van Herwijnen, A., and Schweizer, J.: Comparing measurements of snow mechanical properties relevant for slab avalanche release, J. Glaciol., 65, 55-67, https://doi.org/10.1017/jog.2018.93, 2019.

Reuter, B., Viallon-Galinier, L., Horton, S., van Herwijnen, A., Mayer, S., Hagenmuller, P., and Morin, S.: Characterizing snow instability with avalanche problem types derived from snow cover simulations, Cold Reg. Sci. Technol., 194, 103462, https://doi.org/10.1016/j.coldregions.2021.103462, 2022.

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

Schneebeli, M., and Johnson, J. B.: A constant-speed penetrometer for high-resolution snow stratigraphy, Ann. Glaciol., 26, 107-111, https://doi.org/10.3189/1998AoG26-1-107-111, 1998.

Schweizer, J., and Föhn, P. M. B.: Avalanche forecasting - an expert system approach, J. Glaciol., 42, 318-332, 1996.

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

Schweizer, J., Fierz, C., and Jamieson, J. B.: Assessing the probability of skier triggering from snow layer properties, Proceedings ISSW 2004. International Snow Science Workshop, Jackson Hole WY, U.S.A., 19-24 September 2004, 192-198, 2005.

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.

Schweizer, J., and Jamieson, J. B.: A threshold sum approach to stability evaluation of manual snow profiles, Cold Reg. Sci. Technol., 47, 50-59, https://doi.org/10.1016/j.coldregions.2006.08.011, 2007.

Schweizer, J., and Jamieson, J. B.: Snowpack tests for assessing snow-slope instability, Ann. Glaciol., 51, 187-194, https://doi.org/10.3189/172756410791386652, 2010.

Schweizer, J., Reuter, B., van Herwijnen, A., Richter, B., and Gaume, J.: Temporal evolution of crack propagation propensity in snow in relation to slab and weak layer properties, Cryosphere, 10, 2637-2653, https://doi.org/doi:10.5194/tc-10-2637-2016, 2016.

Schweizer, J., Mitterer, C., Techel, F., Stoffel, A., and Reuter, B.: On the relation between avalanche occurrence and avalanche danger level, Cryosphere, 14, 737-750, https://doi.org/10.5194/tc-14-737-2020, 2020.

Schweizer, J., Reuter, B., Mitterer, C., Monti, F., and Mayer, S.: On stability measurements and modeling, Proceedings ISSW 2023. International Snow Science Workshop, Bend, Oregon, U.S.A., 8-13 October 2023, 313-320, 2023.

Schweizer, J., Knappe, L., Reuter, B., and Mayer, S.: On snow and avalanche climates in the Swiss Alps, Proceedings ISSW 2024. International Snow Science Workshop, Tromso, Norway, 23-27 September 2024, 198-203, 2024.

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

Techel, F., Helfenstein, A., Pérez-Guillén, C., Purves, R., Ruesch, M., Schmudlach, G., Soland, K., Mayer, S., and Winkler, K.: Human vs Machine: who predicts better? Comparing human-made avalanche forecasts with model-driven predictions of snowpack instability and avalanche danger in the Swiss Alps, Proceedings ISSW 2024. International Snow Science Workshop, Tromso, Norway, 23-27 September 2024, 31-38, 2024.

Techel, F., Purves, R. S., Mayer, S., Schmudlach, G., and Winkler, K.: Can model-based avalanche forecasts match the discriminatory skill of human danger-level forecasts? A comparison from Switzerland, Nat. Hazards Earth Syst. Sci., 25, 3333-3353, https://doi.org/10.5194/nhess-25-3333-2025, 2025.

Techel, F., Müller, K., Marquardt, C., and Mitterer, C.: The EAWS matrix, a decision support tool to determine the regional avalanche danger level (Part B): operational testing and use, Nat. Hazards Earth Syst. Sci., 26, 1161-1181, https://doi.org/10.5194/nhess-26-1161-2026, 2026.

van Herwijnen, A., Gaume, J., Bair, E. H., Reuter, B., Birkeland, K. W., and Schweizer, J.: Estimating the effective elastic modulus and specific fracture energy of snowpack layers from field experiments, J. Glaciol., 62, 997-1007, https://doi.org/10.1017/jog.2016.90, 2016.

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, 103596, https://doi.org/10.1016/j.coldregions.2022.103596, 2022.

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

Wever, N., Schmid, L., Heilig, A., Eisen, O., Fierz, C., and Lehning, M.: Verification of the multi-layer SNOWPACK model with different water transport schemes, Cryosphere, 9, 2271-2293, https://doi.org/10.5194/tc-9-2271-2015, 2015.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 4th ed., Elsevier, Amsterdam, Netherlands, 818 pp., 2019.

Zeidler, A., and Jamieson, J. B.: Refinements of empirical models to forecast the shear strength of persistent weak snow layers PART A: Layers of faceted crystals, Cold Reg. Sci. Technol., 44, 194-205, https://doi.org/10.1016/j.coldregions.2005.11.005, 2006.