Dear Editor, dear Simon Horton and Karl Birkeland

Thank you very much for your detailed and constructive review of our, admittedly, rather complex manuscript.
The revised manuscript is certainly still rather complex, due to the description and evaluation of the methodology needed to simulate snow stability distributions and classify the frequency of the most unstable locations. However, we hope that after restructuring the Result section and integrating (or removing) supplementary information previously shown in the Appendix, the manuscript is now easier to read. We did, however, leave the paragraphs in the manuscript, where we present and discuss the strength and limitations of the approach taken. We believe, readers which are not interested in this can now easily skip these sections and focus on the main results, while those interested in the limitations will find these addressed in the manuscript.

Please find following
- A list indicating the major changes
- A point-by-point reply to the reviews
- The manuscript showing all the track changes.

Major changes

- We took up the recommendations by reviewer #1 and restructured large parts of the Result section (Sect 4). However, the actual results (or their interpretation) were not affected by this restructuring of the manuscript. Figures shown in the Appendix were either integrated in the new Results section, or they were removed. In the track-changes document, the entire Sect. 4 is marked, as we had to restructure and rephrase large parts of this section.
- Sect. 4.1 – 4.4 now show the findings based on the Swiss data. Sect. 4.5 presents the ECT data and the data from Norway and compares them to the respective Swiss or Rutschblock data. Sect. 4.6 presents results related to the methodology of bootstrap sampling and frequency classes.
- Fig. 4 is new. It shows the distribution of the frequency classes of very poor stability for the four danger levels.
- For stability classification of ECT and RB, we used a simpler depth criterion (p6 l17-19) than in the original version (p6 l28-31). We followed the approach taken by Techel et al. (*On snow stability interpretation of Extended Column Test results*, NHESS, accepted). Very marginal changes in the proportions observed for RB and ECT (Fig. 3) resulted.

Review by reviewer #1 Simon Horton

Original manuscript (reviewer comment)	**Revised manuscript (changes made)**
Relevance of additional data sets: The key conclusions of the study appear to come from the Swiss Rutschblock test and avalanche data, however while reading the results there are numerous references to patterns between the Swiss/Norwegian and RB/ECT data. I found this distracting from the main research question about the contributing factors to danger ratings. I would consider restructuring some of the sections so the main research question is addressed first, and then perhaps distinct sub-	We have completely restructured the Results section, to address this point. We now show first all the results based on Swiss data (Sect. 4.1 – 4.4), we then compare these in a second step with Norwegian data and/or ECT (Sect. 4.5). Results, which were related more to the methodology rather than linking the observations to avalanche hazard, have been moved to a new Sect. 4.6 together with relevant figures.

sections discussing how the core results differ between SWI/NOR and between RB/ECT results. There are also quite a few Appendix figures with these additional data sets which disrupts the flow while reading the results.	We hope this restructure helps the reader to distinguish more easily between the results relating to the research questions, and those that compare to other data sets or that present additional information regarding the methodology. Furthermore, together with this restructure, we have moved the relevant figures from the Appendix to the main part of the manuscript.
Methods: The beginning of each sub-section could use a bit more context about how that step is relevant to exploring the link between danger ratings and a contributing factor.	We added a sentence in that regard in the subsections: p6 l3, p6 l21
p1 line 7: Although less precise, saying “frequency of unstable locations” may be simpler to understand when reading the abstract only	p1 l7: According to reviewer #2 changed to *…the frequency distribution of snowpack stability and…*
p1 line 13-14: Consider adding “simulated stability distributions” (the snowpack distribution isn’t simulated)	P1 l13: changed to *…simulated stability distributions…*
p2 line7: Preferable to use consistent terminology from the list of key factors, i.e. “probability of avalanche release” instead of “release (or triggering) probability”	We changed the terminology in the list (p2 l3-5) and repeated the term when starting each paragraph (p2 l7, p2 l13, p2 l28)
p2 line 13: Similar to above, starting the paragraph by repeating the term “frequency and location of triggering spots” would make it clearer the paragraph ties back to the list of key factors	
p2 line 23: missing citation	P2 l22-23: we added *…(EAWS, 2017).*
p2 line 24: According to the CMAH spatial distribution also considers spatial density. Statham et al. 2018: “Spatial distribution considers the spatial density and distribution of an avalanche problem and the ease of finding evidence to support or refute its presence.”	P2 l25-27: we changed to *In the CMAH (Statham et al. 2018a), on the other hand, the spatial distribution is related to the spatial density and distribution of an avalanche problem and the ease of finding evidence for it, and is described using the three terms isolated, specific and widespread.*
Table 1: The “data from” column heading isn’t clear if the data is from just a single season or all seasons up to 2018/19 (as explained in footnote). Consider a more precise heading or list season ranges in the table (e.g. 2002-2019)	We changed the year format to 2002 - 2019
p4 line 2-4: These two sentences aren’t necessary, as they are discussed below.	We removed these sentences.
p6 line 4-5: Please be consistent with order of reporting SWI and NOR data, in this sentence NOR is described first.	We changed to always introducing SWI data first.
p6 line 8: It would be helpful to start this section by explicitly explaining the purpose of this step is to relate the snowpack test data to	P6 l3: We added *Snowpack stability is one of the three contributing factors to avalanche hazard and relates to the probability of avalanche release.*

one of the explanatory factors in the study (i.e. probability of avalanche release)	
p6 lines 19-26: This is an example of how the addition of ECT data confuses the reader and distracts from the main point.	P6 l13-16: We shortened this paragraph considerably.
p7 line 2: It would be helpful to start this section by explicitly explaining purpose of this step to relate the snowpack test data to one of the explanatory factors in the study (i.e. frequency of triggering spots)	P6 l21: We added *The second factor contributing to avalanche hazard is the frequency of potential triggering locations, or of snowpack stability.*
p7 line 15: What effect does an equal number of samples for each rating have considering there are likely a higher proportion of days with ratings of 2 and 3. The sample of 10,000 will likely have a skewed number of unstable tests from high danger days. Does this impact the interpretation of the results?	An equal number of samples for each danger level is important, when the danger level for each combination is sought. For instance, if only 1% of the samples would have been 4-High, the danger matrix in Figure 6 would essentially never show a 4-High, as 3-Considerable would dominate these cells due to their larger weight. The definition of the class thresholds changes little, as the median proportion of very poor tests (VP_{med}) drives them. When using a typical distribution of danger levels forecast in Switzerland instead (1-Low to 4-High, 18%, 43%, 36%, 2%, respectively), the variable which defines the class intervals VP_{med} is the same with 0.08.
p9 line 28: Slightly confusing, perhaps add ": : : distribution of observed data for all days at a given danger level represent: : :"	P9 l11-14: We changed to *As we do not have data describing the three factors relating to the same day and region, we used a simulation approach by assuming that the distribution of the observed data represents the typical values and ranges at a specific danger level.*
p10 line 1: Consider different verb than "complemented"	P9 l17: we changed to *…we combined the snowpack stability…*
Sect. 4.1.2: This section has many references to appendix figures, which disrupts the flow because the reader is compelled to flip back and forth to the appendix. The confusion could be reduced by introducing Fig. 4 earlier, which clearly shows the most relevant results, then followed by more discussion about the sensitivities to sample size, etc that reference the appendix figures.	We completely restructured the Result section (Sect. 4). Now, there are no more references to appendix figures as we have moved all the relevant figures to the Results section.
P12 line 10: Are these proportions discussed later? They seem meaningful for interpreting stability test results (e.g. even dangerous days have relatively few sites with very poor stability).	We now discuss these proportions at several locations: P10 l10 – p11l2 and in the Discussion p24 l5-7
p14 lines 2-9: This is an example of where the comparison between countries seems like a	Addressed by restructuring of Results section – new in a separate subsection (Sect. 4.5)

secondary discussion point compared to reporting the main patterns between avalanche size and danger.	
p15 line 9: In this list the percentages reported in brackets could be misinterpreted as proportion of locations with very poor stability. Perhaps the first reported percentage could explain what the percentage means, e.g. "(53% of sample)".	P14 l9: changed as suggested
Fig. 6-8: Good use of figures with a consistent layout showing the lookup table and the supporting data. The idea that Fig 7 and 8 have the exact same matrix structure as Fig 6 wasn't fully clear on the first read, so could perhaps be explained more explicitly in the text.	As suggested, we tried to emphasize that the structure of the figures is the same, in the caption and the text. Additionally, Fig 7 and Fig 8 are now beside each other in Fig 7 (as a and b), thus it should be easier to compare these figures with Fig 6.
P20 line 17: "while observations of natural or artificial: : :"	P22 l6-7: changed as suggested
p20 line 27: Captured "slope stability" or "regional danger"?	P22 l15-17: we meant slope stability, but the reference to this statement was missing and has now been added *However, as shown by Techel et al. (2020), the most favorable and the most unfavorable RB stability classes captured slope stability better than the respective ECT classes, indicating a lower agreement between slope stability and ECT results compared to the RB.*
Sect 5.3.1: Another consideration when comparing with existing methods is the CMAH assesses the frequency of trigger spots for each avalanche problem rather than snowpack as a whole as done in the EAWS matrix. This may make it easier to answer questions about the frequency of unstable locations for a specific problem type but could make it more difficult when combining avalanche problems into an overall danger rating. Just an additional thing to consider when discussing how we can better assess the spatial frequency of instabilities.	We have not taken up this point.
p24 lines 5-9: An updated citation with more comprehensive analysis is Clark (2019), where the influence of many factors on danger ratings are explored (size, likelihood, problem type, region, vegetation band, etc.). The importance of "likelihood" in Clark (2019) still agrees with the main findings in this study.	We now make a reference to Clark, 2019 and Clark and Haegeli, 2018

Comments by reviewer #2 Karl Birkeland

Original manuscript (reviewer comment)	**Revised manuscript (changes made)**
First, the title could be worded more succinctly and less ambiguously. I might suggest	We have changed the title to "*On the importance of snowpack stability, the frequency distribution of snowpack stability, and*

something along the lines of "The importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing avalanche danger". However, the authors might have some other title they prefer. In particular I think they could omit "a data-driven approach" since that can be emphasized in the abstract and the text. Also, in the title and in several places in the paper they write ": : :snowpack stability, its frequency distribution, and avalanche size: : :". I personally find this to be a bit awkward and ambiguous with the use of the term "its". Even though it is slightly longer and involves more words, I think saying ": : :snowpack stability, the frequency distribution of snowpack stability, and avalanche size: : :" states what the authors are trying to say more clearly.	*avalanche size in assessing the avalanche danger level*". We made similar changes at various locations in the manuscript, for instance: P1 l6-7, p7 l21, p9 l10, …
Second, my main criticism of the paper relates to the conclusion by the authors that "avalanche size only has a rather minor influence on the danger level" (bottom of p. 23). Perhaps this is just from the author's choice of words, but in my opinion the data and Figures in the paper do not show a "rather minor influence". Instead, they show an influence that may be less than that of snow stability or frequency, but one that is still clearly evident. An example is in Figure 6 where no matter which letter you get from the combination of stability and frequency on the left side of the Figure, when you go to the right side of the Figure you can see that with all the letters you see an increase in the avalanche danger as the largest avalanche size increases. This is also clearly shown in Figure 8, where going from left to right in the Figure we can see that the proportion of higher danger levels increases as the avalanche size increases. Another example of the influence of avalanche size can be seen in Figure 5. It is true, as the authors state in the Conclusions on p. 24, that "the largest avalanche size – used by itself – had comparably little discriminating power at 1-Low to 3-Considerable". However, while that might be strictly true for "the largest avalanche size", Figure 5 shows that the distribution of avalanche size – particularly of the largest avalanche (Figure 5b) – clearly does play into avalanche danger. The frequency distributions visibly tend toward larger avalanches at higher danger levels, with the proportion of size 3 and	See also our response to this comment. We rephrased at several locations in the manuscript: P11 l15 – p16l4: we rephrased in several locations, also because of the restructuring of the Results section P25 l12-13: *In general, avalanche size had a lesser influence on the danger level, once the cell describing stability has been fixed, as might be anticipated* P26 l6-9: *Considering the largest observed avalanche size per day and warning region was most relevant to distinguish between 3-Considerable and 4-High (Fig. 5 and Tab. 3). For other situations, the largest avalanche size - when used on its own - had less discriminating power to distinguish between danger levels 1-Low to 3-Considerable compared to the other two factors (the lowest stability class present and the frequency of this class; Fig. 5).*

4 avalanches increasing while the proportion of size 1 avalanches decreases. I would tend to disagree with the statement on p. 14, line 10-11, that Figure 5b shows “rather similar size distributions at 1-Low and 2-Moderate”. Comparing the two, we can see a sizable decrease in size 1 avalanches and an almost doubling in the number of size 3 avalanches between Low and Moderate. Given the data presented in the paper, I would argue that the authors should better acknowledge that avalanche size does indeed have an influence on avalanche danger, and is not a “rather minor influence” (as stated on p. 23). I think they could still make an argument that snow stability, and the frequency of snow stability might well have a larger influence on avalanche danger, but avalanche size is also an important part of the avalanche danger assessment process. I would therefore encourage them to revisit various parts of the manuscript where avalanche size is discussed and better acknowledge the influence of size on avalanche danger.	
p. 1, line 2, delete “the”	done
p. 1, line 4, remove the two commas	done
p. 2, line 16, replace “weakest” with “the most unstable” because weakest could	P2 l16: changed to *…lowest…*
p. 2, line 23, what does the “(?)” refer to? Were the authors going to put a reference in there or ??	P2 l22-23: we added *…(EAWS, 2017).*
p. 3, line 2, spell out EAWS completely the first time it is introduced in the text and then refer to it as EAWS afterwards.	done
p. 3, line 7, remove the two commas and replace “work” with “works”	done
p. 3, line 11, replace “but” with “and”	done
p. 3, line 12, replace “And” with “and”	done
p. 3, line 13, delete “does” and change “describe” to “describes”	done
p. 3, line 25, delete “The target variable” and “we want to describe the three factors with”	done
p. 4, line 17, would the authors like to include Foehn, 1987 in addition to Schweizer, 2002 to the RB reference?	We added *Föhn, 1987*
p. 5, line 1, replace “comparably” with “relatively”	done
p. 6, line 1-3. It would be nice if the authors would explain why they removed the	P5 l26-27: we added *…were considered to represent errors in the local estimate of the danger level or of*

upper and lower 2.5% of the avalanche data. I am guessing they did this to filter out possible errors with the extremes or something along those lines? In any event, a single sentence explaining why this was done would be helpful.	*avalanche size. These potentially erroneous data were removed.*
p. 7, line 5. The authors state that they are assuming that "different days with the same danger level exhibit similar stability distributions". I think they probably have to assume this to continue with their analyses. However, although I don't have any concrete data to support this, I feel like stability distributions can certainly vary between days that have the same danger level. This is somewhat built into the Conceptual Model of Avalanche Hazard by the inclusion of "uncertainty" and relates to how large an oval a person might put on the probability/size graph of the CMAH when selecting a danger level. It seems to me that the largest variations in stability distributions fall under "3 - Moderate" and "4 - Considerable" danger levels. For example, sometimes under 4 – Considerable you might have a distribution that is more spread out with the possibility of triggering a larger avalanche, while another time you might have a narrower spread of values, but the size of avalanche expected might be smaller. Both of these could have the same avalanche danger level, but the distribution of stability would vary. I don't think the authors have to make big changes to this paper, but I do think they should acknowledge that this assumption they are making might not always be valid.	P7 l2-4: we changed to *Assuming that a single test result is just one sample from the stability distribution on that day and that different days with the same danger level exhibit a range of similar stability distributions, …* By using the sampling approach, we created a wide range of stability distributions, which we exemplarily describe on p17 l32 – p18 l5 (together with Fig 8c): *When introducing the bootstrap-sampling approach to create a range of plausible stability distributions (Sect. 3.2), we had to assume that a single stability rating is just one sample from the stability distribution on that day and that different days with the same danger level exhibit a range of similar stability distributions. Referring to Fig. 8c, which shows the proportions of very poor and good stability of the 10,000 simulated distributions with n= 25, it can be noted that indeed a range of typical distributions was obtained for the four danger levels. For instance, at 3-Considerable the range of the simulated distributions was wide: 11% of the samples drawn had ≥ 8% (frequency classes several or many) very poor and ≤ 4% (a few or none) good tests results, while 7% of the samples drawn had ≤ 4% (a few or none) very poor and 24% (many) good tests results.*
p. 7, line 18, sentence is a bit awkward and confusing. I would change it to read: "Since nature is not as discrete as the danger levels suggest, we wanted both some overlap between our sampled stability distributions and a reasonably high resolution of	P7 l15-17: done
p. 9, line 12, replace "maximising" with "maximizing"	done
p. 11, Figure 3. This is an interesting and important Figure. One limitation that is noted in the text and also in the figure is the very small N for "4-High" (approximately two orders of magnitude smaller than for 2-Moderate or 3-Considerable). To further	Fig. 3 and p9 l27-28: we added a remark in this regard.

emphasize this, the authors could consider stating something related to this in the Figure caption, possibly something like "Note the small N for 4-High for both tests", or, even better, you could write "Note the N for 4-High is small and is approximately two orders of magnitude less than the N for 2-Moderate or 3-Considerable".	
p. 12, line 11, delete the first "of" in the line.	done
p. 14, line 19, delete "It is of"	Sentence removed
p. 19, line 4. I have seen this under representation of smaller avalanches in most datasets related to ski area snow safety staff in the United States. This isn't written down in too many places, but we do discuss this somewhat in Birkeland and Landry, 2002 (Power-laws and snow avalanches. Geophysical Research Letters 29(11), 49-1 to 49-3).	P24 l29-31: We rephrased and added references in this regard *This frequency-magnitude relation has also been observed for other natural hazards (e.g. Malamud and Turcotte, 1999), and has been described by power laws for avalanche size distributions (Birkeland and Landry, 2002; Faillettaz et al., 2004).*
p. 19, line 6. Replace "As" with "Since" and insert "instead" between "focused" and "on".	done
p. 19, line 8, delete comma	done
p. 19, line 9, replace "weak" with "unstable". I believe the authors are talking an "unstable" snowpack here and not necessarily one that is just structurally weak, correct?	P21 l3 Changed to low
p. 21, line 26, replace "," with "." prior to "For instance,"	
p. 21, line 28, replace "Schweizer et al. (2003) s" with "Schweizer et al.'s (2003)"	done
p. 22, line 25 and 27. The authors refer to the correlations being "strong" or "moderate". What do you mean by this? Are they statistically significant or not? You might want to state whether they are significant and list a p-value. When I refer back to Section 4.1.2 as is suggested on line 27, I believe the authors are referring to p. 12, line 5-8. Is this correct? Here it states that – even with an N = 10 - the correlation is highly significant ($p < 0.001$).	We removed this part of the Discussion, as it is addressed in the Results section (Sect 4.6.2) p18 l23-24
p. 23, line 5. What does the "(?)" refer to? Are the authors planning on adding a reference here?	We added the reference (EAWS, 2017)
p. 30, delete "and tables" from the title of Appendix 2 since this appendix has only figures.	Appendix has been removed
p. 31, in the caption for Figure B1, replace "Fig.s" with "Figs."	Appendix has been removed
p. 34, Figure E1, for the top right part of the Figure (all avalanches for Switzerland),	Appendix has been removed

add “(SWI)” after “all avalanches” to be consistent with the other headers. Also, add the percent number above the bar for size 1 avalanches under Low to match the other graphs in this Figure.	

On the importance of snowpack stability, [..*]the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level[..†]

Frank Techel[1, 2], Karsten Müller[3], and Jürg Schweizer[1]

[1]WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland
[2]University of Zurich, Department of Geography, Zurich, Switzerland
[3]Norwegian Water Resources and Energy Directorate NVE, Oslo, Norway

Correspondence to: Frank Techel (techel@slf.ch)

Abstract. Consistency in assigning an avalanche danger level when forecasting or locally assessing avalanche hazard is essential, but challenging to achieve, as relevant information is often scarce and must be interpreted in [..3]light of uncertainties. Furthermore, the definitions of the danger levels, an ordinal variable, are vague and leave room for interpretation. Decision tools [..4]developed to assist in assigning a danger level [..5]are primarily experience-based due to a lack of data. Here, we address this lack of quantitative evidence by exploring a large data set of stability tests (N = [..6]9,310) and avalanche observations (N = 39,017) from two countries related to the three key factors that characterize avalanche danger: snowpack stability, [..7]the frequency distribution of snowpack stability and avalanche size. We show that the frequency of the most unstable locations increases with increasing danger level. However, a similarly clear relation between avalanche size and danger level was not found. Only for the higher danger levels the size of the largest avalanche per day and warning region increased. Furthermore, we derive stability distributions typical for the danger levels 1-Low to 4-High using four stability classes ([..8]*very poor*, *poor*, *fair* and *good*), and define frequency classes [..9]describing the frequency of the most unstable locations (*none or nearly none*, *a few*, *several* and *many*). Combining snowpack stability, [..10]the frequency of stability classes and avalanche size in a simulation experiment, typical descriptions for the four danger levels are obtained. Finally, using the simulated [..11]stability distributions together with the largest avalanche size in a step-wise approach, [..12]we present a data-driven look-up table for avalanche danger assessment. Our findings may aid in refining the definitions of the avalanche danger scale and in fostering its consistent usage.

*removed: its
†removed: : a data-driven approach
[3]removed: the
[4]removed: ,
[5]removed: ,
[6]removed: 10,125
[7]removed: its frequency distribution
[8]removed: *very poor*, *poor*, *fair* and *good*
[9]removed: (*none or nearly none*, *a few*, *several* and *many*)
[10]removed: its frequency
[11]removed: snowpack
[12]removed: as proposed in the Conceptual Model of Avalanche Hazard, we present an example for

1 Introduction

Consistent communication of regional avalanche hazard in publicly available avalanche forecast products is paramount to avoid misinterpretations by the users (Techel et al., 2018). A key information in public bulletins is the avalanche danger level. The danger levels - from 1-Low to 5-Very High - are described in the European Avalanche Danger Scale (EADS, EAWS, 2018) or its North American equivalent, the North American Avalanche Danger Scale [..[13]](e.g. Statham et al., 2010) with brief definitions of the key factors. The key factors that characterize avalanche danger are (Meister, 1995; EAWS, 2020, 2018):

- the probability of avalanche release,
- the frequency and location of the triggering spots, and
- the expected avalanche size.

These elements are expected to increase with increasing danger level (e.g. Schweizer et al., 2020).

The [..[14]]probability of avalanche release, or 'sensitivity to triggers' as termed in the Conceptual Model of Avalanche Hazard (CMAH, Statham et al., 2018a), is inversely related to snowpack stability, with a higher probability for an avalanche to release with lower stability, and vice versa (e.g. Föhn and Schweizer, 1995; Meister, 1995). Hence, the probability of avalanche release refers to a specific location and relates to the local (or point) snow instability. The latter has recently been revisited and three elements were suggested to describe point snow instability: failure initiation, crack propagation and slab tensile support (Reuter and Schweizer, 2018).

The [..[15]]frequency and location of the triggering spots is typically unknown. So far, it can only be assessed with laborious extensive sampling [..[16]](e.g. Birkeland, 2001; Reuter et al., 2016). However, in a regional avalanche forecast the spatial distribution of snow instability can be described with regard to the frequency and the locations of triggering spots [..[17]]or more generally the locations where [..[18]]snowpack stability is lowest. From these two components, frequency and location, only frequency is relevant when assessing the danger level (Schweizer et al., 2020). The frequency always refers to a specific area, typically a forecast region [..[19]]and/or slope aspects and elevation [..[20]]bands. The frequency distribution describes the question «How often do spots with a certain [..[21]]snowpack stability exist within a region?» – in terms of numbers, proportions or percentages. Typical frequency distributions for the danger levels 1-Low to 3-Considerable were described by Schweizer et al. (2003) using five classes of [..[22]]snowpack stability. Frequency expresses the number of triggering locations assuming a uniform distribution within the reference area and is described using the terms *single*, *some*, *many*, and *most* (EAWS,

[13]removed: (e.g. **?**)
[14]removed: release (or triggering) probability of an avalanche
[15]removed: actual spatial distribution of snow stability
[16]removed: (e.g. Birkeland, 2001; **?**)
[17]removed: (
[18]removed: the snowpack is weakest)
[19]removed: . In addition, in the forecast the
[20]removed: are described where the danger prevails
[21]removed: snow
[22]removed: snow stability.

2017). In contrast, the location of triggering spots or of snowpack stability [..[23]]refers to «Where in the terrain is avalanche release most likely?» [..[24]]It indicates where in the terrain the frequency is slightly higher (e.g. [..[25]]*where the snowpack is shallow*, *close to ridgelines*, *in bowls*, ...). In [..[26]]the CMAH (Statham et al., 2018a), on the other hand, the spatial distribution is related to the [..[27]]spatial density and distribution of an avalanche problem [..[28]]and the ease of finding evidence for it, and is described using the three terms [..[29]]*isolated*, *specific* and *widespread*.

Finally, avalanche size is defined with sizes ranging from 1 to 5 relating to the destructive potential of an avalanche (e.g. CAA, 2014; EAWS, 2019; McClung and Schaerer, 1981).

The EADS descriptions of the key factors for each of the five categories of danger level leave ample room for interpretation and are even partly ambiguous. This may be a major reason for inconsistencies noted in the use of the danger levels between individual forecasters or field observers, and even more prominent between different forecast centers and avalanche warning services (Lazar et al., 2016; Statham et al., 2018b; Techel and Schweizer, 2017; Techel et al., 2018), but also when assessing different avalanche problems [..[30]](Clark, 2019).

The same danger level can be described with different combinations of the three factors. To improve consistency in the use of the danger levels, a first decision aid, the Bavarian Matrix was adopted by [..[31]]the European Avalanche Warning Services (EAWS) in 2005. The Bavarian Matrix, a look-up table, combined the frequency of triggering locations with the release probability. In 2017, an update of the Bavarian matrix, now called the EAWS-Matrix, was presented that additionally incorporates avalanche size (EAWS, 2020). More recently, a so-called Avalanche Danger Assessment Matrix (ADAM, Müller et al., 2016) was proposed, which tries to combine the workflow described in the CMAH with the assignment of the danger levels based on the three factors as suggested in the EAWS-Matrix. Both [..[32]]the current version of the EAWS-Matrix and ADAM [..[33]]are works in progress.

Challenges in the improvement of these decision support tools include the fact that the three key factors characterizing avalanche danger are not clearly defined and hence poorly quantified (Schweizer et al., 2020). Our objective is therefore to address this lack of quantitative evidence by exploring observational data relating to snowpack stability, its frequency distribution and avalanche size. The data originate from different snow climates, [..[34]]and also from different avalanche warning services (Norway, Switzerland). The key questions are: (1) How do the three factors relate to the danger levels? [..[35]]and

[23]removed: describes

[24]removed: Currently, the frequency is described using the terms *single*, *some*, *many*, and *most* (**?**), while terms describing the location are manyfold

[25]removed: *where the snowpack is shallow*, *close to ridgelines*, *in bowls*

[26]removed: contrast, in the CMAH ,

[27]removed: ease of finding evidence of

[28]removed: (Statham et al., 2018a)

[29]removed: *isolated*, *specific* and *widespread*

[30]removed: (Clark and Haegeli, 2018)

[31]removed: EAWS

[32]removed: ,

[33]removed: , are work

[34]removed: but

[35]removed: And

(2) Which combination of the actual value of the three factors [..[36]]best describes the various danger levels? We present a methodology to generate data-driven stability distributions and to obtain class intervals describing the frequency of a given [..[37]]snowpack stability class. Finally, we will compare the findings with currently used definitions in avalanche forecasting, as EADS and CMAH, and make recommendations for improvements towards more consistent usage of the danger scale.

2 Data

All the data described below were recorded for the purpose of operational avalanche forecasting in Norway (NOR; Norwegian Water Resources and Energy Directorate NVE) or Switzerland (SWI; WSL Institute for Snow and Avalanche Research SLF). In the vast majority, these observations were provided by specifically trained observers, belonging to the observer network of either the Norwegian or the Swiss avalanche warning service.

[..[38]]For the analysis, we rely primarily on the Swiss data using the Norwegian data for comparison and validation. Nevertheless, we will occasionally present results for Swiss and Norwegian data side by side.

2.1 Avalanche danger level

The [..[39]]avalanche danger level [..[40]]is an estimate at best, as there is no straightforward operational verification. Whether assessing the danger level in the field or in hindsight, it remains an expert assessment (Föhn and Schweizer, 1995; Techel and Schweizer, 2017).

We rely on the local danger level estimates provided by specifically trained observers. In both countries, this estimate is based on the observations made on the day and on other information considered relevant (Kosberg et al., 2013; Techel and Schweizer, 2017) and can be called a local nowcast. In very few exceptions (19 days during the verification campaigns in the winters 2002 and 2003 in the region surrounding Davos, SWI) a «verified» regional danger rating was available (Schweizer et al., 2003; Schweizer, 2007b).

In this study, we make use of local estimates for dry-snow conditions only. Each stability test or avalanche observation was linked to a danger rating as described below (Sect. 2.2 and 2.3). [..[41]][..[42]]

[36]removed: does best describe

[37]removed: snow

[38]removed: Despite the necessity to harmonize some of the data across warning services, we argue that making use of data from different warning services and snow climates, may highlight potential biases' or differences

[39]removed: target variable, the

[40]removed: , we want to describe the three factors with,

[41]removed: If no local danger level estimates were available, the data were not used.

[42]removed: Throughout this manuscript, we refer to the danger levels using their integer-signal word combination, e.g. 1-Low or 2-Moderate.

Table 1. Data overview.

parameter		country	N	data from*
avalanches	natural	SWI	29,511	[..[43]]2001-2019
	human-triggered	SWI	3,751	[..[44]]2001-2019
	natural	NOR	4,555	[..[45]]2014-2019
	human-triggered	NOR	1,200	[..[46]]2014-2019
RB		SWI	4,[..[47]]439	[..[48]]2001-2019
ECT		SWI	2,745	[..[49]]2007-2019
		NOR	2,[..[50]]126	[..[51]]2014-2019

* - for days between (and including) 1 Dec and 30 Apr.

2.2 [..[52]]Snowpack stability

Operationally available information directly related to snow instability includes simple field observations as well as [..[53]]snowpack stability tests (Schweizer and Jamieson, 2010). Field observations such as recent avalanching, shooting cracks and whumpfs (a sound audible when a weak layer fails due to localized loading) clearly indicate snow instability (Jamieson et al., 2009; Schweizer and Jamieson, 2010). These observations are often made in the backcountry while ski touring and do not require a person to dig a snow pit. [..[54]]Snowpack stability tests, on the other hand, are considered targeted sampling (McClung and Schaerer, 2006) with the aim to assess point snow instability. Here, we used data obtained with two stability tests regularly used to assess snow instability in Switzerland and Norway, the Rutschblock test and the Extended Column Test.

The **Rutschblock test (RB)** is a stability test, ideally performed on slopes steeper than 30°, where a 1.5 m × 2 m block of snow is isolated from the surrounding snowpack and loaded by a person [..[55]](e.g Föhn, 1987; Schweizer, 2002). An observer performing a RB records which of the 6 loading steps, referred to as the [..[56]]*score*, caused failure, and what portion of the block slid (the [..[57]]*release type*: whole block, most of block, edge only). If no failure occurs, RB7 is recorded. [..[58]]*Score* and *release type* provide information on failure initiation and crack propagation, essential components of [..[59]]slab avalanche release (Schweizer et al., 2008b). RB data were only available from Switzerland.

The **Extended Column Test (ECT)** is a stability test that provides an indication on crack propagation propensity (Simenhois and Birkeland, 2006, 2009). In contrast to the RB, the ECT is performed on a [..[60]]relatively small (30 cm × 90 cm) isolated column of snow and loaded by tapping on the block. The observer records the tap at which a crack initiates (1-30) and whether

[52] removed: Snow
[53] removed: snow
[54] removed: Snow
[55] removed: (e.g Schweizer, 2002)
[56] removed: *score*
[57] removed: *release type*
[58] removed: *Score* and *release type*
[59] removed: a
[60] removed: comparably

a fracture propagates across the entire column (ECTP), or not (ECTN; Simenhois and Birkeland, 2009). If no fracture is initiated with 30 taps ECTX is recorded.

Each stability test was linked to a danger rating relating to dry-snow conditions. We considered the danger rating most relevant, which was transmitted together with the snow profile or stability test (in text form, SWI). In the Swiss data set, this danger rating was replaced for stability tests observed on days and in warning regions, for which a «verified» regional danger rating existed (Sect. 2.1). If neither of them was available, the operational database was searched for local danger level estimates reported during the day and in the same region. Often, these local estimates were reported by the same observer who performed the test.

The Swiss RB data set comprised 4,[..[61]]439 RBs, observed mainly on NW-, N-, and NE-facing slopes (67%) at a median elevation of 2,380 m a.s.l. (interquartile range IQR 2,160-2,565 m) and a median slope angle of 35° (IQR: 32-37°). The Swiss ECT data set contained 2,745 ECTs; 67% were observed in NW-, N- and NE-facing slopes at a median elevation of 2,372 m a.sl. (IQR 2,134-2,547 m) and at 34° (IQR 31-36°). The Norwegian ECT data set consisted of 2,[..[62]]126 ECTs, observed at a median elevation of 760 m a.sl. (IQR 730-1,067 m). Consistent information on the slope aspect was not available for Norwegian stability data.

2.3 Avalanches

As part of the daily observations, observers (and occasionally the public) reported avalanches observed in their region. Avalanches can be reported individually, but also by summarizing several avalanches into one observation. While individual avalanches were reported in a similar way in [..[63]]SWI and NOR, the reporting of several avalanches differed. In SWI, observers reported the number of avalanches of a given size. In all reporting forms, information about the wetness and trigger type could be provided. In NOR, observers reported avalanche size, trigger type and wetness, which was typical for the situation, and described the observed number of avalanches using categorical terms (single: 1, some: 2-4, many: 5-10, numerous: ≥11). In [..[64]]either country, avalanche size was estimated according to the destructive potential, and a combination of total length and volume, resulting in avalanche sizes of 1 to 5 (EAWS, 2019). In SWI until 2011, only size classes 1-4 were used.

The analysis was restricted to dry-snow avalanches, where the trigger type was either natural release or human-triggered. These avalanches were linked to a dry-snow local danger rating for the release date of the avalanche(s) and in the same warning region.

To enhance the quality of the data, we filtered observations, which we believe may indicate errors in the local estimate of the danger level or of avalanche size. To this end, we calculated the avalanche activity index (AAI, Schweizer et al., 1998), a dimensionless index summing up avalanches according to their size with weights of 0.01, 0.1, 1, and 10 for avalanche sizes 1 to 4, respectively. We did not assign weights to the trigger type (natural, human-triggered). For NOR, where the number

[61] removed: 698

[62] removed: 682

[63] removed: NOR and SWI

[64] removed: SWI, observers reported the number of avalanches of a given size. In all reporting forms, information about the wetness and trigger type could be provided. In

of observed avalanches is described categorically, we assigned numbers as follows: one = 1, few (2-5) = 3, several (6-10) = 8, numerous (≥11) = 12. For each country, we then rank-ordered the avalanche data and the lowest 2.5% of the days and regions with 2-Moderate, 3-Considerable and 4-High, and the top 2.5% of the days and regions with 1-Low, 2-Moderate or 3-Considerable were considered to represent errors in the local estimate of the danger level or of avalanche size. These potentially erroneous data were removed.

The total number of avalanches that remained was [..[65]]33,262 in Switzerland, observed on [..[66]]6,610 days and regions, and [..[67]]5,755 in Norway, observed on [..[68]]1,618 different days and regions (Table 1).

3 Methods

3.1 Classification of [..[69]]snowpack stability

Snowpack stability is one of the three contributing factors to avalanche hazard and relates to the probability of avalanche release. In the following, we describe how we classified the results of the snow instability tests in the four stability classes ([..[70]]*very poor*, *poor*, *fair* and *good* - stability class names are in italics throughout this manuscript).

Rutschblock test (RB) results were classified [..[71]]in the four stability classes according to Figure 1a using a combination of score and release type, which have been shown to be good predictors of unstable conditions (e.g. Föhn, 1987; Jamieson and Johnston, 1995; Schweizer et al., 2008b). This stability rating is close to the operationally applied stability rating in Switzerland, which includes five classes and in addition considers weak layer properties and snowpack structure (Schweizer, 2007a; Schweizer and Wiesinger, 2001). The classification by Schweizer (2007a) was used in Techel and Pielmeier (2014) for an automatic assignment of stability based on RB score and release type (also five classes). As in Techel et al. (2020), we combined the two classes [..[72]]*very good* and *good* into one class called [..[73]]*good*.

[..[74]]**Extended Column Test (ECT)** [..[75]]results were classified relying on the classification recently suggested by Techel et al. (2020). Using a combination of crack propagation and the number of taps until failure initiation, four stability classes

[65] removed: 5,755 in Norway
[66] removed: 1,618 different
[67] removed: 33,262 in Switzerland
[68] removed: 6,610
[69] removed: snow
[70] removed: *very poor*, *poor*, *fair* and *good*
[71] removed: into
[72] removed: *very good* and *good*
[73] removed: *good*
[74] removed: Recently, a similar classification was proposed for the
[75] removed: (Techel et al., 2020)

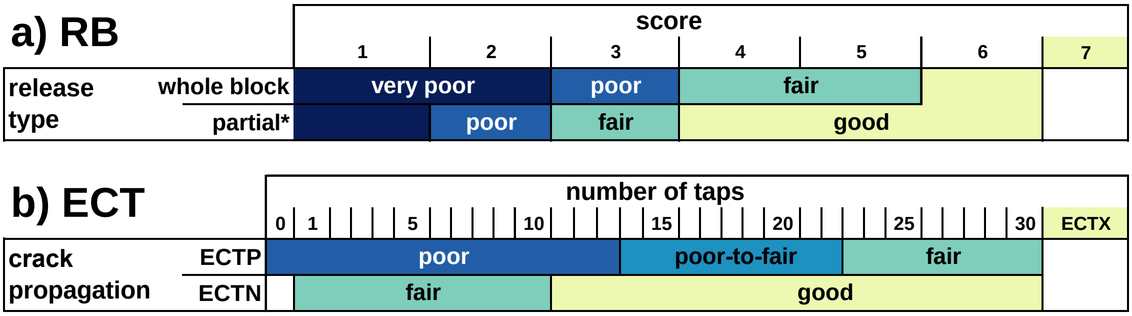


Figure 1. Stability classification of (a) Rutschblock test results (based on Schweizer (2007a); Techel and Pielmeier (2014)) and (b) Extended Column Test results (based on Techel et al. (2020)). * - part of block includes release types most of block and edge only

were defined (Fig. 1b). [..[76]]As the four stability classes for RB and ECT do not exactly line up[..[77]], we assigned the following four class labels to the four ECT classes: [..[78]]*poor*, *poor-to-fair*, *fair* and *good* (as in Techel et al., 2020). If failures in several weak layers were induced in a single stability test, the test results were classified for each failure layer. For this, we considered the failure as not relevant (rating the test result as [..[79]]*good*), if a failure layer was [..[80]]less than 10 cm below the snow surface [..[81]](as in Techel et al., 2020). The lowest stability class was retained for further analysis.

3.2 Simulation of [..[82]]snowpack stability distributions

The second factor contributing to avalanche hazard is the frequency of potential triggering locations, or of snowpack stability.

To determine the distribution of point snow instability within a defined region and at a given danger level many stability test results on a given day are in general needed (e.g. Schweizer et al., 2003). However, as we most often only had one stability test result on a given day, we followed an alternative approach. Assuming that a single test result is just one sample from the stability distribution on that day and that different days with the same danger level exhibit a range of similar stability distri-

[76] removed: Techel et al. (2020) compared the RB- and ECT-classifications shown in Fig. 1 to slope stability classified as either unstable or stable. They showed that with increasing stability class, the proportion of slopes rated as unstable decreased. In a data set with 30% unstable and 70% stable slopes, the four RB stability classes included 76%, 53%, 25% and 11% unstable slopes, while the four ECT stability classes included 57%, 40%, 23% and 15% unstable slopes. This indicates that the four stability classes

[77] removed: : The second RB class had a proportion unstable slopes (53%) similar to the first ECT class (57%), and the second ECT class (40%) had a value in-between the second and third RB classes (53% and 25%). To accommodate this misalignment,

[78] removed: *poor*, *poor-fair*, *fair* and *good*.It is of note, that ECT class *poor* also includes the weakest ECT results , which may be associated with *very poor* stability. To obtain the lowest RB or ECT stability class at each location, we proceeded as follows: If the depth of a weak layer failure was less than 5 cm below the snow surface

[79] removed: *good*

[80] removed: between 6 and

[81] removed: , we increased the stability rating by one step (e. g. from *very poor* to *poor*). If several failure planes were detected in a single stability test, the most unstable stability

[82] removed: snow

butions, we generated stability distributions by random sampling from the entire population of stability tests at a given danger level. Thus, we applied bootstrap sampling (Efron, 1979) and proceeded as follows (see also Fig. 2[..[83]]a and b):

- (i) We randomly selected [..[84]]n stability test results with replacement from the stability tests associated with the same danger level, resulting in a single bootstrap sample. We repeated this procedure [..[85]]B times for each danger level.
- (ii) For each of the [..[86]]B bootstrap samples, we calculated the proportions of [..[87]]*very poor*, *poor*, *fair* and *good* stability tests.

Bootstrap sampling, frequently used to estimate the accuracy of a desired statistic or for machine learning (Hastie et al., 2009), requires a sufficiently large number of replications [..[88]]B to be drawn. We used [..[89]]B = 2,500 for each danger level, resulting in 10,000 stability distributions in total.

The second important parameter when bootstrap sampling is the number [..[90]]n of stability tests drawn in each sample. Small values of [..[91]]n increase variance, and hence overlap between samples drawn from different danger levels, and reduce the resolution of the desired statistic (e.g. for [..[92]]n = 10, the resolution is 0.1, for [..[93]]n = 100 it is 0.01). [..[94]]Since nature is not as discrete as the danger levels [..[95]]suggest, we wanted both some overlap between our sampled stability distributions and a reasonably high resolution of our statistic. Unfortunately, there are no studies we can refer to concerning the amount of overlap that would be appropriate. [..[96]]We tested n={10, 25, 50, 100, 200, 1,000}.

These simulations are compared to a small number of days when more than 6 RB tests (N=41) or more than 6 ECT tests (N=31) were collected in the [..[97]]surroundings of Davos ([..[98]]Switzerland).

[83] removed: , steps 1 and 2
[84] removed: *n*
[85] removed: *B*
[86] removed: *B*
[87] removed: *very poor*, *poor*, *fair* and *good*
[88] removed: *B*
[89] removed: *B*
[90] removed: *n*
[91] removed: *n*
[92] removed: *n*
[93] removed: *n*
[94] removed: We wanted not only some overlap between distributions sampled from different danger levels - Nature
[95] removed: may suggest, but also a preferably
[96] removed: We tested *n*
[97] removed: surrounding
[98] removed: SWI

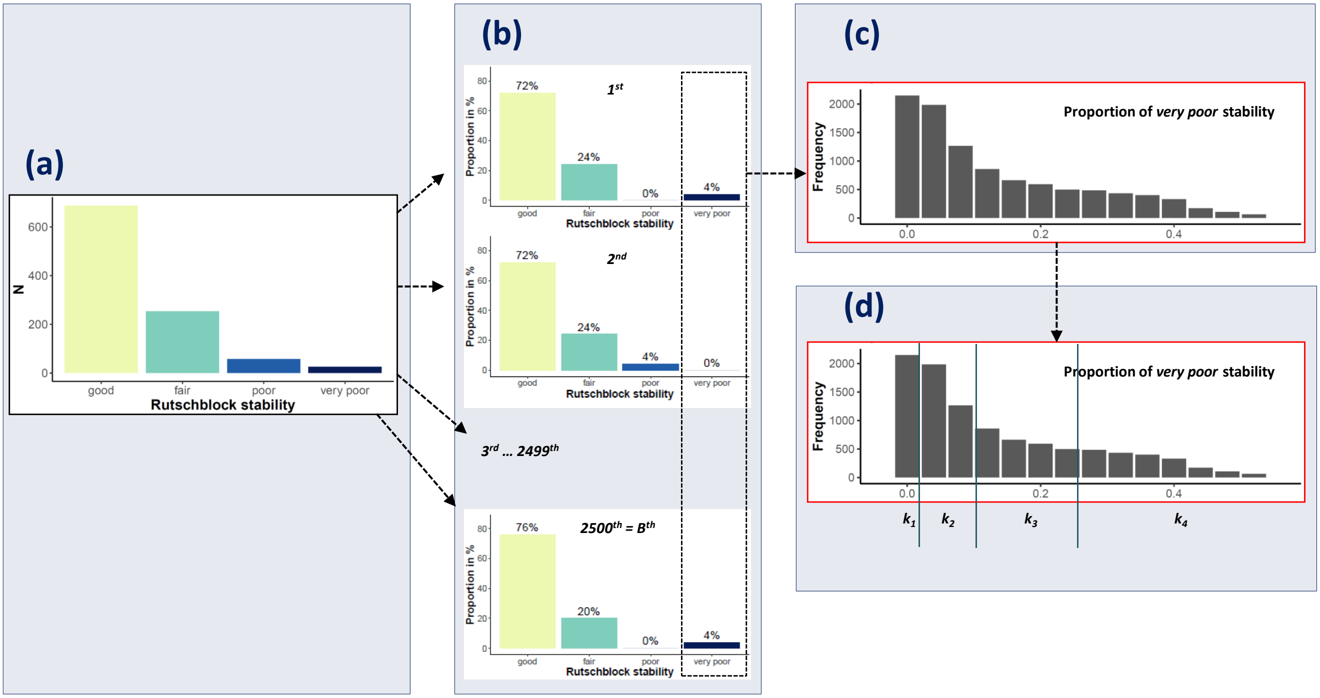


Figure 2. Schematic representation of the workflow for bootstrap sampling and frequency class definition. [..[99]]a - For each danger level, all stability [..[100]]ratings are combined. [..[101]]b - From the [..[102]]observed stability [..[103]]distributions ([..[104]]a), [..[105]]*n* tests are randomly sampled. This is repeated [..[106]]*B* = 2,500 times to obtain typical stability distributions for each of the four danger levels. [..[107]]c - The 4x 2,500 boot-strap samples are merged and the proportion of [..[108]]*very poor* rated stability tests per sample is plotted as a histogram. [..[109]]d - The statistics required for frequency class definitions are calculated and the [..[110]]*k* frequency classes defined. For details refer to the description in the Sections 3.2 and 3.3.

3.3 [..[111]]Snowpack stability and [..[112]]the frequency distribution of snowpack stability - approach to define frequency classes

Currently, [..[113]]neither well defined terms to describe frequency classes (such as *a few* or *many*) nor thresholds to differentiate between the classes exist. In the following, we therefore introduce a data-driven approach to define class intervals that we will use to describe the frequency of a certain [..[114]]snowpack stability class. We considered the following points:

[111]removed: Snow

[112]removed: its

[113]removed: no classification exists that provides thresholds for the frequency a certain snow stability class is present

[114]removed: snow

– Classes should be defined based on the [..[115]]snowpack stability class most relevant with regard to avalanche release, hence the frequency of the class [..[116]]*very poor*. Even though the focus is on the proportion of [..[117]]*very poor* snowpack stability, classes need to capture the entire possible parameter space, i.e. from very rare to virtually all (1 to 99%).

– The number of classes should reflect the human capacity to distinguish between them. We explored 3, 4 and 5 classes only, as these are the number of classes currently used to describe and communicate avalanche hazard and its components (e.g. three spatial distribution categories in the CMAH, four frequency terms in the EAWS matrix, five danger levels, five avalanche size classes).

– Classes must be sufficiently different to ease classification by the forecaster as well as communication to the user. And, if quantifier terms were assigned to these classes, these terms would need to unambiguously describe such increasing frequencies. An example of such a succession of five terms is [..[118]]*nearly none*, *a few*, *several*, *many* and *nearly all* (e.g. Díaz-Hermida and Bugarín, 2010).

Data-driven approaches for defining interval classes are numerous, and are described for instance for thematic mapping (e.g. Slocum et al., 2005) or for selecting histogram bin-widths (e.g. Evans, 1977; Wand, 1997). In general, the choice of class intervals should be appropriate to the observed data distribution. Approaches include, among others, splitting the parameter space into equal intervals, into intervals with an equal number of observations in each bin, or finding natural breaks in the data by minimizing the within-class variance while [..[119]]maximizing the distance between the class centers (e.g. Fisher-Jenks algorithm, Slocum et al., 2005). However, in our case, in which low values of the proportion of [..[120]]*very poor* stability are frequent and higher values rare, we made use of a geometric progression of class widths, considered most suitable for this type of distribution (Evans, 1977). Using this approach, we classified the data into [..[121]]k classes with class interval limits being $\{0, a, ab, ab^2, \ldots, ab^{k-1}, 1\}$, where a is the size (width) of the initial (lowest) class and b is a multiplying factor. According to Evans (1977), a data-driven calculation of b for the closed interval from 0 to [..[122]]1 can be given:

$$b = \left([..^{123}] \frac{1 - \mathrm{VP}_{\mathrm{med}}}{\mathrm{VP}_{\mathrm{med}}} \right)^{\frac{2}{k}}, \qquad (1)$$

[115] removed: snow
[116] removed: *very poor*
[117] removed: *very poor* snow
[118] removed: *nearly none*, *a few*, *several*, *many* and *nearly all*
[119] removed: maximising
[120] removed: *very poor*
[121] removed: *k*
[122] removed: 100

where [..[124]]*VP_med* is the median proportion of [..[125]]*very poor* stability, and [..[126]]k the number of classes preferred. This approach requires a suitable value of the number of classes [..[127]]k to be defined. Given [..[128]]k and b, the initial class width [..[129]]a is (Evans, 1977):

$$a = \frac{VP_{\text{med}}(1-b)}{1-b^{\frac{k}{2}}} \tag{2}$$

To derive a and b, we generated [..[130]]snowpack stability distributions, as outlined in the previous section (see also Fig. 2[..[131]]c and d).

3.4 Combining [..[132]]snowpack stability and [..[133]]the frequency of snowpack stability with avalanche size: a simulation experiment

When assigning a danger level, the information relating to [..[134]]snowpack stability and its frequency distribution needs to be combined with avalanche size. As we [..[135]]do not have data describing the three factors relating to the same day and region, we used a simulation approach by assuming that the distribution of the observed data represents the typical values and ranges at a specific danger level. Randomly sampling and combining a sufficient number of data points results in typical combinations of the three factors according to their presence in the data, but may also produce a small number of less likely combinations. We made use of the simulated frequency distributions of [..[136]]snowpack stability and their respective frequency class (Sect.s 3.2, 3.3). For each danger level, we [..[137]]combined the snowpack stability information with avalanche size by randomly selecting an avalanche size from the empirical avalanche size distribution for the given danger level (which will be shown in Sect. 4.2) .

4 Results

We first present the findings relating to the three contributing factors and their combination making use of Swiss Rutschblock and avalanche data (Sections 4.1 - 4.4). In a second step (Sect. 4.5), the findings regarding snowpack stability and

[124]removed: *VP_med*
[125]removed: *very poor*
[126]removed: k
[127]removed: k
[128]removed: k and b
[129]removed: a
[130]removed: snow
[131]removed: , steps 3 and 4
[132]removed: snow
[133]removed: its
[134]removed: snow
[135]removed: cannot link these three factors using data
[136]removed: snow
[137]removed: complemented the snow

avalanche size are compared with results obtained using different data sources: the ECT to assess snowpack stability and avalanche observations from Norway. Finally, to highlight the influence of the settings used for bootstrap-sampling and frequency classification, a sensitivity analysis is performed (Sect. 4.6).

4.1 [..[138]]Snowpack **stability**

4.1.1 Observed **Rutschblock** [..[139]]test **stability** distributions

We analyzed the [..[140]]stability distributions obtained with the RB test at danger levels 1-Low to 4-High (Fig. 3[..[141]]a). At 4-High, very few RB were observed. The proportion of [..[142]]*very poor* rated RB tests increased monotonically with increasing danger level from 2% at 1-Low to 38% at 4-High [..[143]](Fig. 3a). As a consequence, the combined proportion of [..[144]]*very poor* and *poor* rated tests also increased strongly from [..[145]]7% to 67%[..[146]], while the proportion of tests rated as [..[147]]*good* decreased accordingly ([..[148]]69% to 10%, Fig. 3a). These patterns were also confirmed when exploring the correlation between the RB stability class and danger level (Spearman rank-order correlation; $\rho = 0.4$, $p < 0.001$).
[..[149]][..[150]]

4.1.2 [..[151]]Frequency of *very poor* **stability**[..[152]]

[..[153]]Here, we describe the frequency of *very poor* stability based on sampling 25 Rutschblock tests and four frequency classes. Regarding the sampling and the class definition procedure refer to Sect.s 3.2 and 3.3, regarding the sensitivity of these settings on the results, refer to Sect. 4.6.

---

[138] removed: Snow

[139] removed: and ECT

[140] removed: distribution of RB and ECT results

[141] removed:).

[142] removed: *very poor*

[143] removed: .

[144] removed: *very poor* and *poor*

[145] removed: 8

[146] removed: (Fig. 3a)

[147] removed: *good*

[148] removed: 68

[149] removed: The proportion *poor* rated ECT increased from 11% at 1-Low to 28% at 3-Considerable, while the proportion of the two most unfavorable stability classes combined rose from 19% to 44%. At 4-High only the combined proportion of the two most unfavorable classes showed this increasing trend (61%, Fig. 3b). Again, a positive though weak correlation between stability rating and danger level was noted (ECT: $\rho = 0.22$, $p < 0.001$). ECTs were conducted more frequently at higher danger levels in Switzerland than in Norway (e.g. at 3-Considerable: 39% in SWI and 21% in NOR). The ECT stability class distributions for the two countries are shown in in the Appendix (Fig. **??**).

[150] removed: In both countries, very few RB and ECT were observed at 4-High (for instance ECT in NOR N = 6, in SWI N = 7, see also Fig. **??** in Appendix).

[151] removed: Simulated

[152] removed: distributions and frequency classification

[153] removed: As shown in the previous section, the RB stability classes *very poor* and *good* correlated better with the four danger levels than the ECT. For this reason, and because ECT seems not to separate well between *very poor* and *poor* stability, in the following we present results for RB only. The respective analysis for the ECT is shown as a supplement in the Appendix (Sect. **??**)

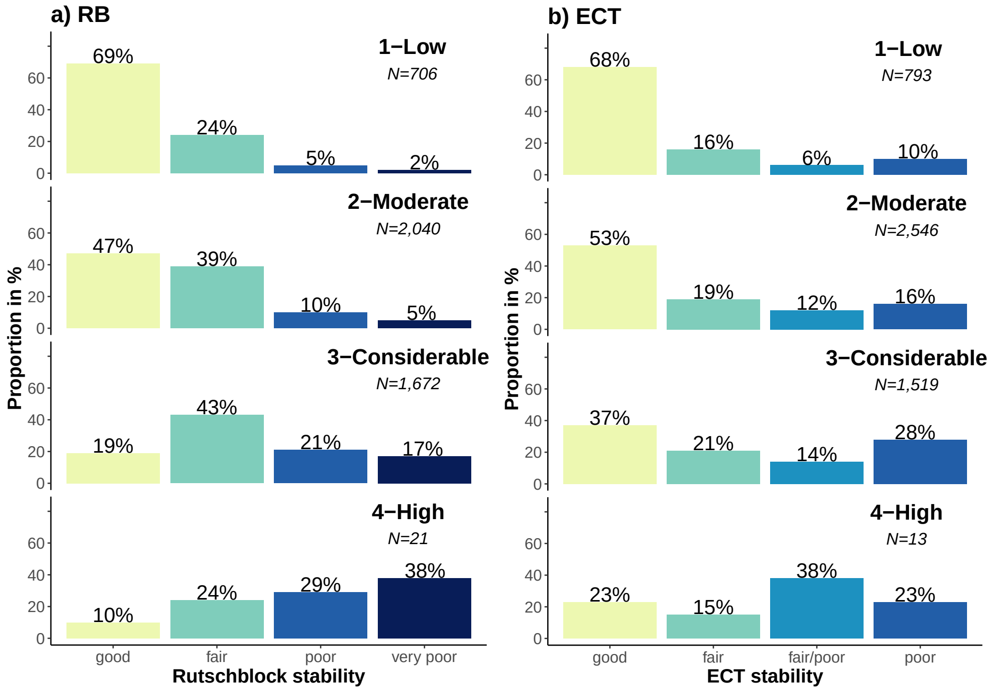


Figure 3. Distribution of stability ratings for the stability tests (a) Rutschblock (RB) and (b) ECT for danger levels 1-Low to 4-High. For the definition of the stability classes refer to Fig. 1 and Sect. 3.1. Note the small *N* for 4-High for both tests.

[..[154]][..[155]]Using four frequency classes, and labeling them *none or nearly none*, *a few*, *several* and [..[156]]*many*, the thresholds in the proportion *very poor* stability between frequency class labels were 0, 0.04 and 0.2, respectively (Tab. 2). This corresponded to a median proportion *very poor* stability observed in each frequency class of 0, 0.04, 0.12, 0.32, or, if expressed in the number of *very poor* Rutschblock test results, in 0, 1, 3 or 8 RB out of 25 drawn.

[154] removed: To obtain a variety of frequency distributions of point snow instability, we sampled stability tests as described in Sect. 3.2. As outlined there, one important parameter affecting such a sampling approach is the number of tests n drawn in each sample. We tested $n = \{10, 25, 50, 100, 200, 1000\}$. We visually checked the resulting histograms for the proportion of *very poor* stability (Figure **??**a-f in the Appendix) and visually checked for clusters in a two-dimensional context by considering the two extreme classes of the stability range, the proportion of *very poor* and *good* tests (Fig. **??**).

[155] removed: The distribution of the proportion of *very poor* stability was skewed towards lower proportions being more frequent than higher proportions (Figure **??**a-f). Increasing n impacted the number of modes detected in the histograms, with two or more modes being present when n reached values of about 50. This decrease of variance with increasing n, which leads to less overlap in samples drawn from different danger levels, is a characteristic of bootstrap sampling. Similar patterns can be noted in the two-dimensional context (Fig. **??**), with clusters not only becoming visually more and more pronounced with increasing n, but the overlap between danger levels reducing particularly at 3-Considerable

[156] removed: 4-High

[..[157]][..[158]]Large proportions of *very poor* stability (e.g. ≥ [..[159]][..[160]]0.5) occurred in less than 1% of the sampled distributions, despite sampling a comparably large number of tests from 4-High, where *very poor* stability test results are more frequent (Fig. 3a), and [..[161]]using a low n in each of the bootstrap samples, which increases the variation in the sampled proportions.

The correlation between the frequency class describing the frequency of *very poor* stability and the danger level was strong ([..[162]]$\rho = 0.81$, $p < 0.001$). [..[163]][..[164]]For instance, the frequency class *none or nearly none* was most frequently sampled from stability tests observed at 1-Low (61% of the cases). Similarly, the frequency class *a few* resulted most often when tests were sampled from 2-Moderate (47%), *several* from 3-Considerable (56%) and *many* from 4-High (86%, Fig. 4). Hence, when the proportion of *very poor* stability was classified as *many*, this was, by itself, a strong indicator that the danger level was 4-High.

[..[210]]

4.2 Avalanche size

Most avalanches [..[211]]in the Swiss data set were size 1 (Fig. 5a), except at 4-High, where a similar proportion of size 1, 2 and 3 avalanches were reported. The proportion of size 1 avalanches decreased with danger level from 64% to 32%, while the combined proportion of size 3 and 4 avalanches was highest at 4-High with 39%. Comparing the distributions at 1-Low

[157] removed: Comparing the bootstrap-sampled distributions with actually observed distributions of stability tests on the same day and in the same region (N = 41), showed that the distribution obtained using bootstrap-sampling reflected the variation in the observed distributions reasonably well (Fig. 9). The influence of a low number n of tests drawn in the bootstrap or tests actually collected in the field, is reflected in the large overlap between danger levels, but also variation within.

[158] removed: Relevant parameters for the definition of class intervals, as introduced in Sect. 3.3, are the respective median proportion of *very poor* stability VP_{med} and the number of classes k desired.VP_{med} showed a minor decrease with increasing resolution of the test statistic defined by n. It decreased from VP_{med} = 0.1 (n = 10) to VP_{med} = 0.08 (n

[159] removed: 25) . The initial (lowest) class width a, which decreased with k, was less than 0.03. Similarly, the factor b, scaling the increase in interval-width from one class to the next, decreased (b = {5.0, 3.4, 2.6}).

[160] removed: The thresholds of the class interval widths therefore depended primarily on k rather than n. The resulting interval bin-widths for an exemplary value of $n = 50$

[161] removed: k = {3, 4, 5} are shown in Table 2. In all cases, an additional class boundary would exist, generally at values between 0.5 and 0.9. As this class would remain empty most of the time, it is not shown in Table 2

[162] removed: n = 50, $\rho > 0.83$,

[163] removed: Even with n = 10, with a large amount of overlap between classes, the correlation between frequency class and danger level was significant (RB: $\rho > 0.7$, $p < 0.001$) The correlation increased with increasing k and individual classes classified best for the respective lowest and highest frequency classes.

[164] removed: Using k = 4 and the respective thresholds in Table 2, the median proportion *very poor* stability observed in each frequency class were 0, 0.04, 0.12, 0.32.

[210] removed: Comparison of observed (points, N = 41) and boot-strap sampled distributions (boxes) for the proportion of *very poor* (a, d), *very poor* and *poor* combined (b, e) and *good* stability tests (c, f), for two settings of the number n of tests drawn. When 7 to 15 RB tests were observed on the same day and within the same region, these are shown together with sampling using n = 10. When more than 16 tests were collected, these are shown together with n = 25. For n = 10 and *good* stability, the observed distributions were significantly different than the sampled distributions at 2-Moderate and 3-Considerable ($p < 0.05$, Wilcoxon rank sum test).

[211] removed: were of

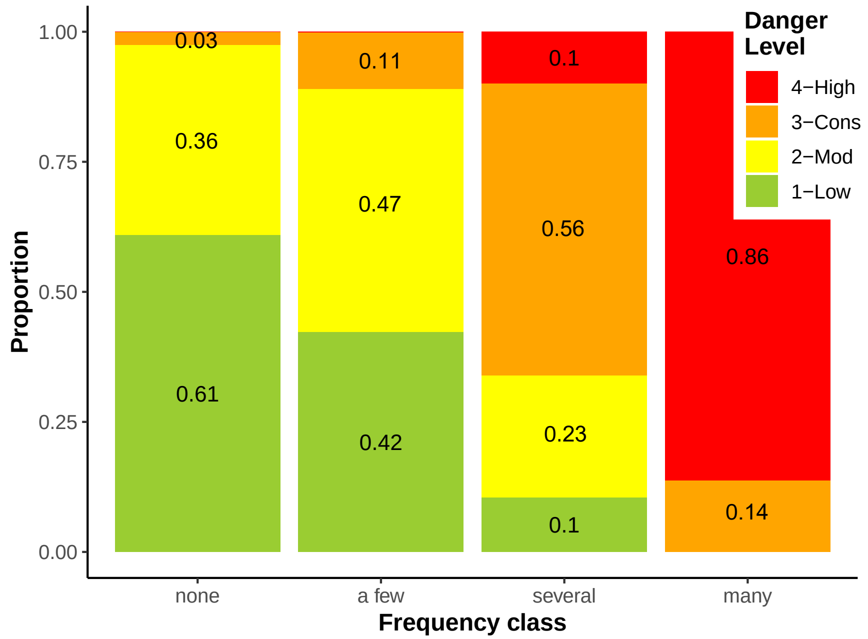


Figure 4. Distribution of *very poor* snowpack stability for 1-Low to 4-High.

to 3-Considerable shows that the most frequent avalanche size has little discriminating power to differentiate between danger levels. The median avalanche size was size 1 at 1-Low and 2-Moderate, [..[212]]size 1 to size 2 at 3-Considerable, and size 2 at 4-High [..[213]][..[214]](Fig. [..[215]]5a).

Considering the size of the largest reported avalanche per day and warning region showed [..[216]]that the largest avalanche per day and region was most frequently size 2 for 1-Low [..[217]]and 2-Moderate, a mix of size 2 and size 3 at 3-Considerable, [..[218]]and size 3 at 4-High [..[219]](Fig. 5b). The proportion of days when size 1 avalanches were the largest observed avalanche

[212]removed: and size

[213]removed: (Fig. 5a).

[214]removed: The size distributions of the reported avalanches differed between the countries: size 1 were proportionally more frequent in SWI than in NOR (30% vs. 17%), while size 4 avalanches had larger proportions in NOR (NOR 2%, SWI 1%; see also Fig. **??**a, b in the Appendix). Even though the proportion of reported size 1 avalanches decreased with increasing danger level, size 1 was clearly the most frequently reported size at danger levels 1-Low to 3-Considerable in SWI, and at 1-Low in NOR

[215]removed: **??**a, b).In Norway, size 2 avalanches were the most frequent size at 3-Considerable and at 4-High. At 3-Considerable in NOR, and at 4-High in both countries, about one third of the avalanches were size 3 or 4. In SWI, the distributions were almost identical for natural and for human-triggered avalanches. In NOR, there were proportionally more human-triggered size 1 avalanches than natural avalanches, for sizes 3 and 4 the opposite was the case.

[216]removed: again rather similar size distributions at 1-Low and 2-Moderate (Fig. 5b). The median

[217]removed: to

[218]removed: except

[219]removed: with size 3. However, the

Table 2. Frequency [..[165]]classification derived from the [..[166]]proportion of [..[167]]*very poor* stability ratings, using four frequency classes. The [..[168]]intervals are shown. $D(1^{st})$ and $D(2^{nd})$ indicate the most and second most frequent danger level the samples were drawn from[..[169]], [..[170]]respectively[..[171]]. Also shown is the classification of the combination of stability class and frequency class based on the two most frequent danger levels, denoted as letters A to F, which will be used in Fig.[..[172]]s 6 and 4.4. For [..[173]]class *none or nearly none* no letter is assigned, as the [..[174]]next higher stability class should be considered.

[..[175]]stability class	[..[176]]frequency class	[..[177]]interval* (n = 25)	danger level $D(1^{st})$	[..[178]]$D(2^{nd})$	letter in [..[179]]stability matrix
[..[180]]*very poor*	[..[181]]*many*	]0.2 - 1]	[..[182]]4	[..[183]][..[184]]3	A
	[..[185]]*several*	[..[186]]]0.04 - 0.2]	3	2	B
[..[187]]	[..[188]]*a few*	[..[189]]]0 - 0.04]	2	[..[190]]1	D
	[..[191]]*none or nearly none*	[0 - 0]	1	2	[..[192]]
poor	[..[193]]*many*	[..[194]]	2	[..[195]][..[196]][..[197]]3	[..[198]]C
[..[199]]	[..[200]]*several*	[..[201]]	2	1	D
	[..[202]]*a few*		1	2	E
	[..[203]]*none or nearly none*	[..[204]]	1	2	
fair	[..[205]]*many*	[..[206]]	1	2	E
	[..[207]]*several*	[..[208]]	[..[209]]1	–	F

* The thresholds indicated in the table are rounded according to the resolution of the test statistic, which depends on the number n of samples drawn in each bootstrap. Rounded to three decimal spaces the interval thresholds for n = 25 were: 0, 0.018, 0.062, 0.21, 1.

decreased [..[220]]significantly with increasing danger level [..[221]](from 33% to 1%, $p < 0.001$), while the proportion of days with at least one size 3 or size 4 avalanche increased [..[222]]significantly (from 20% to 78%, $p < 0.001$). At 4-High, [..[223]]almost 80% of the days had at least one avalanche of size 3 or 4 recorded.[..[224]]

The correlation between the size of the avalanche and the danger level was [..[225]]weak for the median size per day and warning region ($\rho = 0.15$, [..[226]]$p < 0.001$)[..[227]], but somewhat higher for the largest size ($\rho = 0.25$, [..[228]]$p < 0.001$).

[220]removed: considerably
[221]removed: ,
[222]removed: monotonically
[223]removed: more than 75
[224]removed: This proportion was higher in SWI (78%; Fig. **??**d) than in NOR (59%; Fig. **??**c).
[225]removed: weaker
[226]removed: p
[227]removed: than
[228]removed: p

[..[229]]Note that we did not explore days with no avalanches as we were interested in the size of avalanches, not their frequency. The frequency component is addressed using the frequency of locations with [..[230]]*very poor* stability as a proxy.

4.3 Combining [..[237]]the frequency of *very poor* stability and avalanche size

[..[238]] Assuming that the stability class *very poor* corresponds to the actual trigger locations, we combined the snowpack stability class, the frequency of this stability class and avalanche size. Hence, this combination considers all three key factors characterizing the avalanche danger level. [..[239]]

- [..[240]]
- [..[241]]
- [..[242]]

The resulting simulated data set contained the following information: [..[243]]*danger level, frequency class describing occurrence of very poor stability, largest avalanche size*. These data looked like the following, here for 1-Low:

[..[244]]*Sample 1: 1-Low, a few, largest avalanche size 1*

[..[245]]*Sample 2: 1-Low, none or nearly none, largest avalanche size 2*

[..[246]]*Sample 3: 1-Low, a few, largest avalanche size 1*

...

[..[247]]*Sample B: 1-Low - none or nearly none - largest avalanche size 1*

Tab. 3 summarizes the simulated data set. The most frequent combinations of the frequency class and avalanche size for each danger level were:

[229] removed: The number of reported avalanches per day and warning region increased with danger level from 2.5, 4, 5 to 8 for 1-Low to 4-High, respectively. It is of note

[230] removed: *very poor*

[237] removed: snow stability, its

[238] removed: Combining the snow stability class , its frequency

[239] removed: We explored a data set consisting of the Swiss RB and avalanche data only:

[240] removed: The number of frequency classes was set to $k = 4$ with $B = 2{,}500$ repetitions for each danger level. For this example, we selected the largest n with a uni-modal histogram ($n = 25$).

[241] removed: We classified the proportion of *very poor* stability using the thresholds and the four terms (*none or nearly none*, *a few*, *several* and *many*) for the 4 classes (Tab. 2).

[242] removed: Each sample was complemented with an avalanche size, drawn from the distribution of the largest avalanche size per day and warning region, for the respective danger level (Fig. 5b).

[243] removed: *danger level, frequency class describing occurrence of very poor stability, largest avalanche size*

[244] removed: *Sample 1: 1-Low, a few, avalanche size 1*

[245] removed: *Sample 2: 1-Low, none or nearly none, avalanche size 2*

[246] removed: *Sample 3: 1-Low, a few, avalanche size 1*

[247] removed: *Sample B: 1-Low* - none or nearly none - *avalanche size 1*

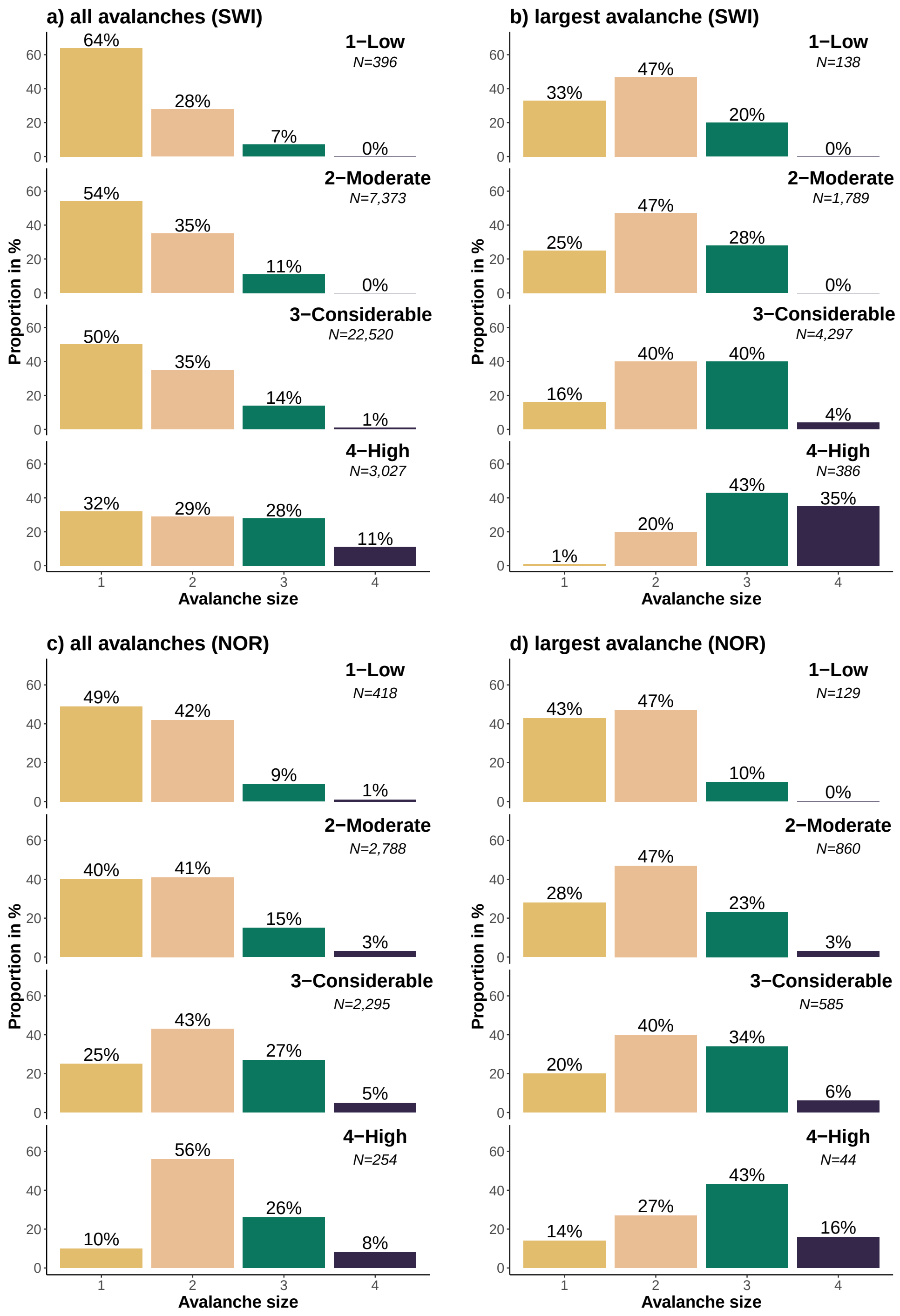


Figure 5. Size distribution of dry-snow avalanches, which released naturally or were human-triggered for danger levels 1-Low to 4-High, showing [..[231]]all avalanches [..[232]]([..[233]]a, c) and the largest reported avalanche per day and warning region [..[234]]([..[235]]b, d) in Switzerland (SWI[..[236]], upper row) and Norway (NOR, lower row).

Table 3. Table showing the combination of the frequency class of [..[261]]*very poor* snowpack stability and the largest avalanche size for the four danger levels. Frequencies are rounded to the full per cent value. Bold values hightlight the most frequent combination, "–" indicates that these combinations did not exist.

	1-Low				2-Moderate				3-Considerable				4-High			
size	*none**	*few*	*several*	*many*	*none**	*few*	*several*	*many*	*none**	*few*	*several*	*many*	*none**	*few*	*several*	*many*
1	17	10	5	–	8	9	7	0	0	2	12	2	–	0	0	1
2	**25**	16	7	–	16	**19**	15	0	1	3	**30**	5	–	0	3	18
3	11	8	3	–	8	9	9	0	1	3	**30**	6	–	0	6	**37**
4	–	–	–	–	0	0	0	0	0	0	3	1	–	0	5	30

* - *none or nearly none*

simulation setting: Rutschblock, avalanches (SWI), n = 25, k = 4, B = 2,500 per danger level

- 1-Low: [..[248]]*None or nearly none* locations with *very poor* stability (53% [..[249]]of sample) existed. The largest avalanches [..[250]]were size 2 (48%).
- 2-Moderate: [..[251]]*A few* locations with *very poor* stability (37%) [..[252]]were present. The typical largest avalanche [..[253]]was of size 2 (50%).
- 3-Considerable: [..[254]]*Several* locations with *very poor* stability (75%) [..[255]]existed. The typical largest avalanches [..[256]]were sizes 2 or 3 (79%).
- 4-High: [..[257]]*Many* locations with *very poor* stability (86%) [..[258]]existed. The typical largest avalanche [..[259]]was of size 3 (43%).

[..[260]]

[248]removed: *None or nearly none* locations with *very poor*
[249]removed:) exist
[250]removed: are
[251]removed: *A few* locations with *very poor*
[252]removed: are present. However, *none or nearly none* or *several* locations are of almost similar frequency (32-31%).
[253]removed: is
[254]removed: *Several* locations with *very poor*
[255]removed: exist
[256]removed: are
[257]removed: *Many* locations with *very poor*
[258]removed: exist
[259]removed: is
[260]removed: Tab. **??** summarizes the simulated stability class - frequency class combinations for all stability classes, and the respective most frequent and second most frequent danger level. The frequency class describing *very poor* stability was closely linked to one or two danger levels, which reflects Tab. 3. *Poor* stability as the most unstable stability class (when *none or nearly none very poor* existed), was generally associated with 2-Moderate or 1-Low. If both *very poor* and *poor* stability fell into the category *none or nearly none*, the resulting danger level was mostly 1-Low. The actual danger level distributions, summarized in Tab. **??**, are shown in the Appendix (Fig. **??**).

[..[262]] [..[263]][..[264]][..[265]][..[266]][..[267]][..[268]][..[269]][..[270]][..[271]][..[272]][..[273]][..[274]][..[275]][..[276]][..[277]][..[278]][..[279]][..[280]][..[281]][..[282]][..[283]][..[284]][..[285]][..[286]][..[287]][..[288]][..[289]][..[290]][..[291]][..[292]][..[293]][..[294]][..[295]][..[296]][..[297]][..[298]][..[299]][..[300]][..[301]][..[302]][..[303]][..[304]]

---

[262]removed: Summary of the simulated RB stability and frequency class combinations, and the respective most frequent danger level D(1st) and the second most frequent danger level D(2nd). Combinations of stability and frequency classes resulting in the same D(1st) and D(^{2nd}) are indicated by the same letters in the *group*. Letters are ordered according to rank-order of D(1st) and D(^{2nd}). If a frequency class is *none or nearly none*, the next higher stability class should be considered. The data behind this summary table is shown in Fig. **??** in the Appendix.

[263]removed: stability

[264]removed: frequency class

[265]removed: D(1st)

[266]removed: D(2nd)

[267]removed: group

[268]removed: *very poor*

[269]removed: *many*

[270]removed: 4

[271]removed: 3

[272]removed: A

[273]removed: *several*

[274]removed: 3

[275]removed: 2

[276]removed: B

[277]removed: *few*

[278]removed: 2

[279]removed: 1

[280]removed: D

[281]removed: *none**

[282]removed: *poor*

[283]removed: *many*

[284]removed: 2

[285]removed: 3

[286]removed: C

[287]removed: *several*

[288]removed: 2

[289]removed: 1

[290]removed: D

[291]removed: *few*

[292]removed: 1

[293]removed: 2

[294]removed: E

[295]removed: *none**

[296]removed: *fair*

[297]removed: *many*

[298]removed: 1

[299]removed: 2

[300]removed: E

[301]removed: *several*

[302]removed: 1

[303]removed: –

[304]removed: F

4.4 Data-driven look-up table for danger level assessment

Finally, we present a data-driven look-up table to assess avalanche danger (Fig. 6) using the simulations presented before. We used a step-wise approach, and two matrices as proposed by Müller et al. (2016) in the so-called Avalanche Danger Assessment Matrix (ADAM).[..[305]] The first matrix (Fig. 6[..[306]]a), which we refer to as [..[307]]*stability matrix*, combines snowpack stability and the frequency class of the most unstable stability class observed. Cell labels (letters A to E) [..[308]]in this matrix were assigned based on similar danger level distributions [..[309]]behind the respective stability class - frequency class combination [..[310]](Tab. 2). The letters reflect combinations with the most frequent and second most frequent danger levels in descending order with A being the highest and E the lowest danger levels. Letter F in Tab. 2, a rare occurrence in our data, was combined with letter E. For class *none or nearly none* no letter is assigned, as the next higher stability class should be considered. The mean simulated RB stability class distributions behind [..[311]]these cells are shown in Figure 4.4[..[312]]a.

The second matrix (Fig. 6[..[313]]b), which we refer to as [..[314]]*danger matrix*, combines snowpack stability and frequency with the largest avalanche size. The *danger matrix* displays the most frequent danger level (bold) and the second most frequent danger level [..[315]]characterizing this combination. If the second most frequent danger level was present more than 30% [..[316]]of the cases, the value is shown with no brackets, if present between 15 and 30% [..[317]]it is placed in brackets. To illustrate the actual danger level distributions behind this matrix, Figure [..[318]]4.4b summarizes the simulated data.

To derive the danger level, these two matrices can be used as follows:

1. In the *stability matrix* (Fig. 6a), the frequency class of *very poor* snowpack stability is assessed. If the frequency class was *none or nearly none*, the frequency class of *poor* snowpack stability is assessed. If the frequency class was again *none or nearly none*, the frequency class of *fair* snowpack stability is assessed.

2. The resulting letter is transferred to the *danger matrix* (Fig. 6b), where it is combined with the largest avalanche size (Fig. 6b).

[305] removed: In a first step, the most unfavorable snowpack stability class is combined with its frequency

[306] removed: , left matrix,

[307] removed: *stability matrix*). The resulting most unfavorable stability class - frequency class combination, which has a frequency greater than *none or nearly none* (>1.8%, Tab. 2), is retained.

[308] removed: shown in the *stability matrix* correspond to

[309] removed: related to this most unfavorable

[310] removed: according to Table **??**. The

[311] removed: the cells A-E

[312] removed: . In asecond step, the most appropriate cell describing stability and its frequency (letter in the *stability matrix*) is combined with avalanche size (

[313] removed: , right matrix,

[314] removed: *danger matrix*). The *danger matrix*

[315] removed: (if

[316] removed: :

[317] removed: : in brackets) characterizing this combination. Again, to

[318] removed: **??**

a) stability matrix

snowpack stability	frequency: none*	few	several	many
very poor	**	D	B	A
poor	**	E	D	C
fair	-	-	E	E
good	-	-	-	-

* none or nearly none

** if none, refer to next higher stability class

\- no data

C cell contains less than 1% of the data

b) danger matrix

stability matrix	largest avalanche size: 1	2	3	4
A	3 -4	4 (-3)	4	4
B	3 (-2/-1)	3 (-2)	3 (-2)	4 -3
C	2 (-3)	2 -3	3 -2	-
D	1 -2	2 -1	2 -1	3 (-2)
E	1	1 (-2)	1 (-2)	-

-3: >30%

(-3): 15-30%

Figure 6. Data-driven look-up table for avalanche danger assessment (similar to the structure proposed by Müller et al. (2016)). [..[319]](a, *stability matrix*) shows the combination of the frequency class of the most unfavorable snowpack stability class (columns) and the snowpack stability class (rows), (b, *danger matrix*) shows the largest avalanche size (columns) and the letters obtained in the stability matrix (rows).

3. The most frequent danger levels that were typical for this combination, are shown.

4.5 Comparison with other data sets

For the main results, presented in Sections 4.1 to 4.4, we relied on stability test results and avalanche data from Switzerland. In the following, we compare these stability and avalanche size distributions to other data sets.

4.5.1 Snowpack stability distributions: comparing RB with ECT results

Additionally to the RB, we explored stability distributions derived from ECT results and performed not only in Switzerland but also in Norway at 1-Low to 4- High (Fig. 3b).

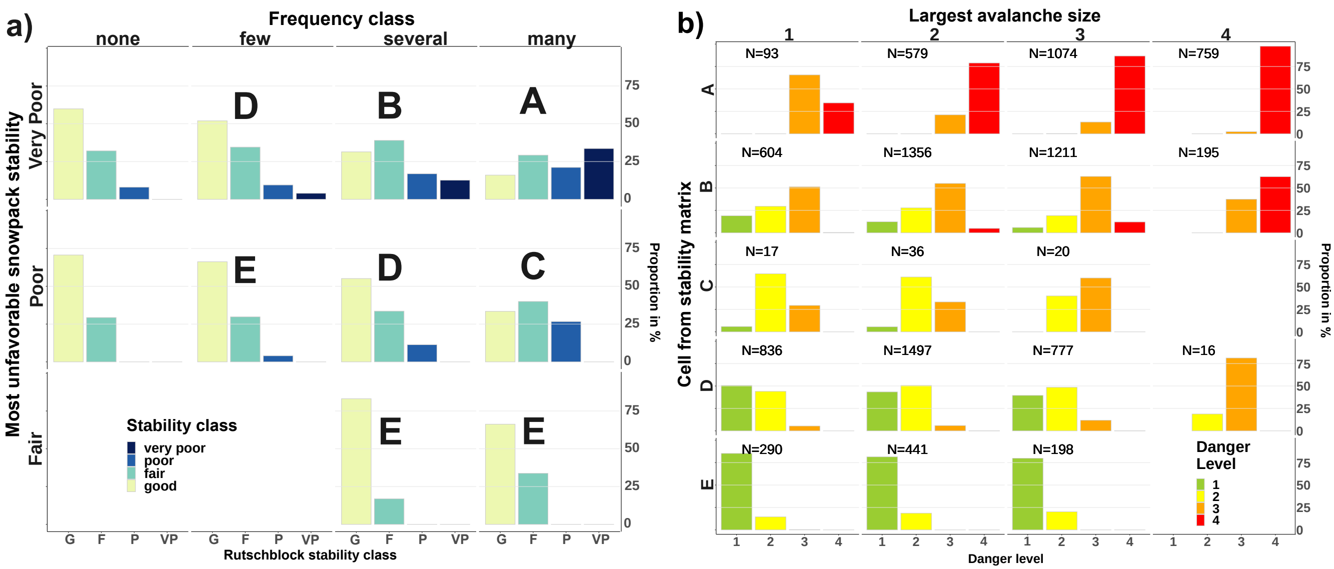
The proportion of *poor* rated ECT increased from 10% at 1-Low to 28% at 3-Considerable, while the proportion of the two most unfavorable stability classes combined rose from 16% to 42%. At 4-High, where very few ECTs were observed, only the combined proportion of the two most unfavorable classes showed this increasing trend (61%, Fig. 3b). Again, a positive though weak correlation between stability rating and danger level was noted (ρ = 0.22, p $<$ 0.001).

In comparison to the RB (Fig. 3a, Sect. 4.1.1), the ECT showed less distinct changes in the frequency of the most unstable and most stable classes between danger levels, and hence the correlation with the danger level was lower (ECT: ρ = 0.22 vs. RB: ρ = 0.4).

4.5.2 Avalanche size: comparing Swiss and Norwegian avalanche size distributions

The avalanche size distributions in Sect. 4.2, based on observations made in Switzerland (SWI; Fig. 5a, b), were compared to observations in Norway (NOR; Fig. 5c, d).

In Norway, size 1 was the most frequently reported size at 1-Low, while size 2 avalanches were the most frequent size



[..320][..321], the distribution of danger levels for combinations of the typical largest avalanche size and the [..322]letters obtained before in the [..323]*stability matrix* (A-E, Fig. 6[..324]a) are shown. The most frequent and second most frequent danger levels in each cell - avalanche size combination are shown in the [..325]*danger matrix* in the right part of Fig. 6b.

[..326][..327], the distribution of danger levels for combinations of the typical largest avalanche size and the [..328]letters obtained before in the [..329]*stability matrix* (A-E, Fig. 6[..330]a) are shown. The most frequent and second most frequent danger levels in each cell - avalanche size combination are shown in the [..331]*danger matrix* in the right part of Fig. 6b.

Figure 7. [..332]Data behind the [..333]matrices shown in Figure 6. The layout of the columns and rows is identical to Fig. 6. The left figure (a) shows the mean simulated stability distributions behind the *stability matrix* (Fig. 6a). Letters describe cells with the corresponding most frequent and second most frequent danger level[..334]. [..335]In the right figure (b)[..336]
[..337][..338], the distribution of danger levels for combinations of the typical largest avalanche size and the [..339]letters obtained before in the [..340]*stability matrix* (A-E, Fig. 6[..341]a) are shown. The most frequent and second most frequent danger levels in each cell - avalanche size combination are shown in the [..342]*danger matrix* in the right part of Fig. 6b.

at 3-Considerable and 4-High (Fig. 5c). The proportion of reported size 1 avalanches decreased with increasing danger level (from 49% to 10% from 1-Low to 4-High), while size 3 and 4 avalanches increased proportionally (from 10% to 34%). Similarities between Switzerland and Norway included a decreasing proportion of size 1 avalanches and increasing proportions of size 3 or 4 avalanches with danger level. Notable differences were primarily related to the proportion value: Considering all reported avalanches, size 1 avalanches were proportionally less frequent in Norway than in Switzerland (NOR 17%, SWI 30%), while size 4 avalanches had larger proportions in Norway (NOR 2%, SWI 1%). This difference is likely linked to a lower reporting rate of smaller avalanches in Norway.

Considering the largest avalanche per day and warning region in Norway (Fig. 5d) showed similar trends in the size distributions as in Switzerland (Fig. 5b). The proportion of size 1 avalanches decreased with increasing danger level, while size 3 and 4 avalanches increased. Size 2 avalanches were the most frequent at 1-Low to 3-Considerable. At 4-High, the largest reported avalanche was typically a size 3 avalanche. Differences between the Norwegian and the Swiss data were again primarily related to the proportion values. For instance, the proportion of size 1 avalanches as the largest reported avalanche decreased from 1-Low to 4-High from 43% to 14% in Norway, compared to 33% to 1% in Switzerland. Differences were also observed for the proportion of size 3 and 4 avalanches as the largest observed avalanche: their proportion increased from 1-Low to 4-High from 10% to 59% in Norway, and from 20% to 78% in Switzerland.

4.6 Bootstrap sampling and frequency class definitions - sensitivity analysis

4.6.1 Bootstrap sampling

To obtain a variety of frequency distributions of point snow instability, we sampled stability ratings as described in Sect. 3.2. As outlined there, one important parameter affecting such a sampling approach is the number of stability ratings n drawn in each sample.

The results shown in Sections 4.1, 4.3 and 4.4 were based on $n = 25$. In addition, we explored the effect of sample size and tested $n = \{10, 25, 50, 100, 200, 1000\}$. Histograms showing the simulated proportion of *very poor* stability for various n (two examples are shown in Fig. 8a and c) were checked for multi-modality (visual inspection and applying the *modetest* (Ameijeiras-Alonso et al., 2018)). Furthermore, the resulting simulations were visually checked for clusters in a two-dimensional context by considering the two extreme stability classes, the proportion of *very poor* and *good* stability ratings (Fig. 8b and d).

The distribution of the proportion of *very poor* stability was skewed towards lower proportions being more frequent than higher proportions (Fig. 8a and c). Increasing n impacted the number of modes detected in the histograms, with two or more modes being present when n reached values of about 50. This decrease of variance with increasing n, which leads to less overlap in samples drawn from different danger levels, is a characteristic of bootstrap sampling. Similar patterns can be noted in the two-dimensional context (Fig. 8b and d). Clusters not only become visually more and more pronounced with increasing n, but the overlap between danger levels decreases particularly at 3-Considerable and 4-High.

When introducing the bootstrap-sampling approach to create a range of plausible stability distributions (Sect. 3.2), we had to assume that a single stability rating is just one sample from the stability distribution on that day and that different days with the same danger level exhibit a range of similar stability distributions. Referring to Fig. 8c, which shows the proportions of *very poor* and *good* stability of the 10,000 simulated distributions with $n = 25$, it can be noted that indeed a range of typical distributions was obtained for the four danger levels. For instance, at 3-Considerable the range of the simulated distributions was wide: 11% of the samples drawn had $\geq$ 8% (frequency classes *several* or *many*) *very poor* and $\leq$ 4% (*a few* or *none*) *good* tests results, while 7% of the samples drawn had $\leq$ 4% (*a few* or *none*) *very poor* and 24% (*many*) *good* tests results.

Comparing the bootstrap-sampled distributions with actually observed distributions of stability ratings on the same day and in the same region ($N = 41$), showed that the distribution obtained using bootstrap-sampling reflected the variation in the observed distributions reasonably well (Fig. 9). The influence of a low number n of tests drawn in the bootstrap or from the distribution of stability ratings actually collected in the field, is reflected in the large overlap between danger levels, but also variation within.

4.6.2 Frequency class definition

Relevant parameters for the definition of class intervals, as introduced in Sect. 3.3, are the respective median proportion of *very poor* stability VP_{med} and the number of classes k desired.

VP_{med} was affected by the resolution of the test statistic for very low values of n. For instance, for $n = 10$, the resolution was 0.1 and VP_{med} was 0.1. For all other n tested, VP_{med} was 0.08 or 0.085, despite large differences in the resolution of the test statistic (e.g. 0.04 for $n = 25$ and 0.005 for $n = 200$). The number of classes k desired, however, influenced the class interval definition as described in Sect. 3.3, as both the initial (lowest) class width a and the factor b, scaling the increase in interval-width, decreased with k. However, for $n \leq 50$ and all k tested, the initial (lowest) class contained only values for the proportion of *very poor* equaling 0. A value of $k = 4$ seemed most suitable, as the resulting three lower class intervals would contain values for sampling with $n > 10$. In all cases, an additional class would exist, generally at values between 0.5 and 0.9. As this class would remain empty most of the time, this class was merged with the respective lower one, thus expanding the upper interval limit of class *many* to 1.

The correlation between the frequency class and the danger level increased with increasing k, and was strong even with $n = 10$, with a large amount of overlap between classes ($\rho > 0.7$, $p < 0.001$).

5 Discussion

In the following, we discuss our findings in the light of potential uncertainties linked to the data (Sect. 5.1) and methods selected (Sect. 5.2). Furthermore, we compare the results to currently used definitions, guidelines and decision aids used in regional avalanche forecasting (Sect. 5.3).

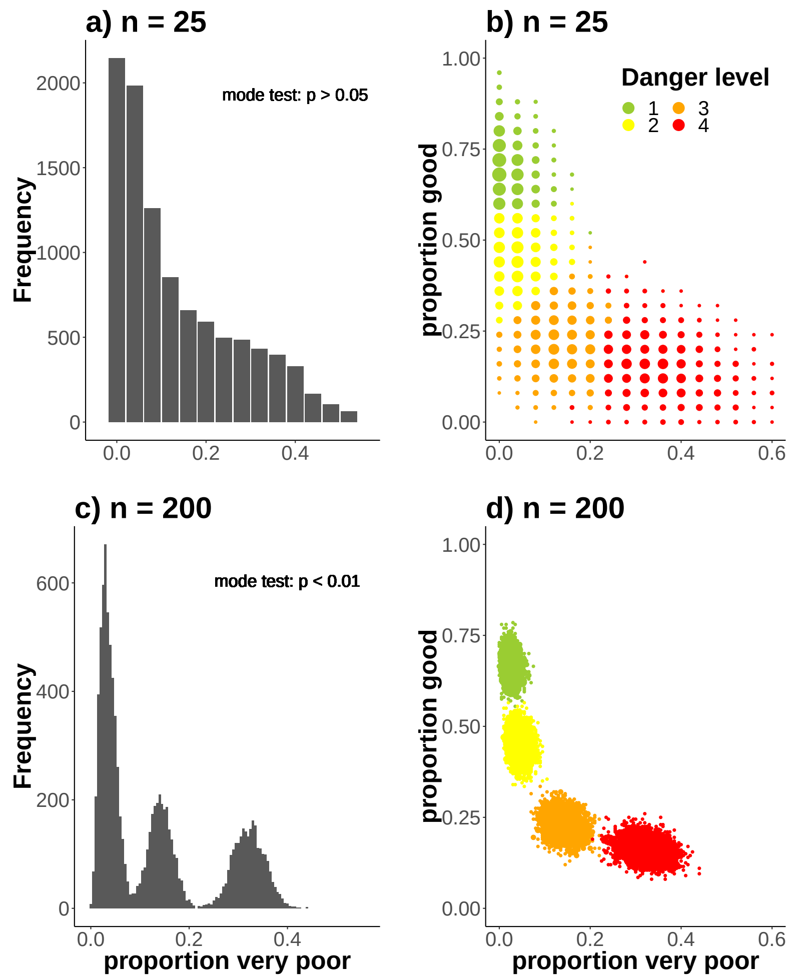


Figure 8. Simulated proportions of *very poor* and *good* derived from RB tests for different number of samples n drawn in each of the bootstraps (upper row a and b: n = 25, lower row c and d: n = 200). In the histograms (a, c) the proportion of *very poor* stability is shown, in the scatterplots (b, d) the most frequent danger level for a combination of *very poor* and *good* stability is shown. - The larger the number of samples n drawn, the more the data became multi-modal and clustered around the means of each danger level. This is indicated by the p-value (*modetest*, median p-value of 10 repetitions, Ameijeiras-Alonso et al., 2018) in a and c.

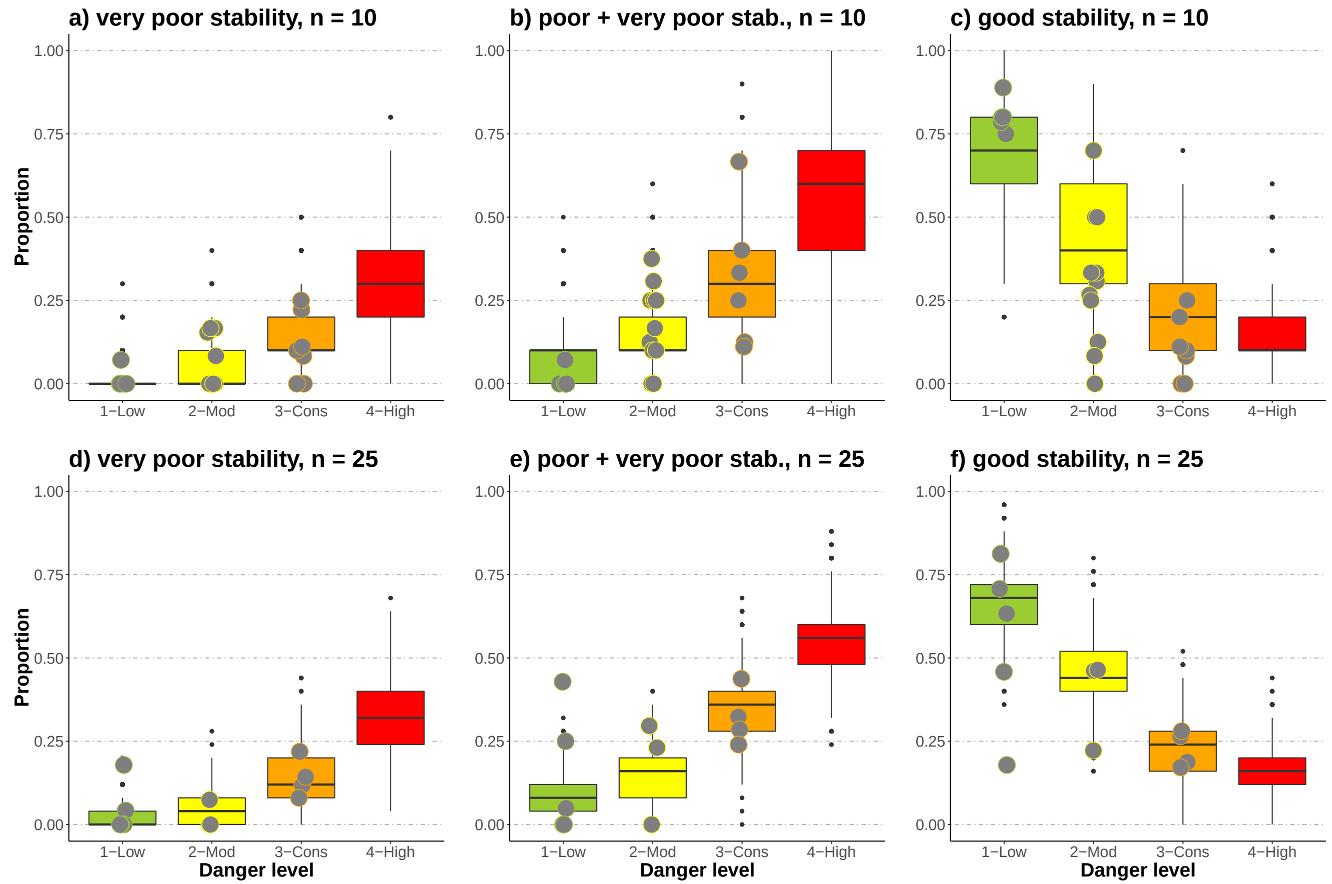


Figure 9. Comparison of observed (points, N = 41) and boot-strap sampled distributions (boxes) for the proportion of *very poor* (a, d), *very poor* and *poor* combined (b, e) and *good* stability tests (c, f), for two settings of the number n of tests drawn. When 7 to 15 RB tests were observed on the same day and within the same region, these are shown together with sampled distributions using n = 10. When more than 16 tests were collected, these are shown together with sampled distributions using n = 25. For n = 10 and *good* stability, the observed distributions were significantly different than the sampled distributions at 2-Moderate and 3-Considerable ($p < 0.05$, Wilcoxon rank sum test).

5.1 Data

5.1.1 Stability tests

Stability tests conducted by specifically trained observers are often performed at locations where the snowpack stability is expected to be low, though in an environment where spatial variability of the snowpack can be high (e.g. Schweizer et al., 2008a). Moreover, in most cases just one stability test was performed by an observer, not permitting us to judge whether this test was representative for the conditions of the day. However, the overall distributions of the stability ratings derived from RB or ECT results (Fig. 3), highlight the increase of locations with low snowpack stability with increasing danger levels.

At 4-High, stability test data were limited, as these situations are not only rare and temporally often short-lived, but also since backcountry travel in avalanche terrain is dangerous and therefore not recommended. As a consequence, not only considerably fewer field observations were made, but these were also dug on less steep slopes at lower elevation, which may potentially underestimate snow instability.

5.1.2 Avalanche observations

We relied on observational data recorded in the context of operational avalanche forecasting. This means that differences in the quality of single observations are possible. For instance, variations in both the estimation of avalanche size (Moner et al., 2013) as well as in locally assessing the avalanche danger level (Techel and Schweizer, 2017) have been noted. Furthermore, observations of avalanche activity often have a temporal uncertainty of a day or more, especially in situations with prolonged storms and poor visibility that often accompany a higher danger level. We addressed these issues by filtering the most extreme 2.5% of the avalanche observations for each danger level.

Completeness of observations is another issue. Avalanche recordings are generally incomplete, in the sense that not all avalanches within an area are recorded as well as that single observations may lack information, e.g. on size. However, the size distributions (Fig. 5) reflect that smaller avalanches are more frequent, which was also observed in previous studies where other recording systems were applied such as recording of avalanches by snow safety staff and the public (Logan and Greene, 2018), manual mapping of avalanches (Hendrikx et al., 2005; Schweizer et al., 2020) or satellite-detection of avalanches (Eckerstorfer et al., 2017; Bühler et al., 2019). Still, smaller avalanches may be underrepresented compared to larger avalanches - as was the case for instance for size 1 avalanches in [..[343]]the Norwegian data set (Fig. [..[344]]5c). This underreporting may depend on the relevance to an observer, but also on the ease of recording or limitations set by the recording of numerous smaller avalanches. [..[345]]Since we did not primarily use the number of avalanches, but instead focused on the largest avalanche per day and warning region, we expect this limitation to be less relevant.

[343] removed: NOR
[344] removed: **??**a
[345] removed: As

[..[346]]To address potential bias in observations linked to [..[347]]Swiss observational standards (e.g. Techel et al., 2018), we [..[348]]compared findings with data from Norway. This brought additional challenges, like a different structure or content of the observational data, which required us to make further assumption (e.g. for counting the number of avalanches reported in forms when several avalanches were reported together in Norway). [..[349]]However, the largest avalanche size per day and warning region (Fig. [..[350]]5b and d) showed similar overall patterns across countries, with increasing frequencies of [..[351]]*very poor* stability and increasing avalanche size with increasing danger level.

[..[352]]

Finally, stability test results, avalanche observations and local danger level [..[353]]estimates are generally not independent from each other, as often the same observer provided all this information. However, as shown by Bakermans et al. (2010), stability test results – compared to other observations - have relatively little influence on a local danger level estimate, while [..[354]]observations of natural or artificially triggered avalanches are [..[355]]unambiguous evidence of instability and may thus raise the quality of the local assessment.

5.2 Methods

5.2.1 Stability classification of RB and ECT

We relied on existing RB and ECT classifications (RB: Schweizer and Wiesinger (2001); Schweizer (2007a); ECT: Techel et al. (2020), Fig. 1). While the RB classification scheme is well-established in the operational assessment of snow profiles in the Swiss avalanche warning service, the classification of ECT into four stability classes has only recently been proposed by Techel et al. (2020). They showed that for a large data set of pairs of ECT and RB performed in the same snow pit, both classifications provided good correlations to slope stability. However, as shown by Techel et al. (2020), the most favorable and the most unfavorable RB stability classes captured slope stability better than the respective ECT classes, indicating a lower agreement between slope stability and ECT results compared to the RB. This was our argument for not fully aligning the four RB and

[346]removed: Stability tests conducted by specifically trained observers are often performed at locations, where the snowpack is expected to be weak, though in an environment where spatial variability of the mountain snowpack can be high (e.g. Schweizer et al., 2008a). Additionally, in most cases just one stability test was performed by an observer, not permitting us to judge whether this test was representative for the conditions of the day. However, the overall distributions of the stability test results, regardless whether RB or ECT were considered (Fig. 3), highlight the increase of locations with low snow stability at higher danger levels.

[347]removed: a specific warning service

[348]removed: used data from two different warning services (NOR, SWI)

[349]removed: The stability distributions of the ECT (Fig. **??**) or the

[350]removed: **??**

[351]removed: *very poor*

[352]removed: At 4-High, stability test data were limited, as these situations are not only rare and temporally often short-lived, but also since backcountry travel in avalanche terrain is dangerous and therefore not recommended. As a consequence, not only considerably fewer field observations were made, but these were also dug on less steep slopes at lower elevation, which may potentially underestimate snow instability.

[353]removed: assessment

[354]removed: numerous

[355]removed: a clear indication for a higher danger level

ECT stability classes and is supported by our findings: The RB stability class distributions changed more [..[356]]prominently from 1-Low ([..[357]]69% *good* stability, 2% [..[358]]*very poor*) to 4-High (10% [..[359]]*good*, 38% [..[360]]*very poor*) than the most favorable and unfavorable ECT stability classes (1-Low: [..[361]]68% *good* stability, 10% *poor*, 4-High: [..[362]]23% *good*, 23% [..[363]]*poor*).

5.2.2 Simulation of stability distributions

We could not rely on a large number of stability [..[364]]tests observed on the same day in the same region, which is a general problem in avalanche forecasting. We therefore generated stability distributions using re-sampling methods (Sect. 3.2) and by selecting sampling settings which lead to considerably overlapping distributions (Fig. 9). We argue that some overlap in stability distributions would characterize the large variability of avalanche conditions. However, [..[365]]we do not know which number n of stability tests drawn captures the variation best[..[366]]. We suppose that a combination of (labour-intensive) field measurements combined with spatial modeling in a large variety of avalanche conditions will be necessary to shed some light on this question [..[367]](e.g. Reuter et al., 2016, for a small basin in Switzerland). Alternatively, spatial modeling of the snowpack, provided that a robust stability parameter can be simulated, would be required.

Repeated sampling from small data sets may underestimate the uncertainty associated with a metric, but more importantly, the question must be raised, whether the sample reflects the population well. While at 1-Low to 3-Considerable, we sampled from between 700 and [..[368]]2000 RB stability ratings per danger level, at 4-High the [..[369]]number of observations was very small ([..[370]]$N = 21$[..[371]]). Hence, both the data shown in Fig. 3 as well as the sampled stability distributions for this danger level are more uncertain than for the other danger levels. While the combined number of locations with [..[372]]*very poor* and *poor* stability increased, and those with [..[373]]*good* stability decreased at 4-High (Fig. 3), judging whether the observed tests reflect the population well is difficult. [..[374]]Unfortunately, we are not aware of other studies, which have explored the

[356]removed: pronounced
[357]removed: 68% *good*
[358]removed: *very poor*
[359]removed: *good*
[360]removed: *very poor*
[361]removed: 60% *good* stability, 8% *poor*
[362]removed: 15% *good*
[363]removed: *poor*
[364]removed: distributions
[365]removed: which *n*
[366]removed: , we do not know. We suspect
[367]removed: (e.g. **?**, for a small basin in Switzerland)
[368]removed: 2800 RB or ECT tests
[369]removed: respective
[370]removed: RB: N
[371]removed: , ECT: N = 13
[372]removed: *very poor* and *poor*
[373]removed: *good*
[374]removed: For instance, when exploring the very small ECT data sets for the two countries individually (NOR: N = 6, SWI: N = 7; Fig. **??**), the uncertainties associated with very small data sets are highlighted.

[..[375]]snowpack stability distribution in a region at 4-High based on many tests, and therefore have no comparison. Even on 7 Feb 2003, one of the days of the verification campaign in the region of Davos/[..[376]]Switzerland (Schweizer et al., 2003), the forecast danger level 4-High was [..[377]]«verified» to be between 3-Considerable and 4-High (Schweizer, 2007b). On this day, 14 Rutschblock tests were observed. 36% of these were either [..[378]]*very poor* or *poor*, thus being close to the average values noted for 3-Considerable (Fig. 3a). [..[379]]We did not consider these data, as we [..[380]]did not analyze data when for intermediate danger levels.

Comparing the distributions of our [..[381]]snowpack stability classes with the characteristic stability distributions obtained during the verification campaign in Switzerland in 2002 and 2003, some differences can be noted (Swiss RB data)[..[382]]. For instance, the proportion of [..[383]]*very poor* and *poor* combined was at 2-Moderate about 15% and at 3-Considerable about 40%, which is lower than [..[384]]findings by Schweizer et al. (2003) (20-25% and about 50%, respectively). At 1-Low, about 70% of the RB tests were classified as [..[385]]*good*, while Schweizer et al. (2003) noted about 90% of the profiles to have [..[386]]*good* or *very good* stability. This suggests a smaller spread in the distribution of our automatically assigned stability classes, compared to the manual classification approach according to Schweizer and Wiesinger (2001).

5.2.3 Classification of [..[387]]snowpack stability frequency distributions

In addition to simulating [..[388]]snowpack stability distributions using a re-sampling approach, we [..[389]]developed a data-driven classification of the proportion of [..[390]]*very poor* stability tests. Our approach shows that the number [..[391]]*n* drawn for each bootstrap has little influence on class interval definitions, as long as the resolution of the test statistic is sufficiently high. Class thresholds are primarily defined by the central tendency of the distribution, in our case the median proportion of

[375] removed: snow
[376] removed: SWI
[377] removed: only
[378] removed: *very poor* or *poor*
[379] removed: It is of note that these datawas not considered in our analysis
[380] removed: analyzed only stability data when only one specific danger level was locally estimated
[381] removed: snow
[382] removed: :
[383] removed: *very poor* and *poor*
[384] removed: Schweizer et al. (2003)'s findings
[385] removed: *good*
[386] removed: *good* or *very good*
[387] removed: snow
[388] removed: snow
[389] removed: attempted for the first time
[390] removed: *very poor*
[391] removed: *n*

[..[392]]*very poor* stability tests VP_{med}, and by the number of classes preferred [..[393]][..[394]][..[395]][..[396]][..[397]][..[398]][..[399]]*k*. Assigning a class to the proportion of [..[400]]*very poor* stability, however, was affected by [..[401]]*n* due to the fact that [..[402]]*n* influences both the resolution of the statistic and the variance[..[403]]. This means that conceptually we can think in frequency classes, as long as class interval boundaries are scaled according to the data used. This need to scale class intervals according to the data-source, however, also implies that there is no unique set of values which could be used. Furthermore, the simulated stability distributions indicate that the focus is on optimizing class definitions to values between 0 and 40% when relying on stability tests, rather than the entire potential parameter space (0-100%).

The preferred number of classes [..[404]]*k* may depend on a number of factors. We suggest that defining [..[405]]*k* should be guided by keeping classes as distinguishable as possible - for instance by addressing the frequently occurring low proportions of [..[406]]*very poor* stability on one side and the rarely observed large proportions of [..[407]]*very poor* stability on the other side, and potentially a class covering the in-between. Furthermore, these terms must be unambiguously understandable to the user, regardless of language.

5.3 Data interpretation

5.3.1 Snowpack stability and its frequency

We showed an increasing frequency (or number of locations[..[408]]) of *very poor* snowpack stability with increasing danger level, [..[409]]in line with previous studies exploring point [..[410]]snowpack stability within a region or small basin [..[411]](Schweizer et al., 2003; Reuter et al., 2016) or the number of natural and human-triggered avalanches within a region (e.g

[392]removed: *very poor* stability tests VP_{med}

[393]removed: *k*.In the case of a low resolution of the test statistic the class interval widths should be scaled according to the number of distinct measurements (Evans, 1977). In other words, with $n = 10$ and $b = 2$, class interval widths would be

[394]removed: 0

[395]removed: ,

[396]removed: 0.1, 0.2

[397]removed: ,

[398]removed: 0.3, 0.4, 0.5, 0.6

[399]removed: , . . .

[400]removed: *very poor*

[401]removed: *n*

[402]removed: *n*

[403]removed: , while the overall class assignment was less dependent on *n*

[404]removed: *k* cannot be defined and

[405]removed: *k*

[406]removed: *very poor*

[407]removed: *very poor*

[408]removed: with *very poor* snow

[409]removed: which is

[410]removed: snow

[411]removed: (Schweizer et al., 2003; **?**)

Schweizer et al., 2020). [..[412]]Furthermore, we showed that high proportions of *very poor* stability (≥ 0.3) were comparably rare (15% of the simulated distributions). Even at 4-High, less than 4% of the distributions had proportions of *very poor* stability ≥ 0.5.

We explored snowpack stability using RB and ECT, which describe the stability at a specific point. However, within a slope or a region, point [..[413]]snowpack stability is variable (e.g. Birkeland, 2001; Schweizer et al., 2008a). [..[414]]In avalanche forecasts this can be expressed by the frequency a certain stability class exists [..[415]]and by additionally describing the locations more specifically. When describing the avalanche danger level in a region, snowpack stability and [..[416]]the frequency distribution of snowpack stability must therefore be considered. We suggest that primarily the frequency of the lowest stability class is relevant for [..[417]]assigning a danger level, as this stability class combined with [..[418]]the frequency of this stability class describes the minimal trigger needed to release an avalanche and how frequent these most unstable locations exist within a region. [..[419]]These two factors must therefore be assessed in combination for all aspects and elevations[..[420]]. Furthermore, the specific description of triggering locations, for instance [..[421]]*at treeline* or *in extremely steep terrain*, may provide an indication where in the terrain these locations may exist more frequently within its frequency class. Even though different terms are used, both the EAWS-Matrix [..[422]](EAWS, 2017) and the CMAH (Statham et al., 2018a) first combine snowpack stability and its frequency distribution, before avalanche size is considered. The respective terms which were used are the 'load' (trigger) and the 'distribution of hazardous sites' in the EAWS-Matrix and the 'sensitivity to triggers' and 'spatial distribution' leading to the 'likelihood of avalanches' in the CMAH.

We explored primarily the frequency of the stability class [..[423]]*very poor*, which is most closely related to actual triggering points. However, as several studies have shown, even when stability tests suggested instability, often only some of the slopes were in fact unstable and released as an avalanche (e.g. Moner et al., 2008; Techel et al., 2020). Thus, depending on the data used to define [..[424]]*very poor* stability, for instance whether stability tests are used or natural avalanches, whether avalanches are observed from one location or using spatially continuous methods like satellite images, an adjustment of class intervals may be necessary [..[425]]to capture the frequency of locations where natural avalanches may initiate or where human-triggered avalanches are possible.

---

[412]removed: This correlation was generally strong, and even when using a sampling setting leading to large variation and overlap (n = 10) and a small number of classes k, the correlation between the frequency class describing *very poor* stability and the danger level was still moderate (Sect.4.1.2).

[413]removed: snow

[414]removed: This

[415]removed: ,

[416]removed: its frequency distribution are therefore inseparable

[417]removed: the assignment of

[418]removed: its frequency

[419]removed: The combination of stability class and its frequency distribution will also define which

[420]removed: should be described with the same danger level

[421]removed: *at treeline* or *in extremely steep terrain*

[422]removed: (**?**)

[423]removed: *very poor*

[424]removed: *very poor*

[425]removed: for it

5.3.2 Avalanche size

The most frequent avalanche size had little discriminating power, with the typical size being of size 1 or size 2, regardless of danger level. This [..[426]]can be explained by the fact that larger events occur normally less frequent than smaller events. This frequency-magnitude relation has also been observed for other natural hazards (e.g. Malamud and Turcotte, 1999), and has been described by power laws for avalanche size distributions (Birkeland and Landry, 2002; Faillettaz et al., 2004).

We showed that considering the largest avalanche per day resulted in a slightly better discrimination between danger levels. This finding is also supported by Schweizer et al. (2020), with the size of the largest avalanche being mostly of size 4 at 4-High. Furthermore, the typical largest expected avalanche is highly relevant for risk assessment and mitigation.

For danger level 5-Very High, for which we had no data, other studies have shown a further shift towards size 4 avalanches. Schweizer et al. (2020) showed that at 5-Very High size 4 avalanches were 15 times more frequent than at 3-Considerable and five times more frequent compared to 4-High. In two extraordinary avalanche situations in January 2018 and January 2019, when danger level 5-Very High was verified for parts of the Swiss Alps, avalanches recorded using satellite data showed that often ten or more size 4 avalanches and/or one size 5 avalanche [..[427]]were observed per 100 km^2 (Bühler et al., 2019; Zweifel et al., 2019).

5.3.3 Combining [..[428]]snowpack stability, [..[429]]the frequency distribution of snowpack stability and avalanche size

In Section [..[430]]4.3 we presented a data-driven look-up table to assess avalanche danger (Fig. 6). As can be seen in this table, the combination of [..[431]]snowpack stability and its frequency that best matches an avalanche situation (A to E), is highly relevant for danger level assessment. In general, avalanche size [..[432]]had a lesser influence on the danger level, once the cell describing stability has been fixed, as might be anticipated. This is in contrast to the original avalanche danger level assessment matrix (ADAM, Müller et al., 2016) that proposed that an increase in either the frequency class or the avalanche size, or a decrease in [..[433]]snowpack stability, should lead to an increase in danger level by one level. Clearly, the presented data-driven look-up table (Fig. 6) highlights that a greater focus must be placed on [..[434]]snowpack stability and its frequency distribution, compared to avalanche size, when assessing avalanche hazard. This was also shown by [..[435]]Clark (2019), who

[426] removed: finding is similar to other studies (Harvey, 2002; Logan and Greene, 2018; Schweizer et al., 2020). All three studies showed that the typical avalanche size did not increase with danger level, except at 4-High in the study by Logan and Greene (2018)

[427] removed: was

[428] removed: snow

[429] removed: its

[430] removed: **??**

[431] removed: snow

[432] removed: only has a rather minor

[433] removed: snow

[434] removed: snow

[435] removed: Clark and Haegeli (2018)

explored the combination of descriptive terms describing the three factors in the data behind the avalanche forecasts in Canada and their relation to the published danger level and avalanche problem. They showed that the 'likelihood of avalanches', which compares to our [..[436]]*stability matrix* (Fig. 6), also had a greater impact on the resulting danger level than avalanche size, even though avalanche size ≤1.5 (considered harmless to people) was often a first split in a decision tree model. Hence, despite using different approaches, partially different terminology and slightly different avalanche danger scales in Europe and North America, the relative importance of the three key contributing factors and the distributions of the danger levels are similar.

Our approach can only provide general distributions observed under dry-snow conditions. The look-up table presented [..[437]]in Fig. 6 should therefore be seen as [..[438]](a) a tool aiding the discussion of specific situations[..[439]], and (b) to improve the definitions underlying the categorical descriptions of the danger levels.

6 Conclusions

We explored observational data from two different countries relating to the three key factors describing avalanche hazard, snowpack stability, [..[440]]the frequency distribution of snowpack stability and avalanche size. We simulated stability distributions and defined [..[441]]four classes describing the frequency of potential avalanche triggering locations[..[442]], which we termed *none or nearly none*, *a few*, *several* and *many*. The observed and simulated distributions of stability ratings derived from RB tests showed that locations with *very poor* stability are generally rare (Fig. 3a, Fig. 8a-d).

Our findings suggest that the three key factors did not distinguish equally prominently between the danger levels:

- The proportion of [..[443]]*very poor* or *poor* stability test results increased from one danger level to the next higher one (Figures 3 and 9). Considering [..[444]]*very poor* snowpack stability and [..[445]]the frequency of this stability class alone, already distinguished [..[446]]well between danger levels [..[447]](Tab. [..[448]]2, Fig. 4).

- Considering the largest observed avalanche size per day and warning region was most relevant to distinguish between 3-Considerable and 4-High (Fig. 5 and Tab. 3). For other situations, the largest avalanche size - [..[449]]when used on

[436] removed: *stability matrix*
[437] removed: here should therefore primarily
[438] removed: a tool stimulating not only
[439] removed: but also when attempting
[440] removed: its frequency distribution
[441] removed: classes summarizing
[442] removed: .
[443] removed: *very poor* or *poor*
[444] removed: *very poor*
[445] removed: its frequency
[446] removed: rather
[447] removed: 2-Moderate, 3-Considerable and 4-High
[448] removed: **??**
[449] removed: used by itself

its own - had [..[450]]less discriminating power to distinguish between danger levels 1-Low to 3-Considerable [..[451]]compared to the other two factors (the lowest stability class present and the frequency of this class; Fig. 5).

In summary, the frequency of the most unfavorable snowpack stability class is the dominating discriminator. At higher danger levels the occurrence of size 4 avalanches discriminates danger level 3-Considerable from 4-High. We further suppose that the occurrence of size 5 avalanches discriminates between 4-High and 5-Very High without [..[452]]a significant additional increase in the [..[453]]frequency of *very poor* stability. This shift in importance between factors is currently poorly represented in existing decision aids like the EAWS-Matrix or ADAM (Müller et al., 2016), but also in the European Avalanche Danger Scale.

To combine the three factors and to derive avalanche danger, we introduced two data-driven look-up tables (Fig. 6), which can be used to assess avalanche danger level in a two step approach. In these tables, only the frequency of locations with the lowest snowpack stability is assessed, with no spatial component, and combined with the largest avalanche size. Spatial information in avalanche forecasts includes the aspects and elevations where the frequency of locations with the lowest stability class exists and possibly terrain features within the frequency class where triggering is particularly likely.

We hope that our data-driven perspective on avalanche hazard will allow a review of key definitions in avalanche forecasting [..[454]]such as the avalanche danger scale.

Data availability. The data will become freely available at www.envidat.org.

Author contributions. FT designed the study, conducted the analysis, wrote the manuscript. KM extracted the Norwegian data. KM and JS repeatedly provided in-depth feedback on the study design and analysis, and critically reviewed the entire manuscript several times.

Competing interests. No competing interests.

Acknowledgements. We thank the two reviewers Simon Horton and Karl Birkeland for their detailed and very helpful feedback, which greatly helped to improve this manuscript.

[450]removed: comparably little discriminating power at
[451]removed: (
[452]removed: an
[453]removed: spatial distribution of *very poor*
[454]removed: , like

References

Ameijeiras-Alonso, J., Crujeiras, R., and Rodríguez-Casal, A.: multimode: An R package for mode assessment, https://arxiv.org/abs/1803.00472, 2018.

Bakermans, L., Jamieson, B., Schweizer, J., and Haegeli, P.: Using stability tests and regional avalanche danger to estimate the local avalanche danger, Annals of Glaciology, 51, 176–186, doi:10.3189/172756410791386616, 2010.

Birkeland, K.: Spatial patterns of snow stability through a small mountain range, Journal of Glaciology, 47, 176–186, 2001.

Birkeland, K. and Landry, C.: Power-laws and snow avalanches, Geophysical Research Letters, 29, doi:10.1029/2001GL014623, 2002.

Bühler, Y., Hafner, E. D., Zweifel, B., Zesiger, M., and Heisig, H.: Where are the avalanches? Rapid SPOT6 satellite data acquisition to map an extreme avalanche period over the Swiss Alps, The Cryosphere, 13, 3225–3238, doi:10.5194/tc-13-3225-2019, https://www.the-cryosphere.net/13/3225/2019/, 2019.

CAA: Observation guidelines and recording standards for weather, snowpack and avalanches, Canadian Avalanche Association, NRCC Technical Memorandum No. 132, 2014.

Clark, T.: Exploring the link between the Conceptual Model of Avalanche Hazard and the North American Public Avalanche Danger Scale, Master's thesis, Simon Fraser University, 115 p., 2019.

Clark, T. and Haegeli, P.: Establishing the link between the Conceptual Model of Avalanche Hazard and the North American Public Avalanche Danger Scale: initial explorations from Canada, in: Proceedings ISSW 2018. International Snow Science Workshop, Innsbruck, Austria, 2018.

Díaz-Hermida, F. and Bugarín, A.: Linguistic summarization of data with probabilistic fuzzy quantifiers, in: Proceedings XV Congreso Español Sobre Tecnologías y Lógica Fuzzy, Huelva, Spain, pp. 255–260, 2010.

EAWS: EAWS Matrix, Tech. rep., https://www.avalanches.org/standards/eaws-matrix/, last access: 2020/01/31, 2017.

EAWS: European Avalanche Danger Scale (2018/19), https://www.avalanches.org/wp-content/uploads/2019/05/European_Avalanche_Danger_Scale-EAWS.pdf, last access: 14 Feb 2020, 2018.

EAWS: Standards: avalanche size, https://www.avalanches.org/standards/avalanche-size/, last access: 09/09/2019, 2019.

EAWS: EAWS Matrix, https://www.avalanches.org/wp-content/uploads/2019/05/EAWS_Matrix_en-EAWS.png, last access 31/01/2020, 2020.

Eckerstorfer, M., Malnes, E., and Müller, K.: A complete snow avalanche activity record from a Norwegian forecasting region using Sentinel-1 satellite-radar data, Cold Regions Science and Technology, 144, 39 – 51, doi:10.1016/j.coldregions.2017.08.004, 2017.

Efron, B.: Bootstrap methods: another look at the jackknife, Annals of Statistics, 7, 1–26, 1979.

Evans, I.: The selection of class intervals, Transactions of the Institute of British Geographers, 2, 98–124, 1977.

Faillettaz, J., Louchet, F., and Grasso, J.-R.: Two-threshold model for scaling laws of noninteracting snow avalanches, Phys. Rev. Lett., 93, doi:10.1103/PhysRevLett.93.208001, 2004.

Föhn, P.: The rutschblock as a practical tool for slope stability evaluation, IAHS Publ., 162, 223–228, 1987.

Föhn, P. and Schweizer, J.: Verification of avalanche danger with respect to avalanche forecasting, in: Les apports de la recherche scientifique à la sécurité neige, glace et avalanche. Actes de Colloque, Chamonix, vol. 162, pp. 151–156, Association Nationale pour l'Étude de la Neige et des Avalanches (ANENA), 1995.

Harvey, S.: Avalanche incidents in Switzerland in relation to the predicted danger degree, in: Proceedings ISSW 2002. International Snow Science Workshop, Penticton, Canada, 2002.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference, and prediction, Springer, 2 edn., 2009.

Hendrikx, J., Owens, I., Carran, W., and Carran, A.: Avalanche activity in an extreme maritime climate: The application of classification trees for forecasting, Cold Reg. Sci. Technol., 43, 104–116, 2005.

Jamieson, B. and Johnston, C.: Interpreting rutschblocks in avalanche start zones, Avalanche News, 46, 2–4, 1995.

Jamieson, B., Haegeli, P., and Schweizer, J.: Field observations for estimating the local avalanche danger in the Columbia Mountains of Canada, Cold Regions Science and Technology, 58, 84 – 91, doi:10.1016/j.coldregions.2009.03.005, 2009.

Kosberg, S., Müller, K., Landrø, M., Ekker, R., and Engeset, R.: Key to success for the Norwegian Avalanche Center: Merging of theoretical and practical knowhow, in: Proceedings ISSW 2013. International Snow Science Workshop, Grenoble - Chamonix Mont-Blanc, France, pp. 316 – 319, 2013.

Lazar, B., Trautmann, S., Cooperstein, M., Greene, E., and Birkeland, K.: North American avalanche danger scale: Do backcountry forecasters apply it consistently?, in: Proceedings ISSW 2016. International Snow Science Workshop, Breckenridge, Co., pp. 457 – 465, 2016.

Logan, S. and Greene, E.: Patterns in avalanche events and regional scale avalanche forecasts in Colorado, USA, in: Proceedings ISSW 2018. International Snow Science Workshop, Innsbruck, Austria, pp. 1059–1062, 2018.

Malamud, B. and Turcotte, D.: Self-organized criticality applied to natural hazards, Natural Hazards, 20, 93–116, 1999.

McClung, D. and Schaerer, P.: Snow avalanche size classification, in: Proceedings of an Avalanche Workshop, Vancouver, BC, Canada, 3-5 November 1980, pp. 12 – 27, 1981.

McClung, D. and Schaerer, P.: The Avalanche Handbook, The Mountaineers, Seattle, WA., 3rd edn., 2006.

Meister, R.: Country-wide avalanche warning in Switzerland, in: Proceedings ISSW 1994. International Snow Science Workshop 1994, Snowbird, UT, pp. 58–71, 1995.

Moner, I., Gavalda, J., Bacardit, M., Garcia, C., and Marti, G.: Application of field stability evaluation methods to the snow conditions of the Eastern Pyrenees, in: Proceedings ISSW 2008. International Snow Science Workshop, Whistler, Canada, pp. 386—392, 2008.

Moner, I., Orgué, S., Gavaldà, J., and Bacardit, M.: How big is big: results of the avalanche size classification survey, in: Proceedings ISSW 2013. International Snow Science Workshop Grenoble - Chamonix Mont-Blanc, 2013.

Müller, K., Mitterer, C., Engeset, R., Ekker, R., and Kosberg, S.: Combining the conceptual model of avalanche hazard with the Bavarian matrix, in: Proceedings ISSW 2016. International Snow Science Workshop, Breckenridge, Co., USA, pp. 472–479, 2016.

Reuter, B. and Schweizer, J.: Describing snow instability by failure initiation, crack propagation, and slab tensile support, Geophysical Research Letters, 45, 7019 – 7029, doi:10.1029/2018GL078069, 2018.

Reuter, B., Richter, B., and Schweizer, J.: Snow instability patterns at the scale of a small basin, Journal of Geophysical Research: Earth Surface, 257, doi:doi:10.1002/2015JF003700, 2016.

Schweizer, J.: The Rutschblock test - procedure and application in Switzerland, The Avalanche Review, 20, 14–15, 2002.

Schweizer, J.: Profilinterpretation (english: Profile interpretation), WSL Institute for Snow and Avalanche Research SLF, course material, 7 p., 2007a.

Schweizer, J.: Verifikation des Lawinenbulletins, in: Schnee und Lawinen in den Schweizer Alpen. Winter 2004/2005, pp. 91–99, Eidg. Institut für Schnee- und Lawinenforschung SLF, 2007b.

Schweizer, J. and Jamieson, B.: Snowpack tests for assessing snow-slope instability, Annals of Glaciology, 51, 187–194, doi:10.3189/172756410791386652, 2010.

Schweizer, J. and Wiesinger, T.: Snow profile interpretation for stability evaluation, Cold Reg. Sci. Technol., 33, 179–188, doi:10.1016/S0165-232X(01)00036-2, 2001.

Schweizer, J., Jamieson, B., and Skjonsberg, D.: Avalanche forecasting for transportation corridor and backcountry in Glacier National Park (BC, Canada), in: Proceedings of the Anniversary Conference 25 Years of Snow Avalanche Research, Voss, Norway, 12-16 May 1998, 203, pp. 238–244, Norwegian Geotechnical Institute, Oslo, Norway, 1998.

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

Schweizer, J., Kronholm, K., Jamieson, B., and Birkeland, K.: Review of spatial variability of snowpack properties and its importance for avalanche formation, Cold Regions Science and Technology, 51, 253–272, doi:http://dx.doi.org/10.1016/j.coldregions.2007.04.009, 2008a.

Schweizer, J., McCammon, I., and Jamieson, J.: Snowpack observations and fracture concepts for skier-triggering of dry-snow slab avalanches, Cold Regions Science and Technology, 51, 112–121, doi:10.1016/j.coldregions.2007.04.019, 2008b.

Schweizer, J., Mitterer, C., Techel, F., Stoffel, A., and Reuter, B.: On the relation between avalanche occurrence and avalanche danger level, The Cryosphere, doi:10.5194/tc-2019-218, 2020.

Simenhois, R. and Birkeland, K.: The Extended Column Test: A field test for fracture initiation and propagation, in: Proceedings ISSW 2006. International Snow Science Workshop, Telluride, Co., pp. 79–85, 2006.

Simenhois, R. and Birkeland, K.: The Extended Column Test: Test effectiveness, spatial variability, and comparison with the Propagation Saw Test, Cold Regions Science and Technology, 59, 210–216, doi:10.1016/j.coldregions.2009.04.001, 2009.

Slocum, T., McMaster, R., Kessler, F., and Howard, H.: Thematic cartography and geographic visualization, Prentice Hall Series in Geographic Information Science, Pearson/Prentice Hall, Upper Saddle River, NJ, 2 edn., 2005.

Statham, G., Haegeli, P., Birkeland, K., Greene, E., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J.: The North American public avalanche danger scale, in: Proceedings ISSW 2010. International Snow Science Workshop, Lake Tahoe, Ca., pp. 117–123, 2010.

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

Statham, G., Holeczi, S., and Shandro, B.: Consistency and accuracy of public avalanche forecasts in Western Canada, in: Proceedings ISSW 2018. International Snow Science Workshop, Innsbruck, Austria., pp. 1491 – 1496, 2018b.

Techel, F. and Pielmeier, C.: Automatic classification of manual snow profiles by snow structure, Nat. Hazards Earth Syst. Sci., 14, 779–787, doi:10.5194/nhess-14-779-2014, 2014.

Techel, F. and Schweizer, J.: On using local avalanche danger level estimates for regional forecast verification, Cold Regions Science and Technology, 144, 52 – 62, doi:10.1016/j.coldregions.2017.07.012, 2017.

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 Hazards Earth Syst Sci, 18, 2697–2716, doi:10.5194/nhess-18-2697-2018, https://www.nat-hazards-earth-syst-sci.net/18/2697/2018/, 2018.

Techel, F., Winkler, K., Walcher, M., van Herwijnen, A., and Schweizer, J.: On snow stability interpretation of Extended Column Test results, Natural Hazards Earth System Sciences, pp. 1–21, doi:10.5194/nhess-2020-50, (accepted), 2020.

Wand, M.: Data-based choice of histogram bin width, The American Statistician, 51, 59–64, doi:10.1080/00031305.1997.10473591, 1997.

Zweifel, B., Hafner, E., Lucas, C., Marty, C., Techel, F., and Stucki, T.: Schnee und Lawinen in den Schweizer Alpen. Hydrologisches Jahr 2018/19, WSL-Institut für Schnee- und Lawinenforschung SLF Davos: 134 pages (WSL Ber. 86), 2019.

1 [..[455]]

[..[456]][..[457]][..[458]][..[459]][..[460]]

1 [..[461]]

[..[462]]

[..[463]][..[464]]

[..[465]]

[..[466]]

[..[467]][..[468]][..[469]]

---

[455]removed: Appendix: ECT - simulated snow stability distributions and frequency classification

[456]removed: As a supplement to the analysis shown for the RB in the main part of the paper, in the following we show the key results for the ECT. As for the RB, we tested $n = \{10, 25, 50, 100, 200, 1000\}$. Besides visual inspection, we additionally tested the *poor* stability distributions for multi-modality using the *modetest* (Ameijeiras-Alonso et al., 2018).

[457]removed: In contrast to the RB class *very poor* stability, the distribution of the proportion of *poor* ECT stability was less skewed towards lower proportions of *poor* stability. Increasing *n* impacted the number of modes detected in the histograms, with two or more modes being present when *n* reached values of about 100 (Fig. **??**g-l). Exploring the bootstrapped-sampled distributions for the most extreme ECT stability classes *poor* and *good* (Fig. **??**) generally showed similar results as for the RB (Fig. **??**). However, while the distributions for the RB also exhibited a logical pattern at 4-High (Fig. **??**f), despite being drawn from a small population (drawn from N = 21), the same cannot be noted for the ECT (Fig. **??**f, drawn from N = 13).

[458]removed: Comparing the sampled distributions with actually observed distributions of stability tests on the same day and in the same region (N = 31), showed that the distributions obtained using bootstrap-sampling reflected the variation in the observed distributions not always well (Fig. **??**). Visually comparing the results for *n* = 10, where there was still a reasonably large number of days with 7 to 15 ECT (N =

[459]removed: 5, 6, 9

[460]removed:), implies that the bootstrap-sampled distributions captured the observed distributions poorly. However, a significant deviation between sampled and observed distributions was only noted for *good* stability at 3-Considerably (p = 0.02, Fig. **??**c). It must be noted, however, that sample sizes are small impacting both the likelihood to obtain unusual data sets in the field as well as for p-values not being the optimal indicator to detect significant differences.

[461]removed: Appendix: Additional figures and tables

[462]removed: Bar plots showing distribution of stability ratings for ECT for (a) Norway and (b) Switzerland. Note the very small number of tests at 4-High. The ECT classification scheme is shown in Fig. 1b.

[463]removed:

[464]removed: Simulated proportions of *very poor* (RB) or *poor* (ECT) stability for different number of samples *n* drawn in each of the bootstraps for (a-f) Rutschblock and (g-l) ECT. The more samples drawn, the more the data becomes multi-modal and clustered around the means of each danger level. This is indicated by the p-value (*modetest*, median p-value of 10 repetitions, Ameijeiras-Alonso et al., 2018). See also Fig.s **??** and **??** for two-dimensional plots.

[465]removed: Simulated proportions of *very poor* (x-axis) and *good* RB-stability (y-axis), for different number of samples *n* drawn in each of the bootstraps (a-f). The colour represents the most frequent danger level for the respective *very poor* - *good* combination. The more samples are drawn, the more the data becomes clustered around the means of each danger level.

[466]removed: Simulated proportions of *poor* (x-axis) and *good* ECT-stability (y-axis), for different number of samples *n* drawn in each of the bootstraps (a-f). The colour represents the most frequent danger level for the respective *poor* - *good* combination. The more samples are drawn, the more the data becomes clustered around the means of each danger level.

[467]removed:

[468]removed:

[469]removed: Bar plots showing the size distribution of all avalanches (upper row) and the largest avalanche *per day and warning region (lower row), for Norway (left column) and Switzerland (right column).

[..[470]]

[..[471]][..[472]]

[470] removed: Comparison of observed (points, N = 31) and bootstrap-sampled ECT distributions (boxes) for the proportion of *poor* (a, b) and *good* stability tests (c, d), for two settings of the number n of tests drawn. Observations with 7 to 15 individual tests on the same day and within the same region are shown together with sampling using $n = 10$. When more than 16 tests were collected, these are shown together with $n = 25$.

[471] removed:

[472] removed: Distribution of danger levels for snowpack stability and frequency class combinations. Combinations with the same most frequent and second most frequent danger level are labelled with the same letter (A to E). If a lower stability class resulted in frequency class *none*, for these cases the distributions for the next higher stability class is shown in the respective row below (i.e. the 2146 cases of *none very poor* are shown in the row *poor*). The letter which comes first in the alphabet is retained and used as a reference for the following matrix (Fig. **??**). This matrix corresponds to the stability matrix in Fig. 6.