





# **On Consistency and Quality in Public Avalanche Forecasting - a Data-Driven Approach to Forecast Verification and to Refining Definitions of Avalanche Danger**

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde  
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Frank Techel

aus

Deutschland

Promotionskommission

Prof. Dr. Ross Stuart Purves (Vorsitz)

Prof. Dr. Jürg Schweizer (Leitung)

Prof. Dr. Reinhard Furrer

Dr. Alec van Herwijnen

Zürich, 2020



# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgments</b>	<b>v</b>
<b>List of abbreviations and notations</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 The thesis in a nutshell . . . . .	2
1.3 Research gap and specific research questions . . . . .	3
1.3.1 Evaluating the quality of avalanche forecasts . . . . .	4
1.3.2 On the importance of the elements characterizing avalanche danger . . . . .	10
<b>2 Avalanche forecasting</b>	<b>13</b>
2.1 Avalanche forecasting - from data to forecast . . . . .	13
2.1.1 Data . . . . .	13
2.1.2 Data analysis . . . . .	17
2.1.3 Workflow and estimation of avalanche danger . . . . .	18
2.2 Public avalanche forecasts . . . . .	19
2.3 Avalanche danger scale . . . . .	22
<b>3 Data</b>	<b>25</b>
3.1 Spatial consistency and bias: regional forecast danger level . . . . .	26
3.2 Reliability of local danger level estimates and quality of forecast danger levels . . . . .	26
3.2.1 Local nowcast estimates of danger level . . . . .	26
3.2.2 Regional forecast danger level and nowcast assessments . . . . .	27
3.2.3 Avalanche occurrence data . . . . .	28
3.3 Characterizing the elements of avalanche danger . . . . .	29
3.3.1 Snow stability tests: Rutschblock and Extended Column Test . . . . .	29
3.3.2 Avalanche observations . . . . .	29

<b>4</b>	<b>Methods</b>	<b>33</b>
4.1	Reliability of ratings by humans . . . . .	34
4.1.1	Agreement rate . . . . .	34
4.1.2	Reliability of individual danger level assessments . . . . .	35
4.2	Categorical forecast verification . . . . .	36
4.2.1	Accuracy (proportion correct) . . . . .	36
4.2.2	Hit rate . . . . .	38
4.2.3	Success rate . . . . .	38
4.2.4	Bias ratio . . . . .	39
4.2.5	Bias . . . . .	39
4.3	Simulation of snowpack stability distributions by bootstrap sampling . . . . .	40
4.4	Snowpack stability and the frequency distribution of snowpack stability - approach to define class intervals . . . . .	40
<b>5</b>	<b>Results</b>	<b>43</b>
5.1	Spatial consistency of forecast danger levels in the Alps . . . . .	44
5.1.1	Spatial consistency in regional forecast danger levels: agreement and bias . . . . .	44
5.1.2	Variations in the use of danger level 4-High in regional forecasts . . . . .	46
5.1.3	Communicating avalanche danger at a regional scale - the potential impact of the size of the warning regions . . . . .	48
5.2	Quality of local danger level estimates . . . . .	52
5.2.1	Variations in local danger level estimates - agreement rate and reliability . . . . .	52
5.2.2	Validity of local danger level estimates - situations representing 4-High . . . . .	53
5.3	Quality of forecast danger levels . . . . .	54
5.3.1	Accuracy (proportion correct) of forecast danger levels . . . . .	55
5.3.2	Accuracy (proportion correct) of forecast danger levels - variations due to individual assessors . . . . .	57
5.3.3	Success rate, hit rate and bias of forecast danger levels . . . . .	58
5.3.4	On the success of forecasting 4-High . . . . .	59
5.4	Elements of avalanche danger - snowpack stability, the frequency distribution of snowpack stability and avalanche size . . . . .	61
5.4.1	Snowpack stability . . . . .	61
5.4.2	Avalanche size . . . . .	63
5.4.3	Combining the frequency of <i>very poor</i> stability and avalanche size . . . . .	64
5.4.4	Data-driven lookup table for danger level assessment . . . . .	65
5.5	On the snowpack stability interpretation of instability tests . . . . .	66
<b>6</b>	<b>Discussion</b>	<b>69</b>
6.1	Consistency and quality in avalanche forecasts . . . . .	69
6.1.1	Key findings . . . . .	70

6.1.2	On the influence of spatial resolution and the way avalanche danger is communicated in avalanche forecasts on consistency and quality . . . . .	72
6.1.3	Forecast quality: overall forecast accuracy, over-forecasting and forecasting 4-High . . .	75
6.1.4	Consistency and quality: potential implications for the value of avalanche forecasts . . .	76
6.1.5	On using local danger level estimates as a data source for forecast verification . . . . .	76
6.2	A data-driven characterization of avalanche danger . . . . .	77
6.3	Data sets and methods - a basis for further data-driven explorations . . . . .	79
6.4	Limitations . . . . .	79
6.4.1	Data . . . . .	79
6.4.2	Methods . . . . .	80
6.5	Communication of findings to a lay audience . . . . .	82
6.5.1	Findings relevant to mountaineering amateurs (recreational forecast users) . . . . .	82
6.5.2	Findings relevant to mountaineering professionals . . . . .	83
6.5.3	Findings relevant to other professional users . . . . .	83
<b>7</b>	<b>Conclusions and Outlook</b>	<b>85</b>
7.1	Improving consistency and quality in avalanche forecasts - challenges and possible ways forward . . . . .	87
<b>A</b>	<b>Publications</b>	<b>91</b>
A.1	Publications and author contributions . . . . .	91
A.2	Spatial consistency and bias in avalanche forecasts - a case study in the European Alps . . .	93
A.3	On using local avalanche danger level estimates for regional forecast verification . . . . .	122
A.4	Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random? . . . . .	142
A.5	On the importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level . . . . .	160
A.6	On snow stability interpretation of Extended Column Test results . . . . .	186
A.7	List of publications and conference contributions . . . . .	206
<b>B</b>	<b>Supplement to Data section</b>	<b>209</b>
B.1	Forecast verification data: Switzerland, Norway, Canada, Colorado . . . . .	209
B.2	Avalanche recordings - mapped avalanches Davos / Switzerland . . . . .	211
	<b>Bibliography</b>	<b>213</b>



# Abstract

In many snow-covered mountain regions, regional avalanche forecasts are disseminated to inform and warn the public about avalanche danger. These forecasts are prepared by human experts (avalanche forecasters), who analyze and interpret relevant data, applying their knowledge and intuition in this process, in an environment where the most relevant data are often sparse in time and space and where spatial variability is high. In these forecasts, information regarding the current and future snow and avalanche conditions are given.

One of the key pieces of information communicated in public avalanche forecasts is an avalanche danger level, according to a five-level, ordinal danger scale. A danger level summarizes the avalanche conditions using an integer-signal word combination (e.g. 1-Low). The avalanche danger scale, the foundation for assessing and communicating avalanche danger in public avalanche forecasts, qualitatively describes the probability of triggering, the frequency and the location of the triggering spots and the destructive size of avalanches for each of the five avalanche danger levels. However, the description of the danger levels, and some of the terms used in the danger scale, are vague and leave room for interpretation.

Efficient and effective avalanche forecasts are necessary to assist recreationists and professionals in their decision-making process when mitigating avalanche risk to prevent potentially life-threatening avalanche accidents. Hence, these forecasts must provide relevant, reliable, and accurate information. Furthermore, avalanche forecasts must cater to a diverse range of users, with a wide scope of skills, experience, and training, and thus also different requirements regarding the depth of information. And lastly, the information must be communicated in an understandable way addressing both recreational and professional forecast users alike.

To explore the goodness of avalanche forecasts, three elements are considered: *consistency*, *quality* and *value*. *Consistency* in avalanche forecasts, or more specifically in the application of the avalanche danger levels by the individual forecasters or forecast centers, is essential to avoid misunderstandings or misinterpretations by users, particularly those utilizing bulletins issued by different forecast centers. The *quality* of the forecast danger level is conceptually difficult to assess, as avalanche danger cannot be measured, and hence cannot truly be verified. Furthermore, verification is challenging, as relevant information is often scarce and must be interpreted in light of uncertainties, and as the definitions of the danger levels are vague and leave room for interpretation. Still, only when factual information regarding deficiencies in forecast quality is available, can the forecast process be improved in a targeted manner. And finally, *value* is not intrinsic to a forecast, but depends on whether users benefit from using the forecast. To do so, a certain level of

consistency and quality is required.

Thus, the two objectives of this dissertation are (1) to provide data-driven insights on consistency and quality in public avalanche forecasts, and (2) to quantitatively describe the three elements characterizing avalanche danger - snowpack stability, the frequency distribution of snowpack stability, and avalanche size. These objectives were achieved by analyzing several newly compiled data sets originating from different warning services and snow climates, collected for avalanche forecasting, and information published in avalanche forecasts.

The first objective was achieved by approaching the topic from different viewpoints:

Firstly, spatial consistency and bias were explored by analyzing the spatially continuous forecasts in the European Alps. A rather low agreement rate of 65% between the forecast danger levels in neighboring warning regions belonging to different forecast centers was noted. Furthermore, considerable variation in the use of danger level 4-High existed. Some of these variations could be linked to operational constraints, like the spatial resolution of the forecasts.

In a second step, the use of local danger level estimates for regional forecast verification was examined. While variations in local danger level estimates existed, and an observer-specific reporting bias was noted for about 10% of the observers, the overall agreement between estimates provided in the same region was relatively high (about 80%). Relying on these nowcast estimates of danger level in Switzerland and Norway, and nowcast or hindcast assessments by forecasters in Canada and Colorado as a reference for forecast verification, showed similar patterns in all four countries: the success rate, a forecast danger level being confirmed by the reference assessment, decreased with increasing danger level from about 90% at 1-Low to less than 60% at 4-High. In fact, at 4-High there was a tendency towards more misses and false alarms rather than correct forecasts. Forecast danger levels that were not confirmed by the reference assessment tended to exhibit a strong over-forecast bias.

And lastly, relying on a Swiss data set where forecasters indicated a sub-level with each forecast danger level, it was shown that forecasters can often forecast avalanche danger at greater detail. Concerning the over-forecast bias, two new findings were noted: Incorporating spatial information, by considering the forecast danger level in immediately neighboring warning regions, showed that an over-forecast bias also existed in a spatial context. The data showed further that in case the forecast danger level did not match the reference assessment, the difference was generally less than a «full » danger level when considering the forecast sub-level.

In summary, these findings highlighted deficiencies in the consistency and quality of avalanche forecasts.

Turning to the second research objective, the three key elements of avalanche danger - snowpack stability, the frequency distribution of snowpack stability, and avalanche size - were quantitatively described for four of the five danger levels based on observational data. To perform this analysis, and relying on a statistical simulation approach, a large number of snowpack stability distributions were obtained, and four frequency classes were defined. The findings showed that the frequency of the most unstable locations is most relevant for the assessment of avalanche danger. This shift in importance between elements characterizing avalanche danger is poorly represented in existing decision aids, but also the European Avalanche Danger Scale. The resulting data-driven lookup table paves the way to refine the definitions of the avalanche danger

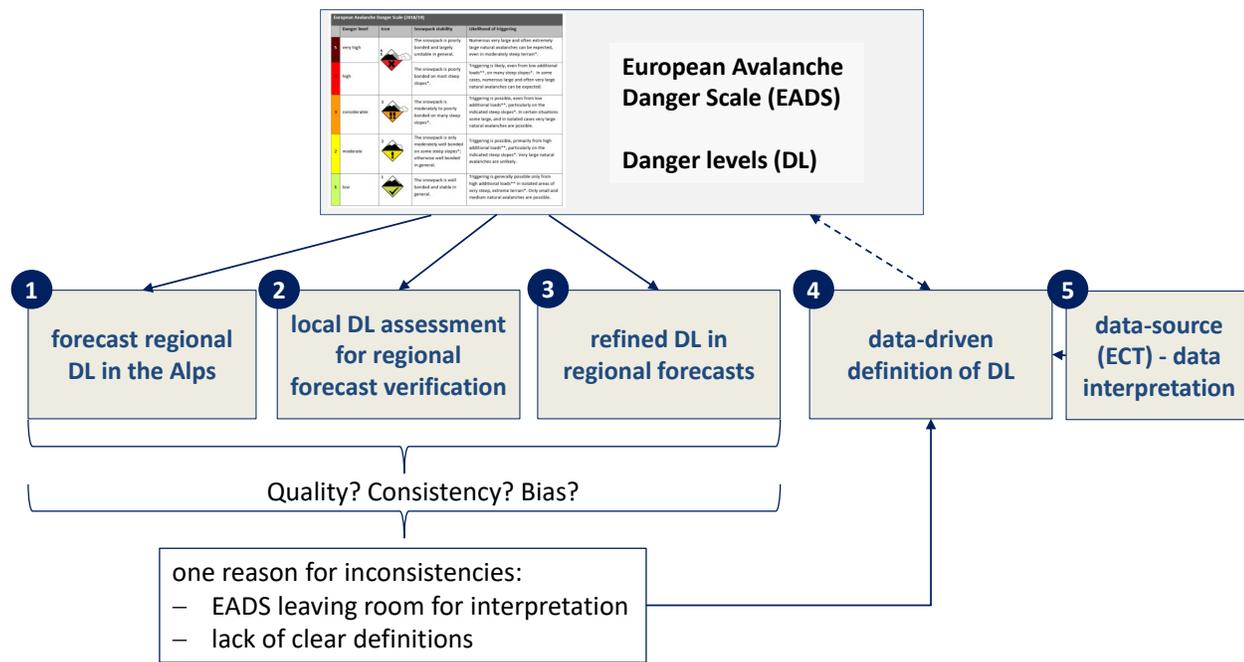
scale and in fostering its consistent usage.

Furthermore, a four-class stability classification scheme was developed for the Extended Column Test, which allowed an objective comparison with the Rutschblock test. While the data indicated a refined way to interpret the test results, key challenges - like the selection of a safe, yet representative site to perform the test in an environment that is spatially highly variable - remain. And finally, the data clearly showed that false-unstable predictions of stability tests outnumber the correct-unstable predictions in an environment where overall unstable locations are rare, as is the case at the lower danger levels 1-Low or 2-Moderate.

The main limitations encountered in this dissertation were related to the lack of independent ground truth. Even though all the analyses were data-driven, the findings related to danger levels are always based on assessments and observations made by humans. To increase the quality of these human assessments, only data provided by specifically trained experts were used. Furthermore, to reduce the influence of individual perspectives on the results, all analyses relied on forecasts, assessments, and observations provided by numerous experts from different forecast centers in different snow climates with a different knowledge base. The findings clearly emphasize the need to revisit and harmonize the way avalanche danger is assessed and communicated to increase consistency and quality, and hence facilitate cross-border forecast interpretation by traveling users. In that respect, the avalanche forecasting community should continue to strive to develop clear and practical guidelines, policies, and definitions on how to assess and communicate avalanche danger. However, improvements should not stop there: timely, relevant, and reliable class 1 data are required permitting not only unbiased, data-based nowcast assessments, but also more accurate and objective forecasts.

The thesis consists of two main parts:

- a synthesis, presenting and discussing a selection of the key data, methods and findings in this dissertation (Chapters 1 - 7), and
- the five research papers (Fig. 1), which are provided in the Appendices A.2 - A.6. In the Appendix, an overview of the respective author contributions (Appendix A.1) and a list of further publications and conference contributions (Appendix A.7) is shown.



*Figure 1:* This thesis consisted of five research papers, in which consistency and quality in avalanche forecasts were explored (forecasts and local nowcasts, publications 1 to 3), an observation-driven characterization of the danger levels was derived (publication 4), and a stability classification scheme was developed (publication 5), which allowed using additional data in publication 4. The five publications are provided in Appendix A of this thesis, together with a list showing the respective author contributions.

# Acknowledgments

I thank my employer SLF<sup>1</sup>, and here specifically Thomas Stucki and Jürg Schweizer, who made it possible that I could step out of the day-to-day forecasting routine - at least for extended periods at a time - and pursue this thesis. This meant that my colleagues in the forecasting team - Beni, Célia, Chris, Gian, Kurt, Lukas, Thomi - had to cover more forecast shifts during winter. Thank you so much for doing this but most of all, for being such a great team to work with.

I wish to thank my main supervisors Ross Purves and Jürg Schweizer. Ross, I am grateful that you took me on as an external Ph.D. student. I thank you for your helpful advice, for your supportive and constructive way of providing feedback, and for bringing in a fresh perspective, sometimes quite different to the avalanche forecasters' way of looking at things. Jürg, thank you for your great support and guidance, for the countless times you provided me with valuable feedback throughout these years, and for letting me such a free rein regarding the research questions I wanted to explore. I also thank my co-supervisors Reinhard Furrer and Alec van Herwijnen for their support during my dissertation.

During this thesis, I approached many forecasters in Europe, North America, and New Zealand regarding data, opinions or feedback. Very often my inquiries were well received. Only through this exchange with forecasters and researchers, and the resulting fruitful collaborations, did I become fully aware of the huge diversity of public avalanche forecasting operations. In this regard, I would like to thank all the forecasters and researchers, who contributed to this thesis.

And finally, I would like to thank Esther, and many other people who were involved in one way or another for their help, support, questions, and feedback.

---

<sup>1</sup>WSL Institute for Snow and Avalanche Research SLF Davos



# List of abbreviations and notations

ADAM	Avalanche Danger Assessment Matrix (Müller et al., 2016)
$BR$	bias ratio (Sect. 4.2.4, p. 39)
CAN	Canada
CMAH	Conceptual Model of Avalanche Hazard (Statham et al., 2018a)
COL	Colorado
$D$	Danger level
$D_{LN}$	Danger level - estimated locally, also called local nowcast (LN)
$D_{reference}$	Danger level - used as a reference assessment
$D_{RF}$	Danger level - issued in a regional forecast (RF)
EADS	European Avalanche Danger Scale (Sect. 2.3, Tab. 2.2)
EAWS	European Avalanche Warning Services
EAWS-Matrix	decision tool developed by EAWS (2017b)
ECT	Extended Column Test
$N$	count (sample size)
NOR	Norway
$P_{agree}$	agreement rate (Sect. 4.1.1, p. 34)
$P_{correct}$	proportion correct according to distribution of forecast $D_{RF}$ (Sect. 4.2.1, p. 36)
$P_{correct.raw}$	proportion correct according to distribution of observations (Sect. 4.2.1, p. 36)
$P_{correct}^*$	proportion correct according to distribution of forecast $D_{RF}$ and considering the upper bound due to $rel_{D,LN}$ (Sect. 4.2.1, p. 36)
$P_{hits}$	hit rate (Sect. 4.2.2, p. 38)
$P_{success}$	success rate (Sect. 4.2.3, p. 38)
$P_{over-under}$	bias, proportion of over-forecasts minus proportion of under-forecasts (Sect. 4.2.5, p. 39)
$P_{v,crit}$	proportion of forecasts with $D_{RF} \geq 4$ -High
RB	Rutschblock test
$rel_{D,LN}$	reliability of $D_{LN}$ estimates
SLF	WSL-Institute for Snow and Avalanche Research SLF Davos/Switzerland
SWI	Switzerland
wr	warning region, smallest spatial units used in public avalanche forecasts



# Chapter 1

## Introduction

### 1.1 Motivation

Public forecasts of regional avalanche danger are disseminated throughout the winter in many mountainous regions. These forecasts - also called advisories, warnings, or bulletins<sup>1</sup> - provide information about the current and forecast snow and avalanche conditions in a specific region. Avalanche forecasts serve as a warning, even when the avalanche situation is rather favorable, as they always specify what trigger is needed to release an avalanche, how frequent these locations are, what destructive potential the expected avalanches may have, and where potentially dangerous locations are most frequent.

There are two key consumer groups of public avalanche bulletins: the first of these groups, are users who undertake activities, such as off-piste riding and backcountry touring in uncontrolled terrain, either as a mountaineering professional or in a recreational setting. In some countries, a second group of users are local, regional, and national risk-management authorities, such as those responsible for the safety of humans in settlements or of users of roads or railways. For recreational users, the information provided in avalanche forecasts is particularly relevant during the planning phase of backcountry tours. Risk-management authorities, on the other hand, may base their risk reduction strategies in part on information given in the forecasts. The provision of clear, consistent, reliable, and accurate information regarding current and future avalanche conditions is underlined firstly by avalanche accident statistics - with on average 100 fatalities each winter in the Alps alone (Techel et al., 2016b), most of whom died during recreational activities. Secondly, very large numbers of individuals recreate in uncontrolled winter terrain, with for example Winkler et al. (2016) reporting that more than two million winter backcountry touring days were undertaken in 2013 in Switzerland alone. And finally, avalanche forecasts are products for the user and must, therefore, meet user needs by providing relevant, but also reliable, accurate, and skillful information (Williams, 1980; Gordon and Shaykewich, 2000). Hence, it is important to assess how good avalanche forecasts are, whether user needs are met, and where improvements in the forecast quality and the forecast product are most necessary. This includes objective, statistical measures assessing forecast quality, but also the public perception of the forecast, which reflects the true value of the forecast to the individual user (Murphy, 1993; Gordon and Shaykewich, 2000). Thus,

---

<sup>1</sup>these terms are used synonymously

trustworthy and data-based facts regarding the quality of the forecast rather than assumptions must be available to inform forecast users, decision-makers, or stake-holders about the quality of the forecast (Gordon and Shaykewich, 2000). Furthermore, only if such factual information were available, can the forecast process or forecast products purposefully be improved.

The avalanche danger scale provides the foundation for the assessment and communication of avalanche danger in public avalanche forecasts. In Europe, all avalanche warning services except Sweden, rely on the the European Avalanche Danger Scale (EADS) in the production and communication of forecasts (EAWS, 2017d). In the EADS, key factors that characterize avalanche danger - the probability of avalanche release, the frequency and location of the triggering spots, and the expected avalanche size - and their values, are described in a qualitative way (Meister, 1995; EAWS, 2020d, 2018). However, the EADS descriptions leave ample room for interpretation and are even partly ambiguous (Schweizer et al., 2020). This may be a major reason for inconsistencies noted in the use of the danger levels between individual forecasters (Lazar et al., 2016; Statham et al., 2018b; Clark, 2019). With the aim to increase consistency in the use of the danger levels between different forecasters and warning services, lookup tables, intended to aid forecasters in the assessment of avalanche danger, were developed (EAWS, 2017b; Müller et al., 2016). These decision-aids incorporate both the categorical descriptions given in the EADS and the experience of the European avalanche forecasters. However, neither the EADS nor these lookup tables have been compared with actual data, which could provide data-driven insights regarding the characterization of the avalanche danger levels.

## 1.2 The thesis in a nutshell

It is the objective of this dissertation to provide data-driven insights regarding consistency and quality in public avalanche forecasts, and the quantitative characterization of the avalanche danger levels.

The two main research questions in the focus of this thesis are: «How *good* are public avalanche forecasts?», and «Can the key elements defining avalanche danger be characterized using a data-driven approach?». These objectives are achieved by analyzing several data sets collected for avalanche forecasting and information published in avalanche forecasts originating from different warning services and snow climates. Relying on well-established statistical approaches, appropriate for the task at hand and comparably easy-to-communicate metrics, consistency and quality are explored as a function of avalanche conditions, but also by taking into consideration the operational constraints setting limits on the spatial and temporal resolution of regional avalanche danger communication. The findings provide insights that allow improving consistency in the forecast production process and the way avalanche danger is communicated, which highlight where deficiencies in forecast quality exist and hence where improvements in forecast quality are most necessary, and allow refining the definitions of the avalanche danger scale and in fostering its consistent usage. These objectives were addressed in five research papers (Fig. 1.1):

1. Techel et al. (2018): Spatial consistency and bias in avalanche forecasts – a case study in the European Alps. *Nat. Hazards Earth Syst. Sci.*

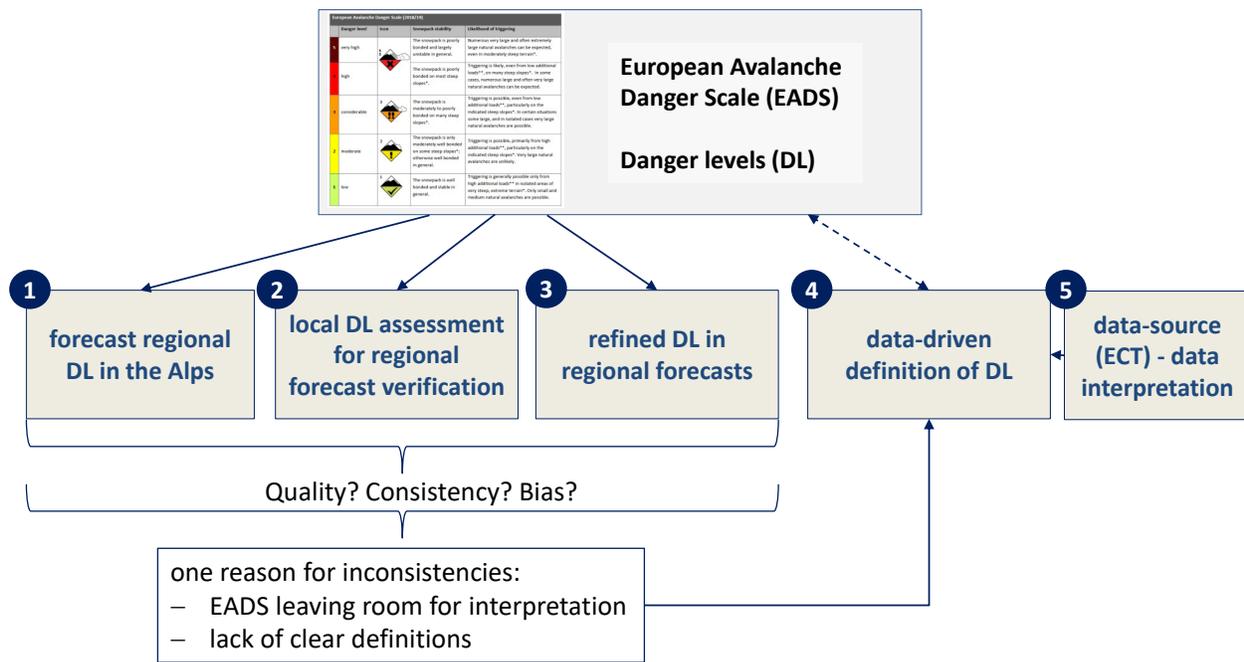


Figure 1.1: The research questions were addressed in five research papers. Publications 1 to 3 addressed issues like consistency and quality in public avalanche forecasts, publication 2 additionally explored the quality of local danger level estimates as a data-source for forecast verification. The European Avalanche Danger Scale (EADS) is the common guideline for danger level assessment. In publication 4, observational data relating to the key elements of avalanche danger were used to provide a data-driven description of the danger levels. In publication 5, a classification scheme was developed for part of the data used in publication 4 (the Extended Column Test ECT, a stability test, see also Sect. 3.3.1).

2. Techel and Schweizer (2017): On using local avalanche danger level estimates for regional forecast verification. *Cold Reg. Sci. Technol.*
3. Techel et al. (2020): Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random? *Cold Reg. Sci. Technol.*
4. Techel et al. (2020): On the importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level. *The Cryosphere*
5. Techel et al. (2020): On snow stability interpretation of Extended Column Test results. *Nat. Hazards Earth Syst. Sci.*

### 1.3 Research gap and specific research questions

Avalanche forecasting is described as «the prediction of the current and future snow instability in space and time relative to a given triggering level» (McClung, 2002a, p. 3). Avalanche forecasting therefore deals with a natural system interacting with humans (McClung, 2002a), where humans are involved in three ways: firstly, avalanches may pose a hazard to humans or their property; secondly, humans may also influence the danger by triggering avalanches themselves; and thirdly, avalanche forecasting is a task performed by

humans.

There are several types of avalanche forecasting operations, with different operational objectives, working at varying spatial and temporal scales. These operations include for instance (Statham et al., 2018a):

- those responsible for the safety of humans, like commercial backcountry skiing operations or snow safety programs of transportation corridors or work sites
- avalanche warning services issuing publicly available regional avalanche forecasts with the main goal of providing warnings and information to the public

The spatial extent covered in these forecasts may range from a few slopes to entire mountain ranges, while time spans covered in forecasts may vary from now to days or even weeks into the future.

In this dissertation, the focus is on public avalanche forecasting at a regional scale: the forecasts cover spatial scales from a few hundred to several thousand square kilometers, also referred to as the drainage scale to the region or mountain range scale (Schweizer and Kronholm, 2007; Statham et al., 2018b), and are generally issued with a validity of several hours or a few days.

### 1.3.1 Evaluating the quality of avalanche forecasts

Forecast validation and evaluation is not only a problem in avalanche forecasting but more generally in forecasting. Murphy (1993), in his classic paper on the nature of a good (weather) forecast, discussed three key elements which he termed *consistency*, *quality* and *value*. Consistency in Murphy's model essentially captures the degree of agreement between a forecaster's understanding of a situation and the forecast they then communicate to the public. Quality captures the degree of agreement between a forecast and the events which occur, and value the benefits or costs incurred by a user as a result of a forecast.

#### 1.3.1.1 Consistency

Murphy (1993) defines consistency with respect to an individual forecaster. However, the concept can be extended to forecast centers, in terms of the degree to which individual forecasters using potentially different evidence reach the same judgment (LaChapelle, 1980), and across forecast centers, in terms of the uniformity of the forecast issued by different forecast centers in neighboring regions. This reading of consistency is both true to Murphy's notion (how reliably does a forecast correspond with a forecaster's best judgment) and broader notions of consistency stemming from work on data quality and information science (Ballou and Pazer, 2003; Bovee et al., 2003).

The European Avalanche Warning Services (EAWS) strive to improve the quality of avalanche forecasting in Europe, providing efficient and effective forecasts (EAWS, 2017d). However, until now, no quantitative evidence existed allowing objective analysis of similarities and differences in the forecasts provided by different avalanche warning services. With a focus on the use of the avalanche danger levels, a data set was compiled consisting of forecast danger levels from many forecast centers in the European Alps, which permitted for the first time to explore spatial consistency and bias between warning services and forecast centers.

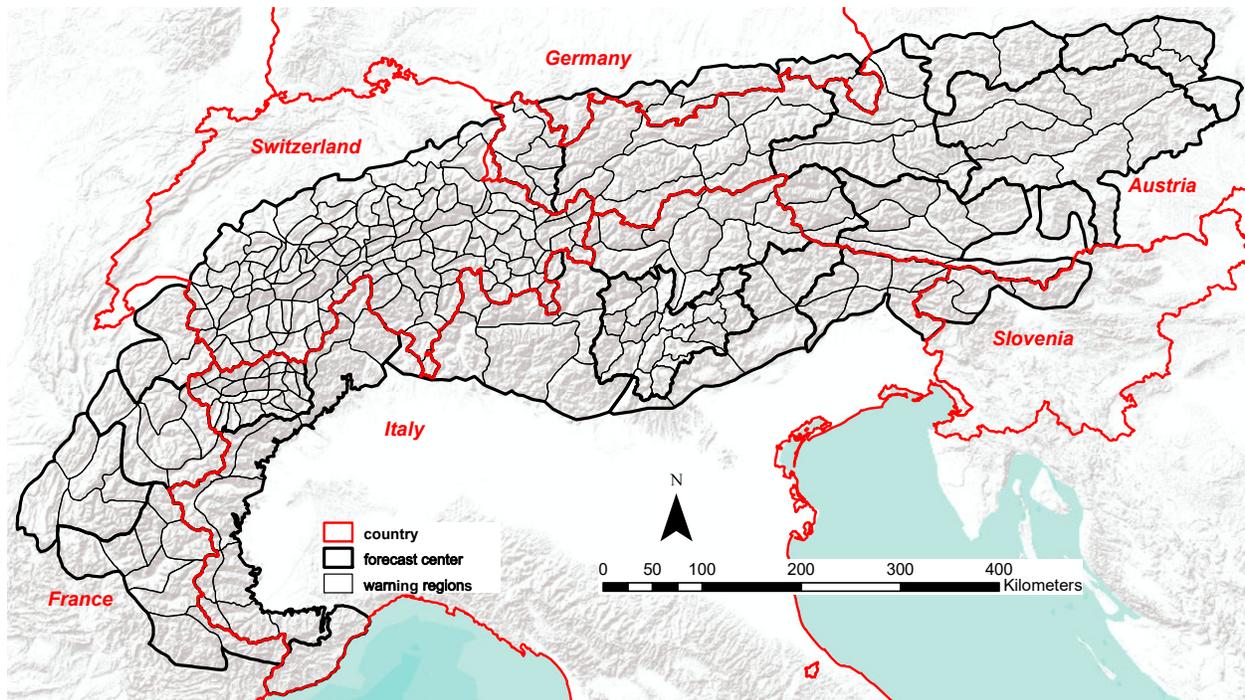


Figure 1.2: Map showing the relief of the European Alps (gray shaded background) with the outlines of the individual forecast centers (bold black polygons) and the warning regions, the smallest geographically defined regions, used by the respective avalanche forecast centers in their products (black polygons). The borders of the Alpine countries are marked red. The map captures the situation and partitioning during the period under study.

The European Alps were particularly suitable for this analysis, as the Alps are one of the few regions in the world, where the forecast domains of neighboring forecast centers border each other, thus providing a spatially continuous forecast region (Fig. 1.2). Furthermore, in the Alps in 2017, avalanche forecasts were provided by thirty different forecast centers in six different countries with a wide range of operational constraints. This additionally allowed to explore whether operational constraints, like the size of the warning regions, the smallest spatial units used in regional avalanche forecasts, impact forecast danger levels. Two research questions were in the center of this study:

1. Do differences in the use of the danger levels between forecast centers exist?
2. Can operational constraints (such as the size of the warning regions) explain these differences?

These, and other, questions were addressed in publication 1 (Fig. 1.1), titled «Spatial consistency and bias in avalanche forecasts - a case study in the European Alps» (Sect. A.2, p. 93ff). A selection of the key findings is presented in this Synthesis (Section 5.1, p. 44ff).

### 1.3.1.2 Quality

Typical questions addressing the quality of a set of forecasts are: Was the forecast correct? Are biases present? How skilled is the forecast, when compared to a reference assessment?

These are relevant questions regarding all kinds of forecasts and are not specific to avalanche forecasts. However, addressing these questions in the context of avalanche forecasting is not straightforward as three

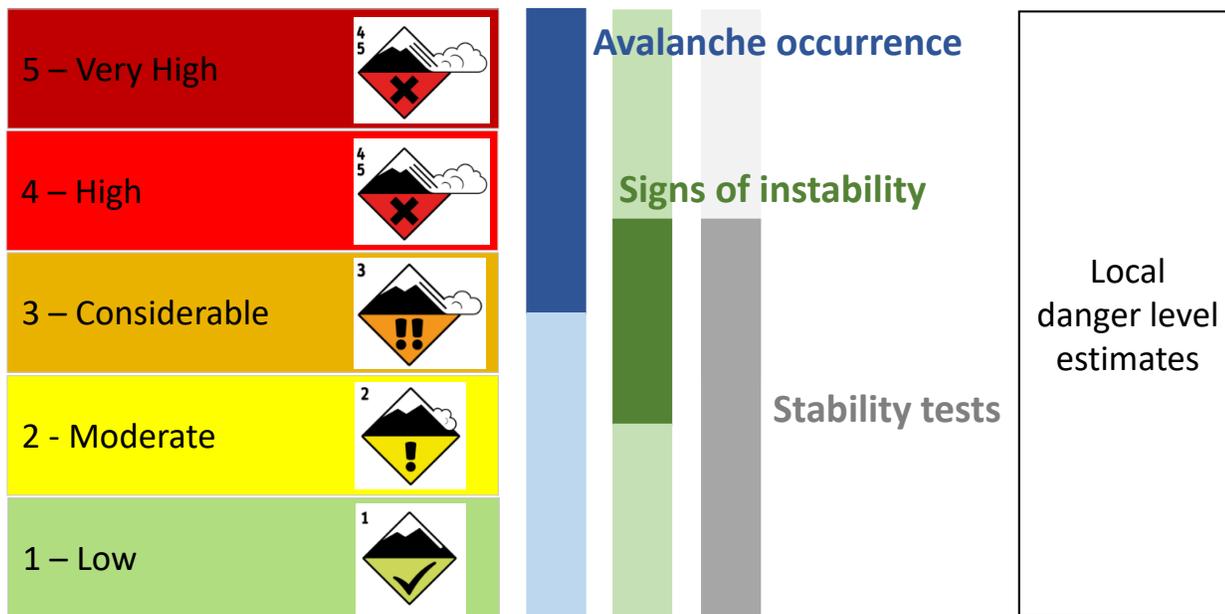


Figure 1.3: Data sources directly related to snow instability (class 1 data, see also Sect. 2.1.1) allowing the verification of avalanche danger, or more specifically the five danger levels. The observation of signs of instability and of stability tests requires people to be in the field. Thus, when conditions are dangerous, no such observations are available. Observations of natural avalanches, on the other hand, can be made from the valley floor. However, natural avalanches occur mostly at higher danger levels. Both, forecasters and local observers use these data to assess avalanche danger.

key problems come to the fore:

Firstly, the target variable - the avalanche danger level - is essentially categorical, since although the European Avalanche Danger Scale (EADS) is an ordinal scale, a real evaluation of a forecast would compare the forecast danger level, qualitatively defined in the EADS, with the prevailing avalanche situation. Secondly, since the target variable captures a state which may or may not lead to an (avalanche) event, verification of forecast quality is only possible in some circumstances and for some aspects of the EADS, for example (Fig. 1.3):

- At higher danger levels, the occurrence of natural avalanches can sometimes be used to verify the danger level (e.g. Elder and Armstrong, 1987; Giraud et al., 1987; Schweizer et al., 2020).
- At lower danger levels, the occurrence of avalanches triggered by recreationists or the observation of signs of instability requires users being present.
- Since the absence of avalanche activity is not alone an indicator of stability, verifying associated danger levels is only possible through digging multiple snow profiles and performing stability tests (Schweizer et al., 2003).

Thus, avalanche danger cannot be fully measured or validated and therefore, it is conceptually difficult to directly measure forecast quality.

And thirdly, the verification itself is considered an expert decision, regardless of whether this assessment is made in the field (local nowcast) or in hindsight when analyzing data, as the same subjective approach

is used when forecasting and when re-assessing forecasts (Elder and Armstrong, 1987). Thus, even if a danger rating is verified using all available information in hindsight, the accuracy of the «verified» danger level depends, for instance, on the data available to perform the assessment and on the skill of the assessor to correctly interpret the data.

The most useful information for verification is the one directly related to snow instability: recent avalanches, signs of instability (whumpfs of shooting cracks) or stability test results (McClung, 2002b) (Fig. 1.3). This so-called Class I data is particularly useful to distinguish between the higher danger levels 3-Considerable and 4-High, and the lower danger levels 2-Moderate and 1-Low (e.g. Jamieson et al., 2008). However, in day-to-day public forecasting, this kind of information is often not readily available due to lacking observations, and other less direct information needs to be considered. Among those are current estimates of the local danger level ( $D_{LN}$ ) reported by observers (e.g. Brabec and Stucki, 1998; Engeset, 2013; Jamieson et al., 2009). In some countries, as in Norway or Switzerland, such local danger level estimates are not only used to review the forecast regional danger level but also to prepare the future forecast (Suter et al., 2010; Kosberg et al., 2013).

### **On the reliability of local danger level estimates**

Imperfect reliability, sometimes also called inconsistency, is part of essentially all human assessments (Stewart, 2001). Thus, when relying on human assessments or observations as a data source, it is important to have an idea about the reliability, or «trust», we can put in a single piece of this information.

In avalanche forecasting, even today, a large part of the most relevant data is provided in the form of observations, estimates, or assessments. One such piece of information is local danger level estimates, provided by experienced and specifically trained observers together with other daily observations. In Switzerland, such local danger level estimates are used since at least 1987 (SLF, 1987). These estimates have become an important data source for the operational daily review of the forecast in some countries like Switzerland (e.g. Suter et al., 2010) or Norway (Kosberg et al., 2013), but have also been used to assess the overall quality of the forecast danger level (e.g. Brabec and Stucki, 1998; Harvey et al., 1998). The advantages of using locally estimated  $D_{LN}$  are, firstly, that a central target variable of an avalanche forecast - the forecast regional danger level - can be reviewed with a similar type of variable - rather than using, for example, avalanche occurrence data. And secondly, the local assessment provides a second opinion of the avalanche danger level incorporating primarily field observations, but also other information not available to an office-based forecaster. However, challenges associated with using this data for forecast verification include differences in the spatio-temporal scale - a regional forecast valid for the day vs. a local nowcast estimated at a certain time - and the subjective nature of the local assessment. Furthermore, if (partly) erroneous observations are used as a reference standard when assessing the overall forecast quality, performance measures would inevitably indicate unreasonably low forecast quality (e.g. Bowler, 2006) as observed forecast quality is bounded by the unreliability of these observations (e.g. Stewart, 2001; Vul et al., 2009). Hence, it is imminent that the quality of local danger level estimates is - at least approximately - known. In publication 2, which had the objective to provide data-driven findings in this regard, this issue was addressed using a large, multi-year data set of local danger level estimates, provided by observers after a day in the field in the

Swiss Alps. The two specific research questions, which are addressed in this Synthesis, are:

1. Do variations in local danger level estimates exist?
2. What implications do these variations have on the reliability of local danger level estimates as a data-source for forecast verification?

These, and further, research questions were explored in detail in publication 2 (Fig. 1.1) titled «On using local avalanche danger level estimates for forecast verification». A selection of the key findings is presented in Section 5.2 (p. 52ff). To compare the findings in this study, which relied on Swiss data, with data from other regions, a similar data set from the Norwegian avalanche warning service NVE (Ekker, 2018) is analyzed and presented together with the Swiss results (Sect. 5.2).

Finally, a highly relevant, yet conceptually difficult to answer question relates to the validity of the local estimates. The validity of these estimates can only be assessed by using additional, and preferably independent, data, which allows a rather direct interpretation of the danger level. Such data are observations on avalanche occurrences, as these allow to validate the local danger level estimates for the higher danger levels. Therefore, local danger level estimates were compared with avalanche occurrence data for the region of Davos (Switzerland). This allowed the discussion of the validity of local danger level estimates, at least for the few days each winter when many large and very large natural avalanches were observed, situations which represent danger level 4-High.

### **Estimating the quality of regional avalanche forecasts**

Until now, studies that have explored the accuracy of forecast danger levels using local danger level estimates did not incorporate the reliability associated with the danger level estimate of the nowcast or hindcast assessor (e.g. Jamieson et al., 2008; Suter et al., 2010; Sharp, 2014; Statham et al., 2018b). However, based on the findings regarding the reliability of local danger level estimates, the observed forecast accuracy - like the proportion of forecasts when the forecast danger level and the local danger level estimate matched - can be put into perspective, allowing an estimation of the reliability of the forecast. Thus, the large data set of local danger level estimates, described in the previous section, was compared with the forecast danger level, integrating the knowledge gained on the reliability of local danger level estimates. The focus was on the following three research questions:

1. What implications do the variations identified between local danger level estimates have for the verification of regional avalanche forecasts?
2. Relying on local danger level estimates, what is the perceived accuracy and bias of forecast danger levels?
3. Can differences between countries with different operational constraints and verification methods be noted?

These questions were addressed in detail in publication 2 (Fig. 1.1). Again a selection of findings is presented in this Synthesis (Section 5.3, p. 54ff) together with verification data sets published in other studies

(Canada: Statham et al., 2018b) or provided by avalanche forecasters in Norway and Colorado (Ekker, 2018; Logan, 2020), for a comparison.

Furthermore, to obtain a different perspective on forecast quality, at least for the most critical days (4-High or 5-Very High), the forecast danger level was compared to avalanche observations in the region of Davos (Switzerland).

### **Refined avalanche danger ratings in regional forecast**

Publications 1 and 2 addressed the consistency and quality of assessing avalanche danger using the resolution of the established five danger levels, as defined in the danger scale (see also Sect. 2.3). Publication 3 (Fig. 1.1) took a different perspective and was motivated by the following two observations: Firstly, summary statistics showing the distribution of published avalanche forecasts indicate that the distribution of forecast danger levels is not very refined: on three of four days, the forecast danger level was either 2-Moderate or 3-Considerable (e.g. Logan and Greene, 2018; SLF, 2017). And secondly, even though assigning and communicating a single danger level may be easier to understand for a user than a probabilistic forecast, categorical values result in the maximum loss of information (Murphy, 1993). This is due to the fact, that the probability assigned to a categorical value (the danger level) is always 100% (Doswell and Brooks, 2020), and the uncertainty related to it can only be expressed in the danger descriptions. Therefore, avalanche warning services emphasize that forecast users refer to the danger description accompanying the forecast to obtain more detailed information. However, the provision of quantitative, higher resolved information is necessary for instance as input for computer models which provide risk assessments based on avalanche forecasts (Schmudlach and Köhler, 2016; Schmudlach, 2016).

This challenge - communicating avalanche danger in a simple and well-established manner on one side, while simultaneously assessing avalanche danger in greater detail on the other side - lead to the question whether sub-levels, assigned to a danger level during the forecast process, actually have skill. Or, in other words:

- Can the forecast regional danger level be refined by assigning a sub-level?
- Are these sub-levels significantly better than randomly assigned ones?

To answer these questions, a newly compiled four-year data set of published avalanche forecasts in Switzerland was analyzed, with forecasts not only including the forecast danger level but also an unpublished sub-level. These were compared with local danger level estimates as a reference standard.

A summary of the findings from publication 3, titled «Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random?», is given in Section 6.1.1). The publication is appended in Sect. A.4 (p. 142ff).

#### **1.3.1.3 Value**

As pointed out by Murphy (1993), forecasts have no intrinsic value. A forecast becomes valuable to a user, if - for instance - the benefits of using the forecast during the decision-making process are greater compared to a situation when the forecast is absent (e.g. Hilton, 1981; Murphy, 1993). Furthermore, value is influenced

by the consistency and quality of the forecast product (Murphy, 1993). Only if forecasts are sufficiently reliable, will they have value for a user.

Even though value was not explored in the sense of a data analysis, the implications inconsistencies or deficiencies in quality may have to users will be briefly discussed (Section 6.1.4).

### **1.3.2 On the importance of the elements characterizing avalanche danger**

More than 25 years ago (in 1993), avalanche forecasters of the warning services of five Alpine countries and Spain relied on their combined experience and knowledge, when developing the five-level ordinal European Avalanche Danger Scale (EADS, introduced in detail in Sect. 2.3). Since then, only minor changes were made to the original version of the EADS. In the EADS, the danger levels are described by snowpack stability, the frequency distribution of snowpack stability, and avalanche size. However, the descriptions of the key factors for each of the five categories of danger level leave ample room for interpretation and are even partly ambiguous. Furthermore, the three key factors characterizing avalanche danger are not clearly defined and hence poorly quantified (Schweizer et al., 2020). These may be reasons for inconsistencies in the use of the danger levels between individual forecasters or field observers, or between different forecast centers and avalanche warning services (Lazar et al., 2016, and as noted in publications 1 and 2).

The objective was therefore to address this lack of quantitative evidence by exploring observational data relating to the three key elements of avalanche danger: snowpack stability, the frequency distribution of snowpack stability, and avalanche size. To achieve this task, a large data set of stability tests and avalanche observations, together with a locally estimated danger level, originating from two countries (Switzerland and Norway) with different snow climates, was compiled. While avalanche observations have been compared to avalanche danger in several studies (e.g. Logan and Greene, 2018; Schweizer et al., 2020), only a few studies have shown snowpack stability distributions, typical at the danger levels, for a small number of days using labor-intensive sampling in the field (Schweizer et al., 2003). Therefore, and based on the large data set of stability tests, stability distributions were simulated using a bootstrap sampling approach. This allowed for the first time a data-driven description of the danger levels, with two research questions being in the center of the study:

1. How do the three elements characterizing avalanche danger - snowpack stability, the frequency distribution of snowpack stability, and avalanche size - relate to the danger levels?
2. Which combination of the actual value of the three elements does best describe the various danger levels?

These questions were addressed in publication 4 (Fig. 1.1), titled «On the importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level». Again, a selection of findings is presented in this Synthesis (Sect. 5.4, p. 61ff). Publication 4 can be found in the Appendix A.5 (p. 160).

However, to include the stability test data from Norway in this analysis, it was necessary to develop a classification scheme for the Extended Column Test (ECT, a stability test, see also Sect. 3.3.1 for details).

This resulted in publication 5 «On snow stability interpretation of Extended Column Test results». From this publication, the resulting classification scheme is briefly introduced (Section 3.3.1, p. 29) and a short summary given in Section 5.5 (p. 66). Publication 5 can be found in Appendix A.6 (p. 186).



# Chapter 2

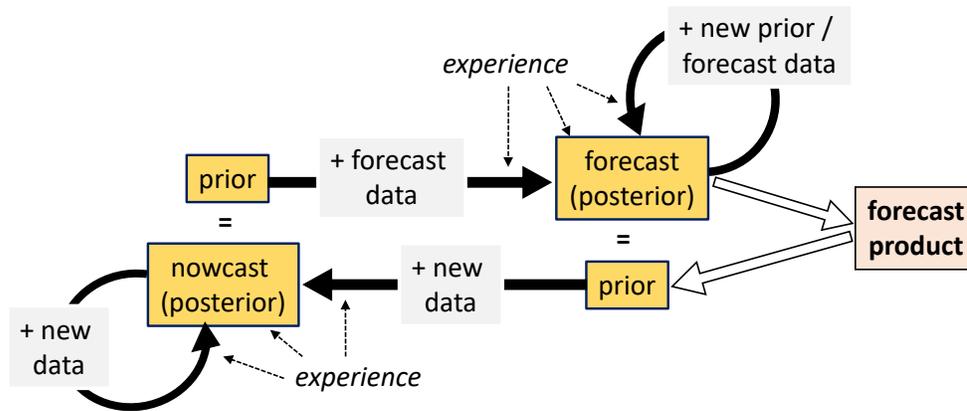
## Avalanche forecasting

### 2.1 Avalanche forecasting - from data to forecast

LaChapelle (1980), in his classic paper on the fundamental processes in conventional avalanche forecasting, describes the forecasting process as the cumulative integration of a widely diverse body of information through time, with final forecast decisions being reached through inductive logic. In this iterative process of data integration (observations, measurements, and other pieces of information representing Nature) an initial hypothesis regarding the current or future state of the snowpack is formulated, which is revised when new pieces of information become available (LaChapelle, 1980) (Fig. 2.1). McClung (2002a), taking the perspective of the human forecaster, describes the goal of forecasting as one in which the perception about the temporal and spatial distribution of instabilities in the snowpack should match reality as closely as possible. This is attained through objective analysis of relevant data, even though objectivity is hard to achieve as the analysis is performed by a human (Fig. 2.2). Hence, a subjective component is always present in data analysis and forecasting. Furthermore, what is considered a relevant piece of information varies depending on avalanche conditions, but also on the forecasters' knowledge and experience (LaChapelle, 1980; McClung, 2002a). For instance, in situations, when the danger is expected to increase due to changing weather and the associated expected changes to snowpack stability, forecast data might be the most relevant. In other situations, the combination of observed evidence concerning the current conditions and comparably (minor) or slow changes in snowpack stability may be of great relevance to predict future avalanche conditions. As forecasters are trying to predict instability, information that directly relates to instability is generally the most sought after (see Section 2.1.1). Furthermore, an office-based forecaster may rely more on a mix of different data sources than someone assessing avalanche danger in the field.

#### 2.1.1 Data

Avalanche forecasters rely on a wide range of data to assess and predict snowpack stability. While some data are more directly related to snowpack instability, other data provide only indirect evidence. (McClung and Schaerer, 2006) According to the ease of deriving information regarding snowpack instability, data have been grouped into three classes, with lower classes indicating a more direct relation to snowpack instability



*Figure 2.1:* Avalanche forecasting is an iterative process. Ideally, starting with an objective prior, for instance, the previous days forecast (McClung, 2002a), a nowcast assessment is formulated and constantly updated with new, relevant information becoming available (LaChapelle, 1980). This nowcast assessment becomes the prior, which the forecast builds on. In a similar iterative process, the forecast may change with changes to the nowcast assessment or with new forecast data becoming available. The final forecast is published in a forecast product. This forecast becomes the new prior for the next nowcast assessment. Data interpretation and forecasting are performed by human experts, and are hence influenced by the forecasters' knowledge and previous experience.

than higher classes (LaChapelle, 1980; McClung and Schaerer, 2006):

- Class 3: meteorological factors
- Class 2: snowpack factors
- Class 1: stability factors

### **Class 3 data: meteorology factors**

Class 3 data are used to predict both the current and future snowpack stability. However, class 3 data requires the forecaster to translate, how the weather influences the snowpack, and thus stability. Class 3 data are available through measurements, often near real-time at automatic weather stations, but also through weather models (Tab. 2.1). Examples include measurements or forecasts of air temperature, wind speed, or precipitation amounts.

### **Class 2 data: snowpack factors**

Class 2 data provide evidence about the layering of the snowpack, as for instance the presence, depth, and strength of weak layers (McClung and Schaerer, 2006). Class 2 data include, for instance, past avalanches - in the sense that the snowpack is different in locations where avalanches have released, past slope use - as by skiers disturbing the continuity of snowpack layering, or information gained from manual snow profiles (Tab. 2.1). As for class 1 data, snowpack information was gained for a long time primarily through observations made in the field. However, nowadays, physical snowpack models have become increasingly important for avalanche danger assessment (Morin et al., 2019). Information gained from snowpack models used in

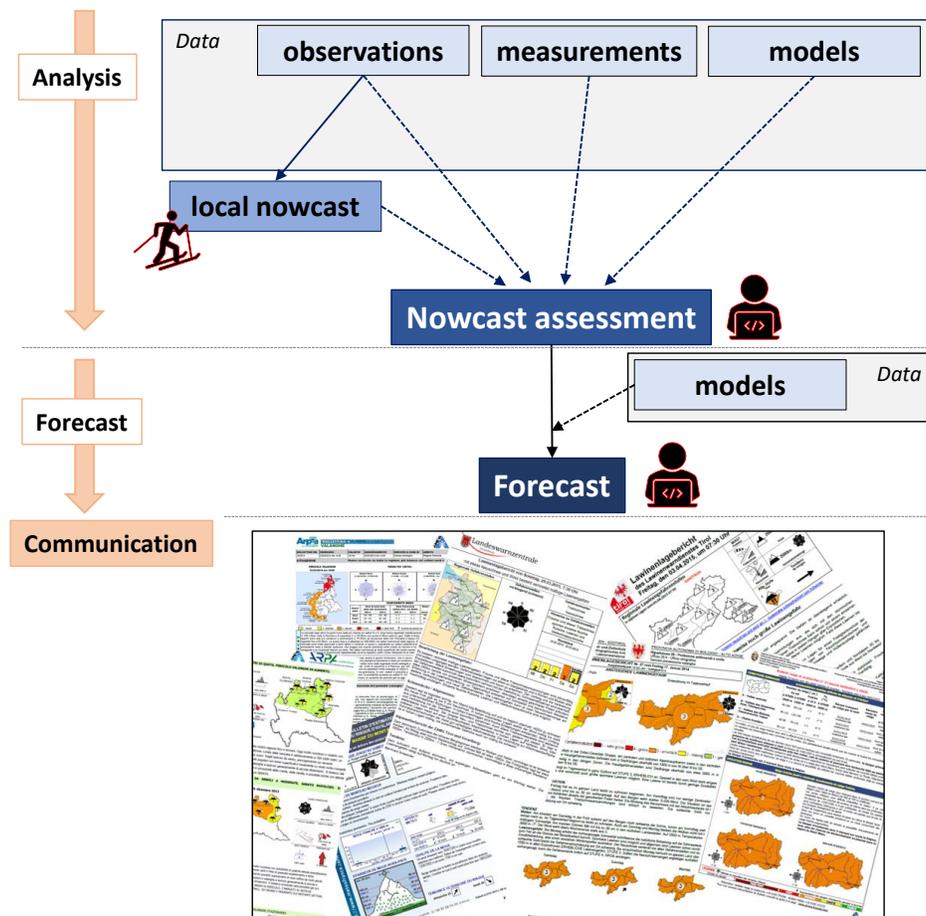


Figure 2.2: Workflow and data sources used in public avalanche forecasting. From the data, that originates from observations, measurements, or models, but may also include expert assessments, a forecaster formulates a regional nowcast assessment of the avalanche conditions. This best assessment of the current state, combined with forecast data, leads to the forecast avalanche conditions. As schematically represented in Fig. 2.1, several iterations may be required to reach the final danger assessment. The final step is the communication of the expected avalanche danger to the public. The human forecaster is directly involved in the process by analyzing and interpreting data and making decisions regarding what to communicate.

some warning services include, for instance, the presence and depth of weak layers (e.g. Monti et al., 2014) and their spatial distribution (e.g. the distribution of surface hoar, Horton et al., 2014), the advance of the melt water front with melt or rain or snowpack wetting in general (e.g. Mitterer and Schweizer, 2013; Wever et al., 2018) or whether snow-drifting due to wind occurs (e.g. Lehning and Fierz, 2008; Vionnet et al., 2018).

### Class 1 data: stability factors

Class 1 data are most closely linked to snowpack instability. Class 1 data include observations of current avalanches, stability tests, but also observations on fracture and cracking of the snowpack, so-called signs of instability (Tab. 2.1). These are the most important pieces of information to be gained in avalanche terrain, and are sometimes considered *obvious clues* (McCammon, 2000; McCammon and Haegeli, 2004). Their recording standards are described in all the operational observation guidelines (e.g. the guidelines in

Table 2.1: Data classes according to McClung and Schaerer (2006), showing the source and an example. Emphasis is on typical data sources used in forecasting. Brackets indicate less typical, but still relevant sources. Local nowcast estimates of avalanche danger ( $D_{LN}$ ), although not part of the data classification by McClung and Schaerer (2006), are also a data source used by some forecasting services.

class	factors	source	parameter (example)
3	meteorology	measurements	air temperature
		models	wind speed
		(observations)	precipitation type
2	snowpack	observations	blowing snow
		measurements	snowpack temperature
		models	snowpack layering
1	stability	observations	avalanches, stability tests
		measurements	avalanches
		(models)	stability indices
–	local $D_{LN}$ estimates		

Canada, the U.S. and Switzerland: CAA, 2014; Greene et al., 2016; Dürri and Darms, 2016).

In many avalanche forecasting operations, class 1 data is primarily obtained through field observations. However, direct evidence of instability - as recent avalanches, shooting cracks, or whumpf sounds - is often lacking. When such clear indications of instability are absent, snowpack instability tests are widely used to obtain information on the stability of the snowpack. Such tests provide information on failure initiation and subsequent crack propagation - essential components for slab avalanche release (Schweizer et al., 2008b). However, performing snowpack instability tests is time-consuming, as they require to dig a snow pit. Furthermore, considerable experience in the selection of a representative and safe site is needed, and the interpretation of test results is challenging (Schweizer and Jamieson, 2010). Alternative approaches such as interpreting snow micro-penetrometer signals (Reuter et al., 2015) are promising, but not sufficiently established yet.

More recently operational systems to automatically detect avalanches using, for instance, infra-sound systems (e.g. Mayer et al., 2020), seismic sensors (e.g. van Herwijnen and Schweizer, 2011) or satellite remote-sensing (Eckerstorfer et al., 2017) have become available, and are increasingly implemented in operational avalanche warning. Furthermore, physical snowpack models - like the *Crocus* (e.g. Brun et al., 1989, 1992) or *SNOWPACK* models (e.g. Lehning et al., 1999) - also provide indices describing the stability of the snowpack. Morin et al. (2019) provides a review on current implementations of snowpack models and their application in operational avalanche forecasting.

Finally, avalanches are the most direct piece of information: if an avalanche releases, there is no doubt that the snowpack was unstable at the location of the avalanche. Such singular pieces of information may

have such a high relevance that they can override a stability estimate based on many other pieces of information before-hand (McClung, 2011).

### **Local nowcast assessments as data source**

And lastly, there is another potential data-source in public avalanche forecasting, namely, using the expert assessments of avalanche danger provided by specifically trained observers.

In some warning services, observers with sufficient experience and presence in avalanche terrain provide an estimate of the avalanche danger level together with their observations (e.g. in Norway and Switzerland Kosberg et al., 2013; Suter et al., 2010). They assess avalanche danger according to the same five-level avalanche danger scale as is used by the forecasters (see Sect. 2.3 regarding the European Avalanche Danger Scale). Observers are advised to integrate all available information into their local estimate of the danger level ( $D_{LN}$ ), including not just the observations from the day of observation, but also prior knowledge concerning the development of the snowpack during the winter or information from third parties. To assure consistent and high-quality feedback, all observers are regularly trained.

The avalanche danger is assessed locally. The area considered is the area of observation during the day in the backcountry or in the ski area, or the area that can be seen from the observation point in the valley floor; this area is approximately 10 km<sup>2</sup> (Jamieson et al., 2008) to 25 km<sup>2</sup> (Meister, 1995).

So far, observational data was described that is provided by specifically trained observers. However, many warning services also rely on observations provided by the public (e.g. Tremper and Diegel, 2014). Most often, this kind of feedback provided by the public is strongly event-driven, that is, people tend to report avalanche occurrences, less severe avalanche involvements or signs of instability, rather than observations where these were absent (Fig. 2.3).

### **2.1.2 Data analysis**

Data analysis in office-based regional avalanche forecasting is challenging for many reasons: Data stems from a variety of data sources, including observations and subjective estimates, with the most relevant data, which are data directly relating to snowpack instability (class 1), often being sparse in time and space. Further, the true state is generally not measurable, as is the case for the avalanche danger level. Therefore, the forecaster is required to make an inference using few data points, in a spatially highly variable environment. Despite the advent of computer technology and snowpack modeling, data aggregation and data interpolation are still predominantly manual tasks in avalanche forecasting (e.g. in Canada, Floyer et al., 2016). Thus, experience and a sound knowledge of the data sources play an important role when interpreting data to estimate the current and future avalanche conditions. While this traditional way of analyzing data may still dominate, some avalanche warning services undertake great efforts to develop tools, which aggregate, assimilate and help analyze data, which model and visualize physical snowpack parameters and aid in avalanche danger determination (e.g. in Canada: Floyer et al. (2016); Horton et al. (2019) or in France: Vernay et al. (2015)).

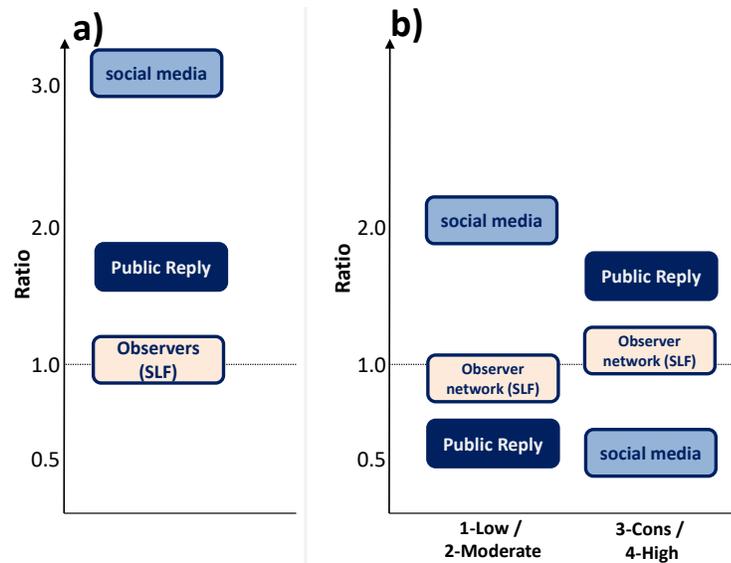


Figure 2.3: Availability of observational data in Switzerland (SLF), depending on the data source. Three data sources are shown: observations reported by the network of trained observers, submitted by the public through the reporting tool *Public Reply*, or entries on social media networks like bergportal.ch and Camptocamp.org, which are also regularly scanned by forecasters for relevant information regarding avalanche conditions. In (a), the ratio of reports submitted on weekend-days compared to week-days is shown. A ratio of 1 means that the same number of reports were received on each of these days, while a ratio of 3 means that reports were three times more frequent on a weekend-day compared to a weekday. In (b), the ratio of reports submitted as a function of the forecast avalanche danger level is shown. A ratio lower than 1 indicates that reports were less frequently submitted compared to the frequency these danger levels were forecast, and vice versa for ratios greater than 1. (Data: Techel (2018))

### 2.1.3 Workflow and estimation of avalanche danger

LaChapelle (1980) discussed the iterative workflow to forecast avalanche danger. He also showed that there is not one specific set of observations, which clearly allows the formulation of how these pieces of evidence are interpreted to assess avalanche danger. Even today, there is no equation, which could be used to calculate avalanche danger. In fact, the same danger level can be described with different combinations of the three elements of avalanche danger - snowpack stability, the frequency distribution of snowpack stability, and avalanche size (EAWS, 2017b; Müller et al., 2016).

Recently, Statham et al. (2018a) formally described the workflow in their Conceptual Model of Avalanche Hazard (CMAH). In the proposed workflow structure (Fig. 2.4), the forecaster answers four sequential questions (Statham et al., 2018a, p. 663):

1. What type of avalanche problem(s) exists?
2. Where are these problems located in the terrain?
3. How likely is it that an avalanche will occur?
4. How big will the avalanche be?

These questions aim at answering three different questions, related to the three elements of avalanche danger:

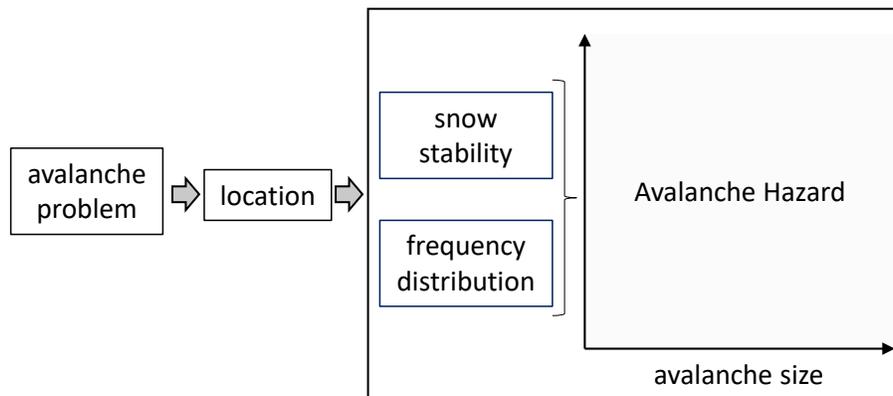


Figure 2.4: Workflow according to the Conceptual Model of Avalanche Hazard (CMAH, Statham et al., 2018a). Here, the terms used in Europe are shown.

1. Snowpack stability: What trigger is required to release an avalanche?
2. Frequency distribution of snowpack stability: How frequent are these most unstable locations?
3. Avalanche size: What is the expected size (or destructive potential) of avalanches?

However, some steps were not described in the CMAH:

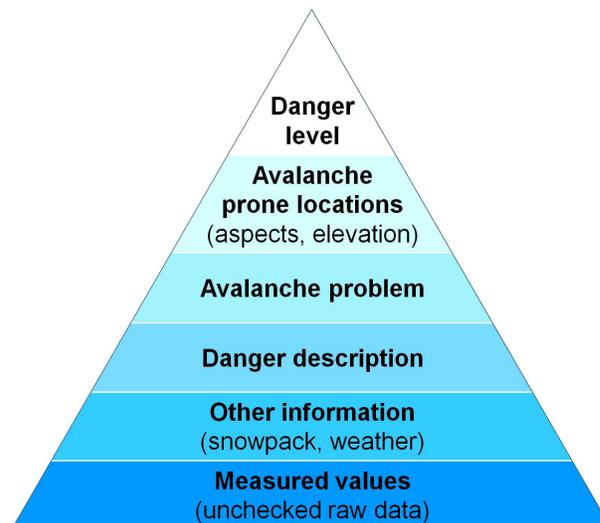
Firstly, given the spatial variability of avalanche conditions, an avalanche forecaster working at the scale of several thousand square kilometers, often has to find spatial patterns in the data, which in turn will define regions with similar avalanche conditions. This spatial analysis has never been formally described.

And secondly, the CMAH lacks the description how a regional forecaster arrives at a specific danger level (Fig. 2.4). This point was addressed by the European Avalanche Warning Services (EAWS): to improve consistency in the use of the danger levels, a decision aid, the *Bavarian Matrix* was adopted in 2005. The *Bavarian Matrix*, a lookup 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, 2020d). More recently, a so-called *Avalanche Danger Assessment Matrix* (ADAM, Müller et al., 2016) was proposed, which combines the workflow described in the CMAH with the assignment of the danger levels based on the three elements as suggested in the *EAWS-Matrix*. Both, the current version of the *EAWS-Matrix* and ADAM, are work in progress.

## 2.2 Public avalanche forecasts

Avalanche forecasts are the primary means for avalanche warning services to provide publicly available information about current and forecast snow and avalanche conditions in their territory. They may take the form of a single advisory, describing the current situation, or an advisory and forecast for one or more days. Typically, avalanche forecasts contain the following information, ranked according to importance (information pyramid, Fig. 2.5; EAWS (2020c)):

1. avalanche danger level (in Europe according to the European Avalanche Danger Scale, Table 2.2)



*Figure 2.5:* In Europe, all avalanche forecast products are structured according to the information pyramid. The most important information - the avalanche danger level - appears first in the avalanche bulletins. The information becomes more detailed, when moving down the pyramid (SLF, 2018).

2. terrain (aspect and elevation) where the danger prevails
3. typical avalanche problems - describing the nature of the cause of instability encountered in avalanche terrain (EAWS, 2020b)
4. danger description - a text description providing information concerning the avalanche situation
5. information concerning snowpack and weather, measured values

These publicly available forecasts of avalanche danger are provided by avalanche warning services, which are national, regional or provincial agencies. These may either be a service with a single forecast center or with several forecast centers in different locations.

Regional avalanche forecasts are issued for a specific time span and region. However, considerable variations exist in the publication frequency, the underlying spatial resolution, and the way spatial variations and temporal changes in avalanche danger are communicated. These operational settings (or constraints) are of importance, as these - together with the availability of data and resources - define the spatial and temporal granularity of the avalanche danger assessment in the production process, but also in the way avalanche danger is communicated in the forecast product.

### **Temporal validity and publication frequency**

The issuing time, temporal validity and publication frequency of the forecasts vary between forecast centers: In 2018 most of the European Avalanche Warning Services (EAWS) members updated their forecasts daily during the main winter season, often in the afternoon or evening with a forecast until the following day (22 out of 28 warning services; Engeset, 2019). Thus, most of the avalanche forecasts published by EAWS members covered 24 hours. In contrast, in North America in 2020, some of the smaller forecast centers

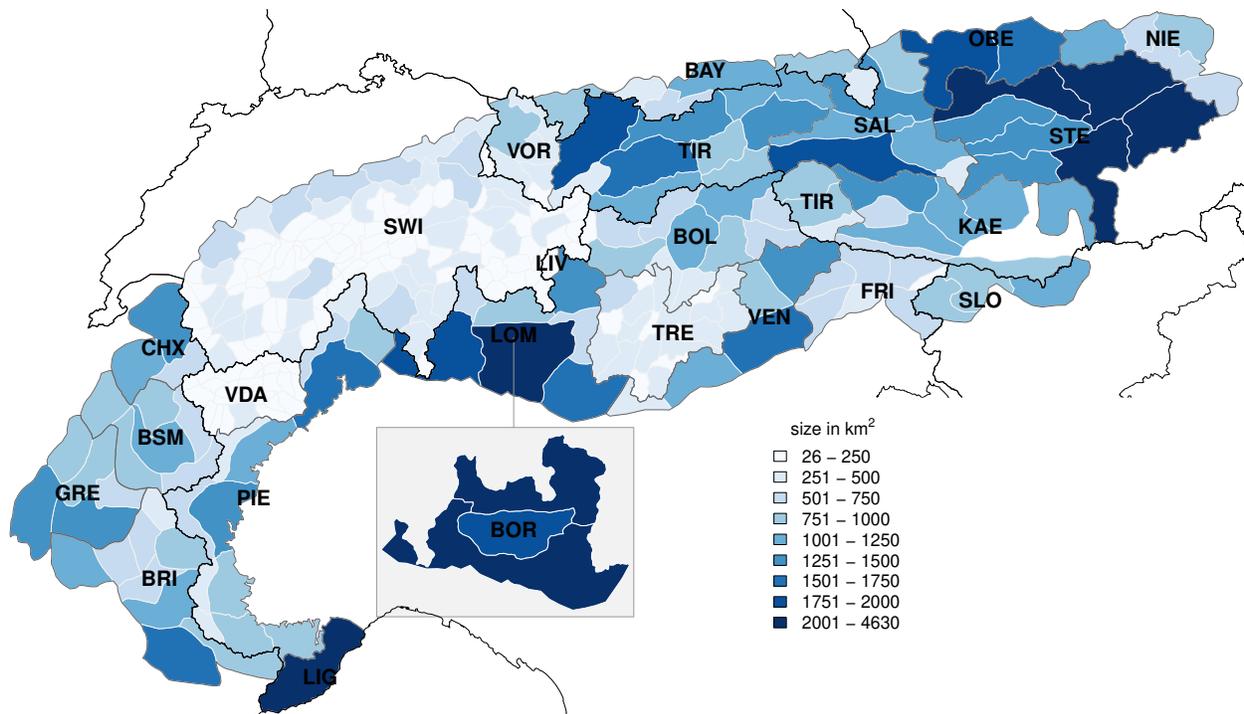


Figure 2.6: Map showing the European Alps with the individual warning regions (white polygon outlines) and their size (color shading of polygon). Three-letter labels correspond to different forecast centers according to Table A.2 in publication 1 (Appendix A.2). Additionally, national (black lines) and forecast center boundaries (grey polygon outlines) are shown. To visualize the (at least partially) overlapping forecast regions in the Italian region of Lombardia, LIV is superposed onto parts of LOM, while BOR is placed as inset to the south of LOM.

issued forecasts only two or three times a week (Canada: J. Floyer, 2020; United States: K. Birkeland, private communications).

### Warning regions

Warning regions, in North America often called forecast zones, are geographically clearly specified areas permitting the forecast user to know exactly which region is covered by the forecast. They may be delineated by administrative boundaries (e.g. between countries, federal states, or regions and provinces), describe climatologically (e.g. in France; Pahaut and Bolognesi (2003)), hydrologically or meteorologically homogeneous regions, or may be based on orographic divisions (e.g. Italy; Marazzi, 2005), or a combination of these (e.g. Valle d'Aosta (Italy); Burelli et al., 2012).

In the Alps (in 2018), the median size of the warning regions was 350 km<sup>2</sup> with considerable variations (Fig. 2.6). The 25% of the smallest warning regions (size < 160 km<sup>2</sup>) were almost ten times smaller than the 10% of the largest regions (size > 1,310 km<sup>2</sup>). In contrast, in other European countries or in North America, warning regions (also called forecast zones) were sometimes even larger than 10,000 km<sup>2</sup> (Fig. 2.7, e.g. in Canada or Norway; Jamieson et al., 2008; Engeset et al., 2018).

The size of the warning regions depends on the approach used by an avalanche warning service to define the warning regions and to externally communicate avalanche danger, but also on the availability of data and resources. In its simplest case, a single danger level is either explicitly communicated for each warning

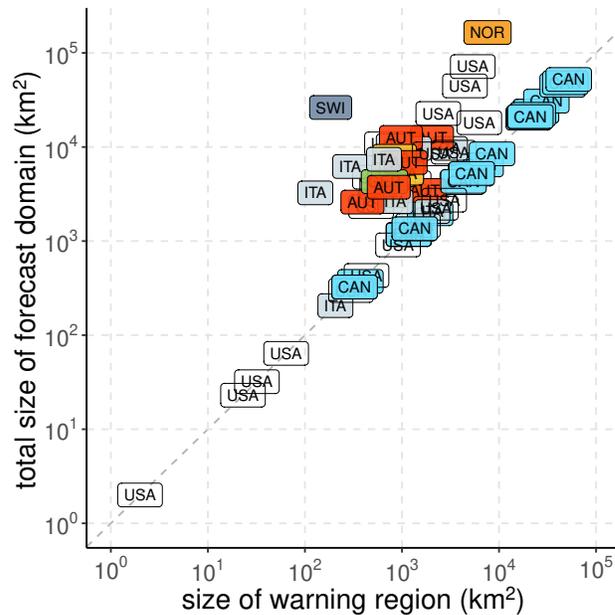


Figure 2.7: (Median) Size of the warning regions (called forecast zones in North America), the smallest spatial units used in the public avalanche forecasts, and the total size of the forecast domain for forecast centers in the Alps (AUT - Austria, FRA - France, GER - Germany, ITA - Italy, SWI - Switzerland; data: publication 1), Canada (CAN; data: AC (2019)), Norway (NOR; data: NVE (2018)) and the United States (USA; data: NAC (2020)). **Both the size of the warning regions and the size of the forecast domains vary by several orders of magnitude between forecast centers. Particularly in Canada and the USA, the forecast domain is often the only warning region (or forecast zone), with no subdivision into smaller regions as is typical in Europe.**

region or may be communicated for an aggregation of warning regions.

## 2.3 Avalanche danger scale

The danger levels - from 1-Low to 5-Very High - are described in the European Avalanche Danger Scale (EADS, Tab. 2.2; EAWS, 2018) or its North American equivalent, the North American Avalanche Danger Scale (e.g. Statham et al., 2010) with brief definitions of the key elements. The key elements that characterize avalanche danger are (Meister, 1995; EAWS, 2020d, 2018):

- the probability of avalanche release (or snowpack stability),
- the frequency and location of the triggering spots (or the frequency distribution of snowpack stability), and
- the expected avalanche size.

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

The 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;

Table 2.2: European avalanche danger scale (EAWS, 2018). The avalanche size classes are described in Tab. 2.3.

Danger level	Snowpack stability	Likelihood of triggering
5-Very High	The snowpack is poorly bonded and largely unstable in general.	Numerous very large and often extremely large natural avalanches can be expected, even in moderately steep terrain <sup>*</sup> .
4-High	The snowpack is poorly bonded on most steep slopes <sup>*</sup> .	Triggering is likely even by low additional loads <sup>**</sup> on many steep slopes <sup>*</sup> . In some cases, numerous large and often very large natural avalanches can be expected.
3-Considerable	The snowpack is moderately to poorly bonded on many steep slopes <sup>*</sup> .	Triggering is possible even from low additional loads <sup>**</sup> particularly on the indicated steep slopes <sup>*</sup> . In certain situations some large, in isolated cases very large natural avalanches are possible.
2-Moderate	The snowpack is only moderately well bonded on some steep slopes <sup>*</sup> ; otherwise well bonded in general.	Triggering is possible primarily from high additional loads <sup>**</sup> , particularly on the indicated steep slopes <sup>*</sup> . Very large natural avalanches are unlikely.
1-Low	The snowpack is well bonded and stable in general.	Triggering is generally possible only from high additional loads <sup>**</sup> in isolated areas of very steep, extreme terrain <sup>**</sup> . Only small and medium-sized natural avalanches are possible.

<sup>\*</sup> The avalanche-prone locations are described in greater detail in the avalanche bulletin (elevation, slope aspect, type of terrain [terrain profile, proximity to ridge, smoothness of underlying ground surface]):

moderately steep terrain: slopes shallower than about 30 degrees,

steep slopes: slopes steeper than about 30 degrees,

very steep, extreme terrain: particularly adverse terrain related to slope angle (more than about 40 degrees)

<sup>\*\*</sup> Additional loads:

low: individual skier / snowboarder, riding softly, not falling; snow-shoer; group with good spacing (minimum 10 m) keeping distances

high: two or more skiers / snowboarders etc. without good spacing; snowmachine; explosives

natural: without human influence

Meister, 1995). Hence, the probability of avalanche release refers to a specific location and relates to the local (or point) snowpack instability. The latter has recently been revisited and three elements were suggested to describe point snowpack instability: failure initiation, crack propagation and slab tensile support (Reuter and Schweizer, 2018).

The frequency and location of the triggering spots are typically unknown. So far, it can only be assessed with laborious extensive sampling (e.g. Birkeland, 2001; Reuter et al., 2016). However, in a regional avalanche forecast, the spatial distribution of snowpack instability can be described with regard to the frequency and the locations of triggering spots or more generally the locations where snowpack stability is lowest. From

Table 2.3: Avalanche size classification according to EAWS (2019).

Avalanche size	Potential damage	Typical length	Typical volume
1 - small	Unlikely to bury a person, except in run out zones with unfavourable terrain features (e.g. terrain traps). In extremely steep terrain, the danger of deep falls prevails the danger of burials.	10 - 30 m	100 m <sup>3</sup>
2 - medium	May bury, injure or kill a person. Corresponds to the typical skier-triggered avalanche.	50 - 200 m	1'000 m <sup>3</sup>
3 - large	May bury and destroy cars, damage trucks, destroy small buildings and break a few trees. When skiers are caught by avalanches of this size, probability for severe consequences are very high.	several 100 m	10'000 m <sup>3</sup>
4 - very large	May bury and destroy trucks and trains. May destroy fairly large buildings and small areas of forest.	1 - 2 km	100'000 m <sup>3</sup>
5 - extremely large	May devastate the landscape and has catastrophic destructive potential.	> 2 km	> 100'000 m <sup>3</sup>

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 and/or slope aspects and elevation bands. The frequency distribution describes the question «How often do spots with a certain 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 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, 2017b). In contrast, the location of triggering spots (or of snowpack stability) refers to «Where in the terrain is avalanche release most likely?» It indicates where in the terrain the frequency is slightly higher (e.g. *where the snowpack is shallow, close to ridgelines, in bowls, . . .*). 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*.

Finally, avalanche size is defined with sizes ranging from 1 to 5 relating to the destructive potential of an avalanche (Tab. 2.3; e.g. CAA, 2014; EAWS, 2019; McClung and Schaerer, 1981).

A danger level never refers to an individual slope, but always describes avalanche danger in a region (EAWS, 2020a).

# Chapter 3

## Data

The research questions introduced in Section 1.3 were addressed in five different publications, which all relied on different - and newly compiled - data sets. Even though the data sets were specifically compiled for this thesis, all the data were either published as part of a regional avalanche forecast, or they were collected for operational avalanche forecasting.

The research questions in publications 1 - 3 explored aspects like the spatial consistency and bias of forecast danger levels across forecast center boundaries (publication 1), the reliability of local danger level estimates (publication 2) or the accuracy of forecast danger levels (publications 2 and 3). Thus, and even though a forecast or estimated danger level was the focus in all of these studies, three completely different data sets were compiled. For publications 1 and 2, for which results are presented in more detail in Section 5, forecast danger levels originated from 23 warning services in the Alps (publication 1) or Switzerland (locally assessed and forecast danger levels) (publication 2). To compare (and discuss) the findings in publication 2 with other warning services/countries, additionally data from Canada, Colorado (U.S.) and Norway (locally assessed and forecast danger levels) were used in this Synthesis. Furthermore, as an additional approach to explore the validity of local danger level estimates and the quality of forecast danger levels for the most critical days, local estimates and forecasts were compared to a large data set of mapped avalanches (Swiss data only). Publication 3 explored a data set of Swiss avalanche forecasts, including not only the forecast danger level but also an internally assigned sub-level. The data set is described in detail in publication 3 (Appendix A.4, p. 142).

In contrast, the research objective in publication 4 was to describe the key elements of avalanche danger, namely snowpack stability, the frequency distribution of snowpack stability and avalanche size, by combining the results from snow stability tests and avalanche observations, considered class 1 data (see Sect. 2.1.1), with a locally estimated danger level. To undertake this analysis, the development of a stability interpretation scheme for the Extended Column Test (ECT), a stability test popular with many snow safety professionals, was necessary. This required the compilation of a fifth data set combining stability test data with observations relating to snow instability described in detail in publication 5 (Appendix A.6, p. 186).

Fig. 3.1 provides an overview of the data used in this Synthesis. In the following, a brief overview is given for the data used in this Synthesis. All of these data sets are described in detail in the respective publications

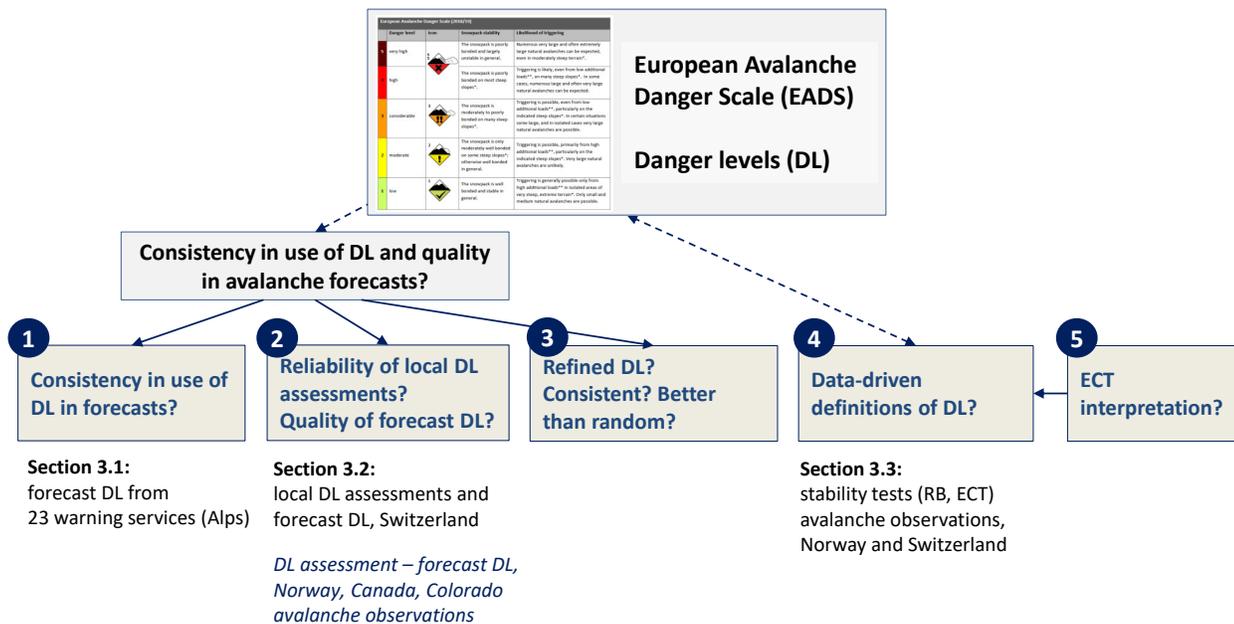


Figure 3.1: Overview showing the five publications with their main research question(s). The data used in publications 1, 2 and 4, which are presented in greater detail in this Synthesis, are briefly introduced in the respective Sections indicated. The data used in publications 3 and 5 are described in detail in the Appendices A.4 and A.6, respectively.

in Appendices A.2 - A.6.

### 3.1 Spatial consistency and bias: regional forecast danger level

To explore the research questions of publication 1, which focused on spatial consistency and bias in forecast danger levels ( $D_{RF}$ ), a data set of  $D_{RF}$  was compiled. For this, all avalanche warning services providing regional forecasts in the Alps were approached regarding data of their forecast danger level  $D_{RF}$ , as together these provide a spatially continuous forecast region (Fig. 1.2).

The data set  $DL_{Alps}$  contains 142,465  $D_{RF}$ , issued by 23 of the 30 warning services in the Alps for the four winters 2011-2012 to 2014-2015 (477 forecast days and 291 warning regions). The data set is described in detail in publication 1 (Appendix A.2, p. 93ff).

### 3.2 Reliability of local danger level estimates and quality of forecast danger levels

#### 3.2.1 Local nowcast estimates of danger level

To explore variations in local danger level estimates ( $D_{LN}$ ) (publication 2), local danger level estimates reported by specifically trained observers were extracted from the operational database at SLF. Updated again in 2019, data set  $DL_{SWI}$  contains 11,760  $D_{LN}$  estimates reported by observers after a day in the field relat-

ing to dry-snow conditions from the winters 2008-2009 to 2018-2019. Even though other parameters were available, like the aspects and elevations where the danger prevailed, only  $D_{LN}$  was analyzed. Location coordinates were available. The data set is described in detail in publication 2 (Appendix A.3, p. 124ff).

To compare the findings based on Swiss data with other data, a similar Norwegian data set was analyzed. These data were provided by Ragnar Ekker, at the time avalanche forecaster at the Norwegian Avalanche Warning Service NVE. The Norwegian data is public and can be retrieved via an API ([www.api.nve.no](http://www.api.nve.no)) or the Python module *varsomdata* ([www.github.com/NVE/varsomdata](https://www.github.com/NVE/varsomdata)). Data set  $DL_{NOR}$  contains 4,511  $D_{LN}$  estimates from the four winters 2013-2014 to 2017-2018, including the danger level estimate, the avalanche problem, the observers' competence level, and the warning region the observer was in (but not the exact location coordinates in the data set extracted by Ragnar Ekker, although these could be retrieved). As for  $DL_{SWI}$ , just the  $D_{LN}$  estimate was analyzed. More details regarding this data set are shown in Appendix B (p. 209).

### 3.2.2 Regional forecast danger level and nowcast assessments

A verification perspective was taken in publication 2 (Fig. 3.1), by comparing forecasts with local nowcasts. To this end, the data described in the previous Section (data set  $DL_{SWI}$ ) was linked with the forecast danger level ( $D_{RF}$ ) relating to dry-snow avalanche conditions on the same day and in the same warning region. The data set is described in detail in publication 2 (Sect. A.3, p. 124ff).

To compare the findings relying on Swiss data, several other data sets were explored, which included local or regional nowcast assessments, or hindcast assessments of avalanche danger (see also Fig. 3.2):

- The nowcast assessments by field observers in Norway ( $DL_{NOR}$ ), introduced in the previous section, also included the forecast regional forecast danger level for the warning region of observation. The data set contains 4,511  $D_{LN} - D_{RF}$  pairs from the four winters 2013-2014 to 2017-2018.
- A data set from Canada ( $DL_{CAN}$ ) consists of 2,774 pairs of the forecast  $D_{RF}$  and regional nowcast assessments of  $D$ , estimated by the forecasters during the preparation of the forecast for the following day and generally after they had been in the field in the morning.  $DL_{CAN}$  was explored and published by Statham et al. (2018b). The forecasts were issued by Parks Canada for the Banff-Yoho-Kootenay National Parks during the winters 2011-2012 to 2017-2018.
- A verification data set from Colorado ( $DL_{COL}$ ) was provided by Spencer Logan, forecaster at the Colorado Avalanche Information Center CAIC for the regional forecasts in Colorado/USA. The data set consists of 2,026 forecast - hindcast assessment pairs for the winter season 2017-2018. Both the forecasts and hindcast assessments were made by CAIC forecasters, but not necessarily the same one.

Additional information regarding these three data sets can be found in Appendix B (p. 209ff).

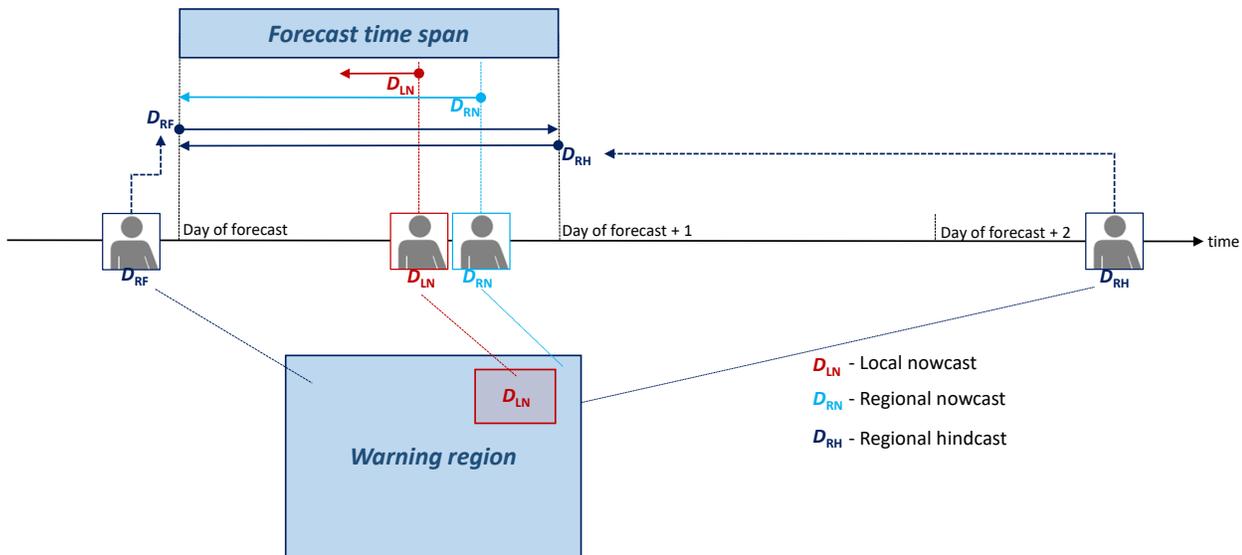


Figure 3.2: A forecast danger level ( $D_{RF}$ ), issued prior to the valid time span of the forecast and valid for the entire forecast time span and an entire warning region, was compared to a reference assessment. Reference assessments included local nowcast ( $D_{LN}$ ), regional nowcast ( $D_{RN}$ ) and regional hindcast ( $D_{RH}$ ) assessments of avalanche danger.  $D_{LN}$  refers to observers providing a danger level estimate after a field day. The assessment often describes avalanche conditions in an area smaller than the warning region, and for a few hours rather than the whole forecast time period. A regional nowcast assessment, for instance an office-based forecaster during the process of preparing next days' forecast, assesses the entire warning region and forecast domain. A regional hindcast assessment assesses avalanche conditions after the fact, for the entire forecast domain and the full forecast time span. At this stage, additional data (for instance avalanche observations) can be incorporated in the assessment .

### 3.2.3 Avalanche occurrence data

At some danger levels, as at 4-High or 5-Very High, avalanche observations may be used to verify a danger level (e.g. Elder and Armstrong, 1987; Schweizer et al., 2020). Therefore, a data set of mapped avalanches was used to assess the validity of local assessments and the quality of the forecast danger level for situations reflecting danger level 4-High.

This data set originates from the region of Davos (Switzerland) from the winters 2004-2005 until 2018-2019. It comprises information about avalanche activity (6,729 primarily natural avalanches) and the forecast avalanche danger level on 2,205 days. For each of these days, an avalanche activity index (AAI) was calculated. The AAI sums up all avalanches by assigning weights to their size (size 1 to size 4, weights 0.01, 0.1, 1, 10, respectively; Schweizer et al., 1998).

This data set of manually mapped avalanches, or subsets of these avalanches, has been used in several publications (e.g. Mitterer et al., 2009; Wever et al., 2018; Harvey et al., 2018; Schweizer et al., 2020) or Master thesis (Völk, 2020).

A more in-depth description of this data set can be found in the Appendix (Chapter B, p. 211).

### 3.3 Characterizing the elements of avalanche danger

The objective of publication 4 was a data-driven characterization of the three key elements describing avalanche danger - snowpack stability, the frequency distribution of snowpack stability, and avalanche size (Fig. 3.3). To achieve this objective, stability tests - to describe snowpack stability and the frequency distribution of snowpack stability - and avalanche observations - to describe avalanche size - were used. These data are most directly related to snowpack instability and are considered class 1 data (class 1 data are explained in Sect. 2.1.1).

#### 3.3.1 Snow stability tests: Rutschblock and Extended Column Test

Two data sets were compiled using two different stability tests popular to assess point snow instability: the Rutschblock test (Swiss data only) and Extended Column Test (data from Norway and Switzerland). Figures 3.4 and 3.5 show the two tests. The matrix in Fig. 3.6 shows how test results were classified into four stability classes. The classification scheme for the ECT was developed in publication 5. Both tests and their classification are described in detail in publication 4 and 5 (Appendices A.5 and A.6).

The Swiss RB data set comprised 4,439 RBs, the combined Swiss and Norwegian ECT data set contained 4,871 ECTs. For each of these test results, the locally estimated danger level relating to dry-snow conditions was available. More details regarding the data used in publication 4 can be found in Appendix A.5. (p. 160ff).

#### 3.3.2 Avalanche observations

The third contributing factor to avalanche danger is avalanche size. Accordingly, avalanche observations collected for the purpose of regional avalanche forecasting in Norway and Switzerland were explored. The data set consists of reported dry-snow avalanches, where the trigger type was either natural release or

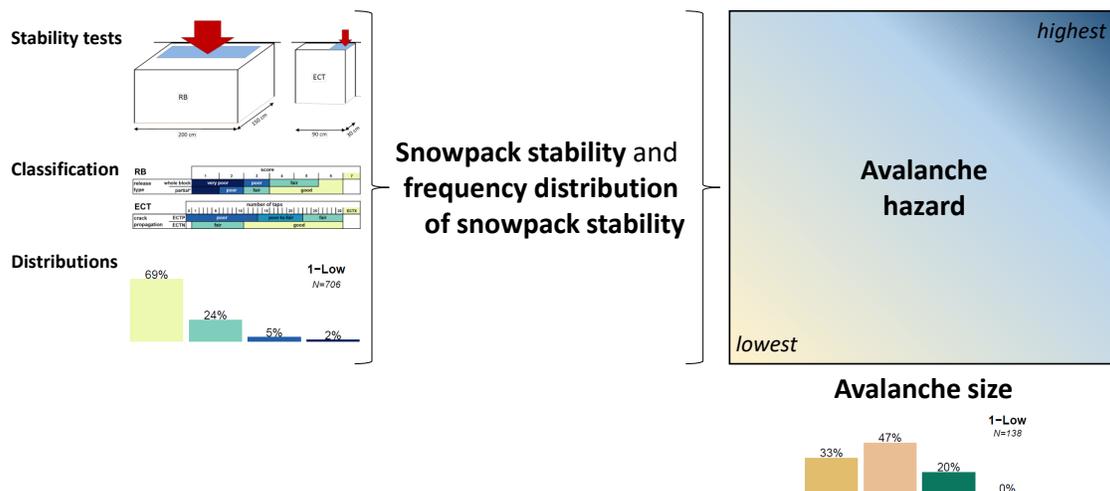


Figure 3.3: Avalanche hazard chart, with the three key elements, and the data used to describe these. Snowpack stability and the frequency distribution of snowpack stability was based on stability tests (Fig. 3.4), with test results classified according to Fig. 3.6.

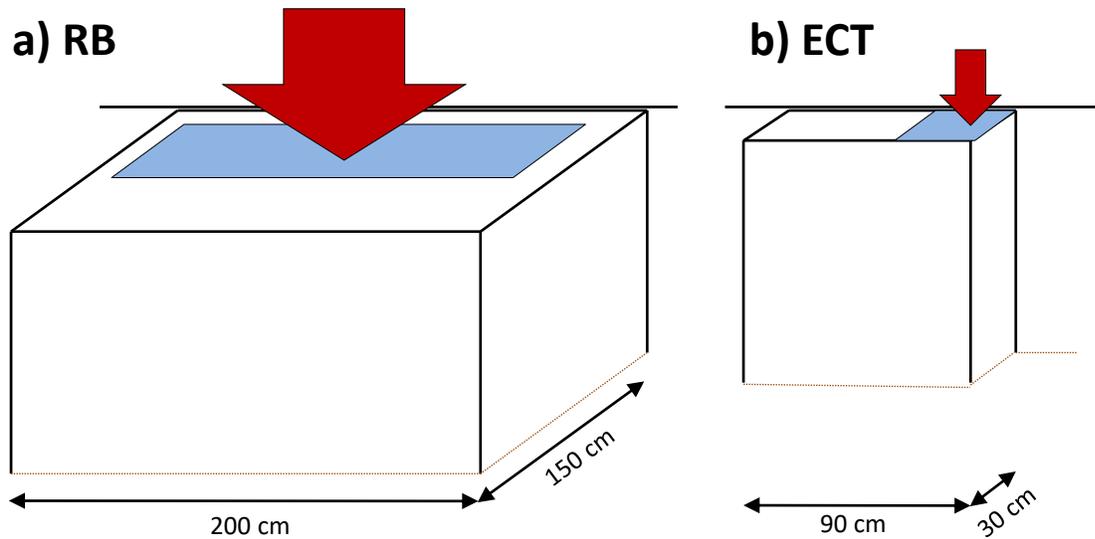


Figure 3.4: Rutschblock (RB, a) and Extended Column Test (ECT, b) according to observational guidelines (e.g. Dürri and Darms, 2016). For both tests, which are described in greater detail in Appendix A.6, blocks of snow are isolated from the surrounding snowpack. The block is then loaded in several steps according to test specifications. The light blue area indicates the approximate area, where the skis (RB) or the shovel blade (ECT) is placed. This area corresponds to the area loaded for the ECT, while the main load under the skis is exerted over a length of about 1 m (Schweizer and Camponovo, 2001). Loading is from above (arrows). The loading step leading to a crack in a weak layer (failure initiation) is recorded, and whether crack propagation across the entire block of snow occurs (crack propagation).



Figure 3.5: Rutschblock, loaded by a skier (left) and Extended Column Test, loaded by tapping on the shovel blade (right). Photos: P. Diener (left), I. Moner (right).

human-triggered, when a  $D_{LN}$  estimate relating to dry-snow conditions was provided for the release date of the avalanche(s) and in the same warning region. The data set contains information on the number of avalanches and their size for a given day and warning region, together with the  $D_{LN}$  estimate. Only days were considered when avalanches were observed.

The total number of avalanches was 39,000, observed on more than 8,000 different days and regions.

More details regarding the data used in publication 4 can be found in Appendix A.5. (p. 160ff).

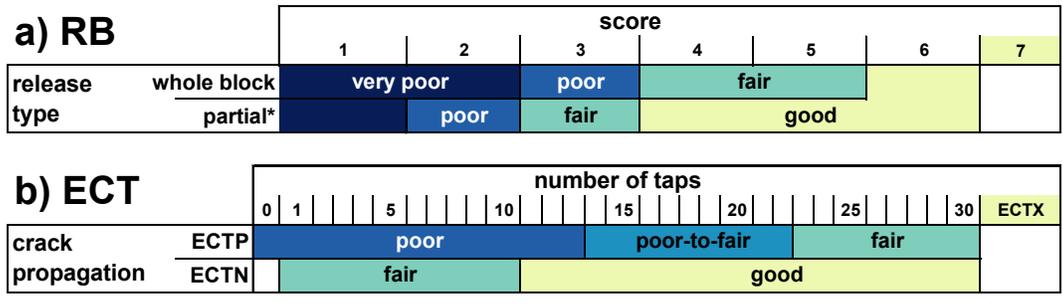


Figure 3.6: Stability classification of (a) Rutschblock test results (based on Schweizer (2007a); Techel and Pielmeier (2014)) and (b) Extended Column Test results (based on publication 5 shown in Appendix A.6 (p. 186)). \* - part of block includes release types most of block and edge only



# Chapter 4

## Methods

The statistical approaches and metrics chosen to answer the research questions were guided by the following criteria:

1. They had to be appropriate for the context of the study, permitting to answer the research questions,
2. apply to the data at hand,
3. with a preference given to metrics which are comparably easy to interpret, also for a non-scientific audience.

In publications 1 to 3 (Fig. 4.1), of which studies 1 and 2 are addressed in detail in this Synthesis, the data were exclusive of rank-ordered type, namely danger levels, either forecast or estimated by a human. The research questions explored the agreement (or correlation) between forecast or locally estimated danger levels as a function of avalanche conditions and distance (publication 1 and 2), or from a verification perspective (publication 2), including the detection of potential biases. Furthermore, in this Synthesis, the reliability of local danger level assessments and the forecast danger level was explored.

There is a wealth of literature on statistical measures exploring the reliability of observations or ratings by humans (e.g. Jacob et al., 1987), on the verification of categorical forecasts (e.g. Murphy, 1993; Wilks, 2011)

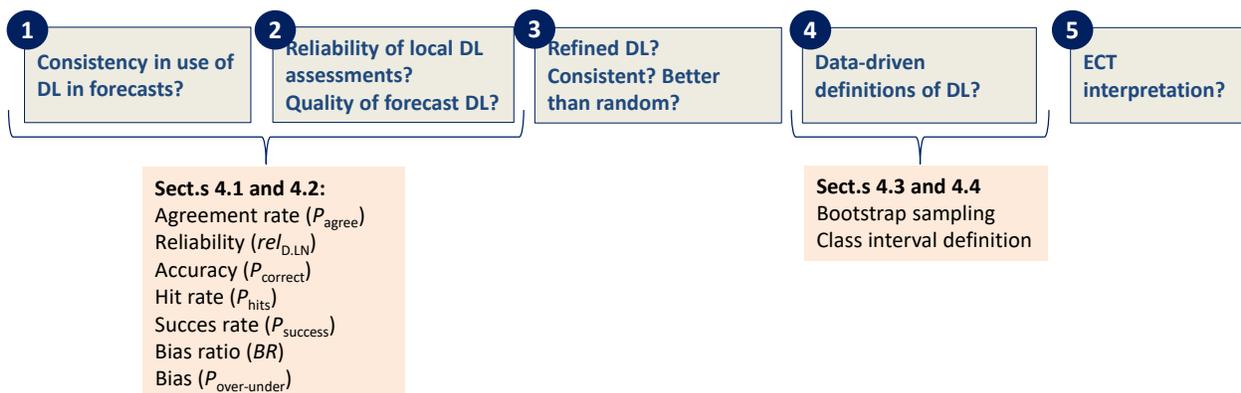


Figure 4.1: Overview showing the five publications (1-5) with their main research question(s), and the Section which describes the metrics and methods used in the Synthesis.

or on categorical data analysis in general (e.g. Agresti, 2007). The methods described in Sections 4.1 and 4.2 were chosen in a way that the metrics were relatively easy to understand; thus, most of the metrics describe proportions.

The research questions and the data in publications 4 and 5 (Fig. 4.1), of which publication 4 is presented in detail in the Synthesis, were different: The research questions aimed at describing the key elements defining avalanche danger by relying on a large data set of observations (stability tests, avalanche observations) together with locally estimated danger levels. To achieve this objective, it was necessary to generate a large data-set of stability distributions from a set of stability test results, for which bootstrap sampling (Efron, 1979) was applied (Section 4.3), and to find an appropriate method to derive frequency classes of these sampled stability distributions (Sect. 4.4).

In the following Sections, the metrics and approaches used in this Synthesis are described in detail.

## 4.1 Reliability of ratings by humans

The reliability of measurements, observations or ratings has been explored in many fields, for instance in psychology (e.g. Jacob et al., 1987), neuro-image analysis (e.g. Vul et al., 2009) or regarding observations used to verify weather forecasts (e.g. Bowler, 2006). In the context of observations or assessments provided by humans, reliability assesses the congruence between the assessment provided by two individuals (Jacob et al., 1987), or, in other words, the repeatability of the class indicated. In the context of this dissertation, reliability is therefore a measure describing the quality of the subjective danger level assessments provided by observers.

### 4.1.1 Agreement rate

Percentage agreement, here referred to as the agreement rate  $P_{\text{agree}}$ , is the most commonly used measure to estimate the reliability of observations or assessments (Jacob et al., 1987).

Regardless whether a forecast danger level is compared with a local assessment (Fig. 4.2, case 3), or whether two local assessments are compared with each other (Fig. 4.2, case 2),  $\Delta D$  describes the difference in the ranks of two danger levels  $D_i$  and  $D_j$ :

$$\Delta D = D_i - D_j. \quad (4.1)$$

$\Delta D = 0$  is considered an agreement,  $\Delta D \neq 0$  a disagreement (as in Jamieson et al., 2008). Hence,  $P_{\text{agree}}$  is the ratio of the number of agreements,  $N(\Delta D = 0)$ , to the number of all comparisons,  $N(\Delta D)$ :

$$P_{\text{agree}} = P_{\Delta D=0} = \frac{N(\Delta D = 0)}{N(\Delta D)}, \quad (4.2)$$

where  $N$  is a counting function. In the context of this Synthesis,  $P_{\text{agree}}$  is used in two ways:

- Firstly, to explore the agreement between forecast danger levels ( $D_{\text{RF}}$ ), forecast for the same day for two immediately neighboring warning regions (publication 1).  $P_{\text{agree}}$  may, therefore, be interpreted as an indicator of the spatial correlation or a measure of spatial continuity in avalanche conditions, which

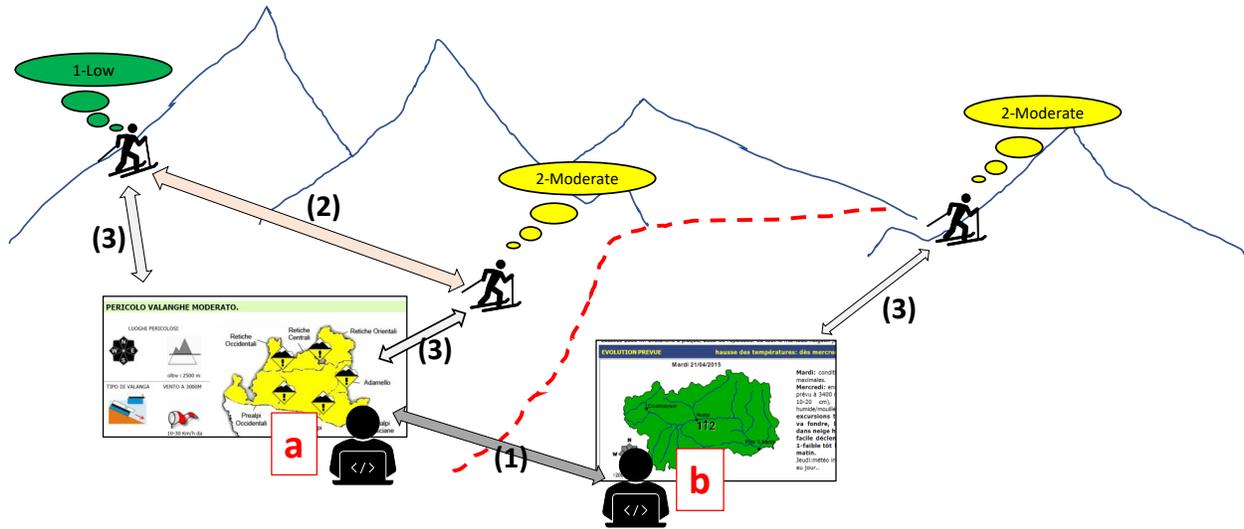


Figure 4.2: Metrics, exploring the spatial consistency and bias in regional avalanche forecasts (1) and the agreement between local danger level estimates provided by two observers within the same warning region (2) are described primarily in Section 4.1, while metrics describing the quality of forecast danger levels (3) can be found in Section 4.2.

also incorporates the imperfect reliability of human judgment when different forecasters issued the forecasts in these warning regions.

- Secondly,  $P_{agree}$  describes the agreement between the danger level estimated locally ( $D_{LN}$ ) of two observers at distance  $z$  from each other (publication 2).

Other metrics, for instance, Cohen’s *Kappa* (Cohen, 1968), are also frequently used to assess interrater-agreement. *Kappa* additionally takes into account the agreement by chance alone.

### 4.1.2 Reliability of individual danger level assessments

The agreement rate between danger level estimates provided by  $K$  individual observers  $P_{agree}(K)$  depends on the reliability of the estimates of each of the  $K$  observers. It may further depend on other factors  $m$ , as for instance the distance  $z$  between the observers or on avalanche conditions ( $D_{RF}$ ).

Here, the reliability of an individual danger level estimate  $rel_{D,LN}$  is the scaling factor required to obtain the agreement rate  $P_{agree}(K)$  between  $D_{LN}$  estimates of  $K$  observers. As it is impossible to derive the reliability for a specific observer, the reliability associated with an individual danger level estimate  $rel_{D,LN}$  is the *geometric* mean/average of individual reliabilities (at conditions  $m$ ):

$$rel_{D,LN} = \sqrt[K]{rel_1 \times rel_2 \times \dots \times rel_K}. \tag{4.3}$$

For the specific case with  $K = 2$  observers,  $P_{agree}(2)$  is the proportion of agreements between  $D_{LN}$  estimates at conditions  $m$ .

It is of note, however, reliability is not a measure of validity, as even several assessments may be wrong at times. Furthermore, the term reliability in the context of (weather) forecast verification generally refers to bias (Murphy, 1993), as described in Sect. 4.2.4, rather than the uncertainty related to the quality of observations.

## 4.2 Categorical forecast verification

Some of the research questions (as in publication 2) aimed at verifying the forecast danger level by using a reference danger level assessment  $D_{\text{reference}}$ . In that sense, the terms  $D_i$  and  $D_j$  in Equation 4.1 refer to the danger level of the forecast  $D_{\text{RF}}$  and the reference assessment, respectively:

$$\Delta D = D_{\text{RF}} - D_{\text{reference}}. \quad (4.4)$$

In the following, metrics describing forecast quality are introduced. The chosen metrics are applied in many fields of science, as for instance, to verify categorical (weather) forecasts (e.g. Murphy, 1993; Gordon and Shaykewich, 2000; Wilks, 2011) or to describe the detection rate of medical tests (e.g. Trevethan, 2017).

### 4.2.1 Accuracy (proportion correct)

Accuracy describes the average correspondence between pairs of forecasts and the events they predicted (Wilks, 2011). With the focus on the danger level  $D$ , this will often be a match between a forecast danger level  $D_{\text{RF}}$  and a reference assessment  $D_{\text{reference}}$  (i.e. a locally assessed  $D_{\text{LN}}$ , Fig. 4.2, case 3). A simple and intuitive measure of accuracy is the *proportion correct* ( $P_{\text{correct}}$ ). Using the notation in Table 4.1,  $P_{\text{correct}}$  is calculated as (Wilks, 2011, p. 308):

$$P_{\text{correct}} = \frac{a + d}{n}, \quad (4.5)$$

where  $n$  is the total sample size.

Equivalently, in contingency tables with three possible outcomes (Table 4.2),  $P_{\text{correct}}$  is the sum of the diagonal where forecast = event, divided by the total sample size (Wilks, 2011, p. 318):

$$P_{\text{correct}} = \frac{r + v + z}{n}. \quad (4.6)$$

$P_{\text{correct}}$  not only reflects the «true » accuracy of the forecast, but can be influenced by errors in the reference assessments  $D_{\text{reference}}$  (e.g. a  $D_{\text{LN}}$  estimate) lowering the observed  $P_{\text{correct}}$  accordingly. Furthermore, if there are variations in forecast accuracy as a function of  $D_{\text{RF}}$  and if the distributions of the forecast danger levels in the forecast data and the verification data are different,  $P_{\text{correct}}$  may be affected. These two points were addressed as follows:

*Table 4.1:* Contingency table cross-tabulating the joint distributions of forecasts and events (observations) with two outcomes. The marginal distributions of forecasts  $y$  and observations  $o$  are shown for the two outcomes 0 and 1.  $n$  is the total number of cases.

		observed		
		1	0	
forecast	1	$a$	$b$	$y_1$
	0	$c$	$d$	$y_0$
		$o_1$	$o_0$	$n$

Table 4.2: Contingency table cross-tabulating the joint distributions of forecasts and events (observations) with three outcomes. The marginal distributions of forecasts  $y$  and observations  $o$  are shown for the three outcomes 1, 2 and 3.  $n$  is the total number of cases.

		observed			
		1	2	3	
forecast	1	$r$	$s$	$t$	$y_1$
	2	$u$	$v$	$w$	$y_2$
	3	$x$	$y$	$z$	$y_3$
		$o_1$	$o_2$	$o_3$	$n$

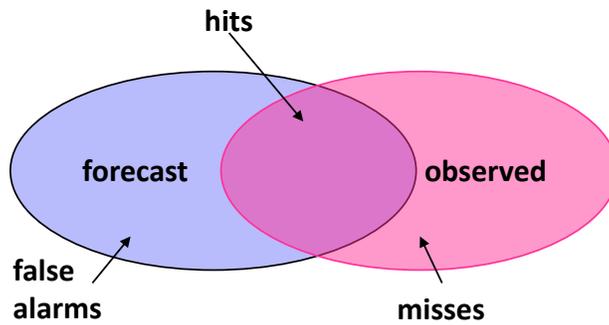


Figure 4.3: In the dichotomous case, a forecast is considered a hit, when the forecast correctly predicted the observed event or non-event; a forecast is considered a false alarm when the event was predicted but not observed, and a forecast is considered a miss, when the event occurred, but was not predicted.

- $P_{\text{correct.raw}}$  describes the observed  $P_{\text{correct}}$  neglecting the impact of differences in the distributions of  $D_{\text{RF}}$  in the forecasts and the verification data set.  $P_{\text{correct.raw}}$  is calculated for the Swiss and Norwegian verification data sets using local nowcast estimates. Both data sets were biased towards less frequently reported local nowcasts when the forecast  $D_{\text{RF}}$  was 1-Low.
- $P_{\text{correct}}$  describes the observed accuracy for data sets, where the distribution of  $D_{\text{RF}}$  in the forecasts and the verification data set is approximately the same. The data sets from Canada and Colorado, where forecasters reassessed the forecast daily and regardless of forecast danger level, are such a case. To reduce the impact of the reporting bias in the Swiss and Norwegian data on  $P_{\text{correct}}$ -values,  $P_{\text{correct}}$  is the weighted mean of  $P_{\text{success}}$  for each of the danger levels (Equ. 4.13), where the weights are assigned according to the proportion of each  $D_{\text{RF}}$  in the forecast data. This weighting approach permits a more realistic comparison of the findings from the four data sets.
- $P_{\text{correct}}^*$  describes the accuracy of the forecast incorporating the reliability of the reference assessment  $D_{\text{reference}}$ . The presence of errors in the reference class (here in  $D_{\text{LN}}$  estimates) reduces the performance of skilled forecasts (e.g. Brenner and Gefeller, 1997; Bowler, 2006), as the strength of the relationship between forecast and observations is not only influenced by their «true» correlation, but also by the reliability of observations (Vul et al., 2009). The reliability of the local assessments

( $rel_{D,LN}$ ) sets an upper limit on the observed accuracy ( $P_{correct}$ ), as accuracy is diminished in proportion to the reliability of the judgments used for verification (Stewart, 2001). Therefore, describing forecast accuracy using solely the proportion correct  $P_{correct}$  may lead to unreasonably low performance values (Bowler, 2006). This means, when individual local danger level estimates are used as a reference standard, an approximation of the accuracy of forecast danger levels ( $P_{correct}^*$ ) in the closed interval from 0 to 1 can be given as:

$$P_{correct}^* \cong \frac{P_{correct}}{rel_{D,LN}}. \quad (4.7)$$

An alternative way to assess forecast quality, which may be less influenced by the (un)reliability of local danger level estimates, is the limitation to cases when two (or more) observers provided the same  $D_{LN}$  estimate within a warning region. Considering such a synonymous opinion a valid approximation of the «true» (yet in reality unknown) danger level in a region, the proportion of forecasts perceived as being correct relying on this data set can be considered an estimate of the accuracy of the forecast regional danger level. Thus,  $P_{correct}$  obtained using this data set should be approximately equal to  $P_{correct}^*$  relying on individual danger level estimates.

It is of note that regardless which approach is used, the question whether  $D_{LN}$  estimates are a valid representation of the «true» avalanche conditions, cannot be answered.

## 4.2.2 Hit rate

The hit rate ( $P_{hits}$ ), also called sensitivity or probability of detection, describes the proportion of the events which are known to have occurred, which were correctly predicted (e.g. Doswell, 2004; Wilks, 2011). For the data explored in this thesis, this translates to: considering all the local danger level estimates, when a certain danger level was reported as the reference, how many of these were correctly forecast. Considering Table 4.1, the hit rate is (Wilks, 2011, p. 310):

$$P_{hits} = \frac{a}{a + c}. \quad (4.8)$$

In Figure 4.3 this corresponds to:

$$P_{hits} = \frac{N(\text{forecast} \cap \text{observed})}{N(\text{observed})}, \quad (4.9)$$

and for multi-category forecasts (Table 4.2):

$$P_{hits}(1) = \frac{r}{r + u + x} = \frac{r}{o_1}; \quad P_{hits}(2) = \frac{v}{s + v + y} = \frac{v}{o_2}; \quad P_{hits}(3) = \frac{z}{t + w + z} = \frac{z}{o_3}. \quad (4.10)$$

## 4.2.3 Success rate

The success rate ( $P_{success}$ ), also referred to as the positive predictive value when describing diagnostic tests in medical literature ( $PPV$ , Trevethan, 2017), describes how many of the events predicted, were correct. In other words, of the cases when a certain danger level was forecast, how often was the danger level perceived as being correct, and are therefore considered a successful forecast. Thus,  $P_{success}$  is:

$$P_{success} = \frac{a}{a + b}. \quad (4.11)$$

In Figure 4.3 this corresponds to:

$$P_{\text{success}} = \frac{N(\text{forecast} \cap \text{observed})}{N(\text{forecast})}. \quad (4.12)$$

For multi-category forecasts, as in Table 4.2, this corresponds to:

$$P_{\text{success}}(1) = \frac{r}{r+s+t} = \frac{r}{y_1}; \quad P_{\text{success}}(2) = \frac{v}{u+v+w} = \frac{v}{y_2}; \quad P_{\text{success}}(3) = \frac{z}{x+y+z} = \frac{z}{y_3}. \quad (4.13)$$

It is of note that  $PPV$  of a diagnostic test not only depends on the accuracy of the test, but also depends strongly on prevalence (e.g. Brenner and Gefeller, 1997; Trevethan, 2017). Applied to the context explored here,  $P_{\text{success}}$  depends on the proportion that a certain danger level  $D$  exists in the data set. In other words, it is much harder to correctly predict rare events compared to frequent events.

#### 4.2.4 Bias ratio

Bias compares the average forecast with the average observation (Wilks, 2011). In dichotomous situations with two outcomes, as in Table 4.1, bias is often expressed as the ratio of the number of forecasts of event 1 to the number of observations of event 1. The bias ratio  $BR$  is then calculated as (Wilks, 2011, p. 307):

$$BR = \frac{a+b}{a+c}. \quad (4.14)$$

Similarly, the bias ratio may be calculated for variables with several outcomes, as shown in Table 4.2. The bias ratio for forecasts predicting events 1 to 3 would be (Wilks, 2011, p. 319):

$$BR_1 = \frac{r+s+t}{r+u+x} = \frac{y_1}{o_1}; \quad BR_2 = \frac{u+v+w}{s+v+y} = \frac{y_2}{o_2}; \quad BR_3 = \frac{x+y+z}{t+w+z} = \frac{y_3}{o_3}. \quad (4.15)$$

A  $BR > 1$  indicates that the respective outcome was more often forecast than observed, vice versa for a  $BR < 1$ . As an example, when comparing forecast  $D_{RF}$  with nowcast  $D_{LN}$ ,  $BR > 1$  indicates over-forecasting and  $BR < 1$  under-forecasting of the respective  $D$ .

#### 4.2.5 Bias

The proportion of forecasts  $D_{RF}$ , which were higher ( $P_{\text{over}}$ ) or lower ( $P_{\text{lower}}$ ) than a local assessment is:

$$P_{\text{over}} = P(\Delta D > 0) = \frac{N(\Delta D > 0)}{N(\Delta D)}; \quad P_{\text{under}} = P(\Delta D < 0) = \frac{N(\Delta D < 0)}{N(\Delta D)}. \quad (4.16)$$

From this, an alternative way of presenting the overall bias ( $P_{\text{over-under}}$ ) can be derived:

$$P_{\text{over-under}} = \frac{N(\Delta D > 0) - N(\Delta D < 0)}{N(\Delta D)}. \quad (4.17)$$

Visualizing  $P_{\text{over-under}}$  together with  $P_{\text{correct}}$  provides an easy visual interpretation of forecast bias. For instance, when  $P_{\text{over-under}}$  is positive and equals  $1 - P_{\text{correct}}$ , all the forecasts would have been too high, while  $P_{\text{over-under}}$  being 0 indicates a balanced distribution of over-forecasts and under-forecasts.

### 4.3 Simulation of snowpack stability distributions by bootstrap sampling

Publication 4 had the objective to use class 1 data (observations directly related to snowpack instability, see also Sect. 2.1.1), like stability tests or avalanche observations, to characterize the three key factors determining avalanche hazard, namely snowpack stability, the frequency distribution of snowpack stability and avalanche size.

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 most often only one stability test result on a given day was available, an alternative approach was followed. 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, stability distributions were generated by random sampling from the entire population of stability tests at a given danger level. Applying bootstrap sampling (Efron, 1979), the process was as follows (see also Fig. 4.4a and b):

- (i)  $n$  stability test results with replacement were randomly selected from the stability tests associated with the same danger level, resulting in a single bootstrap sample. This procedure was repeated  $B$  times for each danger level.
- (ii) For each of the  $B$  bootstrap samples, the proportions of *very poor*, *poor*, *fair* and *good* stability tests were calculated.

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

The second important parameter when bootstrap sampling is the number  $n$  of stability tests drawn in each sample. Small values of  $n$  increase variance, and hence overlap between samples drawn from different danger levels, and reduce the resolution of the desired statistic (e.g. for  $n = 10$ , the resolution is 0.1, for  $n = 100$  it is 0.01). 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 our statistic. Unfortunately, there are no studies which can be referred to concerning the amount of overlap that would be appropriate. Thus a range of values of  $n$  was tested. For their influence on the simulation results refer to publication 4 in Appendix A.5.

Results in this Synthesis are shown for  $n = 25$ .

### 4.4 Snowpack stability and the frequency distribution of snowpack stability - approach to define class intervals

To describe the frequency of triggering locations using a small number of classes, the simulated frequency distributions of snowpack stability (Sect. 4.3), had to be classified (publication 4, see also workflow in Fig.

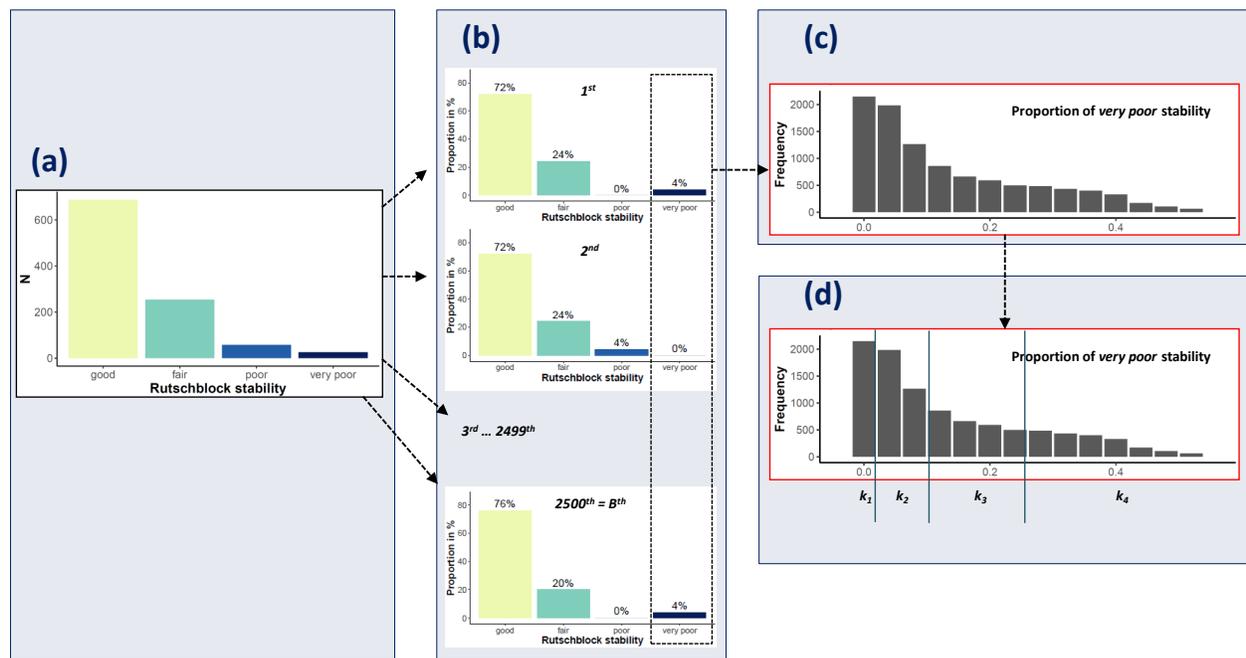


Figure 4.4: Schematic representation of the workflow for bootstrap sampling and frequency class definition. a - For each danger level, all stability ratings are combined. b - From the observed stability distributions (a),  $n$  tests are randomly sampled. This is repeated  $B = 2,500$  times to obtain typical stability distributions for each of the four danger levels. c - The  $4 \times 2,500$  boot-strap samples are merged and the proportion of *very poor* rated stability tests per sample is plotted as a histogram. d - The statistics required for frequency class definitions are calculated and the  $k$  frequency classes defined.

4.4c and d).

The European Avalanche Danger Scale (EAWS, 2018), the North American Avalanche Danger Scale (Statham et al., 2010), but also lookup tables for avalanche danger assessment like the Bavarian Matrix (EAWS, 2005) or the Avalanche Danger Assessment Matrix (Müller et al., 2016), describe the number of avalanches or the frequency of triggering locations using class labels like *some* or *many*. However, none of the before-mentioned provide data-driven quantitative thresholds for these terms when describing the number of avalanches or the frequency of triggering locations, as these depend on the temporal and spatial scale explored.

Therefore a data-driven approach to define class intervals was used, where the choice of class intervals should be appropriate to the observed data distribution.

Regardless which parameter was classified, the following points were considered:

- Classes should be defined with regard to avalanche release. In theory, classes need to capture the entire possible parameter space, i.e. from very rare to virtually all (1 to 99%). However, the parameter space may depend not only on the scaling in time and space, but also on the data-source.
- The number of classes should reflect the human capacity to distinguish between them. 3, 4 and 5 classes are the number of classes 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 *nearly none*, *a few*, *several*, *many* and *nearly all* (e.g. Díaz-Hermida and Bugarín, 2010).

The distribution of the observations relating to the three key elements of avalanche hazard - snow stability, the frequency distribution of *very poor* snow stability and avalanche size - was almost always skewed towards low proportions (e.g. proportion of *very poor* stability) or low numbers (e.g. number of avalanches per day) being much more prominently included in the data set compared to high proportions or high numbers. Therefore, 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  $k$  classes with class interval limits being  $\{0, a, ab, ab^2, \dots, ab^{k-1}, 1\}$ , 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 1 can be given:

$$b = \left( \frac{1 - med}{med} \right)^{\frac{2}{k}}, \quad (4.18)$$

where  $med$  is the median observed value in the data set, and  $k$  the number of classes preferred. This approach requires a suitable value of the number of classes  $k$  to be defined. Given  $k$  and  $b$ , the initial class width  $a$  is (Evans, 1977):

$$a = \frac{(1 - b)med}{1 - b^{\frac{k}{2}}} \quad (4.19)$$

# Chapter 5

## Results

In this chapter, a selection of the main findings from publications 1, 2, 4 and 5 is presented (Fig 5.1). The outline of the Section is as follows:

1. **Consistency** in the use of the danger levels (publication 1) is explored by investigating the spatial consistency of forecast danger levels within and across warning service boundaries (Sect. 5.1.1, Fig. 5.2) and the frequency of the use of danger level 4-High in the Alps (Sect. 5.1.2). In Sect. 5.1.3 the focus is on the impact of the size of the warning regions, the smallest spatial units in the forecasts, and the way a danger level is assigned to a region, on spatial correlation and summary statistics. These are highlighted for forecasts with danger level 4-High.
2. Findings regarding the **reliability of local danger level estimates as a data-source for operational forecast verification** (publication 2, Fig. 5.2) are presented in Section 5.2. Furthermore, the validity of local danger level estimates is exemplarily explored for situations reflecting 4-High using Swiss data (Sect. 5.2.2).
3. A selection of key findings regarding **forecast quality** are presented in Section 5.3 (Fig. 5.2). Forecast quality was explored in several ways: comparing the joint distribution of forecasts and the reference

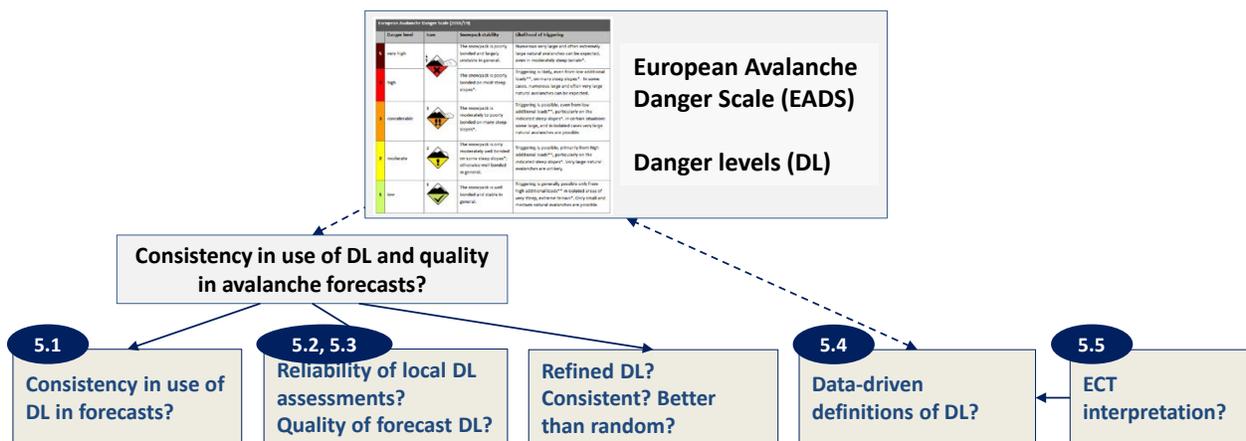


Figure 5.1: Findings from publications 1, 2, 4 and 5 are presented in the indicated Sections.

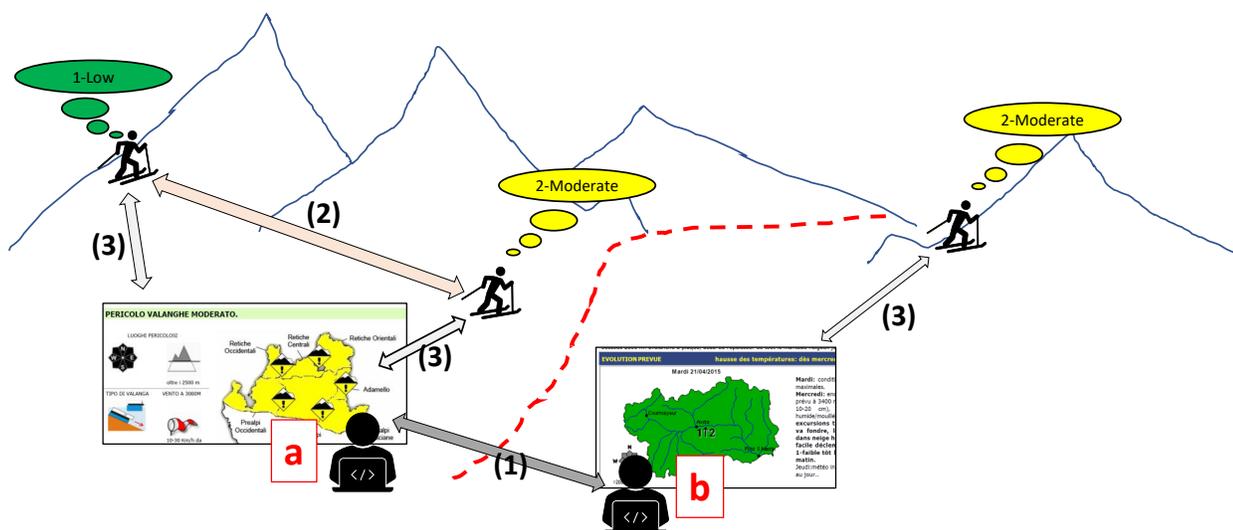


Figure 5.2: (1) Spatial consistency and bias was explored in Section 5.1 by comparing the forecast danger level in immediately neighboring warning regions (warning region boundary = dashed red line), either issued by the same forecast center or by different forecast centers (here a and b). (2) The agreement between local danger level estimates provided by two observers within the same warning region was analyzed in Section 5.2. (3) The forecast danger level was compared with a reference assessment, often a local nowcast estimate after a day in the field (Section 5.3).

assessments (Sect. 5.3.1), depending on forecast avalanche conditions (Sect. 5.3.3) or depending on the assessor (Sect. 5.3.2). And finally, a specific emphasis is put on shedding some light on the accuracy of forecasts of danger level 4-High (Sect. 5.3.4).

4. One argumentation for inconsistencies in the use of the danger levels is the lack of clear, **data-driven characterization of the elements describing avalanche danger** (publication 4). In this respect, observational data relating to the three elements - snowpack stability, the frequency distribution of snowpack stability and avalanche size - are explored to characterize the danger levels (Sect. 5.4). Based on these data, a data-driven lookup table for avalanche danger assessment is introduced (Sect. 5.4.4).
5. Finally, the classification scheme developed to facilitate the **interpretation of Extended Column Test (ECT) results** (publication 5) is briefly presented (Sect. 5.5).

## 5.1 Spatial consistency of forecast danger levels in the Alps

### 5.1.1 Spatial consistency in regional forecast danger levels: agreement and bias

The results presented in this section and in the following Sections 5.1.2 and 5.1.3 rely on data set  $DL_{Alps}$ , which contains spatially continuous forecast danger levels for the Alps during the four winters 2012-2013 to 2015-2016 for a total of 477 forecast days for 291 warning regions, issued by 23 forecast centers. The distribution of forecast danger levels is shown in Fig. 5.3.

The forecast danger level in immediately neighboring warning regions agreed in 83% of the cases (median

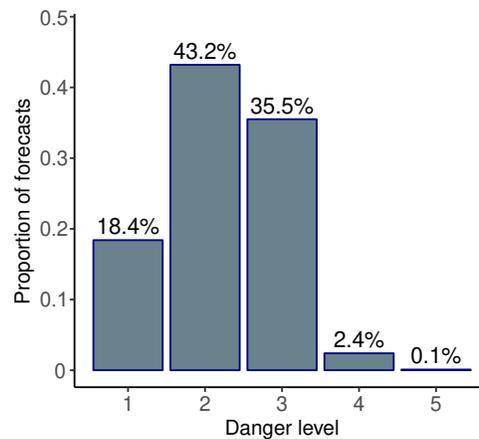


Figure 5.3: Distribution of forecast danger levels  $D_{RF}$  in the Alps (winters 2012-2013 to 2015-2016 for a total of 477 forecast days and 291 warning regions, issued by 23 forecast centers). - **Danger levels 2-Moderate and 3-Considerable were forecast almost 80% of the time.**

$P_{agree}$ ). However,  $P_{agree}$  was significantly higher when comparing warning regions within forecast center boundaries ( $P_{agree} = 0.91$ ) compared to those across forecast center boundaries ( $P_{agree} = 0.63$ ,  $p < 0.001$ , proportion test (R-function *prop.test*, R Core Team, 2017)), or across national borders ( $P_{agree} = 0.62$ ,  $p < 0.001$ ). The two latter values, comparing observed  $P_{agree}$ -values across forecast center boundaries and national borders were not significantly different ( $p > 0.05$ ).

Exploring the agreement rate visually on a map by emphasizing borders with  $P_{agree} \leq 0.8$  essentially captured almost all forecast center boundaries and comparably few boundaries within forecast center domains (Fig. 5.4). This result was confirmed when analyzing only a subset of the warning region pairs, with comparably similar maximum elevation (maximum difference in elevation  $< 250$  m) and similar size of the warning region (size of the larger region  $< 1.5 \times$  the size of the smaller region; Fig. 5.5). For this subset, the median agreement rate was about 0.3 lower across forecast center boundaries ( $P_{agree} = 0.63$ ), than within those ( $P_{agree} = 0.93$ ,  $p < 0.001$ , Fig. 5.5). Similar results were noted for the special case of the three forecast centers in the Italian region of Lombardia, with partially overlapping warning regions.  $P_{agree}$  was 0.63, and thus similar to  $P_{agree}$  across national borders or forecast centers neighboring each other.

Within the boundaries of forecast centers, there was a weak, but significant correlation between  $P_{agree}$  and differences in the elevation of two neighboring regions (Spearman rank-order correlation  $\rho_s = -0.36$ ,  $p < 0.001$ ), with larger differences in elevation corresponding to a lower agreement rate. There was also a weak correlation between  $P_{agree}$  and differences in the size of the warning regions ( $\rho_s = -0.24$ ,  $p < 0.001$ ), where agreement increases as the size difference between warning regions decreases.

In this section  $P_{agree}$  was used as a measure of spatial consistency (or correlation). As shown in Fig. 5.3, in the Alps on four of five days  $D_{RF}$  2-Moderate or 3-Considerable were forecast. Thus, by chance alone, a minimal agreement rate can be expected. This minimal agreement rate was estimated by simulating 10,000 danger levels for two neighboring regions using the danger level distributions shown in Fig. 5.3. Doing so,  $P_{agree}$  values of 0.4 would correspond to a random assignment of  $D_{RF}$ . This compares to  $P_{agree}$  values of about 0.63 across forecast center boundaries in the Alps (Fig. 5.5b).

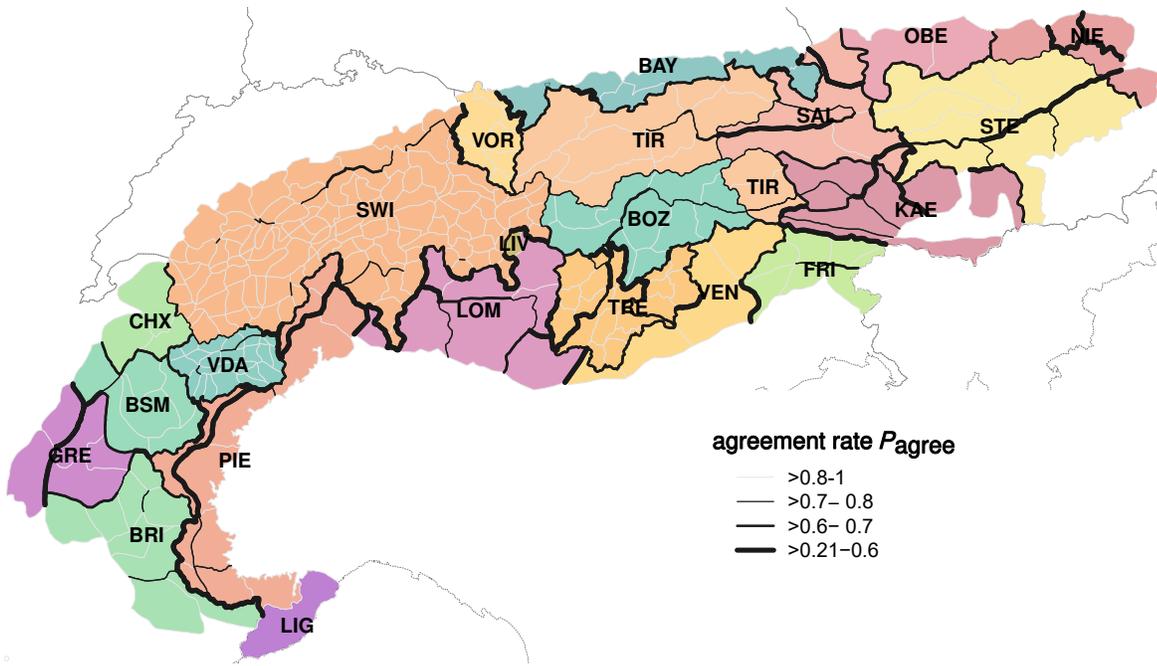


Figure 5.4: Map showing the individual forecast center domains in the European Alps (different colors, for three-letter abbreviations refer to publication 1 (Appendix A.2: Tab. A.2, p. 100)). The borders between warning regions are highlighted depending inversely on the agreement rate  $P_{\text{agree}}$ , with thicker lines corresponding to more frequent disagreements. - **Often, a low(er) agreement rate coincided with the presence of a forecast center boundary between neighboring warning regions.**

Similarly, total agreement ( $P_{\text{agree}} = 1$ ) between neighboring regions implies that subdivisions may be superfluous. Nonetheless, perfect agreement was found for a total of 14 warning region pairs in Switzerland, Italy and Austria. However, to confirm whether this agreement indicates regions which could be merged would require further investigation as to, for example, the nature of typical avalanche problems found, and not only the forecast danger levels.

### 5.1.2 Variations in the use of danger level 4-High in regional forecasts

During the four winters 2012-2013 to 2015-2016, danger level 5-Very High was forecast on less than 0.1% of the 477 forecast days and 291 warning regions, which is equivalent to one day in half the warning regions in the Alps during these four winters. These forecasts were mostly issued during 2013-2014 in the southern part of the Alps. Therefore, forecasts with  $D_{\text{RF}} \geq 4$ -High were combined in the following analysis.

For a specific warning region, the proportion of forecasts with very critical conditions ( $D_{\text{RF}} \geq 4$ -High) is

$$P_{v,\text{crit}} = \frac{N(D_{\text{RF}} \geq 4\text{-High})}{N} \quad (5.1)$$

where  $N$  is the number of forecasts.

Forecasts with  $D_{\text{RF}} \geq 4$ -High were generally rare (median  $P_{v,\text{crit}} = 0.025$ , Fig. 5.6), but were considerably more frequently issued in the warning regions belonging to the four forecast centers in France (Briançon (BRI), Bourg-St-Maurice (BSM), Chamonix (CHX), Grenoble (GRE)) and the Italian forecast centers Piemonte (PIE) and Lombardia (LOM). Visually exploring spatial patterns (Fig. 5.6) showed several forecast

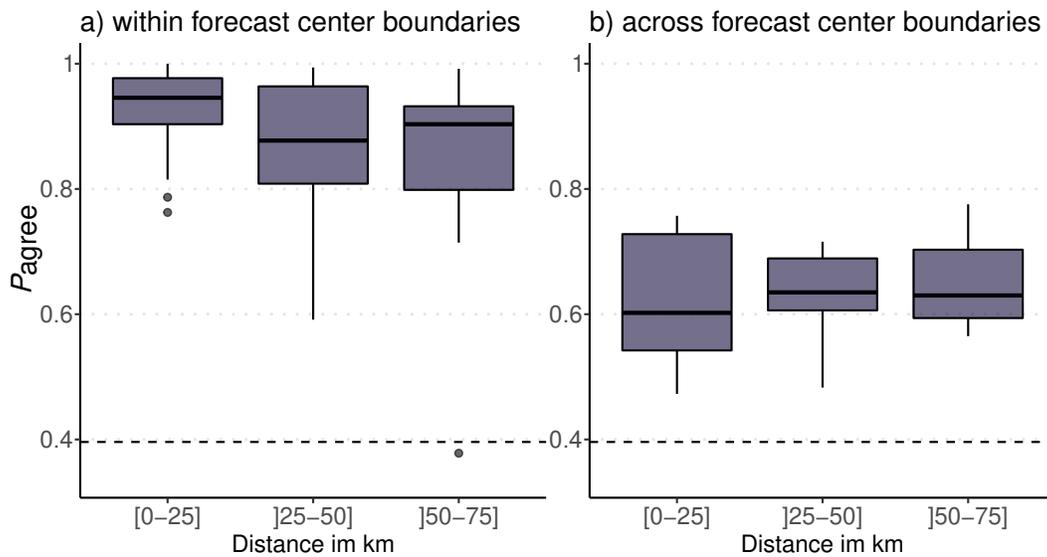


Figure 5.5: Boxplot showing the agreement rate ( $P_{agree}$ ) for neighboring warning region pairs (a) within and (b) across forecast center boundaries, stratified by the distance between the center points of warning regions, with similar maximum elevation ( $\Delta$  elevation < 250 m) and size (the size of the larger warning region is less than 1.5 times the size of the smaller warning region;  $N(\text{within}) = 108$ ,  $N(\text{across}) = 37$ ). The dashed line represents  $P_{agree}$  when randomly drawing 10'000 danger levels for neighboring warning regions using the mean distributions of forecast danger levels in the Alps (distribution as shown in Fig. 5.3). The Boxplots show the median (bold line), the interquartile range (boxes), 1.5 times the interquartile range (whiskers) and outliers outside this range (dots). - **The agreement rate between neighboring warning regions was much lower when warning regions belonged to different forecast centers, and - when comparing across forecast center borders - was not influenced by the size of the warning regions.**

center borders which coincide with large gradients in  $P_{v,crit}$  values. These differences were most obvious when comparing Switzerland (SWI) with its neighbors Chamonix (CHX), Piemonte (PIE), Lombardia (LOM) and Tirol (TIR). In contrast, and with some exceptions, comparably similar values were noted in many of the forecast centers in Austria, Germany, Switzerland and the Italian provinces and regions of Valle d'Aosta (VDA), Bozen-Südtirol/Bolzano-Alto Adige (BOZ) and Trentino (TRE). These variations were also confirmed, when considering only warning regions with a maximum elevation greater than 2500 m: median values for warning regions in Bozen-Südtirol/Bolzano-Alto Adige (BOZ), Switzerland (SWI), Vorarlberg (VOR), Valle d'Aosta (VDA) and Salzburg (SAL) ( $P_{v,crit} < 0.023$ ) were significantly lower than those for Friuli Venezia Giulia (FRI), Bourg-St-Maurice (BSM), Piemonte (PIE), Grenoble (GRE) and Briançon (BRI) ( $P_{v,crit} > 0.076$ ).  $P_{v,crit}$  in Briançon (BRI) was in many cases two or three classes higher compared to its immediate neighbors in Italy (Piemonte (PIE), Liguria (LIG)), but also those in France (Bourg-St-Maurice (BSM), Grenoble (GRE)). The twelve regions with the highest  $P_{v,crit}$  were clustered in the southwest of the Alps, where  $D_{RF} \geq 4$ -High was issued on more than every tenth day ( $P_{v,crit} \geq 0.098$ , max = 0.15). This was most pronounced in the nine regions belonging to the forecast center in Briançon (BRI), which were all among the twelve regions in the Alps with the highest values of  $P_{v,crit}$ .

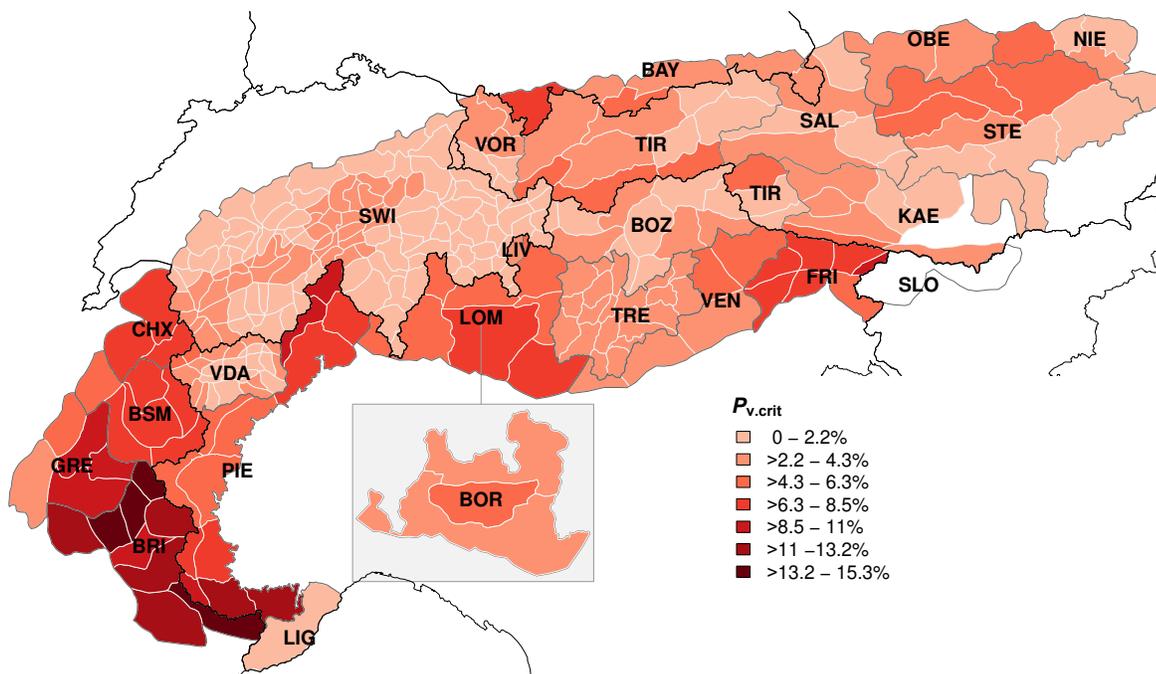


Figure 5.6: Map showing the proportion of days with forecast very critical conditions ( $P_{v.crit}$ ,  $D_{RF} \geq 4$ -High) for each of the warning regions in the European Alps (winters 2012-2013 to 2015-2016). The color shading of the individual warning regions (outlined by white borders) corresponds to the values of  $P_{v.crit}$ . Forecast centers are marked with dark grey polygon borders, national borders with black lines. Three-letter abbreviations describe forecast centers (see publication 1 in Appendix A.2: Tab. A.2, p. 100). To visualize the (at least partially) overlapping forecast regions in the Italian region of Lombardia, LIV is superposed onto parts of LOM, while BOR is placed as inset to the south of LOM. - **4-High was forecast much more frequently in the warning regions belonging to French forecast centers (CHX, BSM, GRE, BRI) or Piemonte (PIE) compared to the immediately neighboring warning regions in Switzerland (SWI) or Valle d’Aosta (VDA), where 4-High was issued with similar frequency as in the rest of the Alps.**

### 5.1.3 Communicating avalanche danger at a regional scale - the potential impact of the size of the warning regions

In the Alps, the size of the warning regions, the smallest spatial units used in the regional avalanche forecasts, varied greatly (Fig. 2.6, p. 21). For instance, forecast centers in Valle d’Aosta (VDA) and Switzerland (SWI) used comparably many, but rather small regions (median size  $< 200 \text{ km}^2$ ). In contrast, in France or in Niederösterreich (Austria), warning regions had a size of about  $800 \text{ km}^2$ . The size of the warning regions in the Alps only rarely exceeded  $2,000 \text{ km}^2$ , as in the two Italian forecast centers in Liguria (LIG) and Lombardia (BOR). This is in contrast to other countries (i.e. Norway, Canada or the U.S.) where warning regions may be as large as several ten thousand  $\text{km}^2$  (e.g. Jamieson et al., 2008; Engeset et al., 2018; Logan and Greene, 2018). A detailed overview regarding the situation in the Alps is provided in publication 2 (Appendix A.2, p. 93ff).

In case warning regions are large and gradients in avalanche danger are expected, a forecaster must decide whether to communicate the highest expected danger level, regardless of its spatial extent, or the danger level representative for the largest part of a region (Fig. 5.7). Other approaches, like a hot-spot approach - focusing on communicating the expected avalanche conditions in the parts of a warning region which are

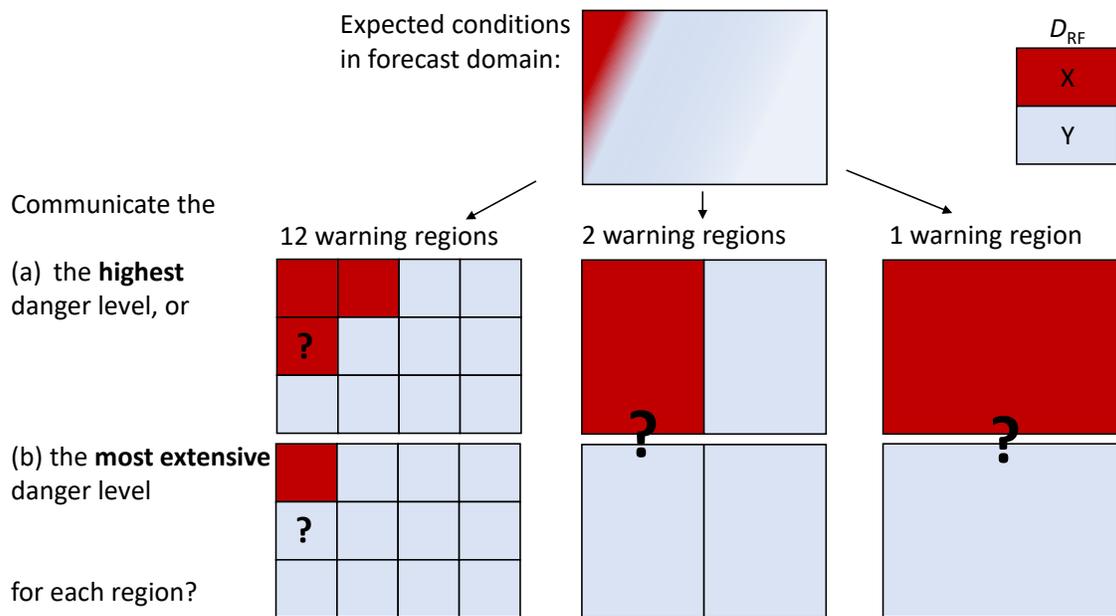


Figure 5.7: The same forecast situation - represented by a higher (dark red, X) and a lower (light blue, Y) danger level ( $D_{RF}$ ) in the upper square - will be communicated differently, depending on the size of the warning regions used by the warning service, and whether the highest or the spatially most extensive avalanche conditions are considered relevant for communication.

most frequented by humans - are also used (e.g. in the very large forecast area of the North Rockies region (Canada), Storm and Helgeson, 2014). Here, it is of note that the EADS lacks a definition in that respect. In the following, the potential impact of the size of the warning regions on the communication of spatial variations in avalanche danger, and thus also its impact on summary statistics like the frequency of forecasts with  $D_{RF} \geq 4$ -High ( $P_{v.crit}$ ) is highlighted using two examples.

Warning services like those in Valle d'Aosta (VDA, Italy) or in Switzerland (SWI) use a comparably fine spatial resolution of the warning regions in the bulletin production process. In the forecasts issued by these warning services, a danger level is not explicitly communicated for each warning region, but for several warning regions aggregated flexibly to a danger region:

- In 2016, the forecast domain in Valle d'Aosta (VDA), with a total area of 3,300 km<sup>2</sup>, was subdivided into 26 warning regions with a median size of 130 km<sup>2</sup>. At a higher-order spatial hierarchy, each of these 26 regions belonged to one of four snow-climate regions (median size 815 km<sup>2</sup>, Burelli et al. (2016, p. 27)).
- In Switzerland (in 2017), the Alpine forecast region (26,400 km<sup>2</sup>) consisted of 117 warning regions with a median size of 180 km<sup>2</sup> (Fig. 5.8). When producing the forecast, a forecaster aggregated a number of warning regions to (generally) three to five regions with the same danger description (with an average size per aggregated region of 5000 - 7000 km<sup>2</sup>; Ruesch et al., 2013; Techel and Schweizer, 2017). Similar to VDA, each of the Swiss warning regions could be linked to three levels of a spatial hierarchy (SLF, 2017, p. 41; Fig. 5.8).

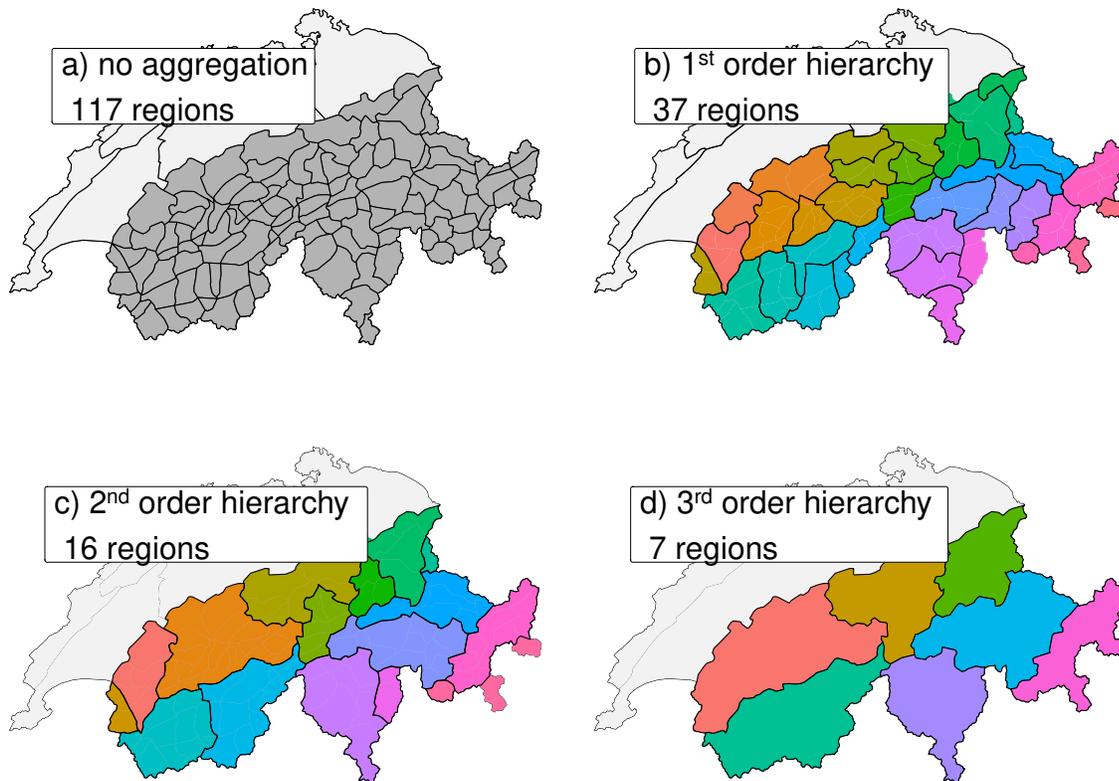


Figure 5.8: Maps showing Switzerland with the 117 warning regions (a), the smallest spatial units used in the forecast in Switzerland (in 2018), and the three spatial hierarchy levels (b-d, according to SLF (2017)). These aggregations were used in the example in Sect. 5.1.3 for Switzerland.

In either case, these predefined regional aggregations were not of great importance anymore in the communication of a regional danger level in the map-based products, due to the flexibility in which the forecaster could assign danger ratings to individual regions (VDA) or aggregates of warning regions (SWI). However, here we use these spatial hierarchy-levels - three for VDA and four for SWI - to explore the variability of the forecast danger level within regions of increasing size, e.g. when the size of the warning regions increases, as in Fig. 5.7a when going from left to right, and the potential implication on summary statistics like the proportion of the most critical forecasts ( $P_{v,crit}$ ).

As shown in Fig. 5.9a, the larger a region, the higher was the variability within these regions (more than one danger level forecast). In other words, a forecaster would not have been able to communicate the spatial variability in danger levels in about 15% of the forecasts, without describing these in text form, if warning regions were five times larger (about 800 km<sup>2</sup>, corresponding to the median size in Niederösterreich or in France), as compared to the currently implemented spatial resolution. Assuming even larger warning regions at the communication level, 3300 km<sup>2</sup>, for instance when considering VDA as one single region, or the seven snow-climate regions in SWI (Fig. 5.8d), and communicating a single danger rating only, would have resulted in about half of the forecasts not reflecting the spatial variability within the respective region.

This shows that variations in the forecast avalanche danger at spatial scales lower than the size of the spatial units used in the production and communication of the forecast are to be expected, particularly if regions are large. In case data would be available to allow the detection of such spatial variations - as in Figure 5.7,

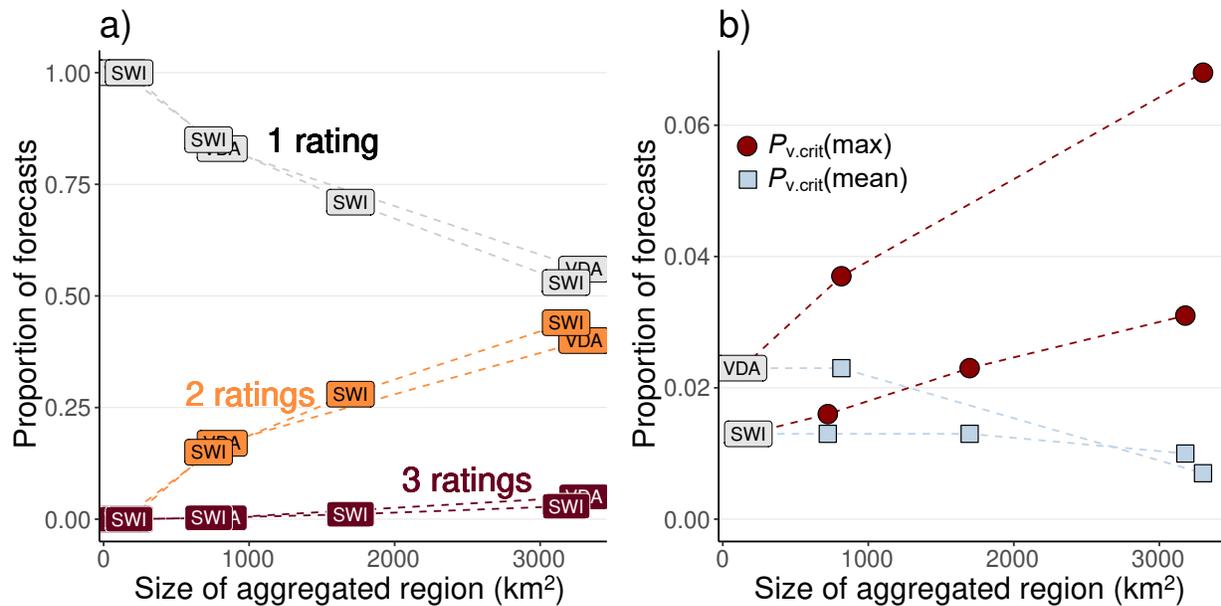


Figure 5.9: (a) Proportion of forecasts with one, two, or three danger ratings per aggregated region. In Valle d'Aosta (VDA) two and Switzerland (SWI) three fixed levels of spatial aggregation were used. The left-most values correspond to the forecasts as published (a single danger rating per warning region), while the right-most value represents the variation in danger levels within the forecast domain of VDA, or within a snow-climate region (SWI). Compare also to Fig. 5.8, which shows the 117 Swiss warning regions (no aggregation) in Fig. 5.8a, and the seven snow-climate regions in Fig. 5.8d. (b) Proportion of forecasts with danger level 4-High or 5-Very High ( $P_{v.crit}$ ). Increasing the size of a region (i.e. going from panel a to panel d in Fig. 5.8),  $P_{v.crit}$  changes depending on the approach used to assign a danger level to a region (see Fig. 5.7).  $P_{v.crit}(\max)$  assumes assigning the highest danger rating per region (Fig. 5.7, case a), and  $P_{v.crit}(\text{mean})$  the spatially most relevant danger rating (Fig. 5.7, case b). Again, the left-most value (text label) shows the respective values for the implemented spatial resolution. - **(a) With an increase in the size of a region, the proportion of days when the spatial variation in avalanche danger cannot be expressed with a single danger level increases. (b) Both the resolution of the spatial units used in the forecasts and the way an avalanche danger level is assigned to a region impact summary statistics like  $P_{v.crit}$ . This explains part of the differences noted in Fig. 5.6.**

a forecaster would need to make a decision whether to focus on communicating the spatially most extensive (or mode), or the highest danger level.

This question can be further explored by taking the proportion of forecasts with very critical conditions  $P_{v.crit}$ . It shows that if the highest danger level expected within a region  $P_{v.crit}(\max)$  was used as a rule for communication, this would increase the absolute values of  $P_{v.crit}$  with increasing size of the warning region (Fig. 5.9b). In contrast, communicating the spatially most widespread danger rating instead ( $P_{v.crit}(\text{mean})$ ), had relatively little influence for smaller regions, but reduced  $P_{v.crit}$  values significantly for the largest-size regions (Fig. 5.9b).

At the current spatial resolution,  $P_{v.crit}$  values for SWI and VDA were comparable, particularly along their joint border (Fig. 5.6). However, already when using the first-order aggregation level as the lowest spatial resolution (an increase from  $< 200 \text{ km}^2$  to about  $800 \text{ km}^2$ ),  $P_{v.crit}(\max)$  values were considerably higher for VDA, and rather similar to those in neighboring warning regions in Chamonix (CHX), Bourg-St-Maurice

(BSM) or Piemonte (PIE). However, even when considering the highest-order hierarchy level, with a median size of the regions of about 3,300 km<sup>2</sup>, and  $P_{v.crit(max)}$  values, these were still only about half to a quarter as high as those observed in some parts of the French Alps (compare Fig.s 5.6 and 5.9b).

## 5.2 Quality of local danger level estimates

In this section the quality of local danger level estimates ( $D_{LN}$ ) as a data-source for forecast verification is assessed (see also Fig. 5.2). The research question *Do variations in local danger level estimates exist?* is addressed in Section 5.2.1. Building on these findings, the reliability of local danger level estimates as a data-source for forecast verification is estimated. And finally, the validity of local danger level estimates is assessed for situations representing danger level 4-High (Section 5.2.2).

### 5.2.1 Variations in local danger level estimates - agreement rate and reliability

Based on the large data set of local danger level estimates in Switzerland ( $DL_{SWI}$ , Sect. 3.2), the agreement rate  $P_{agree}$  between the  $D_{LN}$  estimates of two observers was calculated.

When observers reported  $D_{LN}$  from the same warning region, the smallest spatial unit used in the forecasts,  $P_{agree}$  was 0.78. In these cases, observers were on average only 5 km away from each other (90%-quantile: 11 km). Exploring the agreement rate as a function of distance for cases, when  $D_{LN}$  estimates were reported from warning regions with the same forecast danger level, showed that  $P_{agree}$  decreased from distances less than 5 km ( $P_{agree} = 0.8$ ) up to a distance of about 15 km (equals interval 10-20 km in Fig. 5.10a;  $p < 0.03$ , proportion test (R-function *prop.test*, R Core Team, 2017)). At shorter distances of less than 2.5 km, no further increase in  $P_{agree}$  was noted ( $P_{agree} = 0.8$ , value not shown in Fig. 5.10a). At distances larger than 20 km,  $P_{agree}$  varied between 0.68 and 0.71. This decrease noted in  $P_{agree}$  from about 5 km to 20 km by about 0.1 suggests that spatial variability of avalanche danger already exists at such short spatial scales contributing to the variation in danger level estimates.

At all forecast danger levels,  $P_{agree}$  was higher at distances less than 15 km compared to larger distances (Fig. 5.10a). For observers reporting from the same warning region,  $P_{agree}$  was significantly higher at 1-Low ( $P_{agree} = 0.85$ ,  $p < 0.001$ , Fig. 5.10b) compared to 2-Moderate ( $P_{agree} = 0.8$ ), 3-Considerable ( $P_{agree} = 0.77$ ) or 4-High ( $P_{agree} = 0.68$ ). Whether this points towards increased spatial variability with increasing avalanche danger, or whether this expresses simply greater variability in the estimates due to other causes, is not clear.

Performing the same analysis for the smaller Norwegian data set ( $DL_{NOR}$ , Sect. 3.2) showed that two observers reporting from the same warning region agreed in their estimate in about three quarter of the cases ( $P_{agree} = 0.71$ , Fig. 5.10b). As the exact location of the Norwegian observers was not available for this analysis, it can only be estimated that this may correspond on average to a distance of several dozen kilometers, as the warning regions in Norway generally have a size of several thousand km<sup>2</sup> (Engeset et al., 2018). Hence, this value is comparable to the values observed in Switzerland at distances larger than 20 km ( $P_{agree} \approx 0.7$ ). Furthermore,  $P_{agree}$  decreased with increasing forecast danger level in Norway (1-Low to 4-

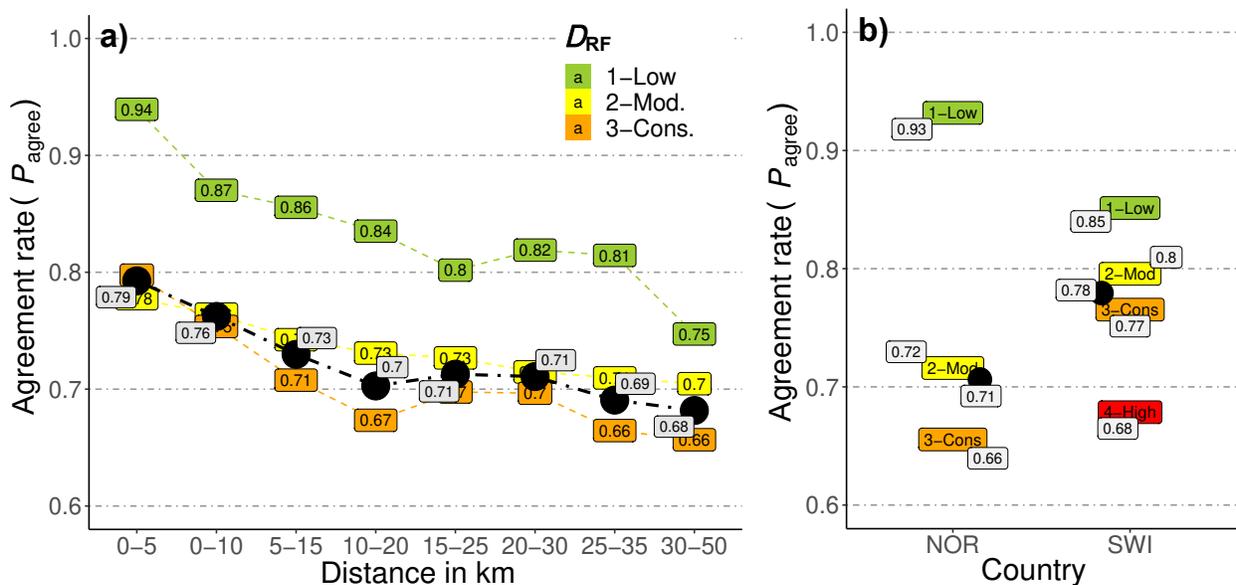


Figure 5.10: Agreement rate  $P_{agree}$  between observer pairs (a) as a function of distance and forecast danger level  $D_{RF}$  (Swiss data) and (b) within the same warning region (Norway (NOR) and Switzerland (SWI)). The black points (a, b) and line (a) show  $P_{agree}$  regardless of forecast danger level, the coloured text labels the respective value of  $P_{agree}$  for the same forecast danger level. In (b) the agreement rate at 4-High in Norway is not shown ( $P_{agree} = 0.37$ ). - **The agreement rate between two local danger level estimates  $D_{LN}$  decreased with distance (up to about 15 km) and forecast danger level  $D_{RF}$ .**

High:  $P_{agree} = 0.93, 0.72, 0.66, 0.37$ , respectively, Fig. 5.10b), which is again a similar pattern as observed in the Swiss data.

The reliability associated with individual  $D_{LN}$  estimates can be interpreted as the factor describing the repeatability of an individual  $D_{LN}$  estimate by a second observer at the conditions  $m$ , where  $m$  can be the distance  $z$  between two observers and avalanche conditions (forecast danger level  $D_{RF}$ ). Based on equation 4.7 (Sect. 4.1.2), the average reliability for individual  $D_{LN}$  estimates ( $rel_{D,LN}$ ) at the scale of the size of a warning region was 0.88 in Switzerland and 0.85 in Norway, where the lower  $rel_{D,LN}$ -value for Norwegian observers is likely due to the larger distances between observers, and hence the increased spatial variation in avalanche conditions. Despite different typical distances  $z$  between observers in these two countries, the same pattern of decreasing  $rel_{D,LN}$  with increasing  $D_{RF}$  was observed: in Switzerland,  $rel_{D,LN}$  decreased from 1-Low ( $rel_{D,LN} = 0.92$ ) to 4-High ( $rel_{D,LN} = 0.82$ ), in Norway from 1-Low ( $rel_{D,LN} = 0.96$ ) to 4-High ( $rel_{D,LN} = 0.61$ ).

### 5.2.2 Validity of local danger level estimates - situations representing 4-High

So far, the reliability of local danger level estimates was explored and no assumption was made regarding the validity of the  $D_{LN}$  estimates. However, this is a highly relevant aspect.

One approach to validate local danger level estimates may be a comparison with recordings of avalanche occurrences. Using such avalanche occurrence data, the days with the highest avalanche activity may be considered to represent danger level 4-High (or higher). These days are often characterized by snowfall and

poor visibility, and conditions for backcountry travel will be dangerous; thus only limited access to avalanche terrain is possible. Therefore it is likely that, at the time of the local assessment, often no or only partial information was available regarding avalanche activity.

Relying on the data set of mapped avalanches in the region of Davos/Switzerland ( $AV_{\text{Davos}}$ , Sect. 3.2.3), local nowcast estimates made in the same region were compared with avalanche activity for the days with the highest avalanche activity defined using the Avalanche Activity Index (AAI, Schweizer et al., 1998). The AAI sums up all avalanches by assigning weights to their size (size 1 to size 4, weights 0.01, 0.1, 1, 10, respectively).

Of the 1% of the rank-ordered days with the highest avalanche activity ( $N = 22$ ),  $D_{\text{LN}}$  estimates were available only from three days, clearly showing the greatly reduced backcountry travel in these hazardous conditions. For the three days with  $D_{\text{LN}}$  estimates, only one of them rated the danger level with 4-High, the other two with 3-Considerable.

Exploring the 2.5% of the rank-ordered days with the highest avalanche activity ( $N = 55$ ),  $D_{\text{LN}}$  estimates were reported on 20% of the days ( $N = 11$ ), which is similar to the average daily reporting rate for the region of Davos (21% of the days with  $D_{\text{LN}}$  estimates). Only one of these 11 days was locally assessed with 4-High. (In)validating local danger level estimates using avalanche observations is not straightforward, even in hindsight. Firstly, there may be errors in the avalanche recordings themselves, like a wrong dating of avalanche occurrence. Furthermore, despite an attempt to record all avalanches in the area, these recordings tend to be incomplete. And lastly, this hindsight assessment also requires a human expert to set thresholds defining 4-High. However, assuming that the avalanche observations are reasonably accurate, a 2.5% threshold corresponded to an  $AAI \geq 16$ , which is equal to at least one size 4 and six size 3 avalanches, or at least 16 size 3 avalanches, and may thus qualify for the description of danger level 4-High: «In some cases, numerous large (size 3) and often very large (size 4) natural avalanches can be expected.» (see EADS in Tab. 2.2, EAWS, 2018). Using 2.5% of the days is also close to the average frequency that danger level 4-High was forecast in the European Alps (Fig. 5.3). Thus, assuming that 2.5% is a reasonable approximation describing situations with 4-High, that avalanche observations were assigned to the correct day and that conditions did not change considerably between avalanche occurrence and the time the local estimate was made, 10 of the 11 days when a local estimate was reported with 4-High were missed by field observers.

### 5.3 Quality of forecast danger levels

In this section, the perspective changes from comparing two forecast danger levels (Sect. 5.1) or two local nowcast assessments (Sect. 5.2), to one where a forecast danger level is compared to a reference assessment with the objective to estimate the quality of the forecast (see also Fig. 5.2). This objective was achieved by exploring data from Switzerland in publications 2 and 3. Here, data sets originating from three other countries are used to complement the Swiss study (publication 2). All the data sets permitted the comparison of a forecast danger level with either a nowcast or hindcast assessment by observers or forecasters (Fig. 5.11). These assessments are considered the reference standard and are referred to as  $D_{\text{reference}}$ . Additionally, for the region of Davos (Switzerland) the forecast was verified using a data set of

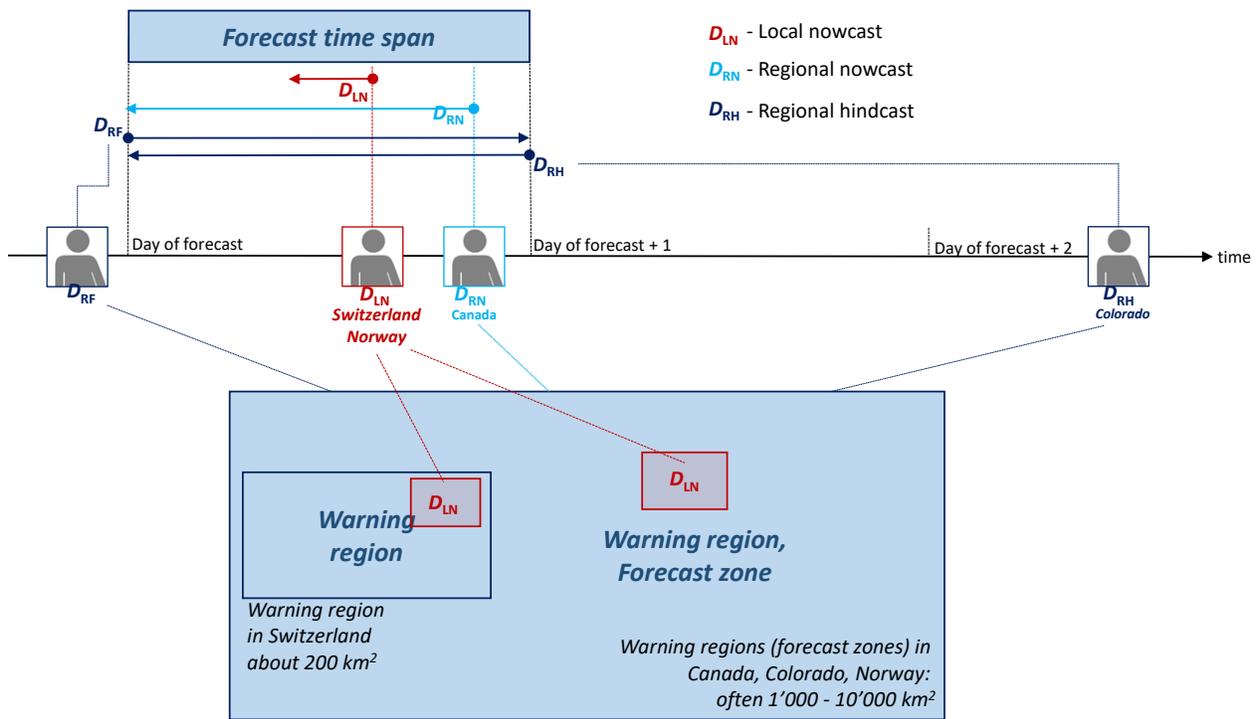


Figure 5.11: Forecast quality was assessed by comparing the published forecast danger level ( $D_{RF}$ ) using nowcast and hindcast assessments from Switzerland, Canada, Colorado and Norway. Depending on the type of assessment, these were made at different times in relation to the forecast time span, and they may refer to parts or the entire warning region. Furthermore, the typical size of the warning regions varied considerably between Switzerland and the other three countries.

avalanches to validate conditions representing 4-High.

### 5.3.1 Accuracy (proportion correct) of forecast danger levels

What is the observed accuracy of a set of forecast danger levels  $D_{RF}$ ? This question was explored using the *proportion correct* ( $P_{correct}$ ) as a statistical measure, and is defined as the proportion of forecasts which matched the reference standard (Sect. 4.2.1).

Relying on individual local nowcast estimates of avalanche danger ( $D_{LN}$ ) as  $D_{reference}$ ,  $P_{correct.raw}$  was 0.77 for Switzerland and 0.72 for Norway (Tab. 5.1). In contrast,  $P_{correct.raw}$  was considerably higher (Switzerland  $P_{correct.raw} \geq 0.85$ , Norway  $P_{correct.raw} = 0.82$ ), when  $P_{correct.raw}$  was explored using only those days, when exactly two  $D_{LN}$  estimates were reported from either the same warning region, or (in Switzerland only) from two immediately neighboring warning regions with the same forecast danger level, and when these indicated the same danger level.

The Swiss and Norwegian verification data sets were biased towards less frequently reported local nowcasts, when conditions were favorable (forecast  $D_{RF}$  1-Low). These days, on which the success rate was comparably high (see Sect. 5.3.3), are under-represented and therefore the overall  $P_{correct}$  is lowered proportionally ( $P_{correct.raw} < P_{correct}$ ). Accounting for this under-representation of forecast days with 1-Low as described in Sect. 4.2.1,  $P_{correct}$ -values were somewhat higher for individual danger level estimates com-

Table 5.1: Proportion correct  $P_{\text{correct}}$  for the four data sets explored.  $P_{\text{correct.raw}}$  refers to the observed values, without correcting for under-reporting at lower danger levels.  $P_{\text{correct}}$  describes the proportion correct accounting for this reporting bias.  $P_{\text{correct}}^*$  is an estimation of forecast accuracy incorporating the reliability of individual  $D_{\text{LN}}$  estimates. For SWI and NOR, individual  $D_{\text{LN}}$  estimates were compared with  $D_{\text{RF}}$ , but also with subsets of days and warning regions (wr) when two observers agreed in their local estimate.

country	assessment by	$N$	$P_{\text{correct.raw}}$	$P_{\text{correct}}$	$P_{\text{correct}}^*$
Switzerland (SWI)	observers $D_{\text{LN}}$ - individual	11,760	0.77	0.81	0.92
	$D_{\text{LN}}$ - two, same wr	842	0.86	0.9	–
	$D_{\text{LN}}$ - two, same / nb wr	1,158	0.85	0.89	–
Norway (NOR)	observers $D_{\text{LN}}$ - individual	4,511	0.72	0.74	0.84
	$D_{\text{LN}}$ - two, same wr	310	0.82	0.83	–
Canada (CAN)	forecasters	2,774	–	0.84	–
Colorado (COL)	forecasters	2,018	–	0.84	–

(abbreviations: nb - neighboring, wr - warning region)

pared to  $P_{\text{correct.raw}}$  (Switzerland  $P_{\text{correct}} = 0.81$ , Norway  $P_{\text{correct}} = 0.74$ ), but also when relying on two danger level assessments (Switzerland  $P_{\text{correct}} \approx 0.9$ , Norway  $P_{\text{correct}} = 0.83$ ). Removing this reporting bias permits a more appropriate comparison with the verification data sets in Canada and Colorado, where assessments were made on a daily basis and regardless of forecast conditions (Fig. 5.11). In Canada, the assessments were made by the forecasters themselves, either while preparing the forecast for the following day, while in Colorado forecasters in the central office re-assessed the forecast a few days later.  $P_{\text{correct}}^*$  values ( $P_{\text{correct}}^* = 0.84$ ) were similar as when considering two local nowcasts in Norway (Tab. 5.1).

$P_{\text{correct}}$  does not incorporate variations (or the unreliability or inconsistency) in the assessment of the reference standard and underestimates the accuracy as the *observed* accuracy is bounded by the reliability of the reference assessment (e.g. Stewart, 2001; Bowler, 2006). An estimation of the accuracy of a forecast danger level  $P_{\text{correct}}^*$ , considering that the reliability of the reference assessment sets an upper bound, is (Sect. 4.2.1):

$$P_{\text{correct}}^* \cong \frac{P_{\text{correct}}}{\text{rel}_{D,\text{LN}}},$$

where  $P_{\text{correct}}$  describes the observed accuracy when relying on individual danger level estimates with reliability  $\text{rel}_{D,\text{LN}}$ .

Incorporating the reliability of individual local assessments derived in Sect. 5.2, the reliability of the forecast danger level was 0.92 in Switzerland and 0.84 in Norway, when using individual  $D_{\text{LN}}$  estimates as a reference. These values are reasonably similar to the  $P_{\text{correct}}$ -values shown in Table 5.1 for Switzerland ( $P_{\text{correct}} \approx 0.9$ ) and Norway ( $P_{\text{correct}} = 0.83$ ), when  $D_{\text{RF}}$  were compared using days and regions when two  $D_{\text{LN}}$  estimates agreed. Considering these values as a more realistic approximation of forecast accuracy, the forecast regional danger level was correct on about six out of seven days in the comparably large regions in Canada, Colorado and Norway, and on about nine out of ten days in the small regions in Switzerland.

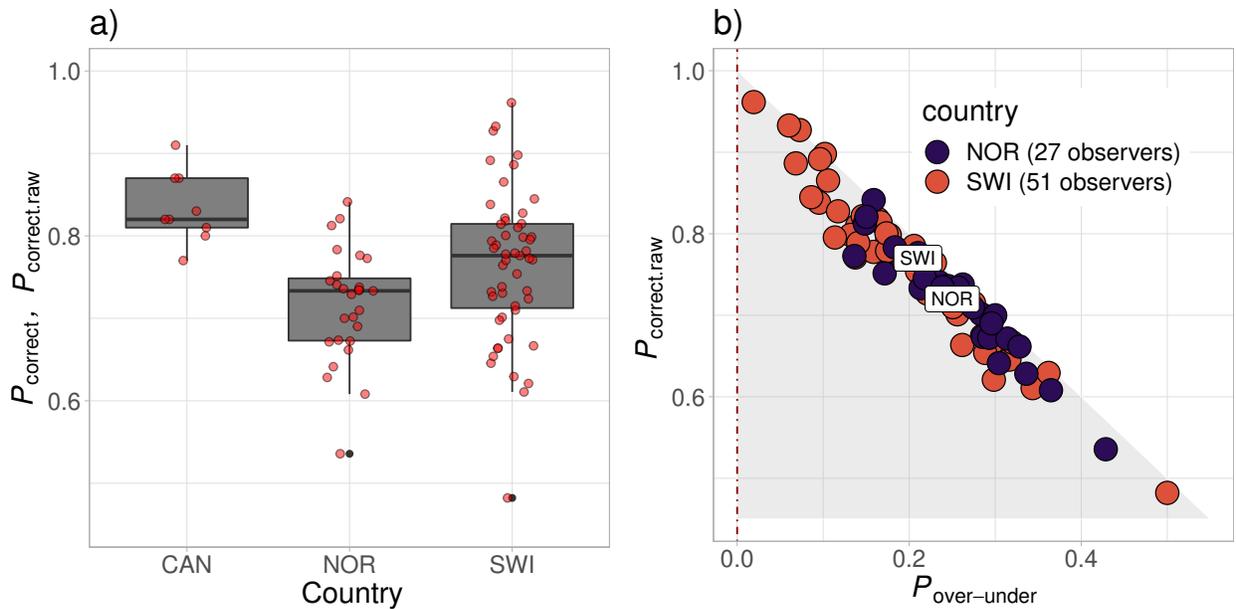


Figure 5.12: a) Proportion correct between sets of forecast  $D_{\text{RF}}$  and nowcast danger levels for each of nine forecasters (in Canada/CAN:  $P_{\text{correct}}$ ) or the local  $D_{\text{LN}}$  estimate for observers with more than 50  $D_{\text{LN}}$  estimates (in Norway/NOR and Switzerland/SWI:  $P_{\text{correct.raw}}$ ). b) For the subset of observers in Norway and Switzerland,  $P_{\text{correct.raw}}$  and the proportion of over-forecast minus the proportion of under-forecasts ( $P_{\text{over-under}}$ ) are shown. Ideally,  $P_{\text{correct}}$  is close to 1 and  $P_{\text{over-under}}$  close to 0. - **Depending on the forecaster or observer, considerable variation in the observed accuracy of forecasts can be noted. If forecasts were perceived as incorrect, these were essentially always considered too high rather than too low.**

### 5.3.2 Accuracy (proportion correct) of forecast danger levels - variations due to individual assessors

The Swiss, Norwegian and Canadian data sets allowed the exploration of  $P_{\text{correct}}$ , as a function of individual forecasters or observers.

As can be seen in Fig. 5.12a, considerable variations in  $P_{\text{correct}}$ -values by different forecasters and observers existed. In Switzerland and Norway, the inter-quartile range of  $P_{\text{correct}}$  for observers with  $> 50$  assessments ranged between 0.71 and 0.82 (Switzerland), and between 0.68 and 0.75 (Norway). Differences in  $P_{\text{correct}}$ -values by individual observers were significant ( $p < 0.05$ , proportion test (R-function *prop.test*, R Core Team, 2017)), when comparing  $P_{\text{correct}}$ -values from either side of the inter-quartile range in Switzerland (Fig. 5.12a), or the respective 15% of the highest and lowest  $P_{\text{correct}}$ -values in Norway. Reasons for these variations can be numerous and may include variations due to a reporting bias, for instance when observers reported more frequently a local assessment when disagreeing rather than agreeing with the forecast, variations linked to some observers assessing the danger level consistently different than others, or due to estimates provided by some observers being simply less reliable in general.

In Canada, where a nowcast assessment by a forecaster was available for all forecast days and regions,  $P_{\text{correct}}$ -values ranged between 0.77 and 0.91 ( $N > 342$  assessments per forecaster). Similar to the Swiss and Norwegian results, variations in  $P_{\text{correct}}$ -values were noted between the Canadian forecasters. These

differences were significant ( $p < 0.05$ ), when comparing the three forecasters with the lowest  $P_{\text{correct}}$ -values with the three forecasters with the highest values. Again, there may be a number of reasons for these variations, including variations in the reliability of the nowcast assessment in general, a different perception of the danger levels, or some forecasters being more prone to a confirmation bias than others (e.g. see McClung, 2002a, on typical biases in avalanche forecasting).

### 5.3.3 Success rate, hit rate and bias of forecast danger levels

So far, the quality of the combined set of forecasts was considered, regardless of the forecast or nowcast avalanche conditions. However, all of these data sets are unbalanced towards two danger levels (2-Moderate and 3-Considerable in Switzerland and Norway, 1-Low and 2-Moderate in Canada and Colorado). Combined, these two danger levels were forecast between 65% and more than 80% of the time. Hence,  $P_{\text{correct}}$  will reflect mainly the forecast performance at these danger levels.

Thus, in the following, forecast performance is explored conditional on the forecast ( $D_{\text{RF}}$ ) or the reference danger level ( $D_{\text{reference}}$ , now- or hindcast) using two statistical measures: the success rate ( $P_{\text{success}}$ ) and the hit rate ( $P_{\text{hits}}$ ).  $P_{\text{success}}$  describes what proportion of the events forecast (a specific  $D_{\text{RF}}$ ), were in fact correct according to  $D_{\text{reference}}$  (see also Sect. 4.2.3). In contrast,  $P_{\text{hits}}$  describes the proportion of a specific  $D_{\text{reference}}$ , which was in fact correctly forecast (Sect. 4.2.2). Ideally,  $P_{\text{success}}$  and  $P_{\text{hits}}$  are close to 1.

The proportion of forecasts, which were confirmed by the reference assessment ( $P_{\text{success}}$ ), was highest at 1-Low, and decreased strongly with increasing danger level in all four data sets (Fig. 5.13a). At 1-Low,  $P_{\text{success}}$  was between 0.85 and 0.93 indicating that if 1-Low was forecast, it was generally also confirmed. At 3-Considerable,  $P_{\text{success}}$  values were still above 0.5. However, at 4-High  $P_{\text{success}}$  was generally lower than 0.5 indicating that when 4-High was forecast, it was often not confirmed by the reference assessment  $D_{\text{reference}}$ . These low  $P_{\text{success}}$  values for Switzerland and Norway at 4-High are also related to the quality of the  $D_{\text{LN}}$  estimates, which were not only less reliably ( $rel_{D_{\text{LN}}} < 0.82$ ) when  $D_{\text{RF}}$  4-High was forecast (Sect. 5.2.1), but were likely often not a correct estimation of avalanche danger (as shown in Section 5.2.2).

In contrast, the proportion of forecasts, which correctly predicted  $D_{\text{reference}}$  ( $P_{\text{hits}}$ ) showed no such pattern of decreasing  $P_{\text{hits}}$  with increasing danger level (Fig. 5.13b). With some exceptions,  $P_{\text{hits}}$  ranged between 0.65 and 0.9 for the four explored danger levels. One such exception was  $P_{\text{hits}} = 0.11$  for 4-High in Colorado, which will be discussed in Sect. 5.3.4, together with the verification of the forecasts in the region of Davos (Switzerland) relying on avalanche observations. The other exception was 1-Low in Switzerland and Norway, where the forecast correctly predicted the reference assessment in only about half of the cases. This compares to Canada and Colorado with a hit rate  $> 0.85$  at 1-Low. A potential explanation for this difference might be that local estimates were provided less often when danger level 1-Low was forecast in Switzerland and Norway, compared to Canada and Colorado, where nowcast and hindcast assessments were made on a daily basis regardless of danger level. For instance, in Switzerland 1-Low was forecast on 20% of the days, while forecasts with 1-Low were present in the joint distributions of forecasts and nowcasts only 11%. A very similar pattern was noted for Norway, with 1-Low forecast on 16% of the days, but forecasts with 1-Low being present comparably less often in the forecast-nowcast comparisons (10%). Thus, one can

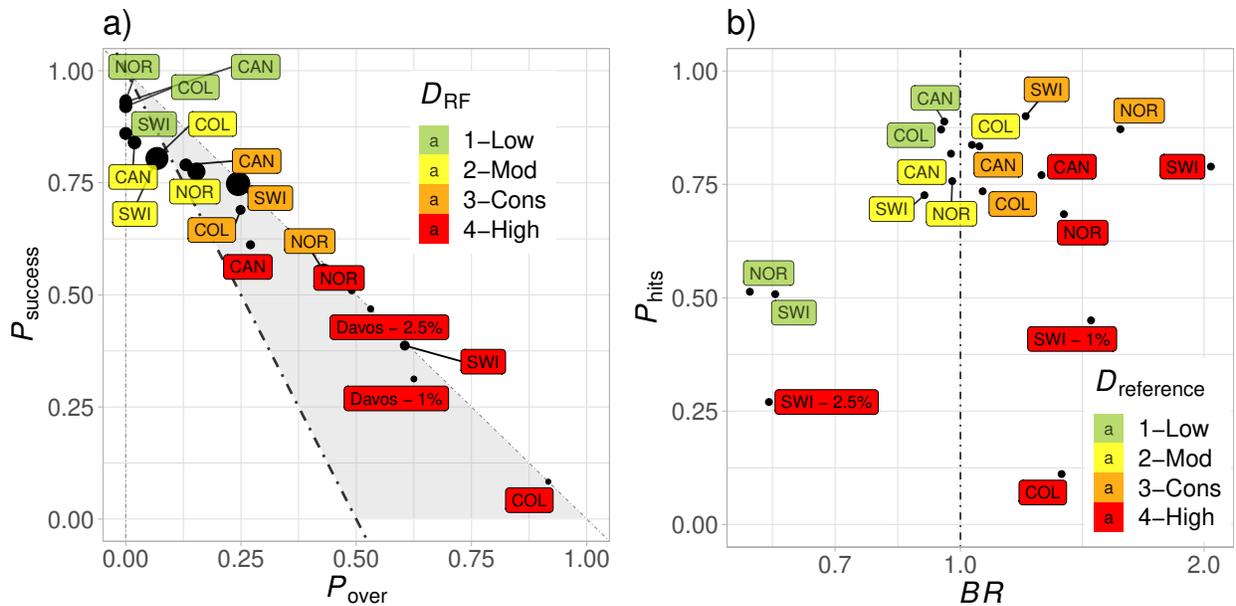


Figure 5.13: Two perspectives on forecast quality: a) the success rate,  $P_{\text{success}}$ , the proportion of forecasts which were confirmed by a reference assessment, and  $P_{\text{over}}$ , the proportion of forecasts assessed as too high (over-forecasts,  $D_{\text{RF}} > D_{\text{reference}}$ ). The light-grey shaded area to the right of the black dashed line represents the parameter space indicating over-forecasting ( $P_{\text{over}} > P_{\text{under}}$ ), while values to the left of this line indicate a higher proportion of under-forecasting ( $P_{\text{under}} > P_{\text{over}}$ ). In b) the hit rate  $P_{\text{hits}}$ , the proportion of  $D_{\text{reference}}$  which were correctly forecast, and the bias ratio  $BR$  are shown. Colours correspond to danger levels and three-letter abbreviations to countries. SWI - 1% and SWI - 2.5% correspond to forecasts verified using avalanche occurrence data (Sect. 5.3.4). - **The success rate decreased with increasing forecast danger level. The hit rate was lowest for 4-High in general.**

surmise that there is a underreporting bias of nowcast estimates confirming a forecast 1-Low.

Now, turning to the forecast bias, consistent patterns showed: all of the 78 Norwegian and Swiss observers with  $> 50$  comparisons between forecast and local assessment, perceived the forecast - if wrong - essentially always as too high rather than too low ( $P_{\text{over-under}} \gg 0$ ; Fig. 5.12b). With the exception of forecast  $D_{\text{RF}}$  1-Low, where over-forecasts are not possible, a pronounced tendency towards over-forecasting was evident for all four countries at danger levels 2-Moderate to 4-High with essentially all forecasts, which were not confirmed by  $D_{\text{reference}}$  being too high ( $P_{\text{over-under}} \gg 0$ , Fig. 5.13a). However, considering the bias ratio  $BR$ , where  $BR > 1$  indicates a more frequent use of the danger level in the forecast compared to the reference assessment, confirmed this bias for 3-Considerable (in Switzerland and Norway) and 4-High (in all countries, Fig. 5.13b). At 2-Moderate an almost balanced distribution was noted, which reflects that errors occurred both towards the lower 1-Low and the higher 3-Considerable.

### 5.3.4 On the success of forecasting 4-High

Beside the comparison between forecasts and nowcast assessments, avalanche observations were used to validate situations likely representing 4-High in the region of Davos (Switzerland). As introduced in Section 5.2.2, two thresholds were used as an indicator for  $D_{\text{reference}} \geq 4$ -High:

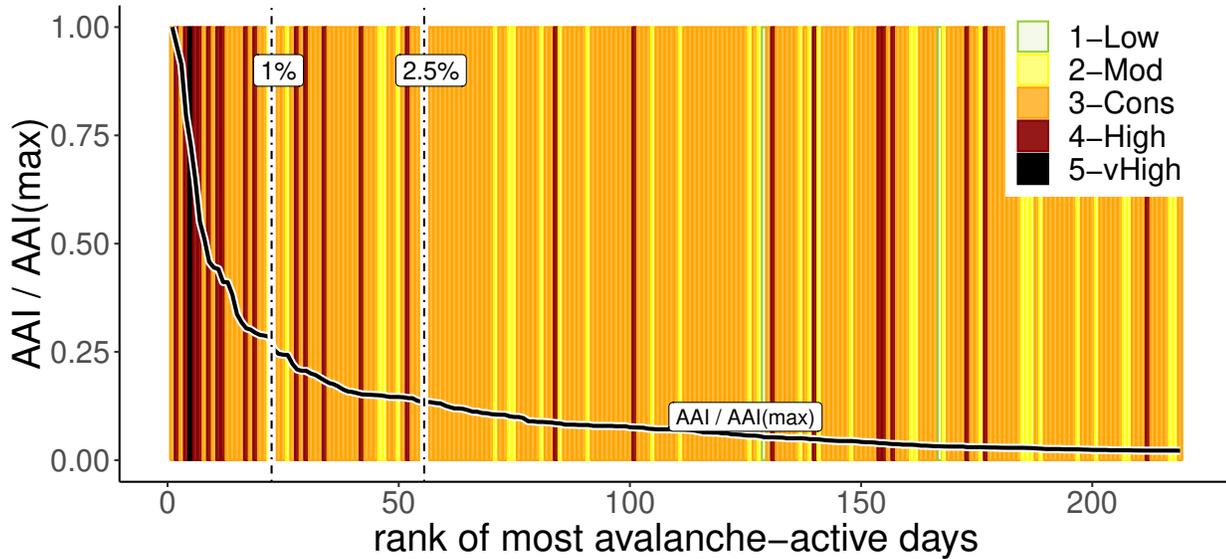


Figure 5.14: Avalanche activity in the region of Davos, expressed as the Avalanche Activity Index (AAI) per day relative to the maximum AAI per day in the 15 year period ( $AAI / AAI(max)$ , black line). Days were rank-ordered, with decreasing avalanche activity with increasing rank. Only the 10% of the most active days are shown ( $N = 221$ ). The colours represent the forecast danger level on the respective day. Thresholds for the most active 1% and 2.5% are shown (discussed in the text).

- The 1% of the rank-ordered days with the highest avalanche activity (of the more than 2200 forecast days) had a minimum AAI of 29 which is equal to two size 4 and nine size 3 avalanches, or 29 size 3 avalanches. Considering these 22 days as 4-High (Fig. 5.14), 10 of the 32 forecasts with 4-High or 5-Very High were within this limit (= success), 12 events were missed and 22 were outside (=false alarms). The hit rate was 0.45, the bias 1.45 indicating a tendency towards over-forecasting (Fig. 5.13).
- Using a 2.5% threshold instead ( $N = 55$  days, Fig. 5.14), 15 of the forecasts would have been correct, 40 events would have been missed, 17 would have been false-alarms. The hit rate was 0.27 and the bias 0.58 indicating more missed events than false alarms (Fig. 5.13). A 2.5% threshold corresponds to an  $AAI \geq 16$  which is equal to one size 4 and six size 3 avalanches, or 16 size 3 avalanches.

In summary, for situations with a re-assessed danger level 4-High, there were more misses and false alarms, compared to successes or hits, regardless which of the thresholds was considered. This is in line with the hindcast assessments made in Colorado, with an almost equal number of false alarms and misses, which were together about ten times more frequent than the hits (see also Appendix B.1, p. 209). Furthermore, when considering a 2.5% threshold as the decisive criteria, there were not only less than 25% of the situations forecast correctly, but there was a rather strong under-forecast bias. This is in contrast to the comparison with the  $D_{LN}$  estimates provided by mountain guides and field observers in Switzerland, which would suggest a strong over-forecast bias instead (Sections 5.2.2 and 5.3.3, Fig. 5.13a). Here, it is of note to mention that the local field estimates also essentially always failed to correctly assess these situations. Now,

turning to the forecasts in Colorado, which also showed low values for  $P_{\text{success}}$  and  $P_{\text{hits}}$  (Fig. 5.13), Logan (2020) provided a potential explanation: in the forecast products, forecasters in Colorado focus on communicating the highest expected danger in the often large regions, while during the re-assessment forecasters tended to rate the average conditions in a region more than comparably small parts of the region which had more unfavorable conditions. However, as shown exemplary for Switzerland and Valle d'Aosta (Section 5.1.3), this change in the way a danger level is assigned to a region - whether the most widespread conditions or the most unfavorable conditions are considered - will by itself cause a comparably lower proportion of 4-High in the re-assessments compared to the forecasts, and hence a lower success rate.

## 5.4 Elements of avalanche danger - snowpack stability, the frequency distribution of snowpack stability and avalanche size

Now, the focus turns to describing the danger levels using observational data related to the three elements characterizing avalanche danger - snowpack stability, the frequency distribution of snowpack stability and avalanche size (publication 4).

### 5.4.1 Snowpack stability

Snowpack stability and the frequency distribution of snowpack stability are two elements defining the danger levels.

#### Observed stability distributions - Rutschblock and Extended Column Test

The stability distributions obtained with the Rutschblock and the Extended Column Test were analyzed at danger levels 1-Low to 4-High (Fig. 5.15). At 4-High, very few stability tests were observed.

**Rutschblock (RB):** The proportion of *very poor* rated RB tests increased monotonically with increasing danger level from 2% at 1-Low to 38% at 4-High (Fig. 5.15a). As a consequence, the combined proportion of *very poor* and *poor* rated tests also increased strongly from 7% to 67%, while the proportion of tests rated as *good* decreased accordingly (69% to 10%, Fig. 5.15a). These patterns were also confirmed when exploring the correlation between the RB stability class and danger level (Spearman rank-order correlation;  $\rho = 0.4$ ,  $p < 0.001$ ).

**Extended Column Test (ECT):** Additionally to the RB, the stability distributions derived from ECT results and performed not only in Switzerland but also in Norway at 1-Low to 4-High were explored (Fig. 5.15b).

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. 5.15b). Again, a positive though weak correlation between stability rating and danger level was noted ( $\rho = 0.22$ ,  $p < 0.001$ ).

In comparison to the RB (Fig. 5.15a), the ECT showed less distinct changes in the frequency of the most

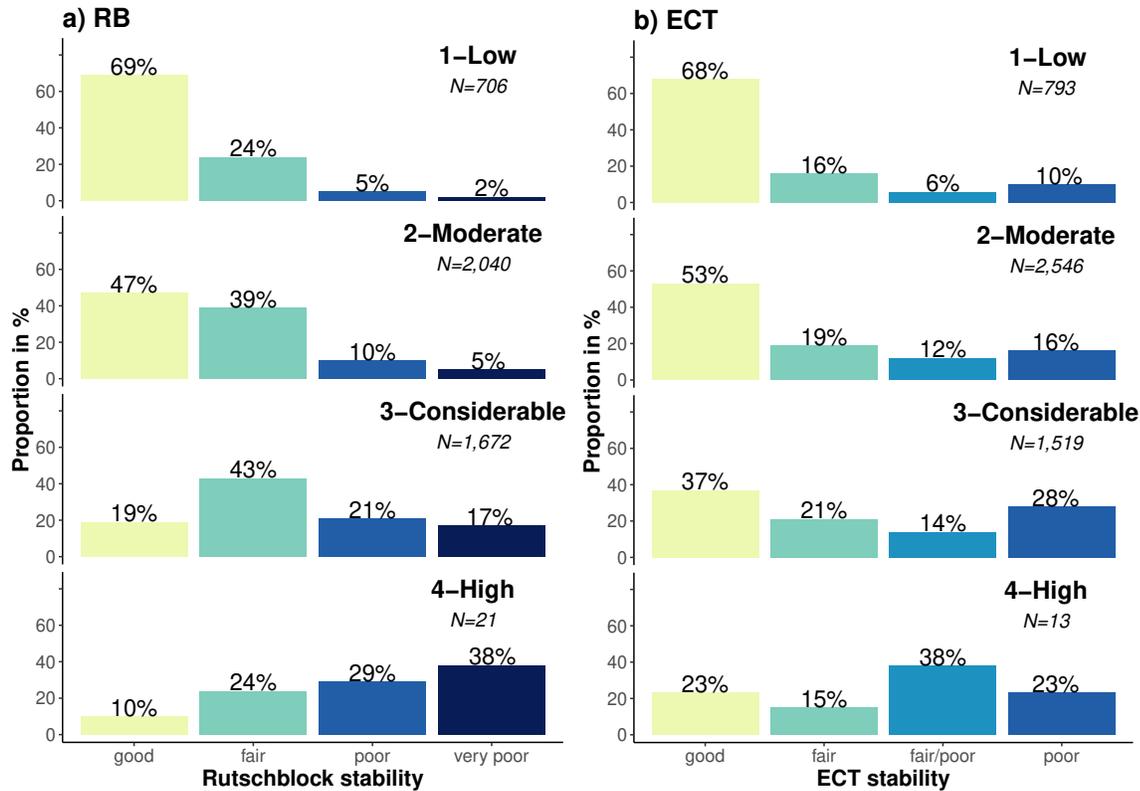


Figure 5.15: Distribution of stability ratings for the stability tests (a) Rutschblock (RB) and (b) ECT for danger levels 1-Low to 4-High. For the definition of the stability classes refer to Fig. 3.6 (Sect. 3.3.1, p. 29). Note the small  $N$  for 4-High for both tests. The blue colours are different for RB and ECT, as these classes do not line up the same (a result observed in publication 5, see Appendix A.6). RB class *poor* corresponds approximately to ECT class *poor* in terms of detecting a similar number of *unstable* slopes given this test result.

unstable and most stable classes between danger levels, and hence the correlation with the danger level was lower (ECT:  $\rho = 0.22$  vs. RB:  $\rho = 0.4$ ).

### Frequency of *very poor* snowpack stability

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

Here, the frequency of *very poor* stability based on sampling 25 Rutschblock tests and four frequency classes is described. Regarding the sampling and the class definition procedure refer to Sect.s 4.3 and 4.4, regarding the sensitivity of these settings on the results, refer to publication 4 (Sect. A.5).

Using four frequency classes, and labeling them *none or nearly none*, *a few*, *several* and *many*, the thresholds in the proportion *very poor* stability between frequency class labels were 0, 0.04 and 0.2, respectively. 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.

Large proportions of *very poor* stability (e.g.  $\geq 0.5$ ) 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

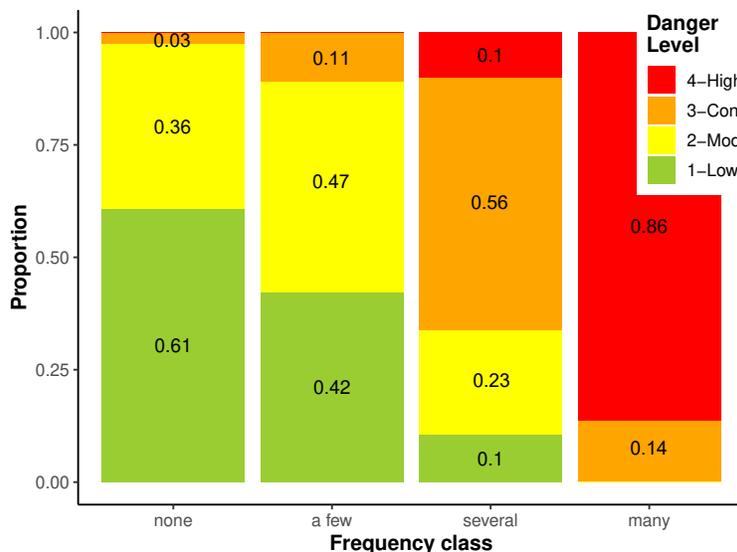


Figure 5.16: Distribution of the danger levels for the four frequency classes describing the proportion of *very poor* snowpack stability, derived from sampling 25 Rutschblock tests (as described in Sect. 4.3). The respective proportions are indicated for each of the four danger levels.

are more frequent (Fig. 5.15a), and 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 ( $\rho = 0.81$ ,  $p < 0.001$ ). 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. 5.16). 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.

## 5.4.2 Avalanche size

Most avalanches in the Swiss data set were size 1 (Fig. 5.17a), 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 to 3-Considerable shows that the most frequent avalanche size had little discriminating power to differentiate between danger levels. The median avalanche size was size 1 at 1-Low and 2-Moderate, size 1 to size 2 at 3-Considerable, and size 2 at 4-High (Fig. 5.17a).

Considering the size of the largest reported avalanche per day and warning region showed that the largest avalanche per day and region was most frequently size 2 for 1-Low and 2-Moderate, a mix of size 2 and size 3 at 3-Considerable, and size 3 at 4-High (Fig. 5.17b). The proportion of days when size 1 avalanches were the largest observed avalanche decreased significantly with increasing danger level (from 33% to 1%,  $p < 0.001$ , proportion test (R-function *prop.test*, R Core Team, 2017)), while the proportion of days with at least one size 3 or size 4 avalanche increased significantly (from 20% to 78%,  $p < 0.001$ ). At 4-High, almost 80%

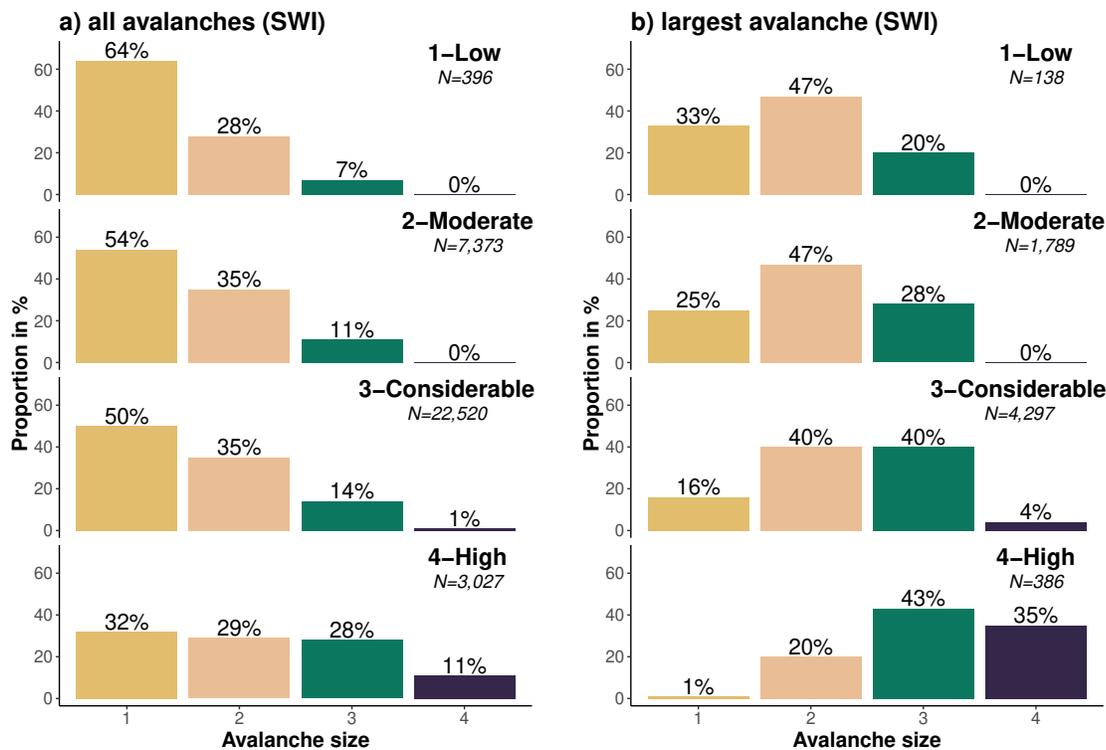


Figure 5.17: Size distribution of dry-snow avalanches, which released naturally or were human-triggered for danger levels 1-Low to 4-High, showing all avalanches (a) and the largest reported avalanche per day and warning region (b) in Switzerland (SWI).

of the days had at least one avalanche of size 3 or 4 recorded.

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

Note that days with no avalanches were not explored as the size of avalanches was of interest, not their frequency. The frequency component is addressed using the frequency of locations with *very poor* stability as a proxy.

### 5.4.3 Combining the frequency of *very poor* stability and avalanche size

Assuming that the stability class *very poor* corresponds to the actual trigger locations, the snowpack stability class, the frequency of this stability class and avalanche size were combined. Hence, this combination considers all three elements characterizing the avalanche danger level. The resulting simulated data set contained the following information: *danger level*, *frequency class describing occurrence of very poor stability*, *largest avalanche size*. These data looked like the following, here for 1-Low:

Sample 1: 1-Low, a few, largest avalanche size 1

Sample 2: 1-Low, none or nearly none, largest avalanche size 2

Sample 3: 1-Low, a few, largest avalanche size 1

...

*Sample B: 1-Low - none or nearly none - largest avalanche size 1*

The most frequent combinations of the frequency class and avalanche size for each danger level were:

- 1-Low: *None or nearly none* locations with *very poor* stability (53% of sample) existed. The largest avalanches were size 2 (48%).
- 2-Moderate: *A few* locations with *very poor* stability (37%) were present. The typical largest avalanche was of size 2 (50%).
- 3-Considerable: *Several* locations with *very poor* stability (75%) existed. The typical largest avalanches were sizes 2 or 3 (79%).
- 4-High: *Many* locations with *very poor* stability (86%) existed. The typical largest avalanche was of size 3 (43%).

#### 5.4.4 Data-driven lookup table for danger level assessment

Finally, a data-driven lookup table to assess avalanche danger is introduced using the simulations presented before (Fig. 5.18). For this, a step-wise approach and two matrices as proposed by Müller et al. (2016) in the so-called Avalanche Danger Assessment Matrix (ADAM) were used.

The first matrix (Fig. 5.18a), which we refer to as *stability matrix*, combines snowpack stability and the frequency class of the most unstable stability class observed. Cell labels (letters A to E) in this matrix were assigned based on similar danger level distributions behind the respective stability class - frequency class combination. 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. For class *none or nearly none* no letter is assigned, as the next higher stability class should be considered.

The second matrix (Fig. 5.18b), which we refer to as *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 characterizing this combination. If the second most frequent danger level was present more than 30% of the cases, the value is shown with no brackets, if present between 15 and 30% it is placed in brackets.

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

1. In the *stability matrix* (Fig. 5.18a), 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. 5.18b), where it is combined with the largest avalanche size (Fig. 5.18b).
3. The most frequent danger levels that were typical for this combination, are shown.

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

\* none or nearly none  
 \*\* if none, check frequency of next higher stability class  
 - no data  
 C cell contains less than 1% of the data

		avalanche size			
		1	2	3	4
snowpack matrix	A	<b>3, 4</b>	<b>4 (3)</b>	<b>4</b>	<b>4</b>
	B	<b>3 (2, 1)</b>	<b>3 (2)</b>	<b>3 (2)</b>	<b>4, 3</b>
	C	<b>2 (3)</b>	<b>2, 3</b>	<b>3, 2</b>	-
	D	<b>1, 2</b>	<b>2, 1</b>	<b>2, 1</b>	<b>3 (2)</b>
	E	<b>1</b>	<b>1 (2)</b>	<b>1 (2)</b>	-

3: >30%  
 (3): 15-30%

Figure 5.18: Data-driven lookup table for avalanche danger assessment (similar to the structure proposed by Müller et al. (2016)). The (a) *stability matrix* combines the frequency class of the most unfavorable snowpack stability class (columns) and the snowpack stability class (rows) to obtain a letter describing specific stability situations, The (b) *danger matrix* combines the largest avalanche size (columns) and the specific stability situations (letter) obtained in the stability matrix (rows) to assess the danger level. In (b): The most frequent danger level is shown in bold. If the second most frequent danger level was present more than 30% of the cases, the value is shown with no brackets, if present between 15 and 30% it is placed in brackets. In (a) and (b): Cells containing less than 1% of the data are marked.

## 5.5 On the snowpack stability interpretation of instability tests

And lastly, a brief overview is given for the findings obtained in publication 5, which focused on the development of a quantitative classification scheme for Extended Column Test (ECT) results (see Section 3.3.1 for details regarding the ECT).

In this analysis, ECT results were compared with observations regarding slope stability (Swiss data). A test location was considered *unstable* if signs of instability or avalanches were reported in the immediate surroundings of the profile location, and *stable* if no such signs were noted. The classification scheme was developed to include this data more easily into operational procedures, but also to be able to integrate ECT results as an additional data-source to characterize snowpack stability and the frequency distribution of snowpack stability (publication 4, Appendix A.6).

The data showed that by combining the results regarding the crack propagation propensity and the number of taps required to initiate the crack, four classes with distinctly different proportions of *unstable* slopes could be derived using a clustering approach. The resulting classification scheme was introduced in the Data section (Fig. 3.6, p. 29) and was applied to rate ECT results in Sect. 5.4.1.

The data showed further that the four derived ECT stability classes correlated with slope stability (Fig. 5.19a), though this correlation was weaker compared to results obtained with a Rutschblock test (RB) performed in the same snow pit (Fig. 5.19b). A similar pattern can be noted in Fig. 5.15b: although the stability distributions obtained with the ECT correlated with danger levels 1-Low to 4-High, the proportions of tests indicating *poor* / *very poor* stability increased in a more distinct way with increasing danger level for the RB (Fig. 5.15a) compared to the ECT.

And finally, a comparison of the Swiss findings (Fig. 5.20a) with results based on ECT data from North America (primarily from the United States) brought very similar results (Fig. 5.20b). Thus, the conclusion obtained using Swiss data is valid in general: Crack propagation propensity, as observed with the ECT, is a

**Proportion of unstable slopes** (signs of instability, avalanches)

a) Extended Column Test (ECT)		ECT stability class				
		<i>poor</i>	<i>poor-to-fair</i>	<i>fair</i>	<i>good</i>	<i>all</i>
$D_{RF}$	<i>1-Low</i>	0.10	0.00	0.00	0.02	0.02
	<i>2-Moderate</i>	0.33	0.11	0.09	0.05	0.10
	<i>3-Considerable</i>	0.70	0.54	0.31	0.22	0.38
	<i>all</i>	0.52	0.30	0.18	0.10	

b) Rutschblock (RB)		RB stability class				
		<i>very poor</i>	<i>poor</i>	<i>fair</i>	<i>good</i>	<i>all</i>
$D_{RF}$	<i>1-Low</i>	0.50	0.00	0.00	0.00	0.02
	<i>2-Moderate</i>	0.48	0.19	0.07	0.05	0.10
	<i>3-Considerable</i>	0.74	0.52	0.29	0.16	0.36
	<i>all</i>	0.65	0.40	0.17	0.07	

Figure 5.19: In publication 5, the stability tests Extended Column Test (ECT, see also Fig. 3.4b) and Rutschblock (RB, see also Fig. 3.4a) were compared to observed signs of instability and avalanches in their surroundings, which are related to the probability that an avalanche can be triggered by a human. In the above matrix, the frequency that such signs were observed conditional on the forecast danger level and the test result is shown. **The four RB classes correlated better with the frequency that signs of instability were observed than the ECT.**

key indicator relating to snow instability. The number of taps required to initiate a crack provides additional information concerning snow instability. Combining crack propagation propensity and the number of taps required to initiate a failure allows refining the original binary stability classification.

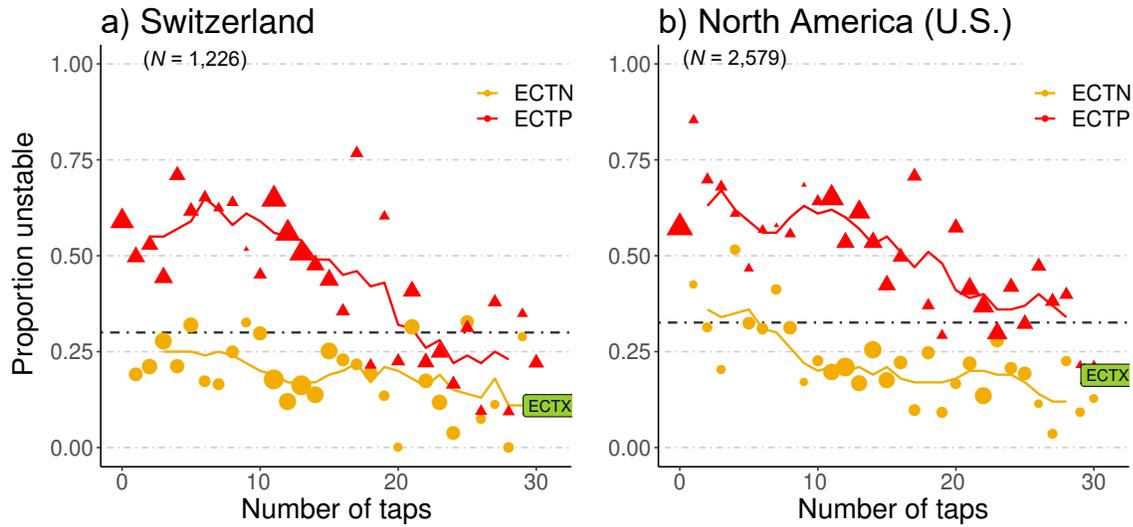


Figure 5.20: Proportion of *unstable* ECT locations for each combination of crack propagation (ECTN - no or partial propagation, ECTP - full propagation, ECTX - no failure) and number of taps until failure. The larger the symbols, the more data points. The respective colored lines represent a running average, calculated over five consecutive number of taps. The black dashed line represents the base rate, the proportion of *unstable* locations in the data set. **ECTP (red triangles) were observed more often in *unstable* locations (above the black dashed line), ECTN (orange circles), and ECTX in *stable* locations. The proportion of *unstable* locations for ECTP $>$ 22 and ECTN $\leq$ 8 neither truly indicated *unstable* or *stable* conditions.** The data for North America will be presented in a manuscript by Techel et al. (2020a).

# Chapter 6

## Discussion

The overarching objective of this thesis was to gain data-driven insights regarding consistency and quality in public avalanche forecasts. With this goal in mind, the focus of this section lies in summarizing the key findings and highlighting potential implications for public avalanche forecasting. The findings relating to consistency and quality in avalanche forecasts are discussed in Sect. 6.1 (see also Fig. 6.1), which is followed by discussing the data-driven characterization of the danger levels (Sect. 6.2). The main limitations encountered during this dissertation are discussed in Sect. 6.4.

### 6.1 Consistency and quality in avalanche forecasts

First, the research questions and some of the key findings are summarized, before discussing these in relation to their application in avalanche forecasting.

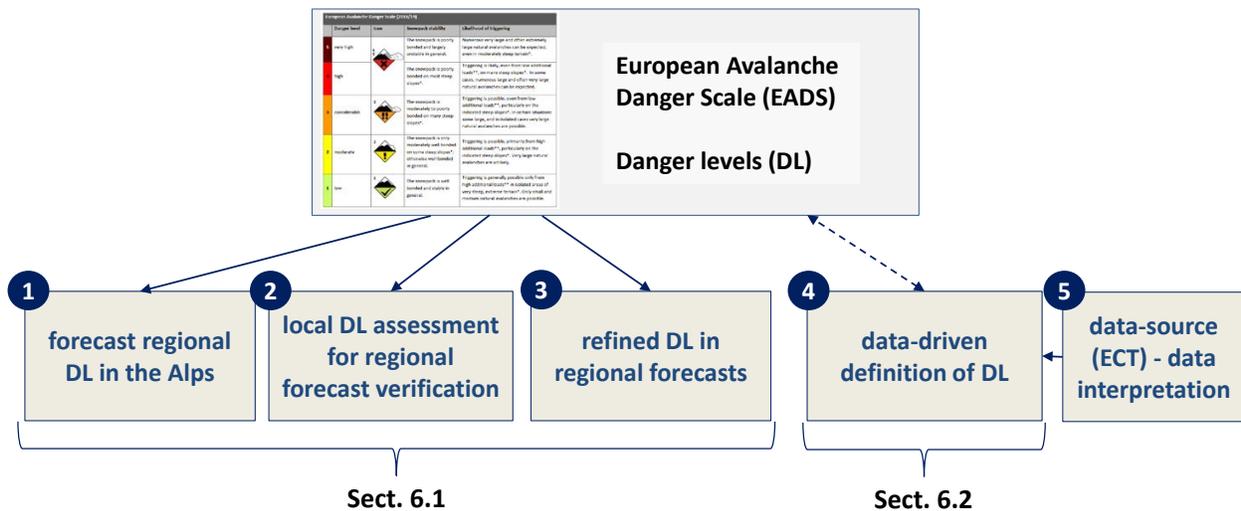


Figure 6.1: Overview showing the five publications, and the respective sections, where some of the key results are discussed.

### 6.1.1 Key findings

In **publication 1**, the focus was on spatial consistency and bias in forecast danger levels. The two research questions in the focus of this study were: *Do differences in the use of the danger levels between forecast centers exist? Can operational constraints (such as the size of the warning regions) explain these differences?*

Considerable differences in the size of the warning regions, the smallest spatial units used in the forecasts by the forecast centers in the European Alps, existed. Their average size differed by a factor of more than 10 between forecast centers, with even larger differences when comparing to the size of warning regions of forecast centers outside of the Alps (Fig. 2.7, p. 22; e.g. Jamieson et al., 2008; Storm and Helgeson, 2014; Engeset et al., 2018). These differences explained, at least partly, the significantly lower agreement rate between forecast danger levels of neighboring warning regions belonging to different forecast centers ( $P_{\text{agree}}(\text{across}) = 0.63$ ), compared to within forecast center boundaries ( $P_{\text{agree}}(\text{within}) = 0.93$ ; Fig. 5.5, p. 47). Furthermore, the analysis of forecast danger levels in the Alps showed that in more than 80% of the forecasts, the forecast danger level was either 2-Moderate or 3-Considerable (Fig. 5.3, p. 45). This narrow dispersion in the forecast danger levels - most often one of two danger levels were used, compared to the five levels available - indicates a lack of refinement (Wilks, 2011). And finally, considerable variations in the use of danger level 4-High was noted, with some regional forecast centers in France using this danger level five times more often compared to most other forecast centers in the Alps (Fig. 5.6, p. 48).

In **publication 2**, the focus was on the quality of local danger level estimates for forecast verification. This study was expanded by additionally exploring a data set from Norway and addressed the following two research questions: *Do variations in local danger level estimates exist? What implications do these variations have on the reliability of local danger level estimates as a data-source for forecast verification?*

In a nutshell, the key findings regarding the reliability of local danger level estimates were:

- Variations in local danger level estimates existed, even at relatively short distances of less than 10 km, with the average agreement rate between two local estimates being about 0.8 (Fig. 5.10, p. 53).
- Hence the reliability, which can be understood as the factor describing the repeatability associated with an individual danger level estimate, and hence the «trust» we can place in a single estimate, was about 0.9.
- The validity of local danger level estimates for avalanche situations considered in hindcast assessments as 4-High was surprisingly very low (Sect. 5.2.2).

Turning to the quality of the forecast danger level, the research questions and their key findings were: *What implications do the variations identified between local danger level estimates have for the verification of regional avalanche forecasts? Relying on local danger level estimates, what is the perceived accuracy and bias of forecast danger levels? Can differences between countries with different operational constraints and verification methods be noted?* The key results were:

- The observed quality of the forecast danger level was lower when compared with individual local

danger level estimates (e.g. in Switzerland:  $P_{\text{correct}} = 0.82$ ), as compared to situations when two observers provided the same estimate ( $P_{\text{correct}} \approx 0.89$ ; Tab. 5.1, p. 56).

- As shown in several studies (e.g. Brenner and Gefeller, 1997; Bowler, 2006) observational errors - or errors in the reference class in general - and assigning these errors randomly penalizes the apparent skill of a forecast further. This also showed when randomly assigning errors to the local danger level estimates: the proportion of forecasts considered correct in Switzerland dropped from  $P_{\text{correct.raw}} = 0.77$  to 0.71 (publication 2, App. A.3). In contrast, using the reliability of local danger level estimates as an upper bound in this calculation suggested an accuracy of the forecast danger level of about  $P_{\text{correct}}^* = 0.9$  (for Switzerland; Tab. 5.1, p. 56).

Thus, an estimate of the overall accuracy of a forecast regional danger level would therefore be somewhere between 0.85 (in Norway) and 0.9 (Switzerland). Or, in other words, on between one out of seven days (Norway) to one out of ten days (Switzerland), a forecast danger level was perceived as wrong. However, such overall accuracy measures do not tell the whole story:

- The success rate, that is the proportion of a forecast danger level which was confirmed by a reference assessment (a nowcast or hindcast assessment), decreased with danger level from a success rate of more than 0.9 at 1-Low to less than 0.6 at 4-High (Fig. 5.13a, p. 59).
- A strong tendency towards over-forecasting was noted when comparing forecasts with the reference assessment. The proportion of forecasts being considered too high increased from forecasts with 2-Moderate to 4-High (Fig. 5.13a, p. 59).
- In contrast, the hit rate, that is the proportion of the danger level estimated in a nowcast which was correctly predicted, ranged between 0.7 and 0.9 for most danger levels (Fig. 5.13b, p. 59).
- Comparing the forecast danger level with avalanche observations for the region of Davos (Switzerland) for situations, which may correspond to a danger level 4-High, showed a tendency towards more misses or false alarms rather than successes (Sect. 5.3.4, p. 59).

In **publication 3**, the balancing act between communicating avalanche danger in a simple and well-established manner using the five danger levels and assessing avalanche danger with greater detail was explored. The specific research objective was whether sub-levels, assigned to a danger level during the forecast process, actually have skill. The two research questions in that regard were: *Can the forecast regional danger level be refined by assigning a sub-level? Are these sub-levels significantly better than randomly assigned ones?* The key findings, based on a four-year data set of avalanche forecasts in Switzerland, where a sub-level was assigned together with the forecast danger level during the production of the forecast, showed that some anomalies in the use of the sub-levels existed (App. A.4: Fig. A.17, p. 148). These could be linked to operational constraints in the production process of the forecast. Despite these anomalies, the forecast sub-levels were clearly better than random (App. A.4: Fig. A.19, p. 152), indicating that forecasters can often forecast avalanche danger at greater detail than the established five danger levels.

### 6.1.2 On the influence of spatial resolution and the way avalanche danger is communicated in avalanche forecasts on consistency and quality

In the following, the influence of variations in the spatial resolution of the forecasts and the way an avalanche danger level is assigned to a warning region on the consistency and quality in forecast products are discussed. In the example in Fig. 6.2 two operational settings are shown: a forecast center communicating avalanche danger at a much lower spatial resolution (one region) compared to a forecast center using a higher spatial resolution (12 regions) and a flexible approach to combine warning regions. Furthermore, two approaches to assign a danger level to a warning region are taken into account, as the avalanche danger scale (Tab. 2.2, p. 23) lacks a definition in this regard: (a) the highest danger level and (b) the spatially most widespread conditions are communicated in the forecast product. And finally, the case that a forecaster can assign a refined danger rating, by using a higher resolution of the danger levels, is considered (case (c) in Fig. 6.2).

#### Influence on consistency

The two schemes - 12 regions vs. 1 region - are clearly inconsistent in terms of their spatial resolution (Fig. 6.2). In the case of one warning region, a greater abstraction is required when translating the expected conditions to a map-based product, compared to a forecast product using 12 regions. Such inconsistencies shown here for the spatial component, will also occur if the temporal resolution of forecasts differs (as noted in publication 1).

A further inconsistency arises, when the approach to assign a danger level to a warning region differs: communicating the highest danger level for a warning region (row a in Fig. 6.2) or communicating the spatially most widespread conditions (row b). Even when the same spatial resolution is used, the resulting forecast will differ. In the case of a one-region resolution, this will impact the forecast danger level for the entire forecast domain. In the 12-region example, only parts of the forecast domain will differ.

Finally, case c in Fig. 6.2, where a forecaster assesses the danger level in greater detail, allows making fewer abstractions - as both the spatial component and the variation within the danger level can be expressed.

The spatial and temporal resolution of a forecast product are relevant components to maintain consistency in forecasts. Defining these should be guided by the availability of relevant and reliable data in a sufficient spatial resolution and temporal frequency allowing the best-possible translation of the expected avalanche conditions in the forecast product. If abstractions are necessary, these should be made at the end of the assessment process, for instance when preparing the forecast product. Thus, a warning service that can refine avalanche danger in space, time and sub-level, regardless whether in the forecast or when re-evaluating the forecast using field observations, measurements and models, should do so - at least for their internal assessment.

These findings indicate a need to revisit and further harmonize the way avalanche danger is assessed and assigned to warning regions. Harmonization should consider

- similar approaches regarding the way a danger level is assigned to a warning region, i.e. whether the highest or the spatially most widespread conditions are communicated, and

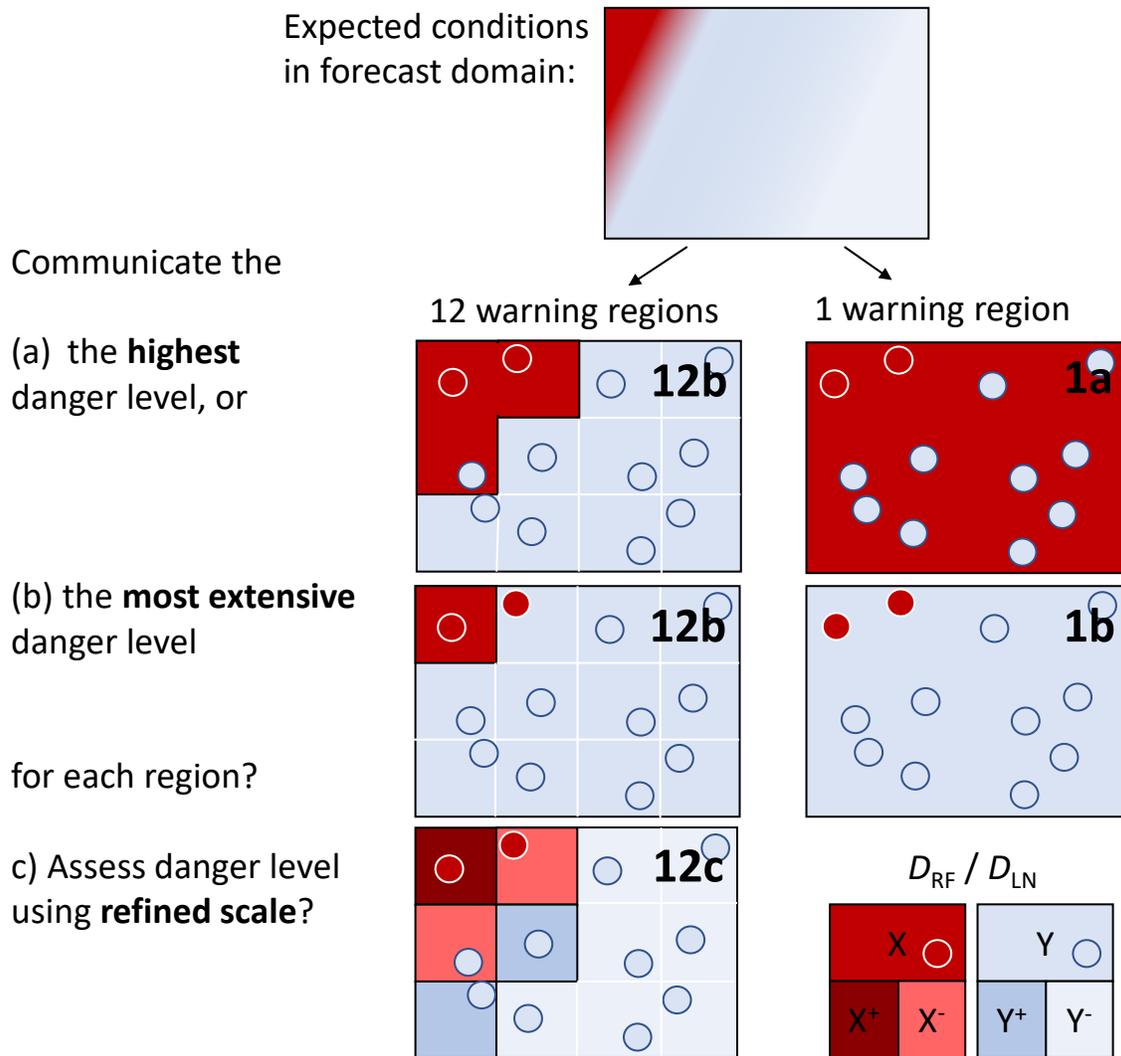


Figure 6.2: In the upper square, showing a forecast domain, the expected avalanche conditions - represented by a higher (dark red,  $X$ ) and a lower (light blue,  $Y$ ) danger level ( $D_{RF}$ ) are shown. This situation will be assessed and communicated differently, depending on the size of the warning regions used by the warning service (12 vs. 1 region), and whether the highest (a) or the spatially most widespread (b) avalanche conditions are considered relevant for communication. Furthermore, in c, forecasters assess avalanche danger  $D_{RF}$  at a refined level of detail, where the respective sub-levels (e.g.  $X^+$  and  $Y^+$ ) are within the definitions of the danger level, with  $X^+ > X^- > Y^+ > Y^-$ . The circles represent local danger level estimates ( $D_{LN}$ ).

- similar approaches regarding the size of warning regions and their aggregation, with a preference towards using a finer spatial resolution.

If different approaches must be used, for instance due to operational constraints or a lack of data, the approach taken should be communicated to forecast users.

### Influence on quality

Spatial consistency in public avalanche forecasts was explored by analyzing the rate that the forecast danger level in neighboring warning regions agreed (publication 1, Sect. 5.1.1).

Within the domain of a forecast center, the agreement rate ( $P_{\text{agree}}(\text{within})$ ) was about 0.95, at distances between 25 and 75 km it was about 0.9 (Fig. 5.5a, p. 47). In this case, it can be assumed that the same forecaster (or forecaster team) issued the danger level for neighboring warning regions. These values can, therefore, be interpreted as an estimate of the typical spatial correlation in avalanche danger level observed within the forecast domains in the Alps. When comparing across forecast center boundaries, however, the agreement rate ( $P_{\text{agree}}(\text{across})$ ) was about 0.6, with no change with distance. In this situation, different forecasters in different forecast centers, with different operational constraints and potentially a somewhat different data or knowledge base, issued the danger level. Findings regarding the quality of the forecast danger level (Sect. 5.3, p. 54) showed that the accuracy of a forecast danger level  $D_{\text{RF}}$  was about 0.85. Assuming this to be a reasonable approximation for the accuracy of the forecast danger level in the Alps, the agreement rate between danger levels issued by different forecast centers is defined by the accuracy of the two forecasts (Sect. 4.1, p. 34), and the distance between these regions (here the distance between the centers of these regions). Excluding the spatial component, the resulting expected agreement rate would be  $0.85 \times 0.85 = 0.72$ . If we now incorporate the spatial variation of the avalanche danger level ( $1 - P_{\text{agree}}(\text{within}) = 0.05$  to  $0.1$ ), we end up with values of about 0.65, rather close to the observed agreement rate  $P_{\text{agree}}(\text{across})$  of 0.6. This simple example suggests that the accuracy related to the forecasts themselves, which also depends on the (un)reliability of the forecaster making this judgment (Stewart, 2001), may not only be of a similar magnitude but may even be larger than the actual spatial variation of avalanche danger. It is therefore well possible that this is the driving contributor to variations noted in the forecast danger levels of immediately neighboring warning regions belonging to different forecast domains, rather than actual spatial variations in avalanche danger.

Returning to Figure 6.2, which also shows 12 local danger level estimates  $D_{\text{LN}}$ , and using these for forecast verification, we can see how the spatial resolution (one or 12 regions) and the way avalanche danger is communicated (the highest or the spatially most widespread danger level) affects the observed accuracy of the forecast: In case 1a (one region), two of twelve  $D_{\text{LN}}$  estimates agreed with the forecast (17%), in 1b the agreement was 83% (10 of 12). While one could argue that communicating the most widespread conditions would be appropriate for most users, it would probably not be in line to provide a warning to the public, as the public would not become aware of the most critical conditions. In contrast, using a higher resolution, the observed accuracy would be 92% in both cases (2a and 2b). Thus, simply by increasing the spatial resolution of the product allowing a more explicit match between expected conditions and the forecast, the observed accuracy is at least as high. Furthermore, inconsistencies in the way avalanche danger is communicated, that is switching from communicating the most unfavorable vs. the spatially most widespread conditions, impacts the overall accuracy to a lesser degree if the spatial resolution is high. This example also highlights challenges, when verification is done for large regions with variations in avalanche conditions. Again, the match between the temporal and spatial resolution of forecasts and re-assessments should be as close as possible, with a preference towards the highest possible resolution given the availability of data.

Already Murphy (1993, p. 288), in his essay on the goodness of (weather) forecasts, wrote: «... it should be evident that placing arbitrary restrictions on the content, format, etc., of forecasts may introduce inconsistencies that detract from their quality».

### 6.1.3 Forecast quality: overall forecast accuracy, over-forecasting and forecasting 4-High

The forecast accuracy was higher in Switzerland ( $P_{\text{correct}} > 0.89$ ) compared to Canada, Colorado, or Norway ( $P_{\text{correct}} \approx 0.84$ ). This difference can probably be explained by the fact that the warning regions are about a factor 10 (or more) smaller in Switzerland compared to those used in Canada, Colorado or Norway. This allows a Swiss forecaster to be more specific when spatial variations in avalanche conditions exist. As was shown using the examples of the forecasts in Valle d'Aosta and Switzerland (Fig. 5.9a, p. 51), the situation that forecast avalanche conditions varied within regions of several thousand km<sup>2</sup> occurred rather often: within regions of this size, two (or more) danger levels were issued on about 40% of the days. Similarly, about 10% of the variation in the forecast danger levels in neighboring warning regions (Fig. 5.5a, p. 47) or in local nowcast estimates (Fig. 5.10a, p. 53) could be attributed to spatial variations in avalanche danger at distances of several dozen kilometers. It is not known which approach of assigning a danger level to a warning region - in case spatial variations in avalanche conditions exist - is used in the forecasts and nowcast or hindcast assessments in Canada, Colorado, and Norway. It is of note, however, that the Canadian values are much higher than a study by Jamieson et al. (2008), who showed  $P_{\text{correct.raw}}$  values for forecasts in Western Canada of about 0.65, and a hit rate of 0.76 for the Rocky Mountain forecasts. Jamieson et al. (2008) showed further that a confirmation bias existed for forecasters, who wrote the forecast themselves. Two findings that hold across the four explored data sets from Switzerland, Canada, Colorado and Norway are (Sect. 5.3, p. 54):

The forecast accuracy was lowest for situations with 4-High, regardless whether the forecast was compared with local estimates, or with nowcast or hindcast assessments by forecasters, or with avalanche observations. Often less than half of these days were correctly forecast. However, these are highly relevant and critical days, as large or very large natural avalanches may release, not only impacting recreational users but also those traveling on highways or living in alpine villages. It is therefore imminent, that these days are more accurately forecast, without causing too many false alarms. This may point towards the necessity to emphasize the need for snowpack models, which accurately predict snowpack stability using grid-based weather forecast data as input. Furthermore, to correctly assess avalanche activity during prolonged storms with often poor visibility, it would be beneficial if real-time automatic avalanche detection methods were available in greater number.

The second point refers to the strong tendency to over-forecast, again seen in all four data sets (Section 5.3), but also noted in publications 2 and 3. Particularly in publication 3 (Appendix A.4), forecast accuracy was also explored as a function of forecast avalanche conditions in neighboring warning regions. The same pattern of over-forecasting was observed in a spatial context: when a warning region bordered another warning region with a lower danger level, the forecast danger level tended to be more often wrong and essentially always too high (e.g. publication 3:  $P_{\text{correct.raw}} = 0.43$ ; Tab. A.15, p. 153), compared to the case when the forecast danger level was lower than in at least one of the neighboring warning regions ( $P_{\text{correct.raw}} = 0.98$ ). While it is understandable to rather «err on the side of caution» (Jamieson et al., 2008) rather than miss events, forecasts should be as accurate as possible to remain credible (Williams, 1980).

### 6.1.4 Consistency and quality: potential implications for the value of avalanche forecasts

A key advantage of the introduction of the EADS in 1993 was seen as the provision of consistent information across the European Alps (Meister, 1995). The forecast danger level is the part of the forecast most known and used in the Alps (Winkler and Techel, 2014; LWD Steiermark, 2015; Procter et al., 2014), influencing the selection of backcountry destinations (Techel et al., 2015b) and local decision-making by recreationists (Furman et al., 2010).

Many users of avalanche forecasts are typically active within warning regions where forecasts are produced by a single regional avalanche forecast center (e.g. in the forecast domains of Voralberg or Tirol in Austria). Such users are likely to become accustomed and calibrated to «their» forecast. Thus, issues are likely to arise when users travel from one forecast center domain to another. For instance, a user will probably assume that the forecast objectively represents the expected conditions given the inherent uncertainty related to forecasts in general. Hence, the user will conclude that spatial variations in the danger level seen in the forecasts of neighboring forecast centers represent spatial variations in the expected avalanche conditions. However, the user will most likely be unaware of variations in the operational constraints or regarding the way avalanche danger is assigned to warning regions between forecast centers, which contributed to a large part to the proportion of disagreements in forecast danger level across borders (Sect. 6.1.2). Furthermore, the differences in the use of danger level 4-High also may have implications: a frequent user of French forecasts traveling to Switzerland may experience some Swiss forecasts with 3-Considerable as a missed alarm, while the opposite may happen when a Swiss user recreates in France, where 4-High is issued much more frequently (Sect. 5.1.2).

In all of these cases, the credibility of the forecasts will be reduced, as they are perceived to be less accurate (Williams, 1980).

### 6.1.5 On using local danger level estimates as a data source for forecast verification

The average agreement rate between two local danger level estimates ( $D_{LN}$ ) at relatively short distances was about 0.8 (Fig. 5.10, p. 53). Several factors contributed to these variations: Spatial variations in avalanche danger exist, even at the relatively small scale of a warning region with an average size of just 200 km<sup>2</sup> (Schweizer et al., 2003); the discrete nature of the avalanche danger scale, where observers have to decide on one specific level, even if they consider the danger level to be somewhere in between two danger levels; and the fact that the avalanche danger scale as well as the process of locally assessing the danger level are not fully defined and can be interpreted differently.

These findings highlight the importance of regular training to ensure common standards. Furthermore, improved and more detailed guidelines on how to locally assess the avalanche danger would be helpful to increase consistency, as by combining a structured workflow as proposed in the Conceptual Model of Avalanche Hazard (CMAH), with lookup tables similar to the ones presented by Müller et al. (2016) or as in Figure 5.18 (p. 66). Furthermore, for situations when the danger is estimated between two levels, it should

be considered whether experienced observers can report intermediate danger levels. In particular, when observers report their local danger level estimate, they should always as well report other observations such as new snow depth, snow drifts, or signs of instability. These additional observations should allow validating the local nowcast. And finally, any reporting tool should guide the observer towards the final danger level estimate.

For forecasters, relying on local danger level estimates as a data-source used in day-to-day avalanche forecasting, it is important to treat these estimates like other data sources they analyze: errors make them all to some extent uncertain. Furthermore, scale issues must be considered as a scale mismatch exists between a local nowcast and a regional avalanche forecast - in both the temporal and the spatial scale (Jamieson et al., 2008). Again, this also applies to all other data used as well. And finally, the time an observer has been staying in the area and the location, from where the estimate was made, should be considered when interpreting the local estimate.

Despite these limitations, the major advantage of using  $D_{LN}$  estimates for verification is the fact that these provide a synthesized interpretation of many local observations, reported in the same unit as the forecast - the danger level, and which cannot be obtained in another way. And there is another good reason, why forecasters should ask experienced, and specifically trained observers with access to backcountry terrain to provide such estimates: they do not only provide their estimate as an observer but also as a user of the forecast. And this latter perspective, how the forecast is perceived by users, is at least as important as assessing the quality of the forecast (Gordon and Shaykewich, 2000).

## 6.2 A data-driven characterization of avalanche danger

**Publication 4** relied on observational data to characterize the key elements of avalanche danger. The research questions were: *How do the three elements - snowpack stability, the frequency distribution of snowpack stability and avalanche size - relate to the danger levels? And: Which combination of the actual value of the three elements does best describe the various danger levels?* Key results included the simulation of snowpack stability distributions, and four classes summarizing the frequency of potential avalanche triggering locations labeled *none or nearly none*, *a few*, *several* and *many*. This allowed a data-driven characterization describing the frequency of potential triggering locations and avalanche size for danger levels 1-Low to 4-High:

- 1-Low: *None or nearly none* locations with *very poor* stability existed. The largest avalanches were size 2.
- 2-Moderate: *A few* locations with *very poor* stability were present. The typical largest avalanche was of size 2.
- 3-Considerable: *Several* locations with *very poor* stability existed. The typical largest avalanches were sizes 2 or 3.
- 4-High: *Many* locations with *very poor* stability existed. The typical largest avalanche was of size 3.

		frequency			
		none	few	several	many
snowpack stability	very poor	(B-D)	C	B	A
	poor	(D)	D	C	B
	fair or good	-	-	D	D

		avalanche size		
		1	2 or 3	4
stability matrix	A	-	4 (3)	4
	B	3 (2, 1)	3 (2)	4, 3
	C	1, 2	2, 1	-
	D	1	1 (2)	-

Figure 6.3: Simplified data-driven lookup table for avalanche danger assessment (compare to Fig. 5.18). (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). - Workflow to use these two matrices: In the *stability matrix* (a), 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. The resulting letter is transferred to the *danger matrix* (b), where it is combined with the largest avalanche size. The most frequent danger levels that were typical for this combination, are shown.

The combination of snowpack stability and the frequency of the most unstable locations was highly relevant for danger level assessment. In general, avalanche size had a lesser influence on the danger level than 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 snowpack stability, should lead to an increase in danger level by one level. The presented data-driven lookup table (Fig. 5.18, p. 66) highlights that a greater focus must be placed on snowpack stability and the frequency distribution of snowpack stability, compared to avalanche size, when assessing avalanche danger. This was also shown by Clark (2019), who explored the combination of descriptive terms describing the three elements in the data behind the avalanche forecasts in Canada and their relation to the published danger level and avalanche problem. He showed that the 'likelihood of avalanches', which compares to the *stability matrix* (Fig. 5.18, p. 66), also had a greater impact on the resulting danger level than avalanche size, even though avalanche size  $\leq 1.5$  (considered harmless to people) was often the 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 elements and the distributions of the danger levels are similar.

Based on these data, a lookup table for danger level assessment was developed. The proposed lookup table can be simplified by removing or incorporating cells, which were supported by very little data, or joining cells which had a similar danger level distribution (Fig. 6.3). However, several challenges remain for the operational application of any such lookup table, namely the lack of data allowing an estimation of the values for each of the three elements. This is not a new problem, but simply more obvious now and does highlight issues that also arise when relying on other decision tools (like the EAWS-Matrix or ADAM; EAWS, 2017b; Müller et al., 2016) or the conceptual model of avalanche hazard (CMAH; Statham et al., 2018a). While a forecaster may have a reasonably good idea about what it might take to trigger an avalanche in the most unstable locations (snowpack stability), the estimation of the frequency of these locations is difficult. Sometimes, observational data, like the occurrence and frequency of signs of instability, may help

to assess this factor. Most often, however, it can only be assessed with laborious extensive sampling (e.g. Birkeland, 2001; Schweizer et al., 2003; Reuter et al., 2016). Particularly forecasters, who have to assess large areas or with limited access to avalanche terrain themselves, will have to rely increasingly on the use of physical snowpack modeling to estimate this parameter. As shown recently by Horton et al. (2020), snowpack modeling can indeed assist to assess the depth of potential weak layers (and thus obtain an indication of avalanche size), and the presence of such weak layers in the terrain.

## 6.3 Data sets and methods - a basis for further data-driven explorations

The outcome from this thesis is not just limited to the results, as for instance those presented in Chap. 5. Both the newly-compiled data sets and the statistical methods applied, although not novel by themselves, may pave the way towards a data-driven exploration of (observational) data related to public avalanche forecasting.

To gain a more profound understanding of factors influencing consistency and quality, or towards characterizing avalanche danger using a data-driven approach, the compiled data sets can be re-analyzed using alternative approaches, or may be expanded with similar data from other warning services. For instance, the data set compiled in publication 1 (Appendix Sect. A.2), contains not only information regarding the forecast danger level, but also on several characteristics describing each warning region (for instance by its location compared to a neighboring region or its location within the Alps, its size or highest elevation, the forecast center it belongs to, ...). This data set was analyzed using primarily a uni-variate approach, sometimes stratifying by a second variable. However, applying, for instance, a generalized linear model, may deepen our understanding on the factors influencing variations in the use of the danger levels.

Well-established statistical approaches were used to explore the data. While these were not novel by themselves, their application to the data at hand was innovative. Furthermore, they were appropriate for the purpose and provided plausible results. Here, of particular note are the use of a bootstrap-sampling approach to obtain a typical range of data distributions (Sect. 4.3), the application of a geometric progression of class widths resulting in data-driven thresholds defining frequency classes (Sect. 4.4), and to combine different pieces of information (Sect. 5.4.3). These methods, which are readily available in statistical software, permit the exploration of similar data in a comparable way.

## 6.4 Limitations

### 6.4.1 Data

Several issues arise considering the data used in this dissertation: the lack of an objective ground-truth, the danger of circular argumentation, and the lack of data for some of the danger levels.

**Lack of an objective ground-truth:** The assessment of the avalanche danger level, regardless whether after a day in the field, in an office-based setting, or when forecasting or re-assessing in a hindcast-setting,

always relies on the same approach of a human expert analyzing available data and making an expert decision to assign a danger level (e.g. Elder and Armstrong, 1987; Föhn and Schweizer, 1995). Furthermore, the most relevant pieces of information - class 1 data, which allow verifying the three elements of avalanche danger, are often sparse in time and space, leaving considerable uncertainty in the assessment of these elements, and hence when estimating the danger level.

**Circular argumentation or non-independence of data:** Avalanche forecasting is a continuous process (e.g. LaChapelle, 1980): a forecast becomes the prior in the verification process, the re-assessed conditions become the prior for the next forecast (see also Fig. 2.1, p. 14). While this is the way avalanche forecasting works, also because of the lack of an objective, measurable ground-truth, this can be problematic in data analysis. For instance, local nowcast danger level estimates will be based on observations made during the day. However, the forecast danger level will often be known to the local assessor as well. Thus, neither a forecast - nowcast comparison nor an observation-based characterization of avalanche danger is fully independent of each other, and some circularity in argumentation cannot be avoided. Similar applies to forecasters re-assessing the forecast (often their own) following study-plot observations: a confirmation bias may be present (e.g. McClung, 2002a; Jamieson et al., 2008). Furthermore, in some situations, no new class 1 data are available for re-assessment, as when an assessment is made early in the morning before new field observations being reported, or in case of rather favorable avalanche conditions when the number of observations reported by field observers and the public often decreased. In this case, an anchoring bias (e.g. McClung, 2002a) towards the forecast product may be present. Seen in this light, all of the findings can only be approximations and considerable uncertainty remains, as a re-assessed danger level is a best guess only with an error rate estimated by Föhn and Schweizer (1995) of about 10 to 20%. However, as these human biases and errors are a characteristic of the data, an effort was made to explore data from different countries with different snow climates, and possibly somewhat different perceptions on avalanche danger assessment.

And finally, in this thesis, it was not examined *how* a forecaster or local observer makes a decision regarding the danger level. Thus, with some exceptions - for instance, regarding the spatial and temporal resolution in the forecasts, the reasons behind variations in the use of the danger levels in the forecasts could not be explored. The findings allow primarily to point out situations *when* inconsistencies were noted. Gaining knowledge regarding the reasons behind such deficiencies will be necessary to purposefully improve forecaster and observer training, within and across forecast center boundaries, to harmonize the way avalanche danger is assessed and communicated.

And lastly, for some danger levels, there was a **lack of data**. For instance, very few stability tests were observed at 4-High, and data for danger level 5-Very High was generally lacking altogether.

## 6.4.2 Methods

Besides these limitations relating to the data, several issues arise from a methodological perspective.

In publications 1 to 3, preference was given to comparably easy-to-understand statistical metrics (often proportions). While these are appropriate statistical approaches used in categorical forecast verification (e.g.

Murphy, 1993; Wilks, 2011), these do not take into account agreement due to chance alone. With only two of the five danger levels being sufficient to describe avalanche conditions on about 80% of the days, already by randomly selecting a danger level according to the base rate frequencies, rather high rates of agreement can be achieved. This was shown, for instance, in publication 1 (see also Sect. 5.1, p. 44), where the agreement rate in the forecast danger levels between neighboring warning regions by randomly choosing a danger level was 40%. This compared to an observed agreement rate of 65% across warning service boundaries.

Particularly publication 4, which aimed at characterizing the elements of avalanche danger, relied strongly on statistical methods to derive snowpack stability distributions from the data. In this case, a bootstrap-sampling approach was used (see Sect. 4.3, p. 40 for details). While the sampling relied on a comparably large number of stability tests at 1-Low to 3-Considerable (between 700 and 2000 RB tests), the number of tests to draw from at 4-High was very low ( $N = 21$ ). Hence, both the stability distributions shown in Fig. 5.15 (p. 62) 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 *very poor* and *poor* stability increased, and those with *good* stability decreased at 4-High (Fig. 5.15), judging whether the observed tests reflect the population well is difficult. Unfortunately, a comparison to other studies that have explored the snowpack stability distribution in a region at 4-High based on many tests is not possible, as such studies are lacking. A second aspect regarding the bootstrap-sampling approach relates to the amount of overlap in the stability distributions drawn from stability tests observed at different danger levels (discussed in detail in publication 4). This overlap depends on the number  $n$  of samples drawn in each bootstrap. While a comparably low value of  $n$  was chosen ( $n = 25$ ), and while this sampling approach brought plausible distributions, it is unclear which sampling setting matches reality best. It must be supposed that a combination of labor-intensive field measurements combined with spatial modeling in a large variety of avalanche conditions will be necessary to shed some light on this question (e.g. Reuter et al., 2016, for a small basin in Switzerland).

Building on these data, the frequency of potential triggering locations was calculated for the four danger levels 1-Low to 4-High. 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 for 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 maximizing the distance between the class centers (e.g. Fisher-Jenks algorithm, Slocum et al., 2005). However, in a case, in which low values of the proportion of *very poor* stability are frequent and higher values rare, the use of a geometric progression of class widths can be considered most suitable for this type of distribution (Evans, 1977). Nonetheless, the data-driven class interval definition required to externally define the number of classes  $k$ . The selection of  $k$  was guided by the human capacity to distinguish between classes and by the number of classes used to describe and communicate avalanche danger 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). Furthermore, consideration was given to the requirement that the number of classes and their terms must be unambiguously understandable to the

user, regardless of language.

## 6.5 Communication of findings to a lay audience

One of the objectives of this dissertation was the provision of data-driven findings regarding the consistency and quality of public avalanche forecasts. In this regard, several important results were obtained. However, communicating the key findings which are potentially relevant to recreational forecast users or avalanche professionals in a comprehensible manner is challenging to achieve considering the complexity of the analysis and the uncertainty related to danger level assessments in general. This challenge, the communication of scientific findings to a broader audience is not only limited to snow and avalanche science but has been addressed in many contexts (e.g. the communication of climate change, Moser, 2010). Even though an attempt was made in this dissertation to rely on comparably simple statistical metrics, like the *proportion correct* to describe forecast accuracy, the key findings must be expressed using clear, simple metaphors and images or mental models to ease the cognitive understanding of an - often - lay audience, together with sufficiently strong recommendations on how to apply these (e.g. Gordon and Shaykewich, 2000; Moser, 2010).

Clearly, there is not *the* typical forecast user, as the forecasts are used by mountaineering amateurs planning recreational activities, but also by mountaineering professionals and avalanche professionals (e.g. Winkler and Techel, 2014; Engeset et al., 2018; St. Clair, 2019). Thus, what is considered a relevant finding, varies not only between these groups but depends also on their skill to interpret and apply the information provided in avalanche forecasts during the decision-making.

### 6.5.1 Findings relevant to mountaineering amateurs (recreational forecast users)

The skill to understand and apply different pieces of information provided in the forecasts in the decision-making process varies greatly between recreational forecast users. According to the avalanche bulletin user typology developed by St. Clair (2019), some users (user *A*) may comprehend a single aspect of the forecast, while others (user *B*) will grasp all the pieces of information provided in the forecast (St. Clair, 2019). This means that user *A* might, for instance, base a *go* or *no-go* decision solely on the forecast danger rating, while user *B* is able to incorporate additional information given in the forecast in the decision-making process (St. Clair, 2019).

In terms of the forecast danger level, the finding that the forecast danger level is on average correct on about six out of seven days is relevant. This means that a danger level and other information given in the forecast provides valuable information regarding the avalanche conditions which have to be expected on a tour. Thus, this information is useful during the planning stage of a tour, provided it is available already in the evening before an intended day recreating in the backcountry. However, the fact that forecast errors are an inherent characteristic of a forecast, which showed in the forecast danger level being wrong on about one out of seven days, clearly shows that a user should allow for a safety margin and consider alternatives or turn-around points already during the planning stage of a tour. Furthermore, this also means that if a user

recreates in avalanche terrain, a continuous re-assessment of avalanche conditions in the field is necessary. Further findings, potentially relevant to recreational forecast users, are related to the variation in the forecast danger level between immediately neighboring warning regions issued by different warning services, and to the differences in the use of danger level 4-High between forecast centers. The first point suggests that, when touring in a border region, the forecast products provided by the forecast centers on either side of the border should be consulted. The second point is relevant particularly for users traveling to other parts of the Alps, and thus to regions where the forecast is issued by a different warning service. Here, a user should be aware that danger level 4-High is used less restrictively in France and some parts of Italy, compared to the rest of the Alps. Therefore, a user accustomed to the forecasts in France or Italy should be extra cautious when incorporating the forecast danger level in the planning of a tour, as the seemingly lower danger level may invite to plan a riskier tour (Jamieson et al., 2009).

### **6.5.2 Findings relevant to mountaineering professionals**

Mountaineering professionals, as mountain guides, are high-end users of avalanche bulletins. Their skill level allows them to «extend the evaluation of bulletin information to a localized assessment of avalanche hazard» (St. Clair, 2019, p. 36). In a recent survey, Landrø et al. (2020a) explored factors and methods used by experts, mostly mountain guides and avalanche forecasters, to assess avalanche danger and to mitigate avalanche risk. Regarding the information provided in the avalanche forecast, an avalanche forecast is much more than the danger level to an expert; it is a source of information providing an optimal overview of the current avalanche situation (Landrø et al., 2020a).

As mountaineering professionals are highly skilled in making their own assessment in the field, the findings presented in this thesis may primarily be of importance in the sense of providing background information. For instance, the fact that variations in the use of the danger levels between warning services exist, particularly at the upper end of the scale, and that the spatial resolution of the forecasts varies, is relevant as these findings influence consistency and quality of the forecast products.

### **6.5.3 Findings relevant to other professional users**

In some countries, the information given in public forecasts is used during the decision-making process by risk-management authorities, as those responsible for the safety of public roads or inhabitants in settlements exposed to avalanches. Risk-mitigation measures will primarily need to be considered in times of increased avalanche danger, when large or very large natural avalanches have to be expected. These conditions correspond to danger level 4-High or higher.

For these forecast users, the comparably low success of forecasting danger level 4-High, with both misses and false-alarms being frequent, is a highly relevant finding. It clearly implies that during times of heightened avalanche conditions a local assessment must not be based exclusively on a regional avalanche forecast, but that a continuous local assessment is paramount.



# Chapter 7

## Conclusions and Outlook

The objectives of this dissertation were two-fold: (1) to obtain data-driven insights regarding consistency and quality in public avalanche forecasts, and (2) to describe the elements characterizing avalanche danger using observational data. These objectives were achieved by analyzing newly compiled data sets originating from different warning services and snow climates, collected for avalanche forecasting, and information published in avalanche forecasts. Applying well-established statistical approaches in an innovative way, and relying on comparably easy-to-communicate metrics, consistency and quality was explored as a function of avalanche conditions, but also by taking into consideration the operational constraints that limit the spatial and temporal resolution of regional avalanche danger communication.

The key achievements in this regard were:

- In publication 1, the spatially continuous forecasts from 23 forecast centers in the European Alps were compared in terms of spatial consistency and bias. Considerable differences in the operational constraints associated with forecast products were noted, when comparing the avalanche forecast products of different forecast centers. Most notably the spatial resolution of the warning regions underlying the forecasts had an impact on biases observed and the agreement rate but also limits at what spatial scale a regional danger level can be communicated in map products. Furthermore, considerable discrepancies in the use of danger level 4-High were detected. And finally, a comparably large proportion of forecasts with different danger levels across forecast center boundaries was noted. The magnitude of these variations was larger than the spatial variability observed in forecast avalanche conditions within forecast center boundaries. Reasons for these inconsistencies may be manifold, but can also be linked to deficiencies in the definition of avalanche danger and the (un)reliability associated with forecasts issued by human forecasters in general.
- Publication 2 focused on the quality of local danger level estimates as a data source for regional forecast verification. This study, originally based exclusively on Swiss data, was complemented with data from Norway. It showed that the agreement between individual estimates was relatively high, but decreased with distance. Additionally, sometimes an observer-specific reporting bias was noted. In a second step, publication 2 and the analysis shown in this Synthesis including data from Canada,

Colorado, and Norway, focused on the accuracy of the forecast danger level. The data showed that the proportion of forecasts, which matched the reference assessment, varied depending on the data used for verification (i.e. a local nowcast assessment or avalanche observations), the assessor providing this assessment and the number of assessors agreeing in their assessment. Whether errors in the reference assessment were considered, and how this was done, was important, as this impacted the observed accuracy of the forecast. For instance, assigning errors randomly decreased the observed performance of the forecast further, while considering the (un)reliability associated with the reference assessment as an upper boundary of observed accuracy provided more realistic values. Generally, a strong over-forecast bias was noted. Furthermore, the forecast accuracy for conditions representing 4-High was low with a tendency towards more misses and false alarms rather than successes.

- In publication 3, which relied exclusively on Swiss data, it was demonstrated that forecasters can forecast avalanche danger in greater detail. Forecast danger levels refined by sub-levels had skill, that is, they were better than a random assignment of sub-levels. This indicates that forecasters, at least when working in a similar setup as the national warning service in Switzerland, can indeed often refine avalanche danger at a higher resolution, but within the five ordinal danger levels. These findings may stimulate a discussion on optimizing the resolution of avalanche danger, last but not least for the internal assessment process and forecaster training, and as a data basis for computer-driven models. While this study confirmed the strong over-forecast bias addressed before, it was noted that, if a forecast danger level was wrong, it was generally by less than a «full» danger level compared to the reference assessment.
- In publication 4, the three key elements describing avalanche danger were characterized using observational data. This included the simulation of stability distributions and deriving four classes describing the frequency of potential avalanche triggering locations. The observed and simulated distributions of stability ratings derived from RB tests showed that locations with *very poor* stability are generally rare. Furthermore, the findings suggest that the three key elements did not distinguish equally prominently between the danger levels:
  - The proportion of *very poor* or *poor* stability test results increased from one danger level to the next higher one. Considering *very poor* snowpack stability and the frequency of this stability class alone, already distinguished well between danger levels.
  - Considering the largest observed avalanche size per day and warning region was most relevant to distinguish between 3-Considerable and 4-High. 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 elements.

In summary, the frequency of the most unfavorable snowpack stability class was the dominating discriminator. At higher danger levels the occurrence of size 4 avalanches discriminated danger level 3-Considerable from 4-High. This shift in importance between elements is poorly represented in existing decision aids, as well as in the European Avalanche Danger Scale.

To combine the three elements and to derive avalanche danger, a data-driven lookup table consisting of two matrices was developed, which can be used to assess the 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.

- Publication 5 focused on the development of a higher-resolved stability interpretation scheme for the Extended Column Test (ECT) than was in use until now. The data confirmed the well-known fact that crack propagation propensity, as observed with the ECT, is a key indicator relating to snow instability. The number of taps required to initiate a crack provided additional information concerning snow instability. Combining crack propagation propensity and the number of taps required to initiate a failure allowed refining the original binary stability classification. Based on these findings, an ECT stability interpretation scheme with four distinctly different stability classes was proposed. This classification increased the agreement between slope stability and test result for the lowest (*poor*) and highest (*good*) stability classes compared to previous classification approaches. However, a comparison with Rutschblock tests (RB), performed in the same snow pit, showed that the RB correlated better with slope stability than the ECT, regardless whether one or two tests were performed.

## 7.1 Improving consistency and quality in avalanche forecasts - challenges and possible ways forward

Data-driven insights regarding the consistency and quality of public avalanche forecasts were gained. These insights highlighted some deficiencies in the way avalanche hazard is assessed and communicated, opening up new directions for research, but also the refinement or development of practical guidelines, recommendations, or definitions:

- **Forecasting the most critical days each winter.** It was noted that situations characterizing 4-High - the few most dangerous days each winter - were less often correctly forecast than they were missed or wrongly forecast. Why? Do forecasters tend to wrongly assess the probability that natural avalanches will occur, and their number and size? Is this due to a lack of relevant and trustworthy data allowing to judge these criteria? Or do externalities such as the consequences of a danger level 4-High for users, and the perception of forecasters of this impact, also influence the forecasters' judgment? Reliable and interpretable data in sufficient spatial and temporal resolution allowing to assess the probability of natural avalanche occurrence and the potential size of avalanches will be necessary to provide increasingly data-based and more objective forecasts. If such data were available, the uncertainty related to forecasting these days would reduce, and the magnitude of the forecast error and bias would likely decrease.
- **Revisit the definitions describing avalanche danger and improve the workflow to assign a danger level.** The data-driven characterization of avalanche danger provides an important step towards a refined characterization of the elements describing the danger levels. Supplementing and comparing

these findings with data from other countries would be beneficial to reduce any potential bias related to a Swiss perception of avalanche danger. Such data-based characterizations of avalanche danger permit already a more objective revision of the danger level definitions. Furthermore, this characterization allows the description of key elements contributing to avalanche danger, and hence, is a step towards the inclusion of the actual assignment of a danger level in a standardized workflow (e.g. as proposed by Müller et al., 2016). However, this requires the provision of understandable and applicable criteria to assign classes to data, and - of course - data that are relevant and interpretable for the task.

- **Forecast verification.** The findings presented in this thesis have shown that the accuracy of a regional forecast danger level was almost similar to the reliability associated with a nowcast danger level, at the spatial scale of the forecast. In both cases, the spatial and temporal resolution of forecast or nowcast, and the reliability of a human expert judging the avalanche danger are influencing factors. Furthermore, the reliability of both the forecast and the local nowcast decreased with increasing danger level. Statistical models, for instance, using Bayesian approaches, could integrate such knowledge to assist office-based avalanche forecasters, who have access to this kind of data, in the verification of the avalanche conditions. Such a tool would allow testing the forecaster's experience-based hypothesis regarding the avalanche conditions using the model as a second opinion (e.g. Purves et al., 2003). Undoubtedly, the process of forecast verification must be as systematic as the forecast, requiring clear definitions regarding the spatial and temporal resolution of the verified forecast. In that respect, guidelines and policies on avalanche forecast verification should be developed. These should not only describe suitable approaches to re-assess the avalanche danger level, but also regarding the verification of other information provided in the forecasts as, for instance, the aspects and elevations, where the danger prevails, or the avalanche problems.
- **Resolution of regional avalanche danger assessment.** Given an increasingly higher demand in the spatial and temporal resolution of forecast products, combined with advances in computer modeling (i.e. grid-based weather forecasts and snowpack models), the time might be ripe to discuss whether avalanche danger can be assessed and forecast at greater detail, at least during the forecast process. Such refinements could include, for instance, the use of intermediate danger levels or by assigning probabilities to danger levels, by increasing the spatial and temporal resolution, but also by keeping track of the key elements of avalanche danger, namely the expected stability of the most unstable locations and their frequency, as well as avalanche size. Making these refinements on the forecaster side should be guided by data availability. Only if these refinements can be made in a sufficiently consistent and accurate manner, should the discussion be initiated whether, and in which form, this information will be useful to the public.
- **Value of avalanche forecasts.** Though only touched on very briefly in this Synthesis, the value of avalanche forecasts to users with different skill levels, training, and background are of great relevance. Recently, this important topic has gained increased attention (e.g. St. Clair, 2019; Finn, 2020). Questions to be answered may include the following: How can the complex information provided in avalanche forecasts be presented in a more efficient way to users of different skill levels? What are

the most relevant pieces of information assisting in the various stages of the decision-making process? And how can these be provided to high-end users, and users with a comparably low level of knowledge? Addressing these questions should involve all user groups of avalanche forecasts equally as targeted improvements in forecast communication should not only focus on the high-end professional users but also the majority of users with a comparably lower level of knowledge. Furthermore, cross-border evaluations may be necessary as users travel across forecast borders.

- **Statistical prediction of avalanche danger.** The development of statistical models, which rely on supervised approaches, require historical data to train, validate and test the model. However, errors and bias in the reference class will influence the observed performance of the model (errors), or lead to biased predictions reflecting the bias in the historical data. The quantitative findings presented regarding the forecast quality and bias, but also regarding the reliability of nowcast estimates of avalanche danger may be combined to form an improved data set for training and testing such models.

In a nutshell, to increase consistency and quality of public avalanche forecasts, timely, relevant, and reliable class 1 data are required permitting not only more objective, data-based nowcast assessments, but also more accurate forecasts. Furthermore, actionable recommendations, as well as improved guidelines and definitions, on how avalanche danger is assessed and communicated, are required. And these must not only focus on forecaster needs but also on the requirements of forecast users.



# Appendix A

## Publications

### A.1 Publications and author contributions

The five research papers, which are part of this dissertation, are appended to the following Sections A.2 - A.6.

Below, the research papers are listed and the respective author contributions are shown:

- 1. Spatial consistency and bias in avalanche forecasts - a case study in the European Alps** (Sect. A.2, p. 93ff)  
*Frank Techel, Christoph Mitterer, Elisabetta Ceaglio, Cécile Coléou, Samuel Morin, Francesca Rastelli, and Ross S. Purves*, published 2018 in Natural Hazards Earth System Sciences  
FT coordinated and designed this collaborative study. FT collected the data from the avalanche warning services in the Alps and conducted the analysis. All co-authors repeatedly provided in-depth feedback regarding previous versions of the manuscript.
- 2. On using local avalanche danger level estimates for forecast verification** (Sect. A.3, p. 122ff)  
*Frank Techel and Jürg Schweizer*, published 2017 in Cold Regions Science and Technology  
FT designed the study, performed data extraction and analysis, and wrote the manuscript. JS provided repeated in-depth feedback on the manuscript.
- 3. Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random?** (Sect. A.4, p. 142ff)  
*Frank Techel, Christine Pielmeier, Kurt Winkler*, published 2020 in Cold Regions Science and Technology  
FT designed the study, conducted the analysis and wrote the manuscript. CP and KW repeatedly provided in-depth feedback on methodology and subsequent versions of the manuscript.
- 4. On the importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level** (Sect. A.5, p. 160ff)  
*Frank Techel, Karsten Müller, Jürg Schweizer*, published 2020 in The Cryosphere  
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.

5. **On snow stability interpretation of Extended Column Test results** (Sect. A.6, p. 186ff)

*Frank Techel, Kurt Winkler, Matthias Walcher, Alec van Herwijnen, Jürg Schweizer*, published 2020 in Natural Hazards Earth System Sciences

FT designed the study, extracted and analyzed the data, and wrote the manuscript. MW extracted and classified a large part of the text from the snow profiles. KW, AvH and JS provided in-depth feedback on study design, interpretation of the results and manuscript.

## A.2 Spatial consistency and bias in avalanche forecasts - a case study in the European Alps

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.*, 2018, 18, 2697-2716, doi: 10.5194/nhess-18-2697-2018

### Abstract

In the European Alps, the public is provided with regional avalanche forecasts, issued by about 30 forecast centers throughout the winter, covering a spatially contiguous area. A key element in these forecasts is the communication of avalanche danger according to the five-level, ordinal European Avalanche Danger Scale (EADS). Consistency in the application of the avalanche danger levels by the individual forecast centers is essential to avoid misunderstandings or misinterpretations by users, particularly those utilizing bulletins issued by different forecast centers. As the quality of avalanche forecasts is difficult to verify, due to the categorical nature of the EADS, we investigated forecast goodness by focusing on spatial consistency and bias exploring real forecast danger levels from four winter seasons (477 forecast days). We describe the operational constraints associated with the production and communication of the avalanche bulletins, and we propose a methodology to quantitatively explore spatial consistency and bias. We note that the forecast danger level agreed significantly less often when compared across national and forecast center boundaries (about 60%), as compared to within forecast center boundaries (about 90%). Furthermore, several forecast centers showed significant systematic differences towards using more frequently lower (or higher) danger levels than their neighbors. Discrepancies seemed to be greatest when analyzing the proportion of forecasts with danger level 4-High and 5-Very High. The size of the warning regions, the smallest geographically clearly specified areas underlying the forecast products, differed considerably between forecast centers. Region size also had a significant impact on all summary statistics and is a key parameter influencing the issued danger level, but also limits the communication of spatial variations in the danger level. Operational constraints in the production and communication of avalanche forecasts and variation in the ways the EADS is interpreted locally may contribute to inconsistencies, and may be potential sources for misinterpretation by forecast users. All these issues highlight the need to further harmonize the forecast production process and the way avalanche hazard is communicated to increase consistency, and hence facilitate cross-border forecast interpretation by traveling users.

### A.2.1 Introduction

In the European Alps, public forecasts of avalanche hazard are provided throughout the winter. These forecasts - also called advisories, warnings, or bulletins<sup>1</sup> - provide information about the current and forecast snow and avalanche conditions in a specific region. In contrast to local avalanche forecasting, e.g. for a transportation corridor or ski area, a regional forecast does not provide information regarding individual

---

<sup>1</sup>we use these terms synonymously

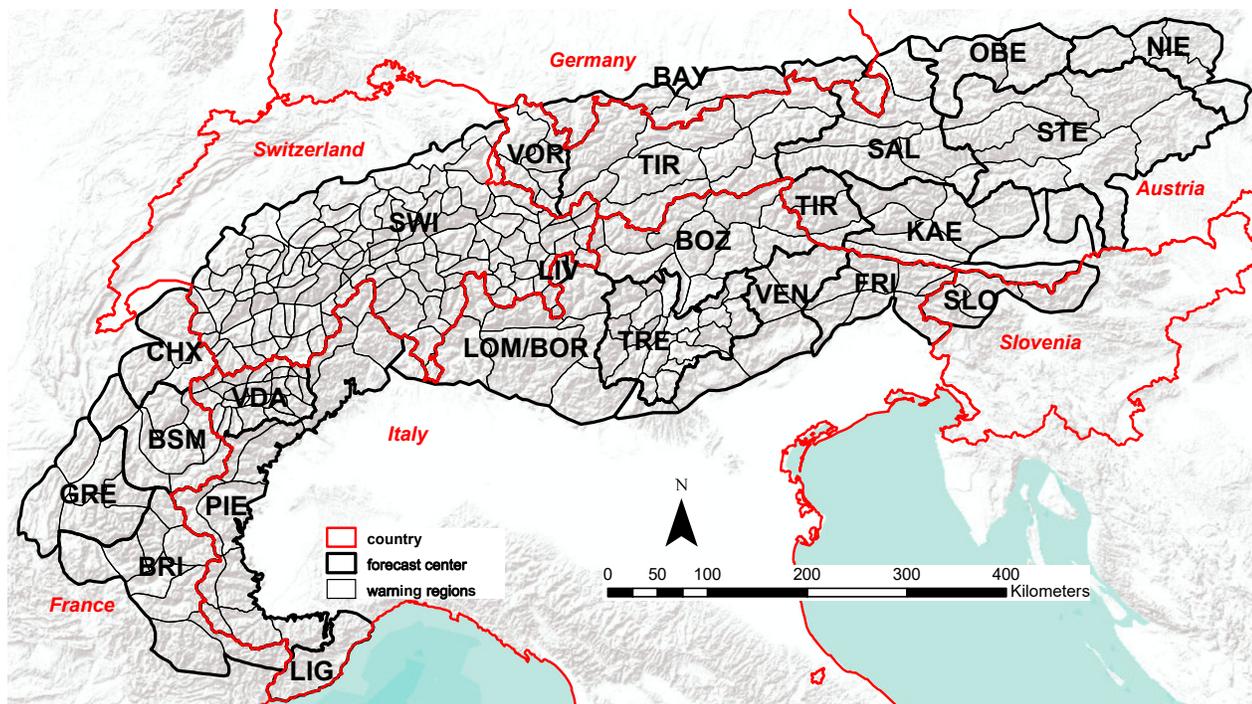


Figure A.1: Map showing the relief of the European Alps (gray shaded background) with the outlines of the individual forecast centers (bold black polygons, three-letter abbreviations) and the warning regions, the smallest geographically defined regions, used in the respective avalanche forecasts (black polygons). The borders of the Alpine countries are marked red. In the Italian Alps, where two avalanche warning services provide forecasts (Associazione Interregionale Neve e Valanghe (AINEVA) and Meteomont Carabinieri), the warning regions generally follow AINEVA. An exception is LIG (avalanche warning service Meteomont Carabinieri). The forecast domains of LOM (AINEVA) and BOR (Meteomont Carabinieri) are identical, however, the three warning regions for BOR are not shown on the map. The forecast domain LIV is superposed onto parts of LOM/BOR (map source: ESRI, 2017). Note that the map captures the situation and partitioning during the period under study.

slopes or specific endangered objects.

One of the key consumer groups are those undertaking recreational activities, such as off-piste riding and backcountry touring in unsecured terrain. The importance of clearly communicating to this group is underlined firstly by avalanche accident statistics - with on average 100 fatalities each winter in the Alps (Techel et al., 2016b), most of whom died during recreational activities. Secondly, very large numbers of individuals recreate in unsecured winter terrain, with for example Winkler et al. (2016) reporting that more than two million winter backcountry touring days were undertaken in 2013 in Switzerland alone. An additional consumer group are local, regional and national risk management authorities, who base risk reduction strategies such as avalanche control measures, road closures, evacuation procedures etc. in part on information provided in regional avalanche forecasts.

In all Alpine countries (Fig. A.1), forecasts are disseminated throughout the entire winter, for individual warning regions, together forming a spatially contiguous area covering the entire Alpine region. Furthermore, in all of these countries the European Avalanche Danger Scale (EADS; EAWS, 2018), introduced in 1993 (SLF, 1993), is used in the production and communication of forecasts (EAWS, 2017d).

Table A.1: European avalanche danger scale (EAWS, 2018).

Danger level	Snowpack stability	Likelihood of triggering
5-Very High	The snowpack is poorly bonded and largely unstable in general.	Numerous very large and often extremely large natural avalanches can be expected, even in moderately steep terrain <sup>*</sup> .
4-High	The snowpack is poorly bonded on most steep slopes <sup>*</sup> .	Triggering is likely even by low additional loads <sup>**</sup> on many steep slopes <sup>*</sup> . In some cases, numerous large and often very large natural avalanches can be expected.
3-Considerable	The snowpack is moderately to poorly bonded on many steep slopes <sup>*</sup> .	Triggering is possible even from low additional loads <sup>**</sup> particularly on the indicated steep slopes <sup>*</sup> . In certain situations some large, in isolated cases very large natural avalanches are possible.
2-Moderate	The snowpack is only moderately well bonded on some steep slopes <sup>*</sup> ; otherwise well bonded in general.	Triggering is possible primarily from high additional loads <sup>**</sup> , particularly on the indicated steep slopes <sup>*</sup> . Very large natural avalanches are unlikely.
1-Low	The snowpack is well bonded and stable in general.	Triggering is generally possible only from high additional loads <sup>**</sup> in isolated areas of very steep, extreme terrain <sup>**</sup> . Only small and medium-sized natural avalanches are possible.

<sup>\*</sup> The avalanche-prone locations are described in greater detail in the avalanche bulletin (altitude, slope aspect, type of terrain): moderately steep terrain: slopes shallower than about 30 degrees; steep slopes: slopes steeper than about 30 degrees

very steep, extreme terrain: particularly adverse terrain related to slope angle (more than about 40 degrees), terrain profile, proximity to ridge, smoothness of underlying ground surface

<sup>\*\*</sup> Additional loads:

low: individual skier / snowboarder, riding softly, not falling; snowshoer; group with good spacing (minimum 10m) keeping distances

high: two or more skiers / snowboarders etc. without good spacing (or without intervals); snowmachine; explosives

natural: without human influence

The EADS is an ordinal, five-level scale, focusing on avalanche hazard, with categorical descriptions for each danger level describing snowpack (in)stability, avalanche release probability, expected size and number of avalanches and the likely distribution of triggering spots (Tab. A.1). The EADS describes situations with spontaneous avalanches but also conditions where an additional load - such as a person skiing a slope - can trigger an avalanche. These categorical descriptions of each danger level aim to inform users on the nature of avalanche hazard at hand. However, individual danger levels capture a wide range of differing avalanche conditions (e.g. EAWS, 2005; Lazar et al., 2016; EAWS, 2017b; Statham et al., 2018b), and therefore, in isolation, are too basic to be used as a stand-alone decision making tool (e.g. Météo France, 2012). Additionally, and in order to describe the avalanche hazard in more detail and to provide better advice to the end users on how to manage these hazards, the EAWS introduced a set of five typical avalanche

problems (EAWS, 2017a). Nonetheless, the EADS provides a consistent way of communicating avalanche hazard. Furthermore, the EADS often forms an important input into basic avalanche education on planning, or decision making heuristics as practiced by many recreationists (e.g. Munter, 1997).

However, the EADS is not only a means of communicating to forecast users. It also impacts on the forecasting process itself, as all forecasters are working to an agreed, common, and at least nominally binding, definition of avalanche hazard.

Forecast validation and evaluation is not only a problem in avalanche forecasting, but more generally in forecasting. Murphy (1993), in his classic paper on the nature of a good (weather) forecast, discussed three key elements which he termed *consistency*, *quality* and *value*. Consistency in Murphy's model essentially captures the degree of agreement between a forecaster's understanding of a situation and the forecast they then communicate to the public. Quality captures the degree of agreement between a forecast and the events which occur, and value the benefits or costs incurred by a user as a result of a forecast.

In avalanche forecasting, two key problems come to the fore. Firstly, the target variable is essentially categorical, since although the EADS is an ordinal scale, a real evaluation of a forecast would compare the forecast danger level, qualitatively defined in the EADS, with the prevailing avalanche situation. Secondly, since the target variable captures a state which may or may not lead to an (avalanche) event, verification of forecast quality is only possible in some circumstances and for some aspects of the EADS, for example:

- At higher danger levels, the occurrence of natural avalanches can sometimes be used to verify the danger level (e.g. Elder and Armstrong, 1987; Giraud et al., 1987; Schweizer et al., 2018).
- At lower danger levels, the occurrence of avalanches triggered by recreationists or the observation of signs of instability requires users being present.
- Since the absence of avalanche activity is not alone an indicator of stability, verifying associated danger levels is only possible through digging multiple snow profiles and performing stability tests (Schweizer et al., 2003).

Thus, avalanche danger cannot be fully measured or validated (Föhn and Schweizer, 1995). This in turn means that, at least at the level of the EADS, it is conceptually difficult to directly measure forecast quality. However, Murphy's notion of considering goodness of forecasts in terms of not only their quality, but also consistency and value, suggests a possible way forward.

Although Murphy defines consistency with respect to an individual forecaster, we believe that the concept can be extended to forecast centers, in terms of the degree to which individual forecasters using potentially different evidence reach the same judgment (LaChapelle, 1980), and across forecast centers, in terms of the uniformity of the forecast issued by different forecast centers in neighboring regions. This reading of consistency is, we believe, both true to Murphy's notion (how reliably does a forecast correspond with a forecaster's best judgment) and broader notions of consistency stemming from work on data quality and information science (Ballou and Pazer, 2003; Bovee et al., 2003).

Inconsistencies in the use of the danger levels between neighboring warning regions and forecast centers may be a potential source of misinterpretations to users traveling from one region to another, unless these differences are only due to avalanche conditions. The main goal of this study is therefore to investigate if

such spatial inconsistencies and biases exist. We do so by quantifying bias between neighboring forecast centers and warning regions in time and space. While we do not expect spatial homogeneity, a stronger bias and a lower agreement rate in neighboring warning regions in different forecast centers, compared to within forecast domains, may indicate such inconsistencies. To do so, we first describe the operational constraints under which avalanche forecasts are produced and communicated. Then, we present methods appropriate to explore spatial consistency and bias in the use of EADS given the operational constraints described above. We address the following three research questions:

1. Does bias between forecast centers exist?
2. Can operational constraints (such as the size of the warning regions) or the elevation of warning regions explain these differences?
3. What implications do the biases identified have for users of avalanche forecasts?

## **A.2.2 Background and definitions**

In the following, we introduce the most important standards, concepts and definitions used in avalanche forecast products in the European Alps. We describe the situation during the winters 2011/2012 until 2014/2015, as these are the years we explore quantitatively in this study.

### **A.2.2.1 Avalanche warning services and forecast centers**

Avalanche warning services (AWS) are national, regional or provincial agencies in charge of providing publicly available forecasts of avalanche hazard (EAWS, 2017d). AWS also have voting rights at the General Assembly of the European AWS (EAWS). An AWS may either be a service with a single forecast center (e.g. the national service in Switzerland or the regional AWS of the federal states in Austria) or with several forecast centers in different locations (e.g. the AWS Météo-France in France with four forecast centers in the Alps or the two AWS in Italy (Associazione Interregionale Neve e Valanghe (AINEVA) and Meteomont Carabinieri) with their provincial and regional centers.

Generally, and with the exception of Italy, a single forecast covering a (number of) warning region(s) is issued by the respective forecast center (Tab. A.2, Fig. A.1). In the case of Italy, forecast centers belonging to AINEVA and Meteomont Carabinieri independently provide forecasts covering the same Alpine regions, while in Livigno (LIV in Fig. A.1) a regional forecast is also issued by the municipality. Even though the forecast products provided by the individual forecast centers may differ in their structure, we assume they adhere to the principles defined by the European Avalanche Warning Services (EAWS, 2017d).

### **A.2.2.2 Avalanche forecasts**

Avalanche forecasts are the primary means for avalanche warning services to provide publicly available information about current and forecast snow and avalanche conditions in their territory. They may take the form of a single advisory, describing the current situation, or an advisory and forecast for one or more days.

Typically, avalanche forecasts contain the following information, ranked according to importance (*information pyramid*; EAWS, 2017c):

- avalanche danger level according to the EADS (Table A.1),
- most exposed terrain - defining the terrain where the danger is particularly significant,
- typical avalanche problems - describing typical situations encountered in avalanche terrain (EAWS, 2017a),
- hazard description - a text description providing information concerning the avalanche situation,
- information concerning snowpack and weather.

In this study, we exclusively explore the forecast regional avalanche danger level. However, we also describe how the danger level is communicated in relation to the most exposed terrain (by elevation) and to its temporal evolution during the day, as this differs between forecast centers and could influence the results.

### A.2.2.3 Temporal validity and publication frequency

The issuing time, temporal validity and publication frequency of the forecasts varies between forecast centers. For the explored four winters, these can roughly be summarized in five groups (the «normal» cases are described, exceptions exist; see also Fig. A.2):

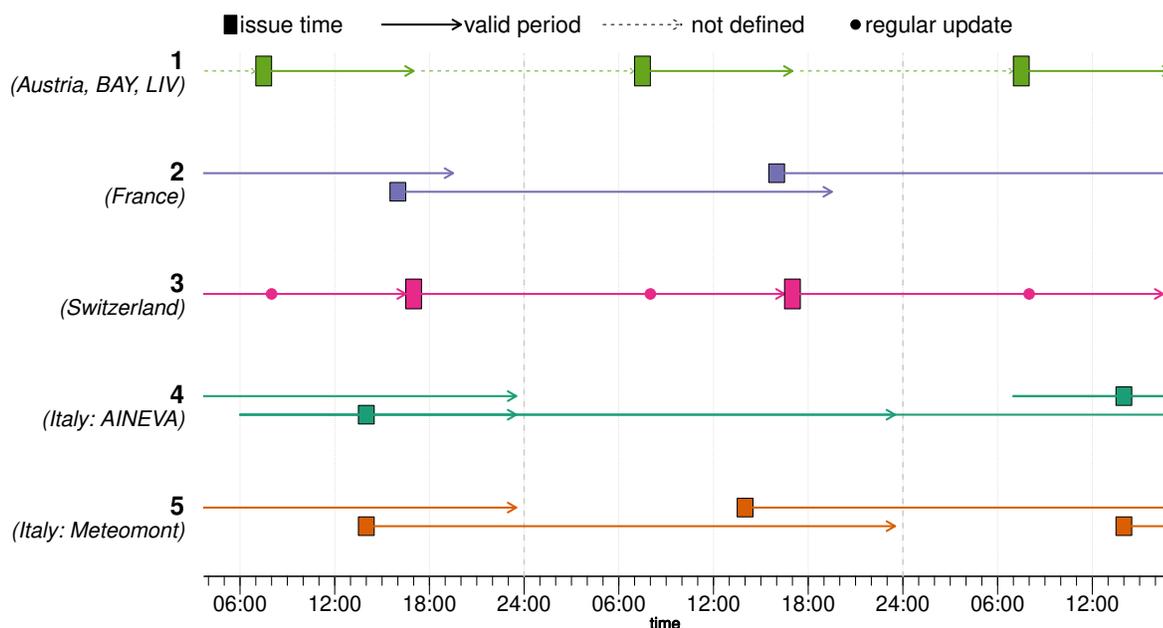


Figure A.2: Schematic summary of the different bulletin publication frequencies, issuing times and periods of validity. In special circumstances, updates during the morning were possible in most forecast centers. Particularly for Italy (AINEVA), it is of note that the exact publication times, valid periods and publication frequencies may differ between forecast centers, but changes may also have been introduced from one season to the next. Forecast centers are labeled according to Table A.2.

1. Bulletins are published daily in the morning (generally around 07:30 CET<sup>2</sup>) and are normally valid for the day of publication (typical for bulletins in Austria, Germany and Livigno (LIV/Italy)).
2. Forecasts are published daily in the afternoon (16:00 CET) and are valid until the following day (France).
3. During the main winter season (often from early December until after Easter), forecasts are published twice daily. The main forecast, published at 17:00 CET valid until 17:00 CET the following day, is replaced by an update the following morning at 08:00 CET (Switzerland).
4. Bulletins are published several times a week (at least on Monday, Wednesday, Friday). Bulletins are issued between 11:00 and 17:00 CET and describe the avalanche conditions on the day of publication, the following day and the day after (typically forecast centers belonging to AWS AINEVA). In more recent years, publication frequency increased towards daily.
5. Bulletins are published at 14:00 CET, describing the current situation and the forecast for the next day(s). Forecasts are published daily, except on public holidays (AWS Meteomont Carabinieri).

Most of the forecast centers can update their forecast product when conditions change significantly.

#### A.2.2.4 Warning regions

Warning regions are geographically clearly specified areas permitting the forecast user to know exactly which region is covered by the forecast. They may be delineated by administrative boundaries (e.g. between countries, federal states, or regions and provinces), describe climatologically (e.g. in France; Pahaut and Bolognesi (2003)), hydrologically or meteorologically homogeneous regions, or may be based on orographic divisions (e.g. Italy; Marazzi, 2005), or a combination of these (e.g. Valle d'Aosta (VDA); Burelli et al., 2012). Generally, warning regions are larger than the minimal spatial resolution of a regionally forecast avalanche danger level, and are therefore recommended to have a size of about 100 km<sup>2</sup> or larger (EAWS, 2017d).

The median size of the warning regions is 350 km<sup>2</sup> with considerable variations (Fig. A.3a, Tab. A.2). The 25% of the smallest warning regions (size < 160 km<sup>2</sup>, all located in Switzerland (SWI), Trentino (TRE) and Valle d'Aosta (VDA)) are almost ten times smaller than the 10% of the largest regions (size > 1310 km<sup>2</sup>). Particularly large spatial units are used by the forecast centers covering the region of Lombardia (BOR) and the Ligurian Alps (LIG, both AWS Meteomont Carabinieri, Italy) and in Oberösterreich (OBE, Austria; size > 1900 km<sup>2</sup>, Table A.2).

The size of the warning regions depends on the approach used by an AWS to define the warning regions and to externally communicate avalanche danger. In its simplest case (see variations introduced in next section), a single danger level is either explicitly communicated for each warning region (e.g. in Austria, France, Germany, often in Italy) or may be communicated for an aggregation of warning regions (Switzerland (SWI), Trentino (TRE) and Valle d'Aosta (VDA)).

---

<sup>2</sup>all times indicated may refer to either CET or CEST

Table A.2: Overview of the forecast centers considered in this study. Italian forecast centers refer to AINENA, except those indicated with subscript <sup>MC</sup> for Meteomont Carabinieri. Forecast centers and warning regions outside the Alps are not shown. Three-letter abbreviations indicate forecast centers. For countries, we use English names, for forecast centers the names in their original language.

country	forecast center	abbreviation	surface area*	number of	size**	max. elevation***
			in km <sup>2</sup>	warning regions	median (min-max) in km <sup>2</sup>	min-max in m.a.s.l.
Austria	Kärnten	KAE	7700	8	1060 (520 - 1300)	2110 - 3740
	Niederösterreich	NIE	3700	5	730 (500 - 1030)	1390 - 2060
	Oberösterreich	OBE	3400	2	1720 (1530 - 1910)	2360 - 2860
	Salzburg	SAL	6800	6	1090 (360 - 1970)	2010 - 3570
	Steiermark	STE	12500	7	2030 (1250 - 2290)	1770 - 2800
	Tirol	TIR	12600	12	980 (380 - 1920)	2460 - 3730
	Vorarlberg	VOR	2600	6	390 (180 - 880)	2080 - 3200
	Switzerland	Schweiz	SWI	26300	117	180 (40 - 660)
Germany	Bayern	BAY	4300	6	660 (450 - 1190)	1870 - 2940
	France	Bourg-St-Maurice	BSM	5100	6	810 (630 - 1220)
Italy	Briançon	BRI	8000	9	840 (450 - 1590)	2760 - 4020
	Chamonix	CHX	3000	3	1070 (580 - 1380)	2700 - 4780
	Grenoble	GRE	5300	5	990 (560 - 1440)	2070 - 3950
	Bozen-Südtirol / Bolzano-Alto Adige	BOZ	7400	11	650 (180 - 1110)	2590 - 3860
	Friuli Venezia Giulia	FRI	3700	7	560 (160 - 690)	1880 - 2740
	Liguria and Toscana <sup>MC</sup>	LIG	2100	1	2060	2140
	Livigno	LIV	200	1	210	3210
	Lombardia	LOM	9700	7	1330 (510 - 2820)	2230 - 3940
Lombardia <sup>MC</sup>	Lombardia	LOM	9700	7	1330 (510 - 2820)	2230 - 3940
	Lombardia <sup>MC</sup>	BOR	9700	3	3120 (1900-4630)	2850 - 3940
	Piemonte	PIE	10300	13	820 (270 - 1630)	2580 - 4530
	Trentino	TRE	6200	21	290 (120 - 540)	2060 - 3620
	Valle d'Aosta	VDA	3300	26	130 (25 - 280)	2620 - 4780
	Veneto	VEN	5500	5	1100 (460 - 1640)	2180 - 3250

\* rounded to nearest 100 km<sup>2</sup>, \*\* rounded to nearest 10 km<sup>2</sup>, \*\*\* rounded to nearest 10 m

The size, as shown here and in Figure A.3a, was calculated using the R-package *raster* (Hijmans, 2016).

The range of the maximum elevations describes the range of the highest elevation calculated using a digital elevation model with 90×90 m cell resolution (Jarvis et al., 2008; SRTM, 2017) per warning region and forecast center (Figure A.3b).

### A.2.2.5 Concepts to communicate temporal changes and elevational gradients in danger level

The communication of the most exposed elevations and slopes, and expected temporal changes are important information provided in avalanche forecasts.

#### Temporal differences in danger rating within forecast period

All forecast centers communicate significant changes (increasing or decreasing danger level) during the valid period of a forecast. In most cases, this is done graphically using either icons or two maps, and only rarely using text.

In cases, when two danger levels are indicated, the first time-step often refers to the avalanche danger in the morning, the second time-step indicates a significant change during the day. Changing danger ratings may refer to either changes in dry- or wet-snow avalanche hazard, or from dry- to wet-snow (or vice versa). However, exceptions to these generalizations exist: In France, but occasionally in forecasts of other forecast centers too, the two time-steps may refer to either day and night, morning or afternoon, or before and after a snowfall. Switzerland is the only warning service where an increase in danger rating for wet-snow situations (typically in spring conditions) is presented using a map product if the wet-snow rating is higher than the dry-snow rating in the morning, but an increase in dry-snow avalanche hazard during the day is exclusively conveyed in text form within the danger description.

#### Elevational differences in danger rating

All forecast centers provide information concerning the most exposed elevations, often in graphical form using icons. The elevational threshold indicated in the bulletin may relate to a difference in danger rating (for instance higher above a certain elevation), or differences in the avalanche problem and the most likely type of avalanche expected (e.g. wet-snow avalanches below and dry-snow avalanches above the indicated elevation), or a combination thereof.

The forecast centers use three different ways to communicate elevational differences in the danger rating. In Switzerland and Italy, the danger rating refers to the most exposed elevations, with no indication of the (lower) danger rating in other elevations. In France, Germany and some regions in Austria, two separate danger ratings are often provided: one above a certain elevation level, and one below, while the forecast center Livigno (LIV) in northern Italy assigns a danger rating to the three elevation bands *below tree line*, *tree line* and *alpine* (as done in North American avalanche forecasts).

### A.2.3 Data

We approached all the warning services in the Alps concerning the forecast danger level for each warning region and day for the four years from 2011/12 to 2014/15 and received data from 23 of the 30 forecast centers.

In most cases, data were provided directly from the warning services or forecast centers. Exceptions were

- Kärnten (KAE, Austria) - Data were extracted from the annual reports ÖLWD (2012)-ÖLWD (2015)
- Bayern (BAY, Germany) - Data were collected from the web archive of the Bavarian warning service

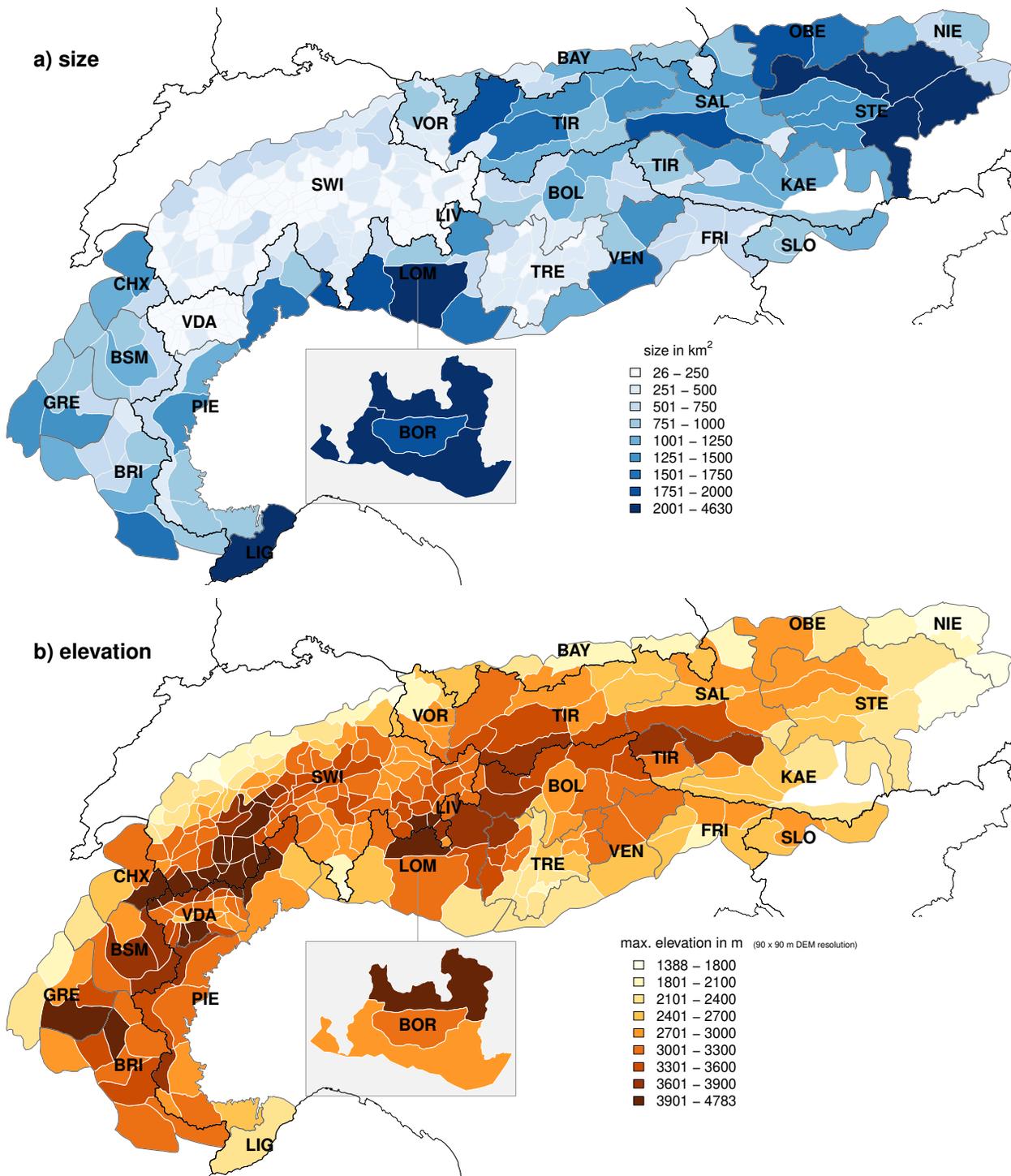


Figure A.3: Map showing the European Alps with the individual warning regions (white polygon outlines) and (a) their size (color shading of polygon) and (b) their maximum elevation (color shading of polygon). Additionally, national (black lines) and forecast center boundaries (grey polygon outlines) are shown. To visualize the (at least partially) overlapping forecast regions in the Italian region of Lombardia, LIV is superposed onto parts of LOM, while BOR is placed as inset to the south of LOM. Forecast centers are labeled according to Table A.2.

- AINEVA forecast centers Friuli Venezia Giulia (FRI), Lombardia (LOM), Veneto (VEN) in Italy - Data were provided by M. Valt/VEN (extracted from the central AINEVA database).

Table A.3: Overview of the data used in this study. Forecast centers are summarized according to data source, format and content.  $D_{t1}$  and  $D_{t2}$  - danger level time step 1 and 2, respectively;  $D_e$  - concept of elevational danger ratings;  $D_{spatial}$  - more than one rating per warning region referring to spatial differences. Danger levels may refer to the day of publication (day+0), the following day (day+1) or the day after (day+2). Forecast centers are labeled according to Table A.2.

country	forecast center	$D_{t1}$	$D_{t2}$	$D_e$	$D_{spatial}$	day+0	day+1	day+2	source
Austria	KAE	no	yes	2	no	100%	–	–	ÖLWD
	NIE, OBE, SAL, STE, TIR, VOR	yes	yes	2	no	100%	–	–	directly
Switzerland	SWI	yes	yes	1	no	–	100%	–	directly
Germany	BAY	yes	yes	2	no	100%	–	–	website
France	BSM, BRI, CHX, GRE	yes	yes	2	yes	–	100%	–	directly
Italy	BOZ, PIE, TRE, VDA	yes	yes	1	no	42%	41%	16%	directly
	FRI, LOM, VEN	yes	(yes)*	1	no	–	–	–	AINEVA
	BOR, LIG	no	yes	1	no	48%	49%	3%	directly
	LIV	yes	yes	3	no	100%	–	–	directly

\* (yes): AINEVA database provided information whether danger level changed, but not to which danger level

$D_e$ : concept of assigning 1, 2 or 3 danger ratings

data source: ÖLWD - from Austrian winter reports ÖLWD (2012) - ÖLWD (2015), directly - directly from respective forecast center, website - from website of Bavarian avalanche warning service, AINEVA - extracted from central AINEVA database (M. Valt (VEN))

The most relevant information concerning differences in raw data analyzed are displayed in Tab. A.3. The danger level was generally valid for the day of publication (d+0, in Austria, Germany, Livigno (LIV), scenario 1) represented essentially a one-day forecast (d+1) in France and Switzerland (although the valid period started already on the afternoon of publication, scenario 2 and 3), but was a mix of current day assessments, and forecasts with one or two days (d+2) lead-time in Italy. In Italy (AWS AINEVA), the most recently published valid danger level was used (e.g. an afternoon update, valid for the current day (d+0) replaced a forecast with a lead time of two days (d+2)). Furthermore, publication frequency increased during the explored time period in some of the AINEVA forecast centers (i.e. in Piemonte (PIE) to weekdays or in Bozen-Südtirol/Bolzano-Alto Adige (BOZ) additionally on Saturdays). Similarly, the validity of the bulletin on the issuing day changed in some Italian forecast centers from a current day assessment to a one-day forecast (i.e. BOZ changed in 2014 from d+0 to d+1), or vice versa (AWS Meteomont Carabinieri: Lombardia (BOR) and Liguria e Toscana (LIG) changed in 2014 from d+1 to d+0).

Temporal differences in danger level within the forecast period were available for all forecasts, except those by BOR and LIG (Italy) and KAE (Austria). In both cases, only the highest danger level per day was available. The data extracted from the AINEVA database (forecast centers Friuli Venezia Giulia (FRI), Lombardia (LOM), Veneto (VEN)) indicated not only the danger level, but also whether the danger rating increased, stayed the same or decreased.

In France spatial variations in the danger level within the same warning region ( $D_{spatial}$ ) were sometimes indicated (e.g. in a bulletin this could read «2-Moderate in the West, 3-Considerable in the East» of a

region).

## A.2.4 Methods

The quantitative part of this study is twofold: first, we make pairwise comparisons of neighboring warning regions, and second, we visualize and detect patterns at larger scales than individual warning regions.

### A.2.4.1 Topological neighbors

We defined warning regions  $i$  and  $j$  as topological neighbors, whenever they shared more than one point of their polygon boundary with each other (rook mode, Dale and Fortin (2014); R-package *spdep* Bivand et al. (2013); Bivand and Piras (2015)). For this purpose, the shapes of the warning regions had to be slightly adjusted so that the coordinates of joint borders matched. This also reflects challenges of working across borders, with different map projections and simplified outlines of warning regions. For the particular case of the three forecast centers in Lombardia (BOR, LIV and LOM), we defined them as neighbors if they either shared a common polygon boundary or at least partially the same territory.

### A.2.4.2 Avalanche danger level statistics

We refer to danger levels  $D$  either using their integer value (e.g.  $D=1$  for 1-Low) or by integer value and signal word combination 1-Low. Similarly to previous studies (e.g. Jamieson et al., 2008; Techel and Schweizer, 2017), we use the integer value of danger levels to calculate proportions and differences.

## Data preparation

We explored the forecast danger levels at the spatial scale of the individual, geographically clearly delineated warning regions. The following cases were treated separately:

- **Austria, Germany, France:** occasional updates during the morning  
In special circumstances, bulletins were updated during the day and the danger level adjusted. These cases were rare (for instance in Bayern (BAY) and Tirol (TIR) twice during the explored four winters). These updates were not considered in the analysis. The data provided by France, where morning updates are also possible until 10:00 CET, already included such updates.
- **France:** spatial gradients within same warning region  
In France, forecasters sometimes communicated two danger ratings for the same warning region expressing a spatial gradient. These cases were rare (0.4% of warning regions and days; Bourg-St-Maurice (BSM) 1%, Briançon (BRI) 0.3%, Chamonix (CHX) 0.1%, Grenoble (GRE) 0%). For these forecasts, we randomly picked one of the two danger levels. The remainder of the forecasts expressed no spatial gradients.
- **Switzerland (SWI):** evening forecast; danger ratings communicated in text form only  
We used the forecast issued at 17:00 CET, rather than the updated forecast the next morning (08:00

CET) as, until the winter 2012/13, the daily morning update was issued only for parts of the Swiss Alps. Furthermore, we only analyzed the danger ratings published on the map product, and not those only described in the forecast text.

- **Italy (AINEVA forecast centers Friuli Venezia Giulia (FRI), Lombardia (LOM), Veneto (VEN)):** forecast danger level changed during valid bulletin period

Data extracted from the AINEVA database provided the danger level valid in the morning, and whether the danger level changed during the day (increase, no change, decrease), but not which danger level was forecast following the change. To supplement this information, we utilized the distributions of the four AINEVA forecast centers, which consistently provided the second danger rating (Bozen-Südtirol/Bolzano-Alto Adige (BOZ), Piemonte (PIE), Trentino (TRE), Valle d'Aosta (VDA)). In these forecasts, changing danger level was by one level in 85% of cases, and by two levels in 15% of cases. For the bulletins in FRI, LOM and VEN we assumed a one-level difference for days with changing conditions, and hence a somewhat more conservative value than in the other Italian bulletins.

### Standardizing the length of the forecasting period during the season

The length of the main forecasting season is considered as being between 14 December and 16 April. During this time, and with the exception of the 2014/15 winter (28 Dec - 16 April), there was a danger rating in at least 95% of the warning regions in the Alps (477 days, 4 winters).

### Danger ratings $D_{\max}$ and $D_{\text{morning}}$

We created two subsets of data ( $D_{\max}$  and  $D_{\text{morning}}$ ), to accommodate the different ways avalanche danger ratings are communicated in forecasts and stored in databases, and to ascertain that no bias was introduced by these differences.

We defined  $D_{\max}$  as the highest danger rating valid during a forecast period, regardless whether this was the only rating provided, whether this was for a first or second time-step, or whether it corresponded to a difference in danger level by elevation. It is of note, that  $D_{\max}$  is sometimes only valid for part of the day or part of the elevation range.

In contrast,  $D_{\text{morning}}$  refers to the maximum danger rating for the first of the two time steps, which in many cases would be considered valid for the morning. Here, it is of note that exact time when a change occurs is never provided in the published forecasts, and only categorically described within the danger description. This was calculated for all forecast centers, except Lombardia (BOR), Liguria (LIG) and Kärnten (KAE), where this information was not available.

#### A.2.4.3 Summary statistics

- **Warning region-specific summary statistics**

For each warning region, we calculated the proportion of forecasts issuing a specific danger level (i.e. forecasts with danger level  $D = 4$ ). Furthermore, for each warning region we calculated the surface area, which we refer to as the *size* of a warning region, using the R-package *raster* (Hijmans, 2016) and the maximum elevation (ArcGIS software). The latter is based on a 90×90 m digital elevation

model (ESRI, 2017).

- **Pairwise comparison of immediately neighboring warning regions**

We compare the forecast danger level in two neighboring warning regions  $i$  and  $j$  by calculating the difference in the forecast danger level  $\Delta D$  for each day  $\Delta D = D_i - D_j$  for all days with  $D_i \geq 1$  and  $D_j \geq 1$ , where  $D$  may refer to  $D_{\max}$  or  $D_{\text{morning}}$ .

The proportion of days when the forecast danger levels agreed  $P_{\text{agree}}$  is then

$$P_{\text{agree}} = P(\Delta D = 0) = \frac{N(\Delta D = 0)}{N(\Delta D)} \quad (\text{A.1})$$

$P_{\text{agree}}$  may be interpreted as an indicator of spatial correlation or measure of spatial continuity in avalanche conditions.

For neighboring warning regions  $i$  and  $j$ , we calculated a bias ratio  $B_{ij}$  similar to Wilks (2011, p. 310):

$$B_{ij} = \frac{N(\Delta D = 0) + N(\Delta D^+)}{N(\Delta D = 0) + N(\Delta D^-)} \quad (\text{A.2})$$

where  $N(\Delta D^+)$  is the number of days with the  $D_i \geq D_j$  and  $N(\Delta D^-)$  the number of days with  $D_i \leq D_j$ .  $B_{ij} > 1$  indicates region  $i$  having more frequently higher danger levels than region  $j$ ,  $B_{ij} = 1$  indicates a perfectly balanced distribution, and  $B_{ij} < 1$  a skew towards more often higher danger levels in region  $j$  compared to  $i$ . We tested whether the bias  $B_{ij}$  was significantly unbalanced, by comparing the observed distribution of the two outcomes ( $N(\Delta D^+)$ ,  $N(\Delta D^-)$ ) to a random distribution using the binomial test (R: *binom.test*, R Core Team (2017)). The resulting p-value depends on the deviation of  $B_{ij}$  from 1, and on the number of days  $N(\Delta D \neq 0)$ . In general, bias values  $B_{ij} < 0.95$  or  $B_{ij} > 1.05$  were statistically significant ( $p < 0.05$ ).

The distance between warning regions refers to the distance between the center points of the respective warning regions.

#### A.2.4.4 Sensitivity and correlation

We tested whether removing subsets of the data (for instance individual years), or using  $D_{\text{morning}}$  compared to  $D_{\max}$  influenced the rank order of the warning regions using the Spearman rank order correlation coefficient  $\rho$ . Similarly, we used  $\rho$  to explore whether the frequency a specific danger level was issued correlated with differences in the size ( $\Delta_{\text{size}}$ ) or in the maximum elevation ( $\Delta_{\text{elevation}}$ ) of two warning regions  $i$  and  $j$ .

We compared populations using the Wilcoxon rank-sum test (Wilks, 2011, p. 159-163). We consider  $p \leq 0.05$  as significant.

## A.2.5 Results

### A.2.5.1 Forecast danger levels

Fig. A.4 summarizes the distribution of issued danger levels across the Alps during the four years (477 forecast days, 281 warning regions). Danger levels 2-Moderate and 3-Considerable are forecast about 80% of the time, regardless whether we consider the forecast danger level valid in the first time-step, often

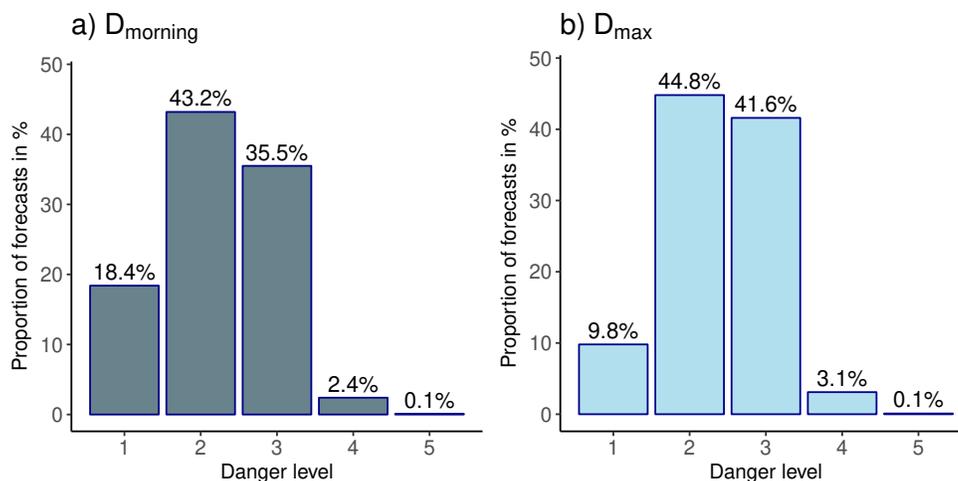


Figure A.4: Distribution of forecast danger levels, for a)  $D_{\text{morning}}$  (danger level valid during first time-step) and b)  $D_{\text{max}}$  (highest danger level). Mean values are shown for all the warning regions in the Alps taken together.

corresponding to the situation in the morning ( $D_{\text{morning}}$ ; Fig. A.4a), or the highest danger level issued ( $D_{\text{max}}$ ; Fig. A.4b). Particularly in spring situations, when avalanche hazard often increases with day-time warming, the afternoon rating is higher than the morning one; hence these two distributions differ significantly ( $p < 0.01$ ). However, as often the results obtained using  $D_{\text{max}}$  and  $D_{\text{morning}}$  were very similar, in the following we only present results if these differed significantly.

In order to address research questions 1 and 2, we explore agreement and bias, the proportion of forecasts at the upper and lower end of the EADS, and the proportion of changing danger ratings during the day. Additionally, we explore the influence of the size of the warning regions on the spatial variability in danger ratings and on the proportion of forecasts with danger levels 4-High and 5-Very High. Finally, we present two case studies to illustrate different aspects of these results in practical situations.

### A.2.5.2 Comparing immediately neighboring warning regions: agreement and bias

The forecast danger level agreed in 83% of the cases (median  $P_{\text{agree}}$ ) between two neighboring warning regions.

$P_{\text{agree}}$  was significantly higher when comparing warning regions within forecast center boundaries (91%, interquartile range IQR 83 - 96%) compared to those across forecast center boundaries (63%, IQR 58 - 70%,  $p < 0.001$ ), or across national borders (62%, IQR 58 - 66%,  $p < 0.001$ ). The latter values were not significantly different. Exploring the agreement rate graphically on a map by emphasizing borders with  $P_{\text{agree}} \leq 80\%$  essentially captures almost all forecast center boundaries and comparably few boundaries within forecast center domains (Fig. A.5). This result is confirmed when using only a subset of the warning region pairs, with  $\Delta_{\text{elevation}} < 250$  m and the size of the larger region being less than 1.5 times the size of the smaller region (Fig. A.6). For this subset, the median agreement  $P_{\text{agree}}$  is about 30% lower across forecast center boundaries, than within those ( $P_{\text{agree}}(\text{same forecast center}) = 93\%$ ,  $P_{\text{agree}}(\text{different forecast center}) = 63\%$ ,  $p < 0.001$ , Fig. A.6). Even when removing the data of the forecast centers in Switzerland (SWI), Trentino (TRE) and Valle d'Aosta (VDA), with median  $P_{\text{agree}}$  values of 95%, the difference remains highly

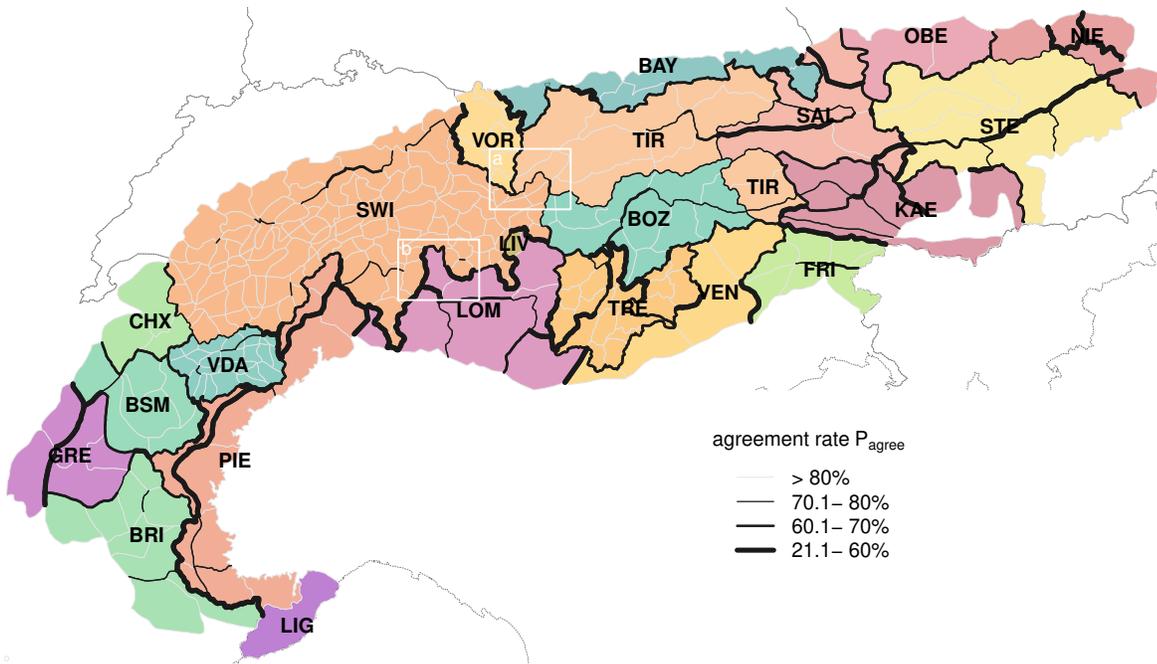


Figure A.5: Map showing the individual forecast center domains in the European Alps (different colors, three-letter abbreviations see Table A.2). The borders between warning regions are highlighted depending inversely on the agreement rate  $P_{\text{agree}}$ , with thicker lines corresponding to more frequent disagreements. The two white boxes (a, b) mark the two regions discussed in more detail in the text.

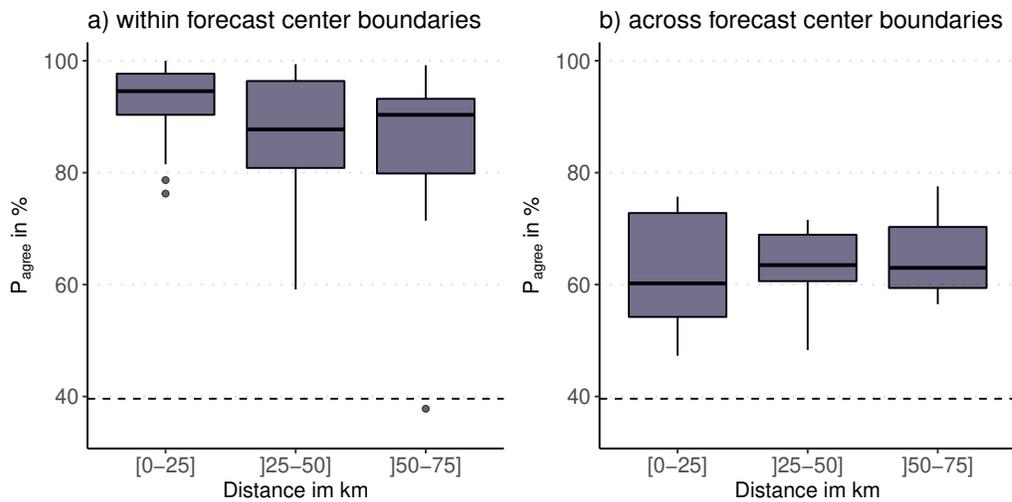


Figure A.6: Boxplot showing the agreement rate ( $P_{\text{agree}}$ ) for neighboring warning region pairs (a) within and (b) across forecast center boundaries, stratified by the distance between the center points of warning regions, with similar maximum elevation ( $\Delta$  elevation < 250 m) and size (the size of the larger warning region is less than 1.5 times the size of the smaller warning region;  $N(\text{within}) = 108$ ,  $N(\text{across}) = 37$ ). The dashed line represents  $P_{\text{agree}}$  when randomly drawing 10'000 danger levels for neighboring warning regions using the distributions shown in Fig. A.4. The Boxplots show the median (bold line), the interquartile range (boxes), 1.5 times the interquartile range (whiskers) and outliers outside this range (dots).

significant ( $P_{\text{agree}}(\text{within forecast center domain}) = 87\%$ ,  $P_{\text{agree}}(\text{across forecast center domains}) = 63\%$ ,  $p < 0.001$ ).

Similar results are noted for the special case of the three forecast centers in the Italian region of Lombardia (BOR, LIV, LOM). For these partially overlapping warning regions  $P_{\text{agree}}$  was 63%, and thus similar to  $P_{\text{agree}}$  across national borders or forecast centers neighboring each other.

Within the boundaries of forecast centers, there was a weak, but significant correlation between  $P_{\text{agree}}$  and differences in the elevation of two neighboring regions ( $\rho = -0.36$ ,  $p < 0.001$ ), with larger differences in elevation corresponding to a lower agreement rate. There was also a weak correlation between  $P_{\text{agree}}$  and differences in the size of the warning regions ( $\rho = -0.24$ ,  $p < 0.001$ ), where agreement increases as the size difference between warning regions decreases.

Within forecast center domains, the bias ratio  $B_{ij}$  correlated weakly with differences in the size ( $\rho = -0.37$ ,  $p < 0.01$ ) and elevation ( $\rho = -0.21$ ,  $p < 0.01$ ), indicating that generally the forecast danger level increased with elevation, but also with the size of the warning region. For the warning regions pairs shown in Fig. A.6, a significant bias existed in 76% of the pairs across forecast center boundaries, compared to 51% within those boundaries.

Compared to warning regions in neighboring forecast centers, the forecast centers Niederösterreich (NIE), Switzerland (SWI) and Bayern (BAY) had the lowest median bias ratios ( $B_{ij} \leq 0.84$ ), indicating that lower danger levels were used more frequently. This is in contrast to Lombardia (LOM), Briançon (BRI) and Salzburg (SAL) with median bias ratios  $B_{ij} \geq 1.19$ . For days and regions where danger levels differed, this corresponded to  $D_{\text{max}}$  being lower on more than two thirds of the pairwise-comparisons for Niederösterreich (NIE), Switzerland (SWI) and Bayern (BAY), and similarly for Lombardia (LOM), Briançon (BRI) and Salzburg (SAL) with more than 60% of forecasts with  $\Delta D \neq 0$  being higher.

### A.2.5.3 Very critical avalanche conditions $D \geq 4$ -High

Danger level 5-Very High was rarely forecast (less than 0.1% of days and regions, mostly during 2013/2014 in the southern part of the Alps; Fig. A.4). Therefore, we explore forecasts with a very critical avalanche situation ( $D = 4$ -High) or a disaster situation ( $D = 5$ -Very High) combined. For a specific warning region, the proportion of forecasts with very critical conditions is

$$P_{\text{v.crit}} = \frac{N(D \geq 4)}{N} \quad (\text{A.3})$$

where  $N$  is the number of forecasts.

Forecasts with forecast danger levels 4-High or 5-Very High were generally rare (median 2.5%, IQR: 1.1 - 4%, Fig. A.7a), but were considerably more frequently forecast in the warning regions belonging to the four forecast centers in France (Briançon (BRI), Bourg-St-Maurice (BSM), Chamonix (CHX), Grenoble (GRE)) and the Italian forecast centers Piemonte (PIE) and Lombardia (LOM). Visually exploring spatial patterns (Fig. A.7a) shows several forecast center borders which coincide with large gradients in  $P_{\text{v.crit}}$  values. These differences are most obvious when comparing Switzerland (SWI) with its neighbors Chamonix (CHX), Piemonte (PIE), Lombardia (LOM) and Tirol (TIR), where two (or more) classes difference often occur. In contrast, and with some exceptions, comparably similar values can be noted in many of the forecast centers in Austria, Germany, Switzerland and the Italian provinces and regions of Valle d'Aosta (VDA), Bozen-Südtirol/Bolzano-Alto Adige (BOZ) and Trentino (TRE). Variations are also confirmed, when consid-

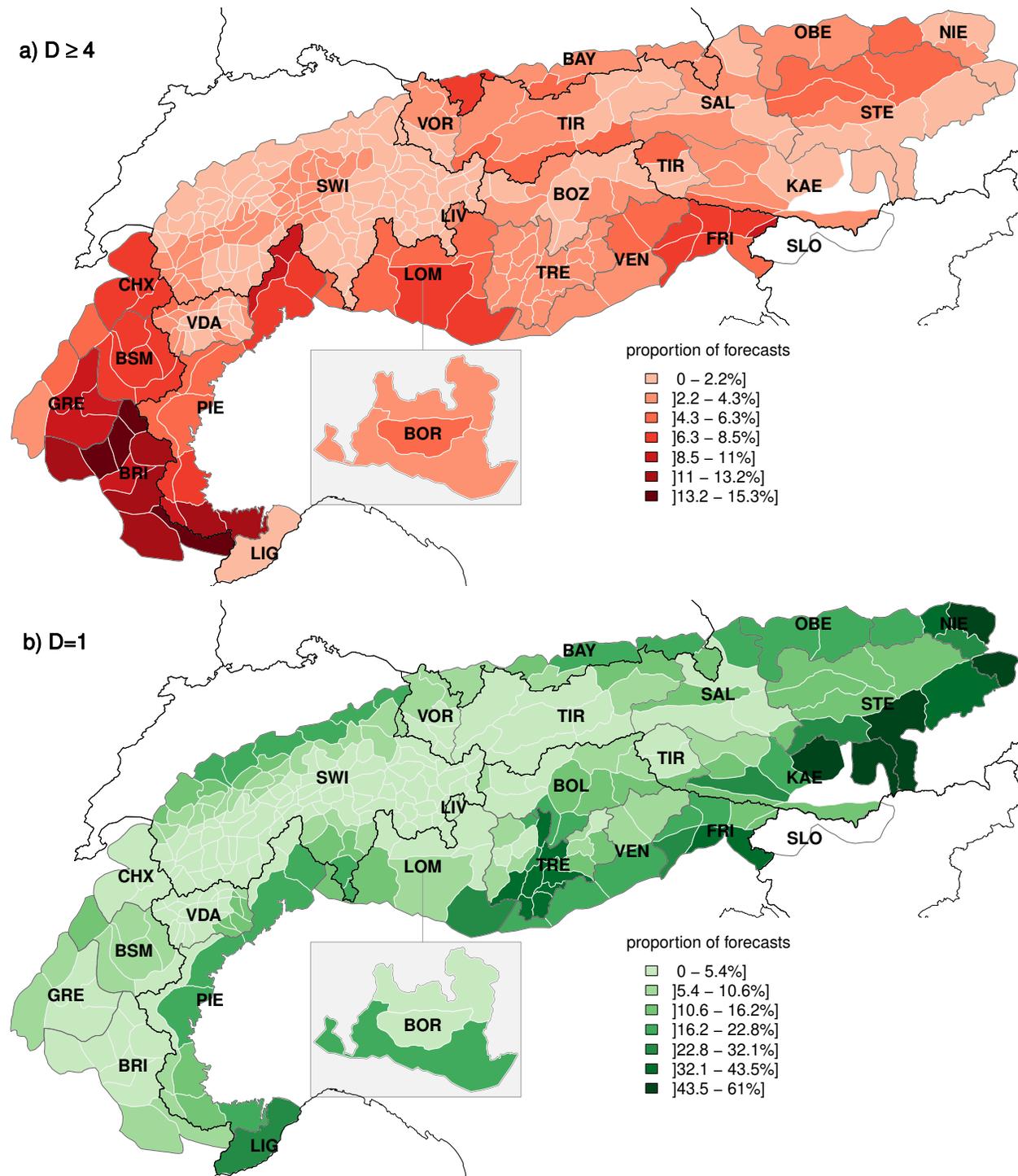


Figure A.7: Map showing the European Alps with (a) the proportion of days with forecast very critical conditions ( $P_{v,crit}$ ,  $D_{max} \geq 4$ ) and (b) with a forecast danger level 1 ( $P_{favor}$ ,  $D_{max} = 1$ ). The color shading of the individual warning regions (white borders) corresponds to the values of  $P_{v,crit}$  and  $P_{favor}$ , respectively. Forecast centers are labeled according to Table A.2 and marked with dark grey polygon borders, national borders with black lines. To visualize the (at least partially) overlapping forecast regions in the Italian region of Lombardia, LIV is superposed onto parts of LOM, while BOR is placed as inset to the south of LOM. Thresholds for the color classes were defined using the Fisher-Jenks algorithm minimizing within-class variation (Slocum et al. (2005); R-package *classInt* Bivand (2017)).

ering only warning regions with a maximum elevation greater than 2500 m ( $N = 222$ ). Median values for warning regions in Bozen-Südtirol/Bolzano-Alto Adige (BOZ), Switzerland (SWI), Vorarlberg (VOR), Valle d'Aosta (VDA) and Salzburg (SAL) (1.6%-2.3%) are significantly lower than those for Friuli Venezia Giulia (FRI), Bourg-St-Maurice (BSM), Piemonte (PIE), Grenoble (GRE) and Briançon (BRI) (7.6%-12%). This can be partly attributed to more frequent occurrence of multi-day continuous periods with  $D \geq 4$ -High. Extended periods with  $D \geq 4$  were comparably frequent in Briançon (BRI) or Piemonte (PIE) (more than 17% of these periods had a length of  $\geq 3$  days), compared to Switzerland (SWI) and Chamonix (CHX) ( $\geq 3$  days: 4%).  $P_{v,crit}$  in Briançon (BRI) was in many cases two or three classes higher compared to its immediate neighbors in Italy (Piemonte (PIE), Liguria (LIG)), but also those in France (Bourg-St-Maurice (BSM), Grenoble (GRE)). The twelve regions with the highest  $P_{v,crit}$  were clustered in the southwest of the Alps (9 in Briançon (BRI), 2 in Piemonte (PIE) and 1 in Grenoble (GRE),  $P_{v,crit} \geq 9.8\%$ , max = 15.3%).

$P_{v,crit}$  correlated very weakly with maximum elevation of a warning region ( $\rho = 0.19$ ,  $p < 0.01$ ). This correlation, however, was much stronger when exploring the proportion of days with  $D \geq 3$ -Considerable ( $\rho = 0.7$ , but also for  $D = 3$  by itself ( $\rho = 0.72$ ; see also the supplement appended to this publication, p. 120).

#### A.2.5.4 Generally favorable avalanche situation $D = 1$ -Low

The proportion of days with a generally favorable avalanche situation  $P_{favor}$  is

$$P_{favor} = \frac{N(D = 1)}{N}. \quad (\text{A.4})$$

Median  $P_{favor}$  across the Alps was 5.3% (IQR: 3.4 - 13.8%), with two regions in Niederösterreich (NIE) having more than 50% of the forecasts with  $D = 1$ -Low. The northern, southern and eastern rim of the Alps, generally regions with lower elevation (Fig. A.3b), often have a larger proportion of days with favorable conditions (Fig. A.7b). For regions with higher elevations, this proportion is lowest. This is also confirmed when correlating the maximum elevation of each warning region with  $P_{favor}$  ( $\rho = -0.75$ ). In contrast, the correlation between  $P_{favor}$  and the size of the warning regions is much weaker ( $\rho = -0.26$ ,  $p < 0.001$ ).

Another obvious difference was the strong gradient between the eastern-most regions, where more than one third of the forecast period had generally favorable conditions, and those in the western and central parts of the Alps with comparably low values of  $P_{favor}$ .

#### A.2.5.5 Elevational gradients and temporal changes within forecast period

Different approaches are used to communicate elevational gradients in danger ratings. Forecast centers issuing two ratings - mostly in France, Austria and Bayern (BAY) - seldom indicated the highest hazard at lower elevations. This is in line with the correlations observed between the maximum elevation of a warning region and  $P_{v,crit}$  (or  $P_{favor}$ ). The same danger rating was issued for all elevations by French forecast centers in two thirds of the forecasts, compared to 60% of the forecasts with an elevational gradient in Tirol (TIR) (Tab. A.4).

All forecast centers, which were technically able to graphically communicate changes in danger level during the forecast period used this option. Most frequently, forecasts indicated no change during the forecast

Table A.4: Elevational differences in danger rating with  $D_{e1}$ , the danger level above an indicated level, and  $D_{e2}$ , the danger rating below this elevation level. Example distributions are provided for some forecast centers.

forecast center	$D_{e1} > D_{e2}$	$D_{e1} = D_{e2}$	$D_{e1} < D_{e2}$
BRI, BSM, CHX, GRE	32%	67%	0.9%
BAY	45%	48%	7.2%
TIR	60%	35%	4.6%

Table A.5: Temporal differences in danger rating within forecast period with  $D_{t1}$ , the danger rating valid for the first time step, and  $D_{t2}$ , for the second time-step. Example distributions are provided for some forecast centers.

forecast center	$D_{t1} > D_{t2}$	$D_{t1} = D_{t2}$	$D_{t1} < D_{t2}$
NIE, OBE	0%	95%	5%
VOR	13%	61%	26%
LOM	6%	72%	22%
FRI, PIE	0.2%	74%	25%
SWI	0%	87%	13%*
BRI, BSM, CHX, GRE	0.9%	84%	15%

\*Switzerland: the proportion of changing danger ratings which were exclusively communicated in the danger description was 2.7%.

period (median 83%). Increasing danger levels ( $D_{t2} > D_{t1}$ ) were communicated regularly by all the forecast centers (median 16%). However, the frequency varied considerably, between 26% in Vorarlberg (VOR) and less than 10% in Niederösterreich (NIE) and Oberösterreich (OBE, Tab. A.5). Of particular note is Switzerland (SWI), the only warning service where increases in danger rating related to dry-snow avalanches were communicated exclusively in the textual danger description. A decrease in danger level during the forecast period was very rarely indicated (median 0.3%). Some forecast centers like Switzerland (SWI) never used this option. Notable exceptions were the forecasts by Vorarlberg (VOR) and Lombardia (LOM), where more than 6% of the forecasts indicated a decreasing danger rating within the forecast period.

#### A.2.5.6 Size of the warning regions, $P_{v,crit}$ and spatial variation in danger level

Varying spatial scales and approaches are used to produce the forecast, and communicate a danger level. One of these approaches relies on a comparably fine spatial resolution of the warning regions in the bulletin production process, as is the case in Valle d'Aosta (VDA, Italy), Switzerland (SWI) and Trentino (TRE).

The forecast center VDA uses 26 warning regions (median size 130 km<sup>2</sup>, Tab. A.2, Fig. A.3). Each of these regions belongs to one of four larger snow-climate regions (median size 815 km<sup>2</sup>, Burelli et al. (2016, p. 27)). In Switzerland, the forecaster aggregates the 117 warning regions in the Swiss Alps (median size 180 km<sup>2</sup>) to (generally) three to five regions with the same danger description (with an average size per aggregated

Table A.6: Variability in danger ratings and the proportion of forecasts with danger levels 4-High or 5-Very High ( $P_{v.crit}$ ) assuming different aggregation levels as the given spatial resolution for danger level communication. The aggregation level *none* indicates the currently used spatial resolution. The aggregated median size and number (N) of regions within the forecast domain are indicated.  $P_{v.crit}(\max)$  assumes the communication of the highest danger rating per region, and  $P_{v.crit}(\text{mean})$  the spatially most relevant danger rating.

forecast center	aggregation	size (km <sup>2</sup> )	N	1 rating	2 ratings	≥ 3 ratings	$P_{v.crit}(\max)$	$P_{v.crit}(\text{mean})$
VDA	none	130	26	100%	-	-	2.3%	2.3%
	first-order	815	4	83%	17%	0.3%	3.7%	2.3%
	second-order*	3300	1	56%	39%	5%	6.8%	0.7%
SWI	none	180	117	100%	-	-	1.3%	1.3%
	first-order	740	35	85%	15%	0.3%	1.6%	1.3%
	second-order	1740	17	71%	28%	1.1%	2.3%	1.3%
	third-order	3260	7	53%	44%	2.9%	3.1%	1%

\* - considering the entire VDA forecast domain as one region

region of 5000 - 7000 km<sup>2</sup>; Ruesch et al., 2013; Techel and Schweizer, 2017). Similar to VDA, each of the Swiss warning regions can be linked to a higher-order spatial hierarchy (SLF, 2015, p. 41)<sup>3</sup>. In either case, these predefined regional aggregations are not of great importance anymore in the communication of a regional danger level, due to the flexibility in which the forecaster can assign danger ratings to regions (VDA) or aggregate regions (SWI). However, here we use these spatial hierarchy-levels - three for VDA and four for SWI<sup>4</sup> - to explore the variability of the forecast danger level within regions of increasing size and the potential implication on summary statistics like the proportion of the most critical forecasts ( $P_{v.crit}$ ).

As shown in Tab. A.6, the larger a region, the higher the variability within these regions (more than one danger level forecast). In other words, a forecaster would not have been able to communicate the spatial variability in danger levels without describing these in text form if warning regions were five times larger (about 800 km<sup>2</sup>, corresponding to the median size in Niederösterreich (NIE) or in France) in about 15% of the forecasts, as compared to the currently implemented spatial resolution. Assuming even larger warning regions at the communication level, 3300 km<sup>2</sup>, for instance when considering VDA as one single region, or the seven snow-climate regions in SWI, and communicating a single danger rating only, would have resulted in about half of the forecasts not reflecting the spatial variability within the respective region.

This shows that variations in the expected avalanche hazard at spatial scales lower than the size of the spatial units used in the production and communication of the forecast are to be expected, particularly if regions are large. In these situations, a forecaster must decide whether to communicate the highest expected danger level, regardless of its spatial extent, or the danger level representative for the largest part of a region. Note that currently the EADS lacks a definition in that respect. Taking the proportion of forecasts with very critical conditions  $P_{v.crit}$  shows that communicating the highest danger level within a region

<sup>3</sup>As an example, the warning region «1121 - Freiburger Alpen» belongs at its highest hierarchy level to the snow-climate region «1 - western part of the Northern flank of the Alps».

<sup>4</sup>no higher hierarchy exists for the warning regions in TRE

$P_{v,crit}(\max)$  increases the absolute values of  $P_{v,crit}$  (Tab. A.6). Communicating the spatially most widespread danger rating instead ( $P_{v,crit}(\text{mean})$ ), has relatively little influence for smaller regions, but reduces  $P_{v,crit}$  values significantly for the largest-size regions (Tab. A.6).

At the current spatial resolution,  $P_{v,crit}$  values for SWI and VDA are comparable, particularly along their joint border (Fig. A.7a). However,  $P_{v,crit}(\max)$  values at the first-order aggregation are already considerably higher for VDA, and rather similar to those in neighboring warning regions in Chamonix (CHX), Bourg-St-Maurice (BSM) or Piemonte (PIE).

### A.2.5.7 Case studies

To make the results more tangible, we present two case studies (Fig. A.8):

The *Silvretta* mountain range, at the border between Austria (Vorarlberg (VOR) and Tirol (TIR)) in the North and Switzerland (SWI) in the South (Fig. A.8a) is split into six warning regions, all including *Silvretta*, and/or *Samnaun* in their region name. These have similar maximum elevations (between 3200 and 3340 m), but differ in size (SWI  $\leq 180$  km<sup>2</sup>, TIR 490 km<sup>2</sup>). According to Schwarb et al. (2001), there is a precipitation gradient during the three winter months December to February with total precipitation amounts decreasing from about 250 - 300 mm (in VOR) to about 150 - 200 mm in the Eastern most regions in TIR.

The agreement rate is high between the Swiss *Silvretta* regions (93%), but considerably lower across forecast center boundaries (SWI - TIR 73%, SWI - VOR 64%). Note further, that between the Swiss *Silvretta* and *Samnaun*  $P_{agree}$  equals 100%. Additionally, there is a significant bias present between SWI and its two Austrian neighbors ( $p < 0.001$ ), with the danger level in Switzerland being lower more often than higher. In contrast, despite a low agreement rate (67%) there is no significant bias between TIR and VOR, implying that differences in forecast avalanche danger are balanced. Note further, that  $P_{agree}$  between VOR and its neighbors in SWI or TIR is 5 to 10% higher when considering  $D_{morning}$  rather than  $D_{max}$ . Danger level 4 was least often forecast in the Swiss warning regions ( $P_{v,crit} < 1.2\%$ ) and most often in the largest of the

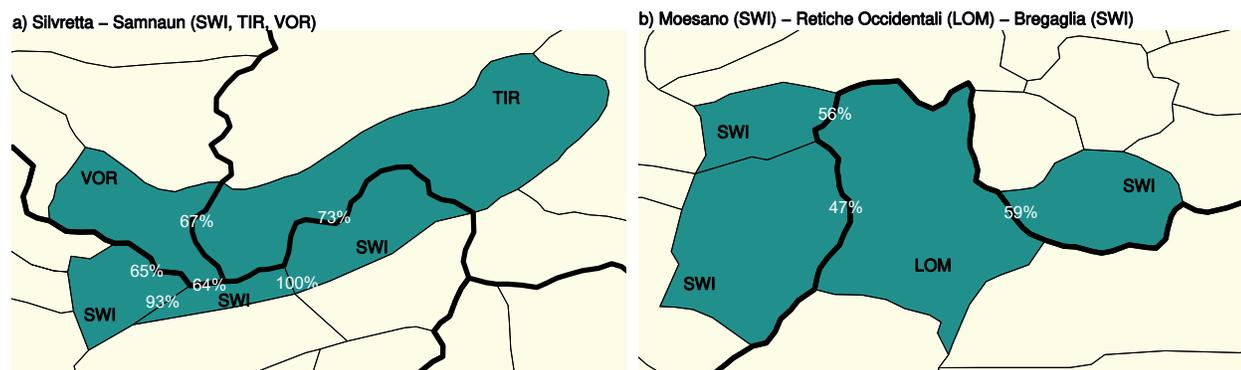


Figure A.8: Example regions: a) Silvretta mountain range with the *Silvretta* warning regions in Vorarlberg (VOR) and Tirol (TIR) and three Swiss warning regions (SWI, from west to east: *Western Silvretta*, *Eastern Silvretta* and *Samnaun*). b) the *Retiche Occidentali* warning region (forecast center Lombardia (LOM)) and the three Swiss warning regions *Alto Moesano*, *Basso Moesano* and *Bregaglia*. Here, the main Alpine divide runs right to the north of the dark-colored regions. - The percentage values show the agreement rate between warning regions ( $D_{max}$ ). The maps show an area of 83 by 45 km. The location of these two example regions in the Alps is marked in Fig. A.5.

five regions, in Tirol (4.7%). In comparison,  $D = 1$  was forecast between 2.4% in Tirol and 4.7% in the two western-most regions in Vorarlberg and Switzerland.

Turning to a location south of the main Alpine divide, where the Italian *Retiche occidentale* warning region in Lombardia (LOM, size 510 km<sup>2</sup>, elevation 3200 m) lies embedded between three Swiss warning regions (SWI, size 120 - 370 km<sup>2</sup>, elevation 2900 - 3300 m). It is an area, which receives most precipitation from southerly air currents. Winter precipitation is generally more abundant in the Southwest (200 - 250 mm) compared to the North and East of these regions (150 - 200 mm, Schwarb et al., 2001). This pattern is more pronounced in spring (March - May). The agreement rate between the three Swiss warning regions was between 79% and 90%, despite them being sometimes separated by the Lombardian warning region. The agreement rate between the Swiss and Lombardian region ranged between 47% and 59%. The bias was very pronounced with Swiss forecasts often being lower than the ones in LOM ( $p < 0.001$ ). This also shows when comparing  $P_{v.crit}$  ( $P_{v.crit}(\text{LOM}) = 5.2\%$  vs.  $P_{v.crit}(\text{SWI}) < 1.2\%$ ) or  $P_{favor}$  ( $P_{favor}(\text{LOM}) = 1.8\%$  vs.  $P_{favor}(\text{SWI}) > 3.8\%$ ).

## A.2.6 Discussion

We explored spatial consistency and bias using published forecast avalanche danger levels by using a comparably large number of real forecasts rather than a small number of hypothetical scenarios, as in the experiment conducted by Lazar et al. (2016). However, using actual forecasts in such a diverse setting as the European Alps, comes at the cost of many confounding factors. Differences between forecast centers in the forecast production and danger level communication required us to make some assumptions prior to data analysis.

In this discussion, we first summarize the main quantitative findings, which we then put into perspective given the data and our methodology. Furthermore, we discuss sources for inconsistencies and bias and potential implications to forecast users.

The main results are:

- The agreement rate  $P_{agree}$  was significantly lower across national and forecast center boundaries (about 60%), compared to within forecast center boundaries (about 90%, Figures A.5 and A.6).
- Significant bias was often observed across national and forecast center boundaries, with several forecast centers showing systematic differences towards lower (or higher) danger levels than their neighbors.
- The proportion of forecasts with danger levels 4-High and 5-Very High showed considerable spatial variability (Fig. A.7a), with pronounced differences across some forecast center boundaries, and was influenced by the size of warning regions.

### A.2.6.1 Dataset: four winter seasons

We explored avalanche forecasts published during four winter seasons (477 forecast days). These included the 2011/2012 winter with extended periods of heavy snowfalls affecting particularly the regions north of the

main Alpine Divide (Northern French Alps, large parts of Switzerland and Austria, Bavarian Alps; Coléou, 2012; ÖLWD, 2012; Techel et al., 2013), but also the 2013/2014 winter, which was one of the snowiest winters on record in the Southern Alps (Italy, southern parts of Switzerland; Goetz, 2014; ÖLWD, 2014; Techel et al., 2015a; Valt and Cianfarra, 2014). These two winters, or removing one of them during data analysis, had an effect particularly on the absolute values of the proportion of forecasts with  $D \geq 4$  ( $P_{v,crit}$ ), while the overall rank order remained comparably similar, regardless of which subset was analyzed (see also the supplement appended to this publication, p. A.2). Removing individual winters also had no significant influence on the agreement rate ( $P_{agree}$ ) or bias ( $B_{ij}$ ) between neighboring warning regions. By comparing with long-term statistics of forecast danger levels (e.g. France, Switzerland, Steiermark; Mansiot, 2016; Techel et al., 2013; Zenkl, 2016), we conclude that our data are generally representative and the four years analyzed cover a typical range of conditions encountered in the European Alps.

### A.2.6.2 Methodology

Danger levels were communicated in different ways in the forecasts. Therefore, we generalized by defining two data subsets which could be applied to most forecast products:  $D_{max}$ , describing the highest danger rating within a forecast period, valid for (part of) the day and the most exposed elevations, and  $D_{morning}$ , where we assumed that time step 1 generally referred to the morning, and time step 2 to the afternoon.

Using  $D_{max}$  or  $D_{morning}$  for analysis influenced absolute values of  $P_{v,crit}$ , but less the rank order, and had little influence on  $P_{agree}$  or  $B_{ij}$ .

We introduced  $P_{agree}$  as a measure of spatial consistency (or correlation). As shown in Fig. A.4, on four of five days  $D = 2$  or  $D = 3$  were forecast. Thus, by chance alone, a minimal agreement rate can be expected. We estimated this minimal agreement rate by simulating 10,000 danger levels for two neighboring regions using the danger level distributions shown in Fig. A.4. Doing so we obtained values of  $P_{agree} = 40\%$  for  $D_{max}$  and  $P_{agree} = 36\%$  for  $D_{morning}$ . Thus, levels of agreement reported in this paper, and in any future work, should be compared with a minimal agreement rate based on realistic values derived from observed danger level distributions.

Similarly, total agreement ( $P_{agree} = 100\%$ ) between neighboring regions implies that subdivisions may be superfluous. Nonetheless, we found 100% agreement for a total of 14 warning region pairs in Switzerland, Italy and Austria. To confirm whether this agreement indicates regions which could be merged would require further investigation as to, for example, the nature of typical avalanche problems found, and not only the forecast danger levels.

The spatial resolution of the warning regions (Tab. A.2, Fig. A.3a), and how these are used in the communication of the forecasts, varied greatly between forecast centers. As we have shown for the forecasts in Switzerland (SWI) and Valle d'Aosta (VDA), this may in turn influence the danger rating communicated to the public. As a consequence, it has an impact on all summary statistics, most notably  $P_{v,crit}$  and  $B_{ij}$ .

We explored a mix of forecasts for the day of publication, the following day, or even the day after. However, forecast accuracy generally decreases with lead time (Jamieson et al., 2008; Statham et al., 2018b). Forecast accuracy may also vary within forecast center domains, as shown by Techel and Schweizer (2017) for the case of Switzerland. We suspect that these may affect primarily the agreement rate  $P_{agree}$ , except if the

forecast bias differs temporally or spatially.

Within forecast center domains, differences in the frequency of the danger levels, the agreement rate  $P_{agree}$ , or the bias  $B_{ij}$  may indicate differences in snow avalanche climate. In all other situations, that is to say when looking at differences between forecast centers, operational constraints must be considered as much as snow-climate, when exploring consistency and bias.

### A.2.6.3 Understanding differences between avalanche warning regions

Our aims in exploring spatial consistency and bias were threefold: firstly to investigate whether differences existed between forecasting centers, secondly to understand potential factors influencing these biases, and finally to consider the influence of these biases on forecast users. Our results clearly demonstrate that spatial inconsistencies and biases exist, above all across forecast center boundaries. In the following we briefly discuss three possible reasons for such differences, two of which suggest limitations in current forecasting approaches.

The size of the warning regions differed considerably between forecast centers (Fig. A.3, Tab. A.2) and had an impact on the issued danger level in general, particularly on  $P_{v,crit}$ . Coarser spatial resolutions of warning regions lead not only to more forecasts with higher danger levels, but also increase variability within warning regions. Such variability cannot be captured with a single value and thus, though it may be expressed within the forecast text, is ignored by our approach. Since differences in warning region size were correlated with both bias and agreement rate, we recommend exploring whether more heterogeneous warning regions - from an avalanche winter regime perspective - might be divided into smaller ones to reduce such bias. We also found correlations between avalanche danger levels, bias, agreement rate and elevation. While higher elevations and higher avalanche dangers are often associated with one another, we suggest the relationship between bias and elevation may result from different ways of communicating avalanche danger for a warning region. In particular, the EADS does not specify whether the highest, or the spatially most representative danger level should be communicated for a warning region. We therefore suggest that the EAWS consider whether being more specific in defining how avalanche danger should be assigned to a warning region may reduce bias.

This lack of specificity in the EADS with respect to avalanche danger is an example of potential differences in the application of the EADS in different forecast centers, which may in turn explain some aspects of inconsistency and bias. Simply put, forecasters must assign a categorical value to a complex forecast, which typically also contains uncertainty. This assignment of an avalanche danger level is not only influenced by conditions, but may also emerge from cultural differences in forecasting practices (McClung, 2000; Greene et al., 2006; Lazar et al., 2016) and explicit or implicit internalization by forecasters of the use and implication of danger levels by local, regional and national risk management authorities. The need to increase consistency in the application of the EADS has been recognized. Efforts made by the EAWS include improvements in the *EAWS matrix*, a tool assisting forecasters in assigning danger levels (Müller et al., 2016; EAWS, 2017b) and the provision of clear definitions of key contributing factors, such as the distribution of dangerous locations and the likelihood of avalanche release. Nonetheless, it is important to recognize that even if the EAWS strive to harmonize practices and production, externalities such as the consequences of

danger levels for users, and the perception of forecasters of this impact, may alter the homogeneity of the product. Furthermore, as observed by LaChapelle (1980) and summarized very recently by Statham et al. (2018a), avalanche forecasts are produced by a forecaster making subjective judgments based on the available data and evidence. Reducing these forecasts to a categorical value neither removes the subjectivity in the process nor does it allow the forecaster to communicate uncertainty.

A third possible reason for differences between warning regions lies not in bias or inconsistency in the use of the EADS, but rather in real differences in the avalanche winter regime (Haegeli and McClung, 2007). Many of the warning region boundaries, especially along national borders, follow the main Alpine divide, which also serves as a main weather divide. Where large differences in avalanche winter regime are observed, a lower correlation in danger ratings would therefore be expected. However, we relied exclusively on forecast danger levels and cannot compare the agreement rate or bias with differences in avalanche winter regime. This is an important limitation in our study. Incorporating avalanche winter regimes in this study, and/or typical avalanche problems - if these were used consistently, would clearly be beneficial for the interpretation of our findings. Such an analysis would require, besides meteorological data, a common database containing snow structure and avalanche information for the entire Alpine mountain range, as already exists for the US and Canada (Mock and Birkeland, 2000; Haegeli and McClung, 2007; Shandro and Haegeli, 2018).

#### **A.2.6.4 Inconsistencies: implications for forecast users**

A final key question is the implications of the potential spatial inconsistencies and biases in the use of danger levels for forecast users. Even though there may be good reasons for such differences, such as the difference in size of warning regions and therefore a need to communicate different information, users are unlikely to appreciate or understand such nuances.

Regional avalanche forecasts are considered an important source of information for backcountry users, particularly during the planning stage, but also on the day of the tour (Winkler and Techel, 2014; LWD Steiermark, 2015; Baker and McGee, 2016). A key advantage of the introduction of the EADS in 1993 was seen as the provision of consistent information across the European Alps (Meister, 1995). Forecast danger level has been shown to be the part of the forecast most known and used in the Alps (Winkler and Techel, 2014; LWD Steiermark, 2015; Procter et al., 2014), influencing backcountry destinations (Techel et al., 2015b) and local decision-making by recreationists (Furman et al., 2010). Many users of avalanche forecasts are typically active within warning regions where forecasts are produced by a single regional avalanche forecast center (e.g. in Voralberg (VOR) or Tirol (TIR)). Such users are likely to become accustomed and calibrated to «their» forecast. Thus, issues are likely to arise when users travel from one forecast center domain to another. For instance, a frequent user of French forecasts traveling to Switzerland may experience some Swiss forecasts with  $D = 3$  as a missed alarm, while the opposite may happen when a Swiss user recreates in France. In both cases this reduces the credibility of the forecasts, as they are perceived to be less accurate (Williams, 1980). We suggest that harmonization efforts should therefore focus not only on the product - an avalanche forecast - but how this product is used and interpreted by different users and their requirements (Murphy, 1993).

## A.2.7 Conclusions

In this study, we explored the avalanche forecast products, and specifically the forecast danger level during four years with 477 forecast days from 23 forecast centers in the European Alps. For the first time,

- (i) we qualitatively described the operational constraints in the production and communication of danger level in avalanche forecast products in the Alps,
- (ii) we developed a methodology to explore spatial consistency and bias in avalanche forecasts,
- (iii) we quantified spatial consistency and bias in forecast danger levels, given operational constraints and the selected methods, and
- (iv) we discuss the implications of spatial consistency and bias for forecasting and forecast users.

We noted considerable differences in the operational constraints associated with forecast products. Most notably the spatial resolution of the warning regions underlying the forecasts had an impact on biases observed and the agreement rate, but also limits at what spatial scale a regional danger level can be communicated in map products. Furthermore, we detected discrepancies in the use of the higher danger levels, as well as a comparably large proportion of forecasts with different danger levels across forecast center boundaries. These findings indicate a need to further harmonize the production process and communication of avalanche forecast products, not just across the Alps but throughout Europe. Harmonization should consider

- (i) similar approaches regarding the size of warning regions and their aggregation, with a preference towards using a finer spatial resolution,
- (ii) focusing not only on forecast products, but also user requirements, and
- (iii) the consistent use of EADS by incorporating the *EAWS-Matrix*, and further developments, and developing a consistent workflow, similar to the approach suggested by Statham et al. (2018a) into the production process.

To carry out our study we had to collect and harmonize data across the Alps. We recommend a development of a centralized system for collecting data, which would enable further studies of forecast properties in the future.

*Data availability:* The data will be made freely available on the data portal [www.enividat.ch](http://www.enividat.ch).

*Author contributions:* Frank Techel coordinated and designed this collaborative study. All co-authors provided repeatedly in-depth feedback regarding previous versions of this manuscript. Authors are listed alphabetically.

*Acknowledgments:* This study would not have been possible without the contributions of numerous avalanche forecasters and researchers, who provided data and/or feedback on the manuscript (in alphabetical order): Igor Chiambretti, Jean-Louis Dumas, Mattia Faletto, Thomas Feistl, Elisabeth Hafner, Fabiano Monti, Patrick

Nairz, Bernhard Niedermoser, Andreas Pecl, Evelyne Pougatch, Jürg Schweizer, Luca Silvestri, Florian Stifter, Thomas Stucki, Arnold Studeregger, Lukas Rastner, Mauro Valt, Romeo Vicenzo, Alec van Herwijnen, Gernot Zenkl.

We greatly appreciate the detailed and constructive comments by the three reviewers Karsten Müller, Rune Engeset and Karl Birkeland.

## A.2.8 Supplements

### Supplement S1: critical avalanche conditions $D = 3$

The proportion of days with a critical avalanche situation, corresponding to a forecast danger level 3-*Considerable* ( $P_{D=3}$ ), is

$$P_{D=3} = \frac{N(D=3)}{N} \quad (\text{A.5})$$

The median  $P_{D=3}$  was 38% ( $D_{\text{morning}}$ , IQR: 29-43%) and 44% ( $D_{\text{max}}$ , IQR: 34 - 50%). Although large variations exist across the Alps, visual inspection of the map shown in Fig. A.9a shows only moderate discrepancies across forecast center boundaries. Differences in  $P_{D=3}$  may be explained largely with the maximum elevation of a warning region: a strong and highly significant correlation was observed ( $\rho > 0.7$ ,  $p < 0.001$ ), regardless whether this was explored for  $D_{\text{morning}}$  (Fig. A.9b) or  $D_{\text{max}}$ . In contrast, a very weak, though significant negative correlation was observed between  $P_{D=3}$  and the size of the warning regions (sign negative,  $|\rho| > 0.16$ ,  $p < 0.01$ ).

The median length of the longest period with consecutive forecasts with  $D = 3$  per winter ( $L_{D=3}$ ) was 11 days ( $D_{\text{morning}}$ , IQR 8-16 days) and 12 days ( $D_{\text{max}}$ , IQR 8.5-16.5 days).  $L_{D=3}$  correlates strongly with elevation ( $\rho > 0.62$ ,  $p < 0.001$ , Figure A.9c) and correlates negatively with the size of the warning region ( $|\rho| < 0.25$ ,  $p < 0.001$ ). The 10% of the regions with the longest continuous periods with  $D = 3$  lie mostly in Switzerland (SWI), Tirol (TIR) and Valle d'Aosta (VDA). Furthermore, the six regions with the highest  $L_{D=3}$  values ( $L_{D=3} > 27$ ) are those immediately surrounding the Swiss forecast center in Davos. Despite  $P_{D=3}$  values being similar in France to some regions in Switzerland, values for  $L_{D=3}$  tend to be lower as these periods are more frequently interrupted by one or several days with  $D \geq 4$  in France.

### Supplement S2: Sensitivity analysis - removing individual years from the data set

Danger level  $D \geq 4$ :

Using  $D_{\text{morning}}$  instead of  $D_{\text{max}}$  decreased  $P_{v,\text{crit}}$  (median change from 2.5% to 2.1%), but resulted in more or less the same order of the warning regions ( $\rho = 0.94$ ), and the same regions with the highest values ( $P_{v,\text{crit}} \geq 7\%$ , max = 12.2%). For the warning region with the highest values of  $P_{v,\text{crit}}$ , using  $D_{\text{morning}}$  rather than  $D_{\text{max}}$ , resulted sometimes in markedly lower values ( $\Delta P_{v,\text{crit}}$  1 to 3%), with especially large differences observed in Piemonte (PIE,  $\Delta P_{v,\text{crit}} = 4.7\%$ ).

Removing the winter 2013/2014 from the data set had the greatest impact on both the absolute values of  $P_{v,\text{crit}}$  as well as the rank order. The largest changes in  $P_{v,\text{crit}}$  were noted when comparing a data set excluding 2013/2014 and one excluding 2011/2012 ( $\rho = 0.59$ ). However, six of the ten regions with the highest values of  $P_{v,\text{crit}}$  remained the same even for those subsets (all belonging to Briançon (BRI)).

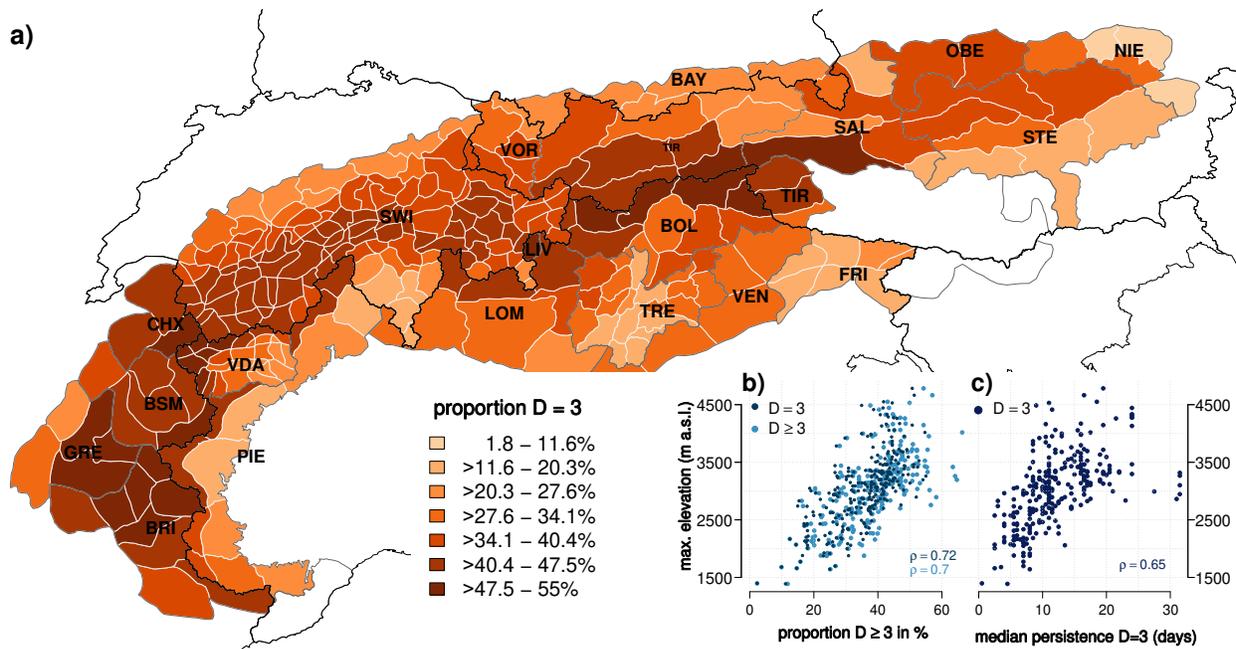


Figure A.9: a) Map showing the proportion of days with a forecast danger level 3 ( $P_{D=3}$  shown for  $D_{\text{morning}}$ ). The color shading of the individual warning regions (white borders) corresponds to the proportion of forecast days with  $D = 3$ . Forecast centers are labeled according to Tab. A.2 (p. 100) and are marked with dark grey polygon borders. National borders are highlighted with black lines. To visualize the (at least partially) overlapping forecast regions in the Italian region of Lombardia, LIV is superposed onto parts of LOM. The inset scatterplots show the relationship between (b) the proportion of forecasts with  $D = 3$  (or  $D \geq 3$ ) and (c) the median length of the longest continuous period with  $D = 3$ .

Danger level  $D = 1$ :

Using  $D_{\text{morning}}$  instead of  $D_{\text{max}}$  resulted in a reduction of the number of forecasts with  $D = 1$ , but the rank order of the warning regions changed only marginally ( $\rho = 0.90$ ). The rank order correlation was most sensitive to removing the 2011/2012 winter, and was lowest when comparing subsets of the data either excluding the 2011/2012 or the 2013/2014 winter. However, even then the correlation was strong ( $\rho = 0.80$ ).

Danger level  $D = 3$ :

The rank order correlations were most sensitive to removing the 2013/2014 winter. However, even then the correlations were generally strong or very strong ( $P_{D=3}$ :  $\rho = 0.84$ ,  $L_{D=3}$ :  $\rho = 0.71$ ).

## A.3 On using local avalanche danger level estimates for regional forecast verification

Techel, F. and Schweizer, J.: On using local avalanche danger level estimates for regional forecast verification. *Cold Reg. Sci. Technol.*, 2017, 144, 52 - 62, doi: 10.1016/j.coldregions.2017.07.012

*Remark: Variable notation and the calculation of  $\Delta D$  (Eqn. A.9, p. 129) in this version of the publication has been adjusted to match the variable definitions in the Methods section of the Synthesis (Eqn. 4.4, p. 36 in Section 4.1.1).*

### Abstract

*Operational verification of regional avalanche forecasts strongly relies on high quality field observations. In addition, specifically trained and experienced observers may provide local danger level estimates - a condensed, but subjective summary of current avalanche conditions. However, these estimates not only reflect local rather than regional conditions, but may also be influenced by, for example, the observers' personal experience and the ease of perceiving the hazard. We explored close to 10,000 local danger level estimates reported by more than 100 trained observers to the national forecasting service in Switzerland. Even at distances less than about 10 km, observers disagreed in their local estimate 22% of the time. Some observers had a bias towards consistently higher or lower local estimates. The proportion correct when comparing local estimates (nowcasts) with the regional forecasts was 76%. It varied considerably between individual observers, but partly also among typical groups of observers (e.g. mountain guides, ski area staff or avalanche forecasters). Taking into account the uncertainty in local estimates and the reporting bias revealed a slightly lower agreement between local nowcast and regional forecast of 71%. These levels of agreement seem rather low, but are in line with previous studies. We conclude that local nowcasts can be used for forecast verification, but substantial uncertainty remains and the «true» avalanche danger level remains unknown.*

### A.3.1 Introduction

In many snow-covered mountainous regions in Europe, North America and New Zealand regional avalanche forecasts are issued to warn the public about the avalanche danger. These bulletins provide information on the current and future state of the snowpack with regard to snow instability and snow structure, the expected likelihood of avalanche triggering and the type and size of the expected avalanches, as well as the likely triggering spots. The area covered by the forecasts strongly varies between several hundred square kilometers, e.g. in Scotland, to more than 30,000 km<sup>2</sup> in some regions in Canada (Bakermans et al., 2010). The bulletins are typically issued in the afternoon or evening with a forecast for the following day (or days), or in the morning. The regional avalanche danger is characterized by one of five danger levels according to a five-level danger scale. Slightly different danger scales are used in Europe (e.g., Meister, 1995) and North America (Statham et al., 2010), but both are essentially based on increasing release probability, increasing frequency and size of avalanches, and increasing frequency of triggering spots with increasing danger level.

The scale of a regional forecast is typically about 100 km<sup>2</sup>, or larger (Zenke, 2013) with a temporal resolution of 6 to 24 hours, or more (Meister, 1995).

The forecast regional avalanche danger level ( $D_{RF}$ ) is the piece of information recreationists remember best after having read the avalanche bulletin (e.g., Winkler and Techel, 2014). The danger level is also an important parameter in decision support tools for winter back-country recreationists such as the Graphical Reduction Method (Harvey et al., 2016) or the Avaluator (Haegeli, 2010). It clearly has an impact on the number of people recreating in the backcountry suggesting that the warnings are effective (e.g., Techel et al., 2015b). Jamieson et al. (2009) concluded that the forecast regional danger level correlated better with the local danger rating, estimated following a day in the field, than any of the field observations made individually during the day. These local ratings (or estimates) for the current day were on the scale of a small drainage or a typical day of winter recreation, i.e. about 10 km<sup>2</sup> (Jamieson et al., 2008); they referred to them as local nowcasts.

In day-to-day public avalanche forecasting, the review of the past forecast is the starting point in the process of preparing the future forecast. In particular, the avalanche danger level is reviewed. However, avalanche danger cannot be measured and hence not be readily verified (Föhn and Schweizer, 1995; Schweizer et al., 2003). In fact, the verification itself is considered an expert decision in hindsight as much as the assessment in the field (local nowcast). Even if a danger rating is verified using all available information in hindsight, the accuracy of the «verified» danger level may not be more than 90% (Schweizer and Föhn, 1996). The most useful information for verification is the one directly related to snow instability: recent avalanches, signs of instability (whumpfs of shooting cracks) or stability test results (McClung, 2002b). This so-called Class I data are particularly useful to distinguish between the higher danger levels 3-Considerable and 4-High, and the lower danger levels 2-Moderate and 1-Low. However, in day-to-day public forecasting this kind of information is often either absent or not readily available due to lacking observations, and other less direct information needs to be considered. Among those are current estimates of the local danger level ( $D_{LN}$ ) by experienced observers (Brabec and Stucki, 1998; Engeset, 2013; Jamieson et al., 2009). In Switzerland, the local danger level estimate is not only used to review the past regional danger level, but also to prepare the future forecast (Suter et al., 2010).

An advantage of using locally estimated  $D_{LN}$  is that a central target variable of an avalanche forecast - the forecast regional danger level - can be reviewed with a similar type of variable - rather than using, for example, avalanche occurrence data. However, challenges include differences in the spatio-temporal scale - a regional forecast valid for the day vs. a local nowcast estimated at a certain time - and the subjective nature of the local assessment. Even though  $D_{LN}$  are subjective interpretations of encountered conditions, they are considered fairly accurate estimates of the avalanche danger (Schweizer, 2010) and avalanche forecasters in Switzerland consider the quality of the estimates by trained observers to be high (Techel et al., 2016a). While situations exist when obvious signs clearly indicate a danger level 3-Considerable (or higher) (Jamieson et al., 2009; Schweizer, 2010), these signs are often lacking. In these situations, when instabilities are highly localized or a high triggering level is needed, McClung (2002a) argues that the human perception of the avalanche hazard will be fair or poor - and consequently the local avalanche danger level estimates may be less reliable.

Our objective is therefore to assess the usefulness and reliability of local avalanche danger level estimates in operational avalanche forecasting. We first analyze the variability in local danger level estimates by trained and experienced observers. Then, we test individual observers and groups of observers for a bias. Finally, we apply these findings to incorporate the uncertainty associated with local danger level estimates when verifying the regional avalanche forecast.

### A.3.2 Data

We analyze the local avalanche danger level estimates - termed «nowcasts» by Jamieson et al. (2008) - and the forecast regional avalanche danger levels, which we both extracted from the Swiss operational avalanche warning service database. Details on these data are given in the following two subsections.

#### A.3.2.1 Local avalanche danger level estimates (nowcast, $D_{LN}$ )

Observers of the Swiss avalanche warning service with sufficient experience and presence in avalanche terrain provide an estimate of the avalanche danger level together with their observations. They use the five-level European avalanche danger scale and in addition may indicate whether or not they expect natural avalanches at danger level 3-Considerable. The observers are advised to integrate all available information into their local estimate of the danger level ( $D_{LN}$ ), including not just the observations from the day of observation, but also prior knowledge concerning the development of the snowpack during the winter or information from third parties. To assure consistent and high quality feedback, all observers are regularly trained.

The avalanche danger is assessed locally. The area considered is the area of observation during the day in the backcountry or in the ski area, or the area that can be seen from the observation point in the valley floor; this area is approximately 10 km<sup>2</sup> (Jamieson et al., 2008) to 25 km<sup>2</sup> (Meister, 1995). In addition to estimating the danger level, the type of avalanche (dry- or wet-snow) is reported. The estimated danger level for dry-snow slab avalanches should reflect the current situation and is therefore a local nowcast, while for wet-snow avalanches the highest expected danger level during the day is reported. Furthermore, the slope aspects and elevations where the danger is most pronounced (danger rose) are indicated by the observer. We used local avalanche danger level estimates of current conditions reported between 11:00 and 22:00 of that day. We considered all local danger estimates related to dry-snow avalanches in the Swiss Alps during the nine winter seasons between 2008-2009 and 2016-2017. This resulted in 9,553 individual avalanche danger estimates. These estimates were reported either via a website (*IFKIS*; Bründl et al., 2004,  $N = 1,774$ , 19%) or a mobile app (*mAvalanche*; Suter et al., 2010,  $N = 6,531$ , 68%). In addition, for observers who did not report their field observations via *IFKIS* or *mAvalanche*, we screened the danger assessments reported with snow profile observations ( $N = 1,248$ , 13%). Observations were not distributed evenly across the Swiss Alps, with the most prominent cluster in the region of Davos (Fig. A.10) where the SLF and the national avalanche warning service is located.

Even though the focus was on analyzing  $D_{LN}$  estimates from the backcountry, we included  $D_{LN}$  estimates made by study-plot observers from the valley floor ( $N = 1,971$ , 55 different observers) or by observers

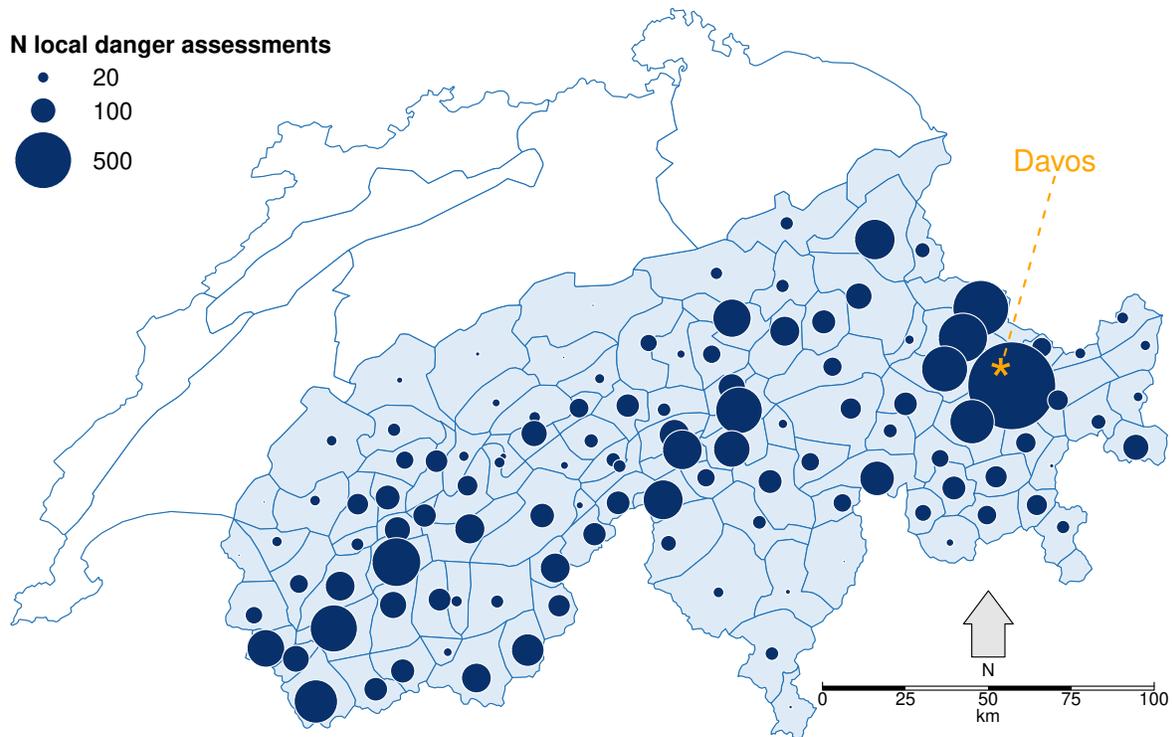


Figure A.10: Map of Switzerland showing number of local danger level estimates per warning region (polygons colored light-blue). The size of the dark-blue circles corresponds to the number of local danger level estimates  $D_{LN}$  for each of the 117 warning regions (total forecast area: 26,400 km<sup>2</sup>; nine winters,  $N = 9553$ ). The national avalanche warning service is located at SLF in Davos.

based in ski areas ( $N = 1,423$ , at least 15 different observers) during the day. In many cases, these groups will rely on different observations when assessing the local avalanche danger. For instance, obvious visual clues such as avalanche activity or blowing snow may be of high relevance to study-plot observers without access to avalanche terrain, while ski area observers will additionally incorporate results obtained through avalanche control by explosives. However, even though ski area observers partly work in avalanche terrain, in many cases they are limited to frequently tracked and controlled terrain. Both, valley floor and ski area observers are attached to a particular place, observing and reporting from the same warning region throughout the winter. In contrast, SLF forecasters and researchers will often, but not always, combine a field day with snow pit observations specifically targeting unstable areas, i.e. provide «roving» information (Jamieson et al., 2008). Mountain guides, on the other hand, are responsible for their clients and may put great emphasis on finding the best skiing in safe conditions. In our data set, mountain guides are the spatially most flexible of the observer groups.

### A.3.2.2 Regional avalanche danger level forecasts (forecast, $D_{RF}$ )

In Switzerland, the public bulletin is issued daily during winter by the avalanche warning service at SLF. Publication frequency is twice per day during the main winter season: in the evening at 17:00 valid until 17:00 the following day, and updated the next morning at 08:00 valid until 17:00 the same day.

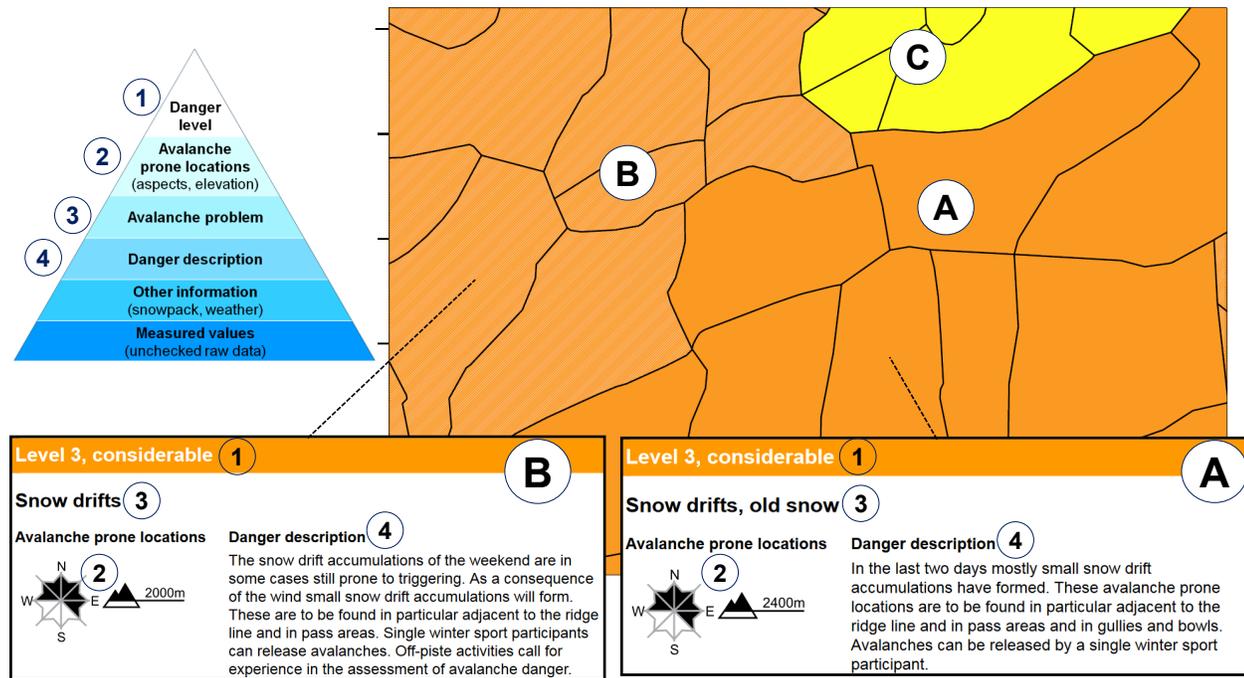


Figure A.11: Map showing an extract of the Swiss avalanche bulletin published on the morning of 7 March 2016 (70 km  $\times$  50 km). The components of the information pyramid (upper left) are highlighted for two regions (A and B) with the same danger level ( $D_{RF} = 3$ -Considerable, different orange shading), but a (partly) different distribution of the avalanche prone locations, avalanche problems and danger description. In this study, we compared observers between warning regions, where the elements 1 to 4 were either the same (e.g. within region A), or where the danger level differed (e.g. between A and C, danger level in C: 2-Moderate). The individual warning regions (the polygons) are normally not visible in the bulletin, but are shown to highlight the aggregation of several warning regions to one region with the same danger rating (elements 1 to 4).

The Alpine warning region comprises an area of 26,400 km<sup>2</sup>, which is considered large according to the classification by Jamieson et al. (2008). This area is divided into 117 sub-areas (hereafter called warning regions) with a mean size of 225 km<sup>2</sup> (Fig. A.10). While a danger level is given for the whole forecast area, i.e. each of the 117 warning regions, the warning regions are not explicitly used in the avalanche bulletin, since they are aggregated to larger areas with similar avalanche conditions (Ruesch et al., 2013; Winkler et al., 2013).

The forecast includes information concerning the forecast regional avalanche danger level ( $D_{RF}$ ), the avalanche problem(s), the slope aspects and elevations where the danger is most pronounced (danger rose) and the danger description (Fig. A.11), which is a text describing the avalanche situation created by using a catalogue of phrases (Winkler and Kuhn, 2017). In addition, a text bulletin describing weather and snowpack conditions and the trend for the following two days is issued.

For this analysis, we used the forecast regional danger level describing the dry-snow avalanche situation for the same nine winter seasons mentioned above. Primarily, we used the morning forecast ( $D_{RF}$ ). Moreover, to assess the difference in forecast performance between the evening forecast and the morning forecast, we also used the danger level issued in the evening forecast.

The forecast danger level  $D_{RF}$  for dry-snow avalanches was level 1-Low on 14% of the days and regions,

level 2-Moderate on 44%, level 3-Considerable on 41% and level 4-High on 1%. Danger level 5-Very high was not forecast during the study period.

### A.3.3 Methods

#### A.3.3.1 Variations in local danger level estimates between observers

We analyzed the variations in local avalanche danger level estimates  $D_{LN}$  between observers on the same day at the scale of the smallest spatial unit used in the Swiss avalanche bulletin. Even though observers report other descriptors of danger such as the avalanche problem or avalanche prone locations, we only considered the danger level. An example is shown in Fig. A.12 where the estimates by observer A inside the polygon with outlines in bold are compared to other nearby observers, first in the same warning region, i.e. to observer B. Moreover, we included observer pairs at distances less than 10 km from each other, but in adjacent warning regions with the same forecast danger level and the same danger description (pair A-C in Fig. A.12, within region A in Fig. A.11). The latter restriction was introduced to ensure the greatest possible consistency in avalanche conditions between neighboring warning regions. To quantify whether variability increased with distance in forecast areas with the same danger, we compared observer estimates at distances between 10 and 20 km (e.g. A-D in Fig. A.12), 20 and 30 km, and between 30 and 50 km from each other. In addition, we compared observer estimates at distances less than 10 km (e.g. A-E in Fig. A.12), and between 10 and 20 km, but between warning regions with a different danger level. The aim of the latter analysis was to explore whether the boundary between regions of different danger level was appropriate.

We calculated the difference in local danger level estimates between all above-mentioned observer pairs using the integer values assigned to the five danger levels (1-Low, 2-Moderate, 3-Considerable, 4-High, 5-Very High) following the approach by Jamieson et al. (2008). The difference in the danger level estimate of an observer  $i$  ( $D_{LN}(i)$ ) compared to other observers' estimates ( $D_{LN}(j)$ ,  $j=1,2,3,\dots,n$ ) is then

$$\Delta D_{LN} = D_{LN}(i) - D_{LN}(j) \quad (\text{A.6})$$

If  $\Delta D_{LN} = 0$ , we called it an agreement, else a disagreement.

The proportion of disagreements ( $P_{\text{disagree}}$ ), which we refer to as the disagreement rate, for an observer  $i$  is therefore the ratio of the number of disagreements  $N(\Delta D_{LN} \neq 0)$  to the number of all comparisons between  $i$  and other observers  $N(\Delta D_{LN})$ :

$$P_{\text{disagree}}(i) = \frac{N(\Delta D_{LN} \neq 0)}{N(\Delta D_{LN})}, \quad (\text{A.7})$$

where  $N$  is simply a counting function.

To explore whether  $\Delta D_{LN}$  was equally often higher or lower or whether a bias existed for observer  $i$  compared to others, we calculated a bias for each observer:

$$P_{LN(\text{higher-lower})}(i) = \frac{N(\Delta D_{LN} > 0) - N(\Delta D_{LN} < 0)}{N(\Delta D_{LN})} \quad (\text{A.8})$$

Furthermore, we explored whether a significant bias towards lower or higher disagreements existed. To this end, we calculated the proportion of equally distributed disagreements ( $N(\Delta D_{LN} > 0) = N(\Delta D_{LN} < 0)$ ),

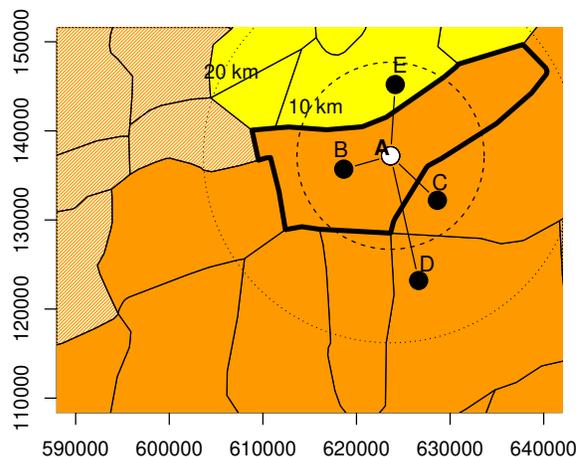


Figure A.12: Map showing an extract of 2000 km<sup>2</sup> of the Swiss Alps (Swiss coordinates in m, 50 km x 40 km) with the color corresponding to the forecast danger level (orange  $D_{RF} = 3$ -Considerable and yellow  $D_{RF} = 2$ -Moderate, forecast issued in the morning of 7 March 2016, example corresponds to bulletin shown in Fig. A.11). The polygons denote the individual warning regions. The polygon with outlines shown in bold is an exemplary warning region with the observer pair combinations we explored.

and the unbalanced disagreements  $\max(N(\Delta D_{LN} > 0), N(\Delta D_{LN} < 0)) - \min(N(\Delta D_{LN} > 0), N(\Delta D_{LN} < 0))$ . While the first, the equally distributed disagreements may be interpreted as random, a significant proportion of unbalanced disagreements may indicate an observer-specific bias. At its most extreme,  $|P_{LN(\text{higher-lower})}| = P_{\text{disagree}}$  would indicate that all disagreements were either higher or lower. We tested whether the larger number of  $N(\Delta D_{LN} > 0)$  or  $N(\Delta D_{LN} < 0)$  deviated significantly from a balanced distribution of disagreements for the combined  $N(\Delta D_{LN} \neq 0)$  using the chi-square based non-parametric proportion test (R Core Team, 2017). This was calculated in two ways: first, using the original data for observers with at least 20 comparisons to others. Here, it is of importance to note that the calculation of the  $p$ -value is sensitive to both the absolute number of disagreements as well as the proportion of unbalanced disagreements. This may result in significant  $p$ -values for observers with a large absolute number of disagreements despite comparably low  $|P_{LN(\text{higher-lower})}|$ , and vice versa. Thus, in addition, we resampled the data with replacement, which we describe below, and calculated the proportion test based on the mean of the resampled data standardized to 100 comparisons for each observer. As the number of comparisons was less than 100 for most of the observers, we are aware that this increases the likelihood to observe a significant  $p$ -value. Therefore, we present these statistics primarily to highlight the differences between both approaches.

### A.3.3.2 Bootstrap sampling

We applied bootstrap sampling techniques to the sample distribution with the aim to infer robust information about the central tendency (mean) and the variability in the sample (standard deviation). From the original sample of size  $N$  we randomly selected  $n$  data units allowing replacement (Wilks, 2011, pp. 172-173). The resampling procedure was repeated 1000 times. For each of the resampled datasets, the selected statistic, in our case the mean or the standard deviation was calculated resulting in a bootstrap distribution of, for

instance, means. The mean of the bootstrap distribution represents a robust mean (and its error) of the original sample; resampled results are marked with two asterisks, e.g.  $P_{disagree}^{**}$ .

### A.3.3.3 Comparing local nowcasts to regional forecasts

Similar to the procedure described above, we calculated the difference between regional forecast  $D_{RF}$  and local nowcast  $D_{LN}$  using the integer values of the danger level

$$\Delta D = D_{RF} - D_{LN} \quad (\text{A.9})$$

If the danger levels agreed ( $\Delta D = 0$ ), we refer to this case as a hit.

The proportion correct ( $P_{\text{correct.raw}}^5$ ) was therefore the ratio of the number of hits  $N(\Delta D = 0)$  to the number of all comparisons  $N(\Delta D)$  between local nowcasts  $D_{LN}$  and regional forecast  $D_{RF}$

$$P_{\text{correct.raw}} = \frac{N(\Delta D = 0)}{N(\Delta D)} \quad (\text{A.10})$$

The forecast bias was calculated

$$P_{\text{over-under}} = \frac{N(\Delta D > 0) - N(\Delta D < 0)}{N(\Delta D)}. \quad (\text{A.11})$$

Again, unbalanced proportions were tested using the proportion test as described above.

As we intended to explore whether  $P_{disagree}$  and  $P_{\text{correct.raw}}$  differed depending on the forecast danger level - the ease of perceiving the hazard -, we divided the data into groups: First, by the forecast danger level (four levels, as danger level 5-Very High was not forecast during the study period), and second whether the forecast danger level had changed to the previous day into the groups increasing danger ( $D_{RF}^{\text{increase}}$ ), no change in danger rating ( $D_{RF}^{\text{no.change}}$ ), and a decreasing danger ( $D_{RF}^{\text{decrease}}$ ).

Splitting  $D_{RF}$  into these groups was motivated by the fact that  $D_{RF}^{\text{increase}}$  is often a forecast in a comparably dynamically evolving situation due to changing weather and the associated expected changes to snow stability, and before changes in the snowpack or weather have been observed or measured (McClung, 2000). In contrast,  $D_{RF}^{\text{no.change}}$  and  $D_{RF}^{\text{decrease}}$ , rely more on the combination of observed evidence concerning the current conditions and comparably (minor) or slow changes in snowpack stability. It is of note that, while  $D_{RF}$  did not change, on 50% of these days the particularly avalanche prone locations, i.e. the slope aspects and elevations where the danger was highest as indicated in the danger rose, changed from one day to the next. Similarly,  $D_{RF}^{\text{decrease}}$  will often be based on information obtained from field observations (and also on  $D_{LN}$  estimates). While the danger level decreases on a particular day by one step from, for instance, 3 Considerable to 2-Moderate, the actual avalanche danger rather decreases smoothly. Hence, the decrease by one step may actually reflect an evolution that took several days. However, the discrete nature of the avalanche danger scale does not allow expressing the gradual decrease. Therefore, the decrease by one step, indicating a jump from one level to a lower one on a particular day, might actually be a rather small decrease from, for instance, a low danger level 3-Considerable to a high danger level 2-Moderate.

We compared  $\Delta D$  using (a) all individual comparisons, (b) days and regions when observers unanimously

<sup>5</sup>The variable notation has been adjusted to match the variable definitions in the Methods section of the Synthesis (Eqn. 4.4, p. 36 in Section 4.1.1).

agreed on  $D_{LN}$  or when a majority  $D_{LN}$  estimate existed, and (c) considering the reporting bias, with proportionally fewer  $D_{LN}$  estimates at lower forecast danger levels and more at higher  $D_{RF}$ , and the disagreement rate  $P_{disagree}$ . To incorporate  $P_{disagree}$ , we made the simplifying assumption that the disagreement was always by one danger level if  $\Delta D \neq 0$  since deviations of more than one danger level were rare (see below).

To consider the reporting bias and the disagreement rate  $P_{disagree}$  we proceeded as follows:

- As outlined in the bootstrap section before, we randomly selected  $n$  data units from the original sample allowing replacement, but using the distribution of the forecast  $D_{RF}$  (1-Low 14%, 2-Moderate 44%, 3-Considerable 41%, 4-High 1%), and whether  $D_{RF}$  had changed from the day before ( $D_{RF}^{no.change}$  80%,  $D_{RF}^{increase}$  and  $D_{RF}^{decrease}$  each 10%) as selection weights for a subset  $M$ .
- For  $M_1$ , for days and regions of  $M$ , when observers unanimously agreed on  $D_{LN}$  or when a majority  $D_{LN}$  estimate existed, we used the majority  $D_{LN}$  estimate. For  $M_2$ , the remaining days and regions of  $M-M_1$ , we again performed bootstrap sampling as outlined above and randomly assigned to  $(100 * P_{disagree})\%$  of  $M_2$  that the  $D_{LN}$  estimate was correct. For example, as will be shown below,  $P_{disagree}$  was 26% for days with  $D_{RF} = 3$  and  $D_{RF}^{increase}$ . In this case, a random 74% of the samples'  $D_{LN}$  estimates would be considered correct. For the remaining proportion of comparisons, we assumed for half of the  $D_{LN}$  estimates that the rating was correct and for the other half that the rating was different. As outlined above, the difference was at most one level to  $D_{RF}$  and one level to  $D_{LN}$ . For cases when a higher or lower deviation from  $D_{LN}$  were possible, we used the observed distributions of  $\Delta D$  shown in the result section. Step 2 was repeated 10 times.
- Step 1 and Step 2 were repeated 10 times and the mean and standard deviations of these repetitions calculated. Results using this approach are described using  $P_{correct}^{**}$ .

Statistical test results were considered significant if  $p \leq 0.05$ . All analyses were performed using the statistics software *R* (R Core Team, 2017).

## A.3.4 Results

### A.3.4.1 Local danger level estimates

#### Observer-specific variations

1673 local danger rating pairs between 118 observers within the same warning region on the same day were analyzed. These comparisons originate from 653 days in 77 out of the 117 different warning regions. In 20% of the cases, more than two observers reported  $D_{LN}$  in the same warning region and for the same day. 45 out of the 118 observers had more than 20 comparisons to other observers and 7 more than 100 comparisons. In 90% of the cases, where the exact location was known, the distance between observers was 11 km or less (median 5.2 km, Table A.7). The disagreement rate  $P_{disagree}$  was lowest within the same warning region (22%) or at distances less than 10 km in neighboring warning regions with the same danger rating (23%, Table A.7). If observers were in neighboring warning regions with the same danger rating, but at distances greater than 10 km, the disagreement rate was around 30% with no further decrease with increasing distances.

Table A.7: Disagreement rate  $P_{\text{disagree}}$  between observer pairs with respect to the location of the observers (within the same warning region or a neighboring warning region, or the distance between observers) and the forecast danger. Same danger means that danger level, avalanche prone locations, avalanche problems and the danger description were identical (see Fig. A.11). Different danger means different danger level. The number of pairs ( $N$ ) and the median distance is given.

Warning region or distance between observers	Danger	$P_{\text{disagree}}$	$N$	distance (km)
same warning region	same	22%	1673	5.2
neighboring warning region	same	28%	3385	15.5
distance < 10 km	same	23%	2326	5.8
distance 10-20 km	same	30%	2139	15.1
distance 20-30 km	same	28%	2295	25.2
distance 30-50 km	same	31%	3383	39.9
neighboring warning region	different	51%	395	19.1
neighboring warning region, distance < 10 km	different	40%	65	7.1
neighboring warning region, distance 10-20 km	different	49%	144	15.5

As can be noted in Figure A.13,  $P_{\text{disagree}}$  varied considerably between observers. In fact, 8 out of the 40 observers with more than 20 comparisons to other observers had a disagreement rate  $P_{\text{disagree}} \geq 30\%$ . 37% of the disagreements were unbalanced for observer pairs within the same warning region (39% for the resampled data, Fig. A.13). For some observers all the disagreements were unbalanced ( $|P_{\text{LN}(\text{higher-lower})}| = P_{\text{disagree}}$ ), corresponding to the points on the dotted lines in Figure A.13). Testing whether the disagreements were significantly unbalanced, compared to an equal distribution of disagreements, showed that 2 observers exhibited a significant bias towards either higher or lower  $D_{\text{LN}}$  estimates (Fig. A.13a). Including comparisons with observers in neighboring warning regions with the same danger rating, 9 (or 12%) out of the 75 observers with more than 20 comparisons to others had a significant bias (Fig. A.13b).

Due to the considerable differences in the number of comparisons for each observer, the bias was significant for some observers with a comparably lower absolute bias compared to others. As an example, the observer marked with an A had a comparably large number of comparisons to others ( $N = 292$ ) and a disagreement rate relatively close to the overall mean ( $P_{\text{disagree}} = 25\%$ ) with a  $P_{\text{LN}(\text{higher-lower})}$  of 13% (Fig. A.13a). While this bias was significant for observer A ( $p = 0.002$ ), a similar or larger  $P_{\text{LN}(\text{higher-lower})}$  was not significant for 8 of 10 other observers with (considerably) fewer comparisons to others. However, testing the unbalanced proportion of disagreements on samples standardized to 100 observations for each observer, observer A would not be considered biased (Fig. A.13c). Using this latter approach, 13 (or 17%) out of the 75 observers would be considered as being significantly biased ( $p \leq 0.05$ ).

### Group-specific variations

Exploring the disagreement rate within groups of observers and for observer pairs within the same warning region, showed no significant differences in  $P_{\text{disagree}}$  within the group SLF ( $P_{\text{disagree}} = 22\%$ ,  $N = 86$ , em-

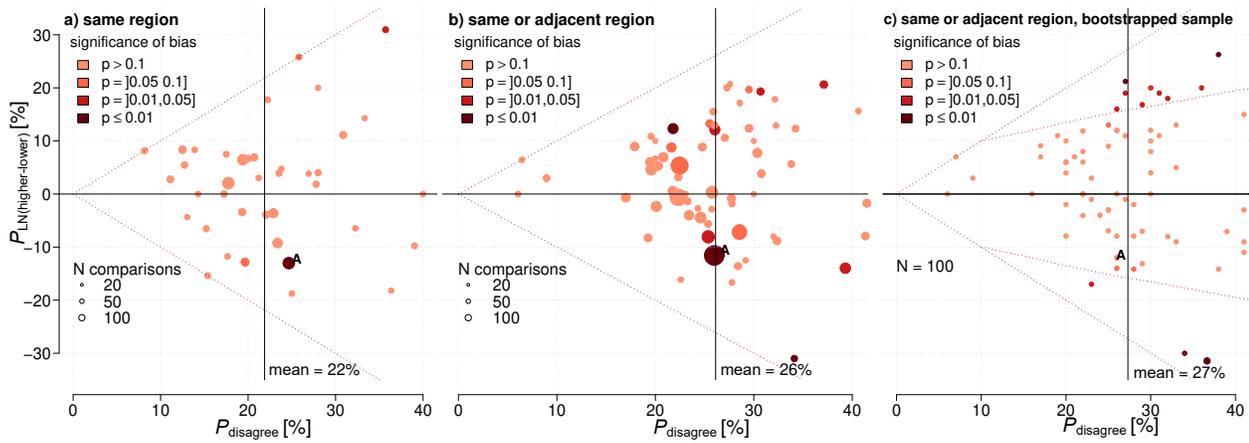


Figure A.13: The disagreement rate  $P_{\text{disagree}}$  and the bias ( $P_{\text{LN}(\text{higher-lower})}$ ) between local danger level estimates for each observer. (a) within the same warning region, (b) and (c) within the same or an immediately neighboring region with the same danger level and description. In (a) and (b) the size of the circles corresponds to the number of comparisons, whereas in (c) the resampled data are standardized to 100 comparisons. Points on the dotted lines indicate that all disagreements are either higher or lower for this observer ( $|P_{\text{LN}(\text{higher-lower})}| = P_{\text{disagree}}$ ). Color coding corresponds to significance levels.

ployees at SLF, forecasters and researchers, mostly in the surroundings of Davos) compared to the group guides (*mAvalanche* network, without SLF), regardless whether this was compared for the region of Davos ( $P_{\text{disagree}} = 24\%$ ,  $N = 55$ ) or the whole Swiss Alps ( $P_{\text{disagree}} = 22\%$ ,  $N = 516$ ).

Comparing the disagreement rate between the estimates made after a day in the backcountry and those by observers in the valley floor ( $N = 201$ ) or in ski areas ( $N = 325$ ), showed very similar values (22% and 23%, respectively). However, valley floor  $D_{\text{LN}}$  estimates were significantly more often higher than those made based on observations in the backcountry (18% higher and 4% lower;  $p < 0.01$ ). Estimates made by ski area staff also tended to be lower than those by observers from the backcountry; however, the difference was not significant (14% higher, 8% lower).

#### A.3.4.2 Variations with regard to the forecast regional danger level

In addition to the group-specific variations, we explored whether  $P_{\text{disagree}}$  varied with the forecast danger level  $D_{\text{RF}}$  and the change in the forecast to the previous day. Local danger level estimates were reported significantly less often on days with forecast danger level 1-Low (7% vs. 14%, observer and forecast, respectively,  $p < 0.001$ ) and more often at danger level 3-Considerable (51% vs. 42%,  $p < 0.001$ ).  $D_{\text{RF}}$  did not change in 80% of the days and warning regions from one day to the next.  $D_{\text{RF}}$  decreased by one level on 10% and by two levels on 0.03% of the days, while it increased by one level on 9% of days and by two levels 0.4% of the days.

As shown in Tab. A.8, and using comparisons within the same warning region or at distances less than 10 km from each other in regions with the same danger description,  $P_{\text{disagree}}$  was highest on days when  $D_{\text{RF}}$  increased ( $27\% \pm 3\%$ , mean and standard deviation) and on days with a  $D_{\text{RF}}$  4-High ( $27\% \pm 7\%$ ), and lowest on days when  $D_{\text{RF}}$  decreased ( $14\% \pm 2\%$ ).  $P_{\text{disagree}}$  was particularly low on days when the danger

Table A.8: Disagreement rate  $P_{\text{disagree}}$  within the same warning region or in neighboring warning regions with the same danger rating at distances less than 10 km, with respect to the forecast regional danger level  $D_{\text{RF}}$  and whether  $D_{\text{RF}}$  changed from the previous day. The arrow-symbols indicate whether  $D_{\text{RF}}$  increased ↗, stayed the same → or decreased ↘. The mean and the standard deviation of the disagreement rate  $P_{\text{disagree}}$ , and the number of pairs  $N$  are given.

$D_{\text{RF}}$	mean( $P_{\text{disagree}}$ )				standard deviation ( $P_{\text{disagree}}$ )				$N$			
	↗	→	↘	all	↗	→	↘	all	↗	→	↘	all
1-Low	–	13%	14%	13%	–	4%	7%	3%	–	70	28	98
2-Moderate	30%	24%	15%	22%	7%	2%	3%	1%	43	736	168	947
3-Considerable	26%	25%	6%	24%	4%	1%	4%	1%	137	1228	35	1400
4-High	22%	37%	–	27%	10%	12%	–	7%	18	16	–	34
all	27%	24%	14%	23%	3%	1%	2%	1%	198	2050	231	2479

level was lowered from level 4-High to level 3-Considerable ( $6\% \pm 1\%$ ).  $P_{\text{disagree}}$  was significantly different between days when  $D_{\text{RF}}$  increased and those when  $D_{\text{RF}}$  decreased (27% vs. 14%,  $p = 0.04$ ) and when  $D_{\text{RF}}$  was 1-Low and 4-High (13% vs. 27%,  $p = 0.02$ ).

In situations, when two observers were in close proximity but in neighboring warning regions with differing  $D_{\text{RF}}$ , the disagreement rate was 40% for distances less than 10 km and 49% for distances between 10 and 20 km (Table A.7).

### A.3.4.3 Comparing local nowcasts to regional forecasts

In total, 9543 individual comparisons between local danger level estimates  $D_{\text{LN}}$  and regional danger level forecasts  $D_{\text{RF}}$  were analyzed. The estimates were provided by 137 different observers on 1076 days and in 115 warning regions.

The proportion correct ( $P_{\text{correct.raw}}$ ) was 76% (Fig. A.14a). If the forecast was different from the local estimate, then generally the difference was one danger level. In only 0.5% of the comparisons  $D_{\text{RF}}$  was two levels too high or too low.  $D_{\text{LN}}$  was more often lower than  $D_{\text{RF}}$  with 20%  $D_{\text{RF}}$  too high vs. 4%  $D_{\text{RF}}$  too low.  $D_{\text{RF}}$  was most frequently considered too low when  $D_{\text{RF}}$  decreased (11%, Tab. A.9). In contrast,  $D_{\text{RF}}$  was most often considered too high when  $D_{\text{RF}}$  increased (37%). The proportion correct was lowest on days when  $D_{\text{RF}}$  increased ( $P_{\text{correct.raw}} = 61\%$ ) or when the forecast danger level was 4-High ( $P_{\text{correct.raw}} = 28\%$ , Tab. A.9). The latter would indicate that the forecast danger level was perceived mostly as being incorrect. On the opposite side,  $P_{\text{correct.raw}}$  was highest when the danger level decreased or generally at lower danger levels of 1-Low and 2-Moderate.

For days, when observers were in a warning region which was neighboring one with a different danger level, observers disagreed often with  $D_{\text{RF}}$  when in the region with the higher danger rating ( $P_{\text{correct.raw}} = 51\%$ ). In contrast, when in the region with the lower rating, observers frequently estimated  $D_{\text{LN}}$  the same as  $D_{\text{RF}}$  ( $P_{\text{correct.raw}} = 84\%$ ).

Considering each observer individually revealed large scatter (Fig. A.15). While almost all observers tended to estimate the local danger to be lower than forecast, the frequency on which they considered  $D_{\text{RF}}$  to be

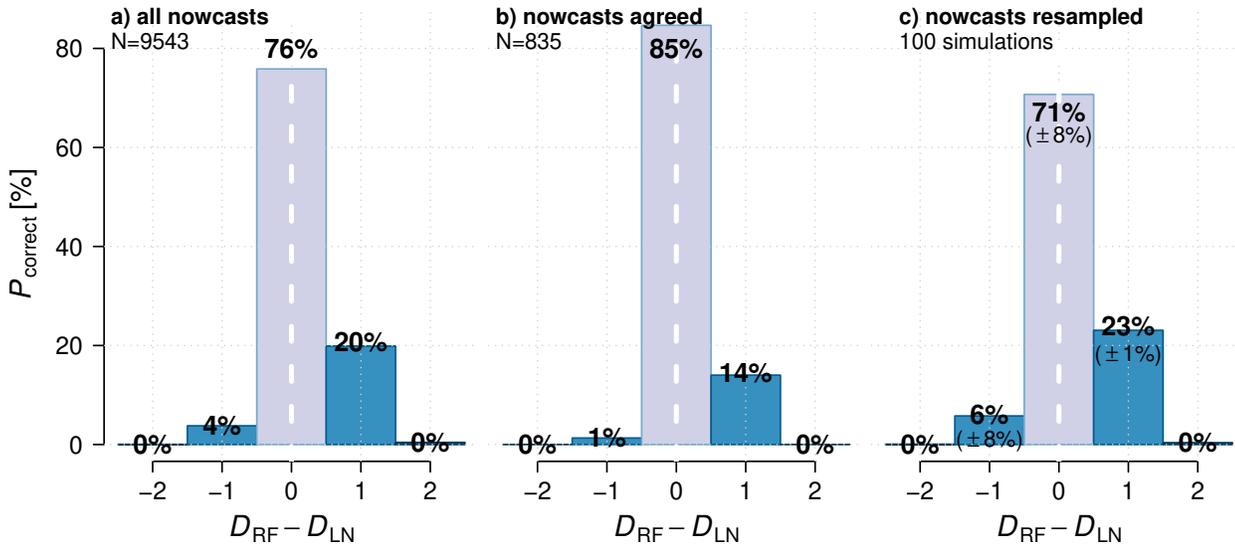


Figure A.14: Distributions of the differences between the local danger level estimate ( $D_{LN}$ , nowcast) and the regional danger level forecast ( $D_{RF}$ , forecast) for (a) all estimates individually compared to the forecast, (b) for days and regions when the observers in the same region agreed on the danger level, and (c) for a resampled dataset of the observed comparisons, but incorporating (1) the reporting bias and the proportion of days with a higher or lower  $D_{RF}$  (Table A.9) and (2) the disagreement rate  $P_{\text{disagree}}$  between observers (Table A.8).

Table A.9: Proportion correct  $P_{\text{correct.raw}}$  between  $D_{RF}$  and  $D_{LN}$ , proportion of  $\Delta D > 0$  and  $\Delta D < 0$  in relation to the forecast regional danger level and whether  $D_{RF}$  changed to the day before ( $N = 9543$ ). The arrow-symbols indicate whether  $D_{RF}$  increased  $\nearrow$ , stayed the same  $\rightarrow$  or decreased  $\searrow$ .

$D_{RF}$	$P_{\text{correct.raw}}$				$P(\Delta D > 0)$				$P(\Delta D < 0)$			
	$\nearrow$	$\rightarrow$	$\searrow$	all	$\nearrow$	$\rightarrow$	$\searrow$	all	$\nearrow$	$\rightarrow$	$\searrow$	all
1-Low	–	89%	80%	86%	–	–	–	–	–	11%	20%	14%
2-Moderate	47%	79%	87%	79%	39%	16%	2%	14%	14%	5%	11%	6%
3-Considerable	67%	73%	92%	73%	33%	27%	7%	27%	0%	0%	1%	0%
4-High	36%	20%	–	28%	64%	80%	–	72%	0%	0%	–	0%
all	61%	76%	86%	76%	37%	21%	3%	20%	2%	3%	11%	4%

wrong by one danger level varied considerably.

The proportion correct was almost identical for those working at SLF ( $P_{\text{correct.raw}} = 74\%$ ,  $N = 1,047$ ) as for other observers and mountain guides ( $P_{\text{correct.raw}} = 76\%$ ,  $N = 8,489$ ). However, if just the avalanche forecasters at SLF were considered as a group, a slightly higher proportion correct was noted (80%,  $N = 417$ ). Expanding the comparison to the estimates made during the day in the valley floor ( $P_{\text{correct.raw}} = 87\%$ ,  $N = 1971$ , 55 different observers) or by observers working in ski areas ( $P_{\text{correct.raw}} = 82\%$ ,  $N = 1423$ , at least 15 different observers) confirmed the variation between observer groups as much as between individual observers. Comparing just the days and regions, when estimates made in the valley floor and after a day in the backcountry were available ( $N = 201$ ), valley floor observers estimated  $D_{LN}$  significantly often higher than the field observers ( $p < 0.01$ ). Although ski area observers were also more often lower

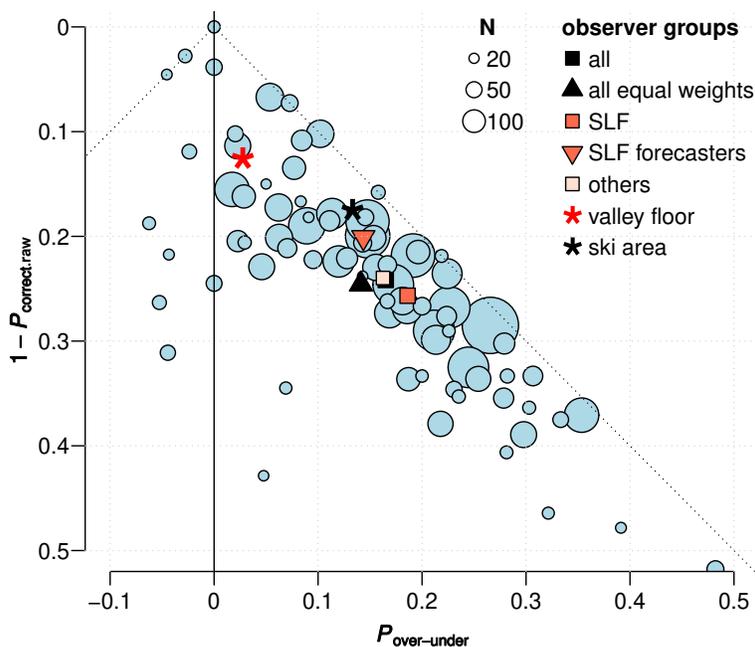


Figure A.15: For each observer, the proportion of days with a local estimate being different than the regional forecast  $1 - P_{\text{correct,raw}}$  and the bias  $P_{\text{over-under}}$  is shown. The dotted lines correspond to  $P_{\text{over-under}} = (1 - P_{\text{correct,raw}})$  indicating that all differences between ( $D_{\text{LN}}$  and  $D_{\text{RF}}$  would either be lower or higher. Values for the mean of all afternoon backcountry observers (weighted by the number of observations = «all» and with equal weight for each observer «all equal weight»), for the subset of SLF employees, SLF forecasters and backcountry excluding all SLF staff «other» are shown. For comparison, mean values for estimates made from valley floor observers and ski area staff are added.

in their local danger level estimate than observers reporting from the backcountry, this difference was not significant ( $N = 325$ ). Regardless, which of these groups was considered, the tendency towards lower local estimates compared to  $D_{\text{RF}}$  was confirmed.

The forecast danger level changed on 19.7% of the days and regions in the afternoon forecast (17:00), compared to 2.7% in the morning forecast (08:00; 1.7% up, 1% down). The local estimates made after a day in the backcountry showed a marginally, and not significantly higher agreement with the morning forecast ( $P_{\text{correct,raw}} = 75.9\%$ ) than with the evening forecast of the previous day ( $P_{\text{correct,raw}} = 75.3\%$ ).

Considering only days and regions when two or more observers agreed in their nowcast estimate or when there was a majority opinion on  $D_{\text{LN}}$ , the agreement with the forecast  $D_{\text{RF}}$  was higher ( $P_{\text{correct,raw}} = 85\%$ ,  $N = 835$ , Fig. A.14b). However, as can be seen in Table A.10, the values are rather extreme and the proportion correct ranges from 0% to 100%. We attribute this to the relatively small sample size in some of the cells in Table A.10.

Incorporating the reporting bias (Tab. A.9) and the disagreement rate in the calculation (Tab. A.8) and using the full sample ( $N = 9,543$ ),  $P_{\text{correct}}^{**}$  was 71% ( $\pm 8\%$ , Fig. A.14c, Tab. A.11). Comparing Tables A.9 and A.11 shows that the proportion correct increased for days when the proportion correct in Table A.9 and  $P_{\text{disagree}}$  in Table A.8 were low (for instance for days with  $D_{\text{RF}} = 4\text{-High}$ ). In contrast, comparably high  $P_{\text{correct,raw}}$  values (e.g. for  $D_{\text{RF}} = 2\text{-Moderate}$  or  $1\text{-Low}$ , Table A.8) decreased somewhat.

Table A.10: Proportion correct  $P_{\text{correct,raw}}$  between  $D_{\text{RF}}$  and  $D_{\text{LN}}$  for days and regions, when observers either unanimously agreed on  $D_{\text{LN}}$  or when a majority opinion existed and depending on the forecast regional danger level and whether  $D_{\text{RF}}$  changed to the day before in the same region ( $N = 835$ ). The arrow-symbols indicate whether  $D_{\text{RF}}$  increased ↗, stayed the same → or decreased ↘.

$D_{\text{RF}}$	↗	→	↘	all
1-Low	–	100%	100%	98%
2-Moderate	33%	88%	98%	88%
3-Considerable	82%	82%	100%	82%
4-High	17%	0%	–	7%
all	67%	84%	99%	85%

Table A.11: Proportion correct  $P_{\text{correct}}^{**}$  between  $D_{\text{RF}}$  and  $D_{\text{LN}}$ , proportion of  $\Delta D > 0$  and  $\Delta D < 0$  incorporating the disagreement matrix (Tab. A.8) and the reporting bias, the frequency of  $D_{\text{RF}}$  and whether  $D_{\text{RF}}$  changed to the day before in the same region. The arrow-symbols indicate whether  $D_{\text{RF}}$  increased ↗, stayed the same → or decreased ↘. Cell values represent the mean of 100 repetitions.

$D_{\text{RF}}$	$P_{\text{correct}}^{**}$				$P(\Delta D > 0)^{**}$				$P(\Delta D < 0)^{**}$			
	↗	→	↘	all	↗	→	↘	all	↗	→	↘	all
1-Low	–	83%	73%	81%	–	–	–	–	–	17%	27%	19%
2-Moderate	52%	72%	80%	72%	36%	22%	6%	20%	15%	7%	10%	7%
3-Considerable	62%	67%	88%	67%	37%	33%	11%	33%	0%	0%	1%	0%
4-High	41%	32%	–	39%	59%	68%	–	61%	0%	0%	–	0%
all	58%	71%	79%	71%	39%	24%	5%	23%	3%	5%	16%	6%

## A.3.5 Discussion

### A.3.5.1 Local danger level estimates: variability and bias

Even though the observers were often in relatively close proximity (in 90% of the cases less than 11 km from each other), 22% of the local danger level ratings disagreed within the same warning region. There may be several explanations for this variability.

Avalanche conditions may vary even at the relatively small scale of a warning region with an average size of just 200 km<sup>2</sup> (Schweizer et al., 2003). Variations may also be due to where the observations were made. For instance, if some of the observations were made in frequently tracked terrain (for instance, close to ski areas) and some in less frequently tracked terrain (for instance, a forecaster or researcher searching for instability), or if some observers traveled in more favorable aspects and others in more unfavorable aspects and elevations, variation in the perception of the hazard may be expected resulting in different ratings. In fact, Schweizer et al. (2003) showed that the danger level differs between slope aspects and elevations where the danger was most prominent and the rest of the terrain by often half a danger level, sometimes

even one danger level. Accordingly, often a one-step lower danger level may be assumed in frequently tracked terrain when, for example, applying the Graphical Reduction Method (Harvey et al., 2012).

Moreover, as shown by Haladuick (2014), even if several observers worked together and used the same observations, they disagreed on the danger level in 7% of the cases. This discrepancy may be attributed to the discrete nature of the avalanche danger scale where observers have to decide on one specific level in their reporting form, even if they consider the danger level to be somewhere in between two danger levels. We therefore suggest considering that experienced observers can report intermediate danger levels. However, the discrepancy might also be due to the fact that the avalanche danger scale as well as the process of locally assessing the danger level are not fully defined and can be interpreted differently - even by experienced forecasters (Müller et al., 2016).

We noted the highest disagreement rate at danger level 4-High (27%), and on days when the danger level was forecast to increase (27%). This finding was rather surprising since we assumed that in particular at danger level 4-High, clear evidence of the prevailing danger exists so that ratings should rather agree - in accordance with McClung (2002a) who argued that in situations with widespread instability human perception of the hazard is expected to be good and variations small. We attribute the low agreement rate in these situations to the dynamic nature of the avalanche situation, i.e. to a temporal mismatch as the danger changes during the day. Furthermore, some of the differences may be related to poor visibility and limited access to terrain. In contrast, the agreement rate was somewhat higher at lower danger levels, which we attribute to a less dynamic evolution of the avalanche conditions in these situations.

The disagreement rate was lowest within a warning region (22%). It increased when comparing local estimates in neighboring warning regions with the same forecast danger to about 30% (distance  $\geq 20$  km). At greater distances, no further increase was noted indicating that conditions were rather similar and confirming the spatial aggregation of warning regions to a region with the same forecast danger level and description. In contrast, we noted a disagreement rate of about 50% at distances between 10 and 20 km between  $D_{LN}$ -estimates in neighboring warning regions with different forecast danger levels. In these cases, a 100% disagreement rate may be expected. However, observers estimated the danger level as being one level lower 50% of the time when in the region with the higher forecast danger level, partly explaining why the disagreement rate is lower than 100%. This also suggests that the boundary between regions with a different danger rating is reasonably well located, with a bias towards over-forecasting in the warning region with the higher danger level ( $P_{\text{correct.raw}} = 51\%$ ) and a higher accuracy in the region with the lower danger level ( $P_{\text{correct.raw}} = 84\%$ ). Hence, the boundary was rather somewhere within the warning region with the higher danger level than at its actual boundary towards the warning region with the lower danger level.

More than one third of the disagreements were, considering observers individually, unbalanced towards either higher or lower danger level estimates. This highlights that at least some observers had a tendency towards consistently lower or higher  $D_{LN}$  than others in the same region. However, due to relatively small numbers we can only assume that the proportion of significantly biased observers is somewhere between 5 and 15%.

### A.3.5.2 Using local danger level estimates for forecast verification?

Assessing the quality of a forecast involves the comparison of matched pairs of a forecast with corresponding observations (Wilks, 2011), in our case the local nowcasts  $D_{LN}$ . Brabec and Stucki (1998) who also explored local danger level estimates, stated several requirements for forecast verification: the data source should be independent of the product (the forecast) to be verified and the person undertaking the verification should be independent of the forecast, the approach should be applicable to any region and in any avalanche situation. Moreover, the forecast and the corresponding observations should represent a similar spatial scale and have similar temporal resolution.

Clearly, these requirements are almost impossible to fulfill in the case of avalanche forecasts. We can certainly not assume full independence between the forecast danger level and the nowcast - even if rated by different, independent people. We expect that observers read, or were at least roughly aware of, the avalanche bulletin prior to their field day. On the other hand, observers are expected and trained to report local conditions and their local estimate of the avalanche danger level - independent of the forecast.

As pointed out by Jamieson et al. (2008) a scale mismatch exists between a local nowcast and a regional avalanche forecast - in both the temporal and the spatial scale. In the case of the Swiss avalanche bulletin, the smallest spatial forecast unit is approximately one order of magnitude larger than the size of a local observation. In fact, this scale mismatch is often much larger as generally several warning regions are aggregated to one area with a unique danger description. The mean size of these areas is about 7000 km<sup>2</sup>, hence more than two orders of magnitude larger than the area of a local observation. This means that we compare local estimates at the drainage to regional scale (about 1 to 100 km<sup>2</sup>) to forecasts at the mountain range scale (about 1,000 to 10,000 km<sup>2</sup>, Schweizer and Kronholm, 2007). In addition, there is a temporal scale mismatch - a forecast valid for a 12 to 24 hour period is compared to a local assessment, which is often based on (part of the) day spent in the field (often less than 6 hours; e.g. Meister (1995)).

Despite these scale issues, the major advantage of using  $D_{LN}$  estimates for verification is the fact that it has the same unit as the forecast, the danger level. The danger level represents a synthesized interpretation of many local observations that cannot be reported independently. However, it is important that observers are specifically trained to assess the danger level according to common standards.

Local danger level estimates may also be influenced by the time period an observer has been staying in the area. For instance, a mountain guide who just arrived in a new area may have less information to base the local estimate on compared to a ski patroller who works at the same ski resort the whole season. In fact, the observers reporting via the *mAvalanche* network may provide this information concerning the quality of their assessment as either «neutral» - for instance when they were for the first day in an area or had limited access to terrain - or «certain» when they had lots of information. However, this quality information was neither correlated with the disagreement rate, the locally estimated danger level nor the proportion correct, but it strongly varied between observers. Some observers never indicated that they were certain, others reported that they felt almost always certain (96%). In situations, when two observers indicated «neutral» quality, the disagreement rate was slightly higher compared to two observers being «certain» ( $P_{\text{disagree}} = 26\%$  and  $P_{\text{disagree}} = 19\%$ , respectively). It is therefore questionable, whether such information provides added value when interpreting danger ratings, since it seems to primarily reflect individual preferences. Similarly, whether

these observers travelled in frequently tracked terrain or not, was neither correlated with the disagreement rate nor the proportion correct.

We quantified the variability (the disagreement rate) in local danger level estimates at relatively small distances and detected some observers who deviated from the overall mean. However, as the avalanche danger level is not measurable, we do not know which observer is closest to the actual situation. Still, we argue that the mean of a diverse group of trained observers might provide a good estimate of the accuracy of the forecast, particularly when the sample is quite large. The diversity of observers, we used local estimates reported by more than 100 observers, supports this assumption since, for instance, Page (2007) states that the error in a group is smallest when the group's diversity is large.

Some groups of observers had a significantly higher agreement rate with the forecast than others (Fig. A.15). In our study, valley-floor and ski area observers as well as SLF forecasters were closer to the forecast danger level than other observers confirming previous research (Jamieson et al., 2008; Suter et al., 2010). This finding may reflect residence time (as these observers are particularly familiar with their region) or an anchoring bias towards the forecast danger level. In any case, we suggest using local danger level estimates for forecast verification from a diverse group of trained observers, and obtained results must be interpreted in view of the observers and observer groups used.

#### A.3.5.3 Estimating the accuracy of the forecast regional danger level

We presented three approaches to obtain a best estimate of the accuracy of the forecast. Comparing all assessments individually with the forecast has the advantage of a large number of comparisons. With sufficiently large numbers and a diverse range of observers, we expect that the overall estimate is a first good approximation of forecast accuracy ( $P_{\text{correct.raw}} = 76\%$ , Fig. A.14a).

A higher proportion correct ( $P_{\text{correct.raw}} = 85\%$ , Fig. A.14b) was obtained using only danger level estimates reported on days and in regions when several observers agreed on a danger level, or when a majority opinion existed. These combined estimates of independent observers are likely less influenced by observer-bias and more accurate, even though misperceptions by several observers are still possible as shown in an example by Techel et al. (2016a). The overall higher proportion correct can be expected, as situations with less obvious danger ratings are likely excluded using this sample.

Finally, the third approach, yielding a proportion correct ( $P_{\text{correct}}^{**}$ ) of  $71\% \pm 8\%$  (Fig. A.14c) incorporated the uncertainty in the  $D_{\text{LN}}$  estimate (the disagreement rate) and the reporting bias for the comparison with the forecast.

Although we do not know which of the approaches comes closest to reality, we consider the results from this last approach for the remainder of the discussion, as the standard deviation around the mean highlights the considerable variation that may exist.

This study confirmed the trend observed in almost all studies towards higher regional forecasts compared to local danger level estimates (e.g. Cagnati et al., 1997; Jamieson et al., 2008; Schweizer and Föhn, 1996; Schweizer et al., 2003; Suter et al., 2010). The only exception we are aware of is the study by Brabec and Stucki (1998); they reported the forecast to be more often lower than estimates in the field. Otherwise, all studies suggest that the forecast tends to «err on the side of caution» (Jamieson et al., 2008). This

«over-forecast bias» (Wilks, 2011) was also noted when comparing neighboring regions which differed by one danger level. While the proportion correct in the region with the lower danger level was generally high, the danger level was confirmed only in about half the cases in the region with the higher danger level.

The proportion correct of the forecast was higher at lower danger levels, and particularly high in situations with the forecast danger level not changing or decreasing to the previous forecast. In contrast, the forecast  $D_{RF}$  was frequently perceived as too high when the danger level was 4-High, or when the danger level increased. In these situations, the forecast strongly relies on weather predictions, in particular forecast precipitation, which may be erroneous. Furthermore, the lower proportion correct may be related to the fact that observers may only have limited access to avalanche terrain.

With the beginning of the winter season 2012-2013, the avalanche forecast changed from a primarily text-based to a primarily map-based product (Winkler et al., 2013). This allowed a more flexible aggregation of warning regions to larger areas with the same danger description. As a result, the average number of areas with the same danger level and description per forecast increased from 3.3 to 4.3 indicating a reduction in the average size of the forecast areas from 7,900 km<sup>2</sup> to 6,000 km<sup>2</sup>. However, the average number of different danger levels used in each forecast increased only marginally from 2 to 2.3. As our analysis only considers the avalanche danger level, it is not surprising that we noted only a marginal and not significant increase in the forecast accuracy (from  $69.6 \pm 3\%$  to  $70.8 \pm 9\%$ , excluding the area «central part of the southern flank of the Alps», which did not have a morning forecast until 2012). This finding is comparable to the results of a survey conducted among bulletin users. They estimated the mean accuracy to be 83.2%, compared to 82.6% prior to the introduction of the new bulletin (Winkler and Techel, 2014).

The avalanche warning service is located in Davos in the eastern Swiss Alps. In the surroundings of Davos the proportion correct was marginally higher ( $71.8\% \pm 8\%$ ,) than the Swiss average ( $70.8 \pm 9\%$ ). In other areas the proportion correct was comparable or even higher, for instance in the Lower Valais in the western Swiss Alps ( $74.8\% \pm 9\%$ ). In contrast, a significantly lower proportion correct ( $66.9\% \pm 4\%$ ,  $p < 0.001$ ) was observed for the region south of the main Alpine ridge. Reasons for this difference might be a higher persistence of danger levels in the inner-alpine regions of Valais and Grisons due to an often existing persistent weak layer problem, but also the considerably greater number of regular field observations allowing forecasters a daily verification and correction of the forecast. This supports the conclusion by Winkler and Techel (2014) that the forecast accuracy may not necessarily decrease with increasing distance from the forecast center, as long as a sufficient number of high quality field observations are regularly available.

### A.3.6 Conclusions

We analyzed a large number of local danger level estimates in view of verifying the forecast regional avalanche danger level. To this end, we first explored variations and bias between local estimates of trained observers in the same warning region.

In general, the locally estimated danger level is a condensed and interpreted summary of observations, prior knowledge and other information an observer may have. The assessment may also depend on the observer's experience, the location when assessing the danger, and may be influenced by the time spent in

a region as well as the forecast danger level.

While the agreement between individual estimates was relatively high (78%), we sometimes noted an observer specific reporting bias. These findings highlight the importance of regular training to ensure common standards and the fact that even experienced observers disagree in their rating. The disagreement rate of 22% clearly shows the difficulty of assessing the avalanche situation, and describing it with a single danger level. Part of the difficulty is related to the fact that the avalanche danger is not well defined - and cannot be fully defined as it cannot be measured.

Nevertheless, improved and more detailed guidelines on how to locally assess the avalanche danger would be helpful and increase consistency. In particular, when observers report their local danger level estimate, they should always as well report other observations such as new snow depth, snow drifts or signs of instability. These additional observations should allow validating the local nowcast. Any reporting tool should guide the observer towards the final danger level estimate.

In addition, public forecasters may make better use of local nowcasts if they have access to additional objective information such as the residence time an observer has spent in an area, but also if intermediate ratings are reported. While the latter suggestion will not decrease the disagreement rate, it will give the observer an opportunity to communicate such intermediate situations, while at the same time, facilitating the data interpretation by public forecasters.

The agreement rate between local nowcasts and regional forecasts varied considerably between different observer groups and was 76% if all individual ratings following a day in the backcountry were considered. Incorporating the reporting bias and the disagreement rate between local nowcasts into the verification analyses yielded an agreement rate of  $71\% \pm 8\%$ . The forecast was biased towards over-forecasting, in time and space. These values of forecast accuracy, based on estimates by a large and diverse group of observers, are in line with results from previous studies. Given the agreement rate between individual observers, the above mentioned values of forecast accuracy seem plausible. It seems rather questionable whether the accuracy of the avalanche forecast can be higher than the agreement rate between individual estimates in a specific warning region.

Overall, the rule of thumb that the forecast avalanche danger level may not appropriately describe the avalanche situation on 1-2 days per week has been confirmed. This finding highlights the importance that anyone travelling in avalanche terrain needs to be capable of locally assessing the avalanche danger and cannot simply rely on the forecast danger level only.

The local estimates must clearly be considered a best guess only, but we are not aware of any other method that allows a more objective verification - unless, in the future, there would be a method available to readily measure avalanche danger.

*Acknowledgments:* We thank all the observers who reported avalanche conditions, as well as Lukas Dürr and the two anonymous reviewers for their constructive comments that helped to improve the paper.

## A.4 Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random?

Frank Techel, Christine Pielmeier, Kurt Winkler: Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random? *Cold Regions Science and Technology*, 2020

### Abstract

In public avalanche forecasts, avalanche danger is summarized using a five-level ordinal danger scale. However, in Switzerland - but also in other countries - on about 75% of the forecasting days, only two of the five danger levels are actually used, indicating a lack of refinement in the forecast danger level. A refined classification requires the forecasters to assess the avalanche danger in greater detail than the established danger levels. This leads to the fundamental question, whether a reasonable accuracy and consistency of refined danger ratings can be achieved at all. We address this question relying on a data set from Switzerland, where forecasters of the national avalanche warning service have refined the forecast danger level using three sub-levels (*minus*, *neutral*, *plus*) during four forecasting seasons. These sub-levels, which describe where within a danger level the danger was estimated, were not provided to the public. With the goal to assess whether the forecast sub-levels were better than a random assignment of sub-levels, we compared these forecasts with local nowcast estimates of avalanche danger, for days when two observers reported such an estimate ( $N = 1146$ ), as ground truth. The agreement between the forecast regional danger level and the local danger level estimate was 81%, with a distinct over-forecast bias in cases when forecast and nowcast disagreed. This tendency towards over-forecasting also showed in a spatial and temporal context. Furthermore, some anomalies in the use of the sub-levels were noted, particularly for sub-level *plus* in combination with danger level 2–Moderate. Despite these anomalies, the forecast sub-levels were clearly better than a randomly assigned sub-level, resulting in a lower misclassification cost. Furthermore, in case of over-forecasting, the forecast sub-level was in 70% of the cases the sub-level closest to the local estimate, and thus the difference between forecast and nowcast danger level was likely less than one «full» danger level. This indicates that forecasters can often forecast avalanche danger at greater detail than the established danger levels, provided that relevant and reliable data is available in sufficient spatial and temporal density, and that the warning regions, the smallest spatial units used in the forecast are sufficiently small. Therefore, we argue, such refinements of the danger level should be made whenever possible, last but not least for an improved internal assessment of avalanche danger.

### A.4.1 Introduction

Avalanche forecasts, providing avalanche warnings to the public, are issued in many snow-covered mountain regions. An important component of these forecasts is the publication of a regional avalanche danger level  $D_{RF}$ , assigned according to a five-level, ordinal danger scale (EAWS, 2018; Statham et al., 2010).  $D_{RF}$  uses an integer-signal word combination (e.g. danger level 4–High) to summarize the expected avalanche

conditions.

The forecast danger level is a relevant parameter particularly during the planning phase of back-country tours, and it is used in decision support tools for back-country recreationists (e.g. McCammon and Hægeli, 2007; Landrø et al., 2020b). Furthermore,  $D_{RF}$  also impacts the behaviour of recreationists undertaking tours in backcountry terrain (Furman et al., 2010), that is in terrain without organized avalanche mitigation. In addition, in Switzerland, the forecast danger level correlated highly with the avalanche risk of backcountry recreationists (Techel et al., 2015b; Schmudlach et al., 2018), and a decrease of touring activities on days and in regions with danger level 3–Considerable has been noted (Zweifel et al., 2006; Techel et al., 2015b). And finally, in some countries, as in Switzerland, risk-management authorities incorporate information provided in the forecast in their planning of risk-mitigation measures.

However, two problems come to the fore: Firstly, summary statistics of published avalanche forecasts indicate that the distribution of the forecast danger levels is not very refined: on three of four days the forecast danger level was either 2–Moderate or 3–Considerable (e.g. Logan and Greene, 2018; Techel et al., 2018). And secondly, even though assigning and communicating a single danger level may be easier to understand for a user than a probabilistic forecast, categorical forecasts result in the maximum loss of information (Murphy, 1993). This is due to the fact that the probability assigned to a categorical value (the danger level) is always 100% (Doswell and Brooks, 2020), and the uncertainty related to it can only be expressed in the danger descriptions. Therefore, avalanche warning services emphasize that forecast users refer to the danger description accompanying the forecast to obtain more detailed information.

This challenge - communicating avalanche danger in a simple and well-established manner on one side, while simultaneously assessing avalanche danger in greater detail on the other side - lead to the question whether sub-levels, assigned to a danger level during the forecast process, actually have skill. In other words, if a forecast regional danger level  $D_{RF}$  was refined by assigning a sub-level by a forecaster, were these sub-levels significantly better than a randomly assigned one?

To answer this question, we explored a four-year data set of published avalanche forecasts in Switzerland, and compared the forecast  $D_{RF}$ , including an unpublished sub-level ( $D_{RF,sub}$ ), with local nowcast danger level estimates ( $D_{LN}$ , LN = local nowcast). As a danger level cannot be measured, and hence not truly be verified, such nowcast estimates have been used in several studies to «verify» the avalanche danger level (e.g. Brabec and Stucki, 1998; Jamieson et al., 2008; Sharp, 2014; Techel and Schweizer, 2017). Furthermore, we discuss potential benefits and challenges associated with  $D_{RF,sub}$ , taking the viewpoint of an avalanche forecaster as well as the bulletin user.

## A.4.2 Data

### A.4.2.1 Regional forecast danger level and sub-level

In Switzerland, the national avalanche warning service WSL Institute for Snow and Avalanche Research SLF (SLF) issues a public avalanche forecast covering the Swiss Alps and the Jura mountains (SLF, 2019) (Fig. A.16a). The main forecast is published at 17:00 CET<sup>6</sup>, valid until 17:00 the following day. For the main

<sup>6</sup>the forecast is always published in local time, therefore all times refer to either CET or CEST

part of the winter, the forecast is updated every morning at 08:00. The forecast product is map-based (Fig. A.16a) and contains information on the danger levels, most critical aspects and elevations, the avalanche problems and a danger description. Furthermore, a snowpack and weather summary is provided with a short, two-day outlook.

The forecast domain is split into warning regions. More than 130 static spatial warning regions (polygon boundaries in Fig. A.16b) form together the forecast areas of the Swiss Alps (26,400 km<sup>2</sup>, in 2018/2019 subdivided into 117 warning regions with a median size of 183 km<sup>2</sup>) and the lower elevation Jura mountains (2,900 km<sup>2</sup>, 2018/2019: 12 warning regions with median size 255 km<sup>2</sup>). For the Jura, the daily publication of a forecast started in winter 2017/2018. No forecast is issued for the lowlands between the Alps and the Jura (white area in the map in Fig. A.16a). Avalanche danger is communicated for dynamically aggregated warning regions, so-called *danger regions* (for instance regions A, B, C<sub>1</sub> and C<sub>2</sub> in Fig. A.16a; Ruesch et al. (2013)). Warning regions are aggregated to a single danger region when the expected avalanche danger can be described with the same avalanche danger level, valid for the same aspects and elevations and with identical avalanche problems and danger description (SLF, 2019). Danger regions may be spatially continuous (e.g. regions C<sub>1</sub> or D in Fig. A.16a), or may be disconnected from each other (e.g. regions A or C<sub>1</sub> in Fig. A.16a).

Forecasters assign a regional danger level according to the danger level definitions provided in the European Avalanche Danger Scale (EAWS, 2018), by considering snowpack stability, the frequency of triggering locations and the expected avalanche size. Since January 2017, avalanche forecasters have assigned one of three ordinal sub-levels to each forecast danger level  $D_{RF}$ : *plus*, *neutral*, *minus*. The intention of assigning these sub-levels was to indicate where within the danger level avalanche danger was estimated. Therefore, the avalanche conditions described by the sub-levels are within the corresponding danger levels' definitions:

- *plus* means that the danger tends towards the next higher danger level, e.g. a *plus* assigned to 3–Considerable (notation  $D_{RF,sub} = 3\text{--}plus$ ) tends towards 4–High

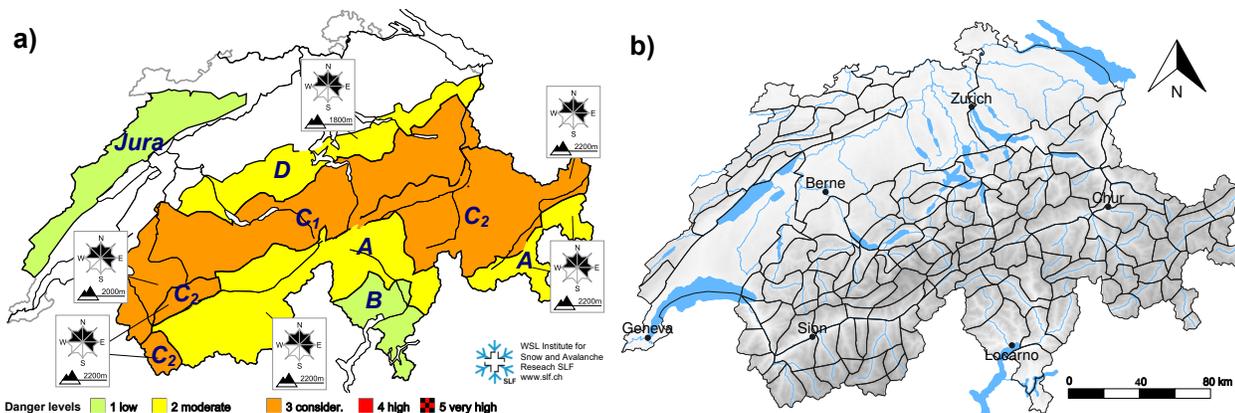


Figure A.16: Maps showing (a) the Swiss avalanche bulletin, issued in the morning of 10/03/2018 at 08:00 CET, and (b) a relief map (relief = grey shading) of Switzerland showing the major rivers and lakes (blue) and the more than 130 individual warning regions, the smallest spatial units used in the forecast (black polygons). (a) Letters A to D are explained in the text. (b) Reproduced by permission of swisstopo JA100118.

- *neutral* means that the danger is approximately in the middle of the level, e.g.  $D_{RF,sub} = 3\text{--}neutral$ , and
- *minus* means that the danger is at the lower end of its level, e.g.  $D_{RF,sub} = 3\text{--}minus$  tends towards 2–Moderate.

Neither a numerical value nor a probability distribution was associated with the sub-levels.  $D_{RF,sub}$  was assigned to each danger region, but was not published.

To distinguish between the «full» danger level  $D_{RF}$  and  $D_{RF,sub}$ , we use the integer-signal word combination for  $D_{RF}$  (e.g. 3–Considerable) and a combination of the integer and sub-level-term (e.g. 3–*plus*) for  $D_{RF,sub}$ . An evaluation after the first winter showed that the seven forecasters at SLF were generally comfortable assigning a sub-level to a danger level. However, to increase consistency the following rules were defined:

- In case of spatial gradients, for instance, a region bordering regions with a lower and a higher danger level, for the sub-level the approximate center of the region should be assessed. An example is shown in Fig. A.16a, where region *A* lies between a region with a lower (region *B*) and a higher danger level (regions  $C_1 - C_2$ ).
- In case of temporal variations during the valid period of the forecast, the morning is assessed. This is standard practice in the avalanche forecast in Switzerland.
- For the lowest danger level 1–Low, no sub-level is assessed.

In this study, we relied on forecasts issued in the morning (at 08:00 CET), valid until 17:00 CET for the Swiss Alps, and relating to dry-snow avalanche conditions. We limited this analysis to forecasts describing dry-snow avalanche conditions to allow a comparison with local nowcast estimates of avalanche danger, which are provided for dry-snow conditions only (see following Sect. A.4.2.2). We made use of the forecast danger levels  $D_{RF}$  and the respective sub-levels  $D_{RF,sub}$ .

Between January 2017 and April 2020, 439 avalanche bulletins were published in the morning with a total of 2,173 different danger regions describing dry-snow avalanche conditions.

#### A.4.2.2 Nowcast danger level estimates

In Switzerland, specifically trained observers assess and report the avalanche danger level in their region (e.g. Suter et al., 2010; Techel and Schweizer, 2017). These danger level estimates describe current conditions, and can therefore be considered a local *nowcast* (Jamieson et al., 2008), where local does not refer to an assessment of a single slope, but to an area of observation, estimated as 10 to 25 km<sup>2</sup> (Jamieson et al., 2008; Meister, 1995). Observers are advised to incorporate all information considered relevant for the assessment, including observations made during the day in the field, but also prior knowledge they may have regarding, for instance, the development of the snowpack or information from third parties (for more details refer to Techel and Schweizer, 2017). Observers reporting a  $D_{LN}$  estimate (LN = local nowcast) are advised to assess current conditions for dry-snow situations and the expected highest  $D_{LN}$  for wet-snow conditions.  $D_{LN}$  is assessed according to the European Avalanche Danger Scale (EAWS, 2018). Additionally, when estimating 3–Considerable, observers reported whether natural avalanches were expected or not

(SLF, 2002).

As in Techel and Schweizer (2017), we limited the analysis to  $D_{LN}$  estimates describing current conditions. Therefore, we relied exclusively on  $D_{LN}$  estimates which referred to dry-snow avalanche conditions, which were reported between 10:00 and 17:00 CET from observers who were in the field.

Variations in  $D_{LN}$  estimates between observers in the same warning region have been noted (Techel and Schweizer, 2017), but also when relying on the same set of observations (Haladuick, 2014). To incorporate this uncertainty, we considered only  $D_{LN}$  estimates reported on days and in warning regions when two or more observers were in the same warning region. When two observers indicated the same  $D_{LN}$  estimate, we considered this as a sufficiently robust estimate of avalanche danger for the day and region. In contrast, when two observers differed in their assessment by one danger level, we considered this as an indication that the danger was likely somewhere between the two reported levels.

After applying the selection criteria and merging forecasts with nowcasts by date and warning region, the data set consisted of  $D_{RF}$ ,  $D_{LN}$  pairs for which either two  $D_{LN}$  estimates resulted in the same  $D_{LN}$  ( $N = 891$ ), or for which  $D_{LN}$  differed by one danger level ( $N = 255$ ). Furthermore, 210  $D_{LN}$  estimates for 3–Considerable were available, where two or more observers provided the same indication whether natural avalanches were expected or not.

### A.4.3 Methods

Danger levels ( $D_{RF}$ ,  $D_{LN}$ ) are ranked ordinal data with five levels.  $D_{RF.sub}$ , which additionally describes a rank order within each danger level, increases the resolution of the forecast  $D_{RF}$  compared to  $D_{LN}$ . Accounting for this difference in resolution, and whether  $D_{LN}$  estimates showed agreement or not, we proceeded step-wise to explore whether the forecast  $D_{RF.sub}$  had skill:

- For the 891 cases, when  $D_{LN}$  estimates agreed:
  1. We calculated the difference between the forecast and the nowcast danger levels  $\Delta D = D_{RF} - D_{LN}$ .
  2. When forecasts and nowcasts agreed ( $\Delta D = 0$ ), the skill of  $D_{RF.sub}$  could not be explored, as  $D_{RF.sub}$  was within the same danger level as  $D_{LN}$ . For these cases, we assigned a misclassification cost of 0 (Tab. A.12).
  3. For all other cases, that is when forecast and nowcast disagreed ( $D_{RF} \neq D_{LN}$ ), we calculated the difference in sub-level ranks between  $D_{RF.sub}$  and  $D_{LN}$  and considered this difference as the misclassification cost (Tab. A.12). For ordinal classification approaches, a misclassification cost equal to the difference of ordinal levels between the diagonal and the event is considered reasonable (Galimberti and Soffritti, 2012).
- For the 255 cases, when two  $D_{LN}$  estimates disagreed by one danger level:
  1. We considered these cases to indicate that observed avalanche conditions were likely somewhere in between two danger levels (e.g. when one  $D_{LN}$  estimate was 2–Moderate and another

Table A.12: Misclassification cost assigned to forecast-nowcast pairs, for cases when two  $D_{LN}$  estimates were the same. The misclassification cost increases by 1 with each increase in the difference in sub-level ranks ( $D_{RF,sub}$ ) for cases when  $D_{RF} \neq D_{LN}$ .  $D_{RF} = 1$ -Low is not shown, as no sub-levels were forecast for this danger level. Values shown bold have a misclassification cost of 0

$D_{RF,sub}$	$D_{LN}$				
	1-Low	2-Mod	3-Cons	4-High	5-vHigh
2-minus	1	<b>0</b>	3	6	9
2-neutral	2	<b>0</b>	2	5	8
2-plus	3	<b>0</b>	1	4	7
3-minus	4	1	<b>0</b>	3	6
3-neutral	5	2	<b>0</b>	2	5
3-plus	6	3	<b>0</b>	1	4
4-minus	7	4	1	<b>0</b>	3
4-neutral	8	5	2	<b>0</b>	2
4-plus	9	6	3	<b>0</b>	1
5-minus	10	7	4	1	<b>0</b>
5-neutral	11	8	5	2	<b>0</b>

Table A.13: Misclassification cost assigned to forecast-nowcast pairs, for cases when two  $D_{LN}$  estimates differed by one danger level.  $D_{RF} = 1$ -Low is not shown, as no sub-levels were forecast for this danger level. Values shown bold have a misclassification cost of 0

$D_{RF,sub}$	$D_{LN}$		
	1-Low/2-Mod	2-Mod/3-Cons	3-Cons/4-High
2-minus	<b>0</b>	2	5
2-neutral	1	1	4
2-plus	2	<b>0</b>	3
3-minus	3	<b>0</b>	2
3-neutral	4	1	1
3-plus	5	2	<b>0</b>
4-minus	6	3	<b>0</b>
4-neutral	7	4	1
4-plus	8	5	2
5-minus	9	6	3
5-neutral	10	7	4

3-Considerable). We then assigned a misclassification cost of 0 to the respective highest and lowest sub-levels of these two danger levels (e.g. when  $D_{LN}$  2-Moderate and 3-Considerable,

the misclassification cost was 0 for 2–plus and 3–minus, Tab. A.13).

- For all other cases, the misclassification cost increased by one according to the difference in ranks (Tab. A.13).

With the goal to explore whether  $D_{RF,sub}$  was better than a random sub-level, we randomly assigned a sub-level to each  $D_{RF}$ , thus obtaining a  $D_{RF,sub,random}$ . This random assignment of sub-levels, however, was not fully random as we sampled according to the distributions of the forecast sub-levels for each of the danger levels (as shown in Fig. A.17b). This approach already introduces some skill in the random assignment of sub-levels. Proceeding as described before, we obtained the difference in sub-level ranks and thus the misclassification cost for  $D_{RF,sub,random}$  according to Tab.s A.12 and A.13.

## A.4.4 Results

We present the results in two steps: To detect potential anomalies in the use of the sub-levels, we first explore the use of the danger levels and sub-levels in the forecasts in Sect. A.4.4.1 by exploring overall distributions, temporal changes and spatial gradients in danger ratings between immediately neighboring warning regions. And secondly, we focus on the quality of the forecast sub-levels, that is, the agreement between forecast and local estimate and whether forecast sub-levels were better than random (Sect. A.4.4.3).

### A.4.4.1 Forecast danger levels and sub-levels

#### Overall distributions

Figure A.17a shows the distribution of forecast danger levels  $D_{RF}$  for dry-snow conditions in the Swiss Alps during the four-year period. 2-Low and 3–Considerable were forecast about 80% of the time. Avalanche danger was not explicitly communicated for each of the more than 100 warning regions in the Alpine forecast area, but warning regions were aggregated to, on average, five danger regions (for instance regions A, B, C<sub>1</sub>, C<sub>2</sub> in Fig. A.16a). These differed in at least one of the forecast parameters - danger level, aspects, elevation range, avalanche problems or danger description. However, most often only two different danger levels  $D_{RF}$  (mean 2.4) and three sub-levels  $D_{RF,sub}$  (mean 3.4) were used to describe dry-snow avalanche

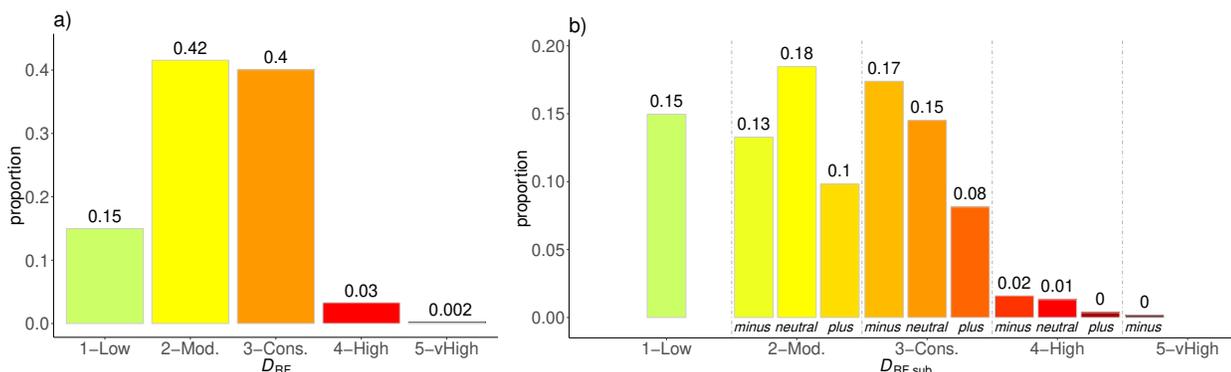


Figure A.17: Distribution of the forecast danger level  $D_{RF}$  (a) and the sub-levels  $D_{RF,sub}$  (b) during the four winters 2016/2017 to 2019/2020 for dry-snow avalanches, as issued in the morning forecast.

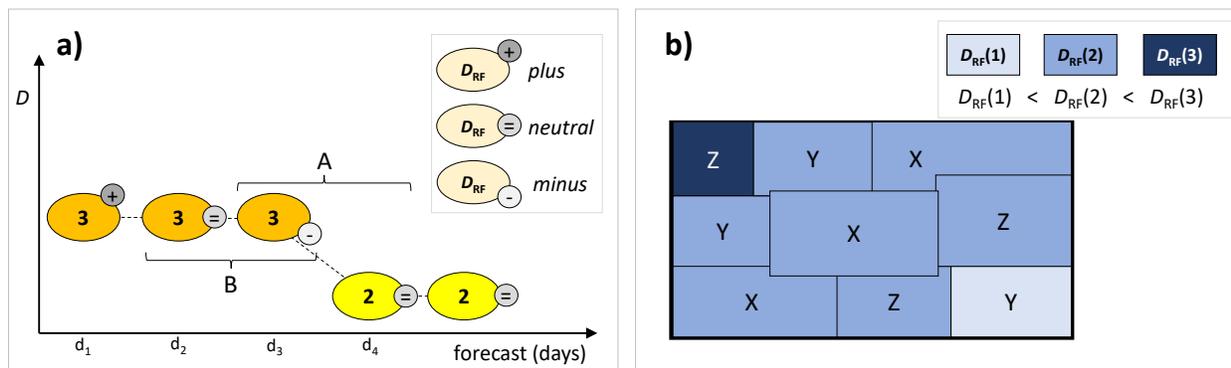


Figure A.18: (a) Schematic representation of the temporal evolution of the forecast danger level  $D_{RF}$  (coloured ellipses) and sub-levels  $D_{RF.sub}$  (grey circles). The cases A,  $D_{RF}$  decreased from 3–Considerable (3) to 2–Moderate (2), and B, the day prior to A when  $D_{RF}$  was 3–Considerable, are described in the text. (b) Schematic representation of a forecast domain with nine warning regions with three different danger levels  $D_{RF}$  indicated by different blue colors. The following neighbor relations are described in the text: Regions marked with an X: all neighboring regions have the same  $D_{RF}$ , regions marked with a Y: at least one neighboring warning region had a higher  $D_{RF}$ , regions marked with a Z: at least one neighboring warning region had a lower  $D_{RF}$ .

danger in the Alps in a forecast, highlighting that forecasters not solely communicated the variations in the danger level between regions but that almost as often differences in the aspects and elevations where the danger prevailed and/or in the avalanche problems encountered and/or in the danger description were reason to aggregate warning regions to a separate danger region.

The proportion of the forecast sub-levels  $D_{RF.sub}$  decreased monotonically from 3–minus to 5–minus (Fig. A.17b). At 2–Moderate, no such pattern showed. Of note was the comparably low proportion of 2–plus (10%), used less often than both the immediately lower 2–neutral (18%) and higher (3–minus) sub-levels (17%). Expecting an approximately similar usage of these  $D_{RF.sub}$ , we note that these proportions were significantly different (proportion test (R Core Team, 2017):  $p < 0.001$ ). This pattern also showed when comparing 2–plus with 2–minus and 3–neutral (13% and 15%, respectively,  $p < 0.001$ ), suggesting some anomaly in the use of 2–plus.

### Temporal changes

For the same warning region,  $D_{RF}$  stayed the same from one day to the next about 75% of the time, while  $D_{RF.sub}$  changed about every second day (51%). If  $D_{RF.sub}$  changed, it was 52% of the time within the same danger level  $D_{RF}$ . Increases in  $D_{RF.sub}$  were generally by one sub-level (52%) or two sub-levels (24%), decreases were often more gradual (by one sub-level 65%, by two sub-levels 26%). Thus, pronounced changes - more than two sub-levels change from one day to the next - were significantly more often forecast when danger increased rather than decreased (24% vs. 10% of the cases,  $p < 0.001$ ). An exception to the generally rather gradual decrease of  $D_{RF.sub}$  during times when avalanche conditions returned towards stability, were days, when  $D_{RF}$  was lowered from 3–Considerable to 2–Moderate (case A in Fig. A.18a). On these days, the decrease in sub-levels was 71% of the cases by two or more sub-levels. In contrast, on days immediately prior to these days, most often no change in  $D_{RF.sub}$  was noted. When considering only days

when a decrease within 3–Considerable was forecast for the day before this change in  $D_{RF}$  (case *B* in Fig. A.18), this was by one sub-level in 75% of the cases. Hence, the forecast decrease in avalanche danger was clearly more distinct in case *A* compared to *B*, indicating some anomaly in the use of the danger levels (i.e. staying comparably long on 3–*minus* and then decreasing straight to 2–*neutral*). Likely, this is linked to the general tendency to over-forecast as will be addressed in more detail in Sect. A.4.4.3.

In the course of the winter, it is common that periods with very slow changes in avalanche conditions occur, which will often be forecast with the same danger level. Exploring periods, when 2–Moderate or 3–Considerable were forecast on at least 10 consecutive days, showed that  $D_{RF,sub}$  changed on about one of three days (32%) expressing variations in avalanche conditions.

### Spatial gradients

On average each warning region shared borders with five to six neighboring warning regions. Therefore, gradients in  $D_{RF}$  or  $D_{RF,sub}$  between a warning region and at least one of its neighbors were comparably frequent occurrences:  $D_{RF}$  differed in 35% and  $D_{RF,sub}$  in 52% of the cases.

In the 84% of the cases when differences in  $D_{RF,sub}$  existed, they were within the same danger level  $D_{RF}$ . Excluding situations when 1–Low was forecast, differences were primarily by one sub-level (44%) or two sub-levels (38%).

Spatial gradients of two or more sub-levels were observed most often between the chain of the northernmost warning regions (the lower elevation Pre-Alps, region *D* in Fig. A.16a) and the next chain of warning regions further into the Alps (region  $C_1$  in Fig. A.16a).

Considering the issued danger level, no clear patterns showed: spatial gradients of two or more sub-levels were observed for all  $D_{RF,sub}$  combinations.

#### A.4.4.2 On the agreement rate of local nowcasts

1,146 comparisons between forecast  $D_{RF}$  and nowcast estimates  $D_{LN}$  provided by two observers were analyzed. The two nowcast estimates agreed 78% of the time.

The proportion of disagreements between two  $D_{LN}$  estimates increased with increasing  $D_{RF}$  from 16% at 1–Low to 33% at 4–High.

Considering the forecast sub-level, disagreements occurred significantly more often when the sub-level was *minus* (29%), rather than *neutral* (19%,  $p < 0.001$ ) or *plus* (16%,  $p < 0.001$ ).

Regardless of the forecast sub-level, nowcasts disagreed also significantly more often when estimates were made in a warning region where the forecast danger level  $D_{RF}$  was higher than in at least one of the immediately neighboring warning regions (32%, case *Z* in Fig. A.18b), compared to cases when the same  $D_{RF}$  was forecast in all neighboring warning regions (22%,  $p < 0.01$ , case *X*), or when at least one neighboring warning region had a higher  $D_{RF}$  (14%,  $p < 0.001$ , case *Y* in Fig. A.18b).

In summary, differences between two local danger level estimates were most frequent when the forecast sub-level was *minus* or when  $D_{RF}$  was higher than in a neighboring warning region. This indicates that such disagreements were not just due to random variations in the local assessments, but may in fact represent to some extent that the danger was probably somewhere in between two danger levels.

### A.4.4.3 On the quality of forecast danger levels and sub-levels

In the following, we compare forecasts with nowcasts, first for the cases when nowcasts agreed, and then when nowcasts disagreed.

#### On the quality of forecast sub-levels when local estimates agreed

When two observers reported the same  $D_{LN}$  estimate ( $N = 891$ ), the forecast danger level  $D_{RF}$  and the locally estimated danger level  $D_{LN}$  agreed 81% of the time ( $N = 718$ , Tab. A.14). In these cases, and ignoring situations with forecast danger level 1–Low, when no sub-level was indicated ( $N = 648$ ), the sub-level was most often *neutral* ( $N = 272$ , 42%) with almost equal proportions of *plus* ( $N = 195$ , 28%) and *minus* ( $N = 181$ , 29%, bold values in Tab. A.14; Fig. A.19a).

Whenever  $D_{RF} \neq D_{LN}$  and  $D_{RF} \neq 1\text{--Low}$  ( $N = 171$ ), and not considering  $D_{RF,sub}$ , differences were essentially always by one danger level ( $N = 168$ , 98%). Deviations indicated a clear tendency towards over-forecasting ( $D_{RF} > D_{LN}$ ), which was 23 times more frequent than under-forecasting. In the 164 cases of over-forecasting, most often the sub-level rating  $D_{RF,sub}$  was the sub-level closest to the  $D_{LN}$  estimate ( $N = 115$ , 70%), suggesting that the difference would often be less than a «full» danger level (Tab. A.14, Fig. A.19a). In contrast, randomly assigned sub-levels showed a less pronounced pattern (Fig. A.19b). For the rare situation, when  $D_{RF} < D_{LN}$  the forecast sub-level ratings  $D_{RF,sub}$  showed no better performance than the randomly assigned sub-levels. As a consequence, the misclassification cost was significantly lower for

Table A.14: Contingency table showing the forecast  $D_{RF,sub}$  and  $D_{LN}$ , for cases when two local estimates agreed. Values shown bold have a misclassification cost of 0. In addition, the proportion of agreements between  $D_{RF}$  and  $D_{LN}$  ( $P(D_{RF} = D_{LN})$ ) is shown for the respective  $D_{RF,sub}$ .

$D_{RF,sub}$	$D_{LN}$					$P(D_{RF} = D_{LN})$
	1–Low	2–Mod	3–Cons	4–High	5–vHigh	
1–Low	<b>70</b>	2	0	0	0	0.97
2–minus	36	<b>88</b>	1	0	0	0.70
2–neutral	16	<b>123</b>	3	0	0	0.86
2–plus	4	<b>95</b>	1	0	0	0.95
3–minus	1	55	<b>101</b>	0	0	0.64
3–neutral	2	13	<b>143</b>	0	0	0.91
3–plus	0	3	<b>81</b>	2	0	0.94
4–minus	0	0	23	<b>6</b>	0	0.21
4–neutral	0	0	8	<b>6</b>	0	0.42
4–plus	0	0	2	<b>5</b>	0	0.71
5–minus	0	0	0	1	<b>0</b>	0
5–neutral	0	0	0	0	<b>0</b>	–

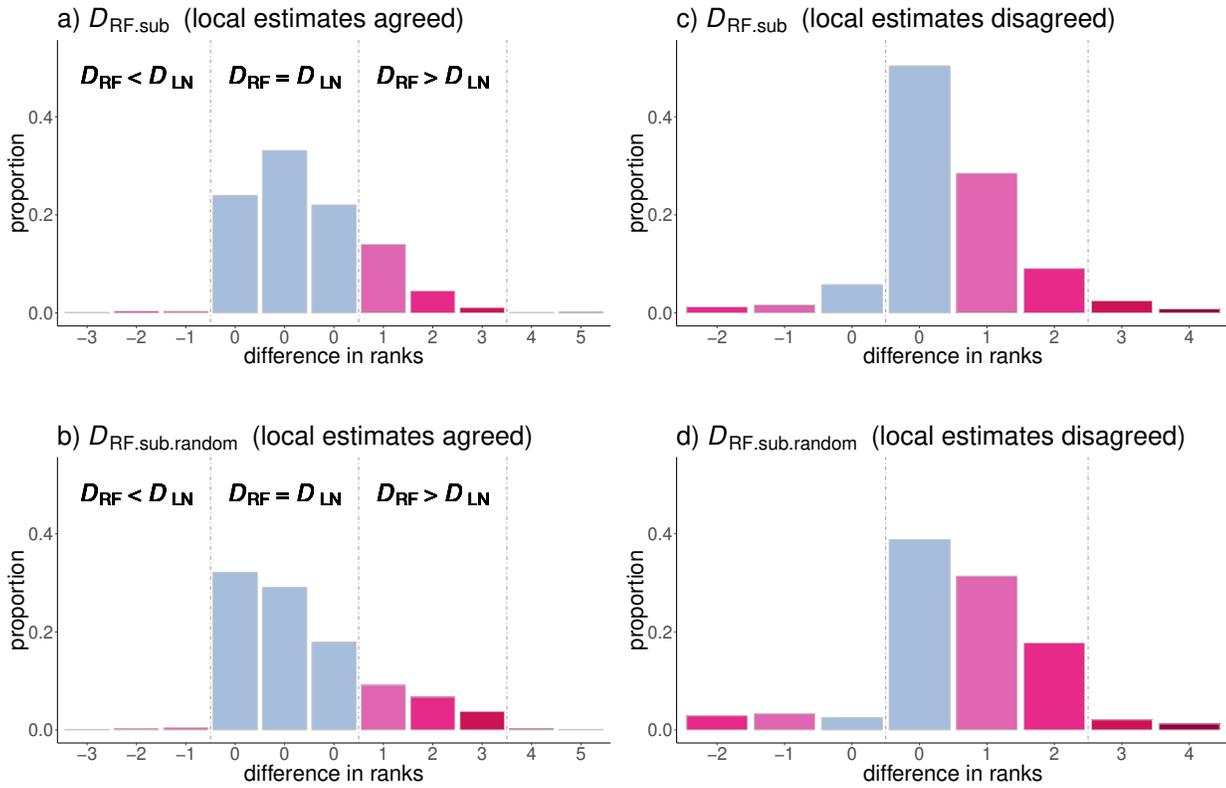


Figure A.19: Distance in sub-level ranks between the forecast  $D_{RF.sub}$  (upper row, a and c) and the randomly assigned sub-level  $D_{RF.random}$  (lower row, b and d) for cases with local estimates  $D_{LN}$  agreeing (left column, a and b) and local estimates disagreeing (right column, c and d). Absolute values of the distance in ranks correspond to the misclassification costs as in Tab.s A.12 and A.13. Light-blue colours indicate cases when no misclassification occurred.

$D_{RF.sub}$  (median = 1, mode = 1) than  $D_{RF.sub.random}$  (median = 2, mode = 1), when considering cases with  $D_{RF} \neq D_{LN}$  (Wilcoxon rank-sum test (R Core Team, 2017):  $p < 0.001$ ).

The proportion of forecast-nowcast pairs with the same danger level ( $P(D_{RF} = D_{LN})^7$ ) increased from sub-level *minus* (0.63,  $N = 181$  of 312), to *neutral* (0.87), to *plus* (0.94,  $N = 181$  of 193; Tab. A.15), regardless of  $D_{RF}$ , but decreased with increasing  $D_{RF}$  (for instance from 2–*minus* (0.7) to 4–*minus* (0.21), Tab. A.14). Avalanche danger does not change abruptly at the border from one warning region to another. Therefore, the proximity to a region with a higher (or lower) danger level can serve as an indication that the danger is in the upper (lower) part of the danger level. Regardless of  $D_{RF.sub}$ , when a warning region bordered at least one other warning region with a lower  $D_{RF}$  (case Z in Fig. A.18b), the proportion of forecasts which matched the local nowcasts ( $P(D_{RF} = D_{LN})$ ) was 0.43 (Tab. A.15). In contrast, when all neighboring warning regions had the same  $D_{RF}$ , the proportion was 0.86 (case X in Fig. A.18b), and 0.98 when at least one neighboring warning region had a higher  $D_{RF}$  (case Y in Fig. A.18b). The agreement between  $D_{RF}$  and  $D_{LN}$  was lowest in case Z, when additionally the forecast sub-level was *minus* ( $P(D_{RF} = D_{LN}) = 0.33$ ; Tab. A.15). Thus, not only high sub-levels, but also proximity to areas with higher danger levels correlated with the proportion that  $D_{RF}$  and  $D_{LN}$  matched.

Exploring the forecast-nowcast pairs, when observers estimated 3–Considerable (Tab. A.16), showed that

<sup>7</sup>corresponds to  $P_{success}$  as described in the Synthesis

when an observer indicated that natural avalanches were expected ( $N = 34$ ), the forecast  $D_{RF.sub}$  was 3-plus 53% ( $N = 18$ ) and 3-minus only 9% of the time ( $N = 3$ ). In contrast, when natural avalanches were not expected ( $N = 176$ ),  $D_{RF.sub}$  was more frequently estimated as 3-minus or 3-neutral (36% /  $N = 64$  and 44% /  $N = 78$ , respectively).

**On the quality of forecast sub-levels when local estimates disagreed**

Typically, when two estimates disagreed ( $N = 255$ ), the forecast danger level matched the higher of the two estimates ( $N = 212$ , 83% of the cases, Tab. A.17). 14% ( $N = 35$ ) of the time it matched the lower of the two estimates, and in 3% ( $N = 8$ ) of the cases there was no match between the individual  $D_{LN}$  and  $D_{RF}$ . This confirms the tendency towards forecasting a higher danger level than was locally estimated. In the 212 cases of over-forecasting, the sub-level was minus 57% of the time ( $N = 121$ , Fig. A.19c). For the comparably rare cases, when  $D_{RF}$  matched the lower of the two estimates, the forecast sub-level was most often plus (67%,  $N = 14$  of 21, excl. 1–Low).

56% of the comparisons between  $D_{RF.sub}$  and  $D_{LN}$  had a misclassification cost of 0 ( $N = 135$  of 241, excl. 1–Low, Tab. A.17), a significantly larger proportion compared to 41% for  $D_{RF.sub.random}$  ( $p < 0.001$ , Fig. A.19c and d). Hence, the misclassification cost was significantly higher for  $D_{RF.sub.random}$  (median = 1, mode = 0) than for  $D_{RF.sub}$  (median = 0, mode = 0,  $p < 0.001$ ).

Table A.15: Proportion of agreements between  $D_{RF}$  and  $D_{LN}$  ( $P(D_{RF} = D_{LN})$ ) as a function of sub-level and spatial variations in  $D_{RF}$  between neighboring warning regions. The three cases X, Y, Z are shown in Fig. A.18b.

sub-level	$D_{RF}$ neighboring region			all
	higher (case Y)	same (case X)	lower (case Z)	
minus	1*	0.78	0.33	0.63
neutral	0.98	0.91	0.65	0.87
plus	0.97	0.94	1*	0.94
all	0.98	0.86	0.43	

\* $N \leq 6$ .

Table A.16: Contingency table showing whether natural avalanches were expected, for cases when both two  $D_{LN}$  estimates and the forecast danger level were 3–Considerable. The number of cases  $N$  is shown.

$D_{RF.sub}$	natural avalanches		all
	not expected	expected	
3–minus	64	3	67
3–neutral	78	13	91
3–plus	34	18	52
all	176	34	210

Table A.17: Contingency table showing the forecast  $D_{RF,sub}$  and  $D_{LN}$ , for cases when two local estimates differed. Values shown bold have a misclassification cost of 0.

$D_{RF,sub}$	$D_{LN}$			
	1–Low/2–Mod	2–Mod/3–Cons	3–Cons/4–High	4–High/5–vHigh
1–Low	<b>14</b>	0	0	0
2–minus	<b>40</b>	2	0	0
2–neutral	28	3	0	0
2–plus	9	<b>10</b>	0	0
3–minus	6	<b>66</b>	1	0
3–neutral	2	33	1	0
3–plus	0	10	<b>3</b>	0
4–minus	0	0	<b>14</b>	0
4–neutral	0	0	8	0
4–plus	0	0	3	<b>1</b>
5–minus	0	0	0	<b>1</b>
5–neutral	0	0	0	0

## A.4.5 Discussion

In this section, we first debate the quality of the local nowcast estimates, the ground-truth we relied on (Sect. A.4.5.1). Following, we discuss the findings regarding forecast accuracy and bias (Sect. A.4.5.2), and the results related to our main research question: «Can avalanche danger be forecast in greater detail than the five levels of the European danger scale?» (Sect. A.4.5.3). Finally, we comment on the operational requirements which need to be fulfilled to assess avalanche danger at greater detail (Sec. A.4.5.4), and we take the perspective of the forecast user, considering the potential benefit of providing more detailed information in the forecast product (Sect. A.4.5.5).

### A.4.5.1 On the reliability of local danger level estimates

As ground-truth, we relied on local nowcast estimates provided by specifically trained observers (Sect. A.4.4.2). In Switzerland, these are the most reliable data-source, when assessing avalanche danger (Techel and Schweizer, 2017). However, it is conceptually difficult to truly verify avalanche danger, as there is not one unique set of observations describing a specific danger level (e.g. Bakermans et al., 2010). Furthermore, local nowcasts rely on the same subjective approach to assess avalanche danger as forecasters do, and are therefore a best guess only (Föhn and Schweizer, 1995). Hence, it is important to be aware of uncertainties and potential biases introduced when relying solely on local danger level estimates for verification:

- Two studies showed that mountain guides in Switzerland assess the avalanche danger level more often lower than other observer groups (for instance when compared to recreational forecast users

or observers in ski areas; Winkler and Techel, 2014; Techel and Schweizer, 2017). Most of the  $D_{LN}$  estimates used in this study were provided by mountain guides, which are part of the observer network.

- It must be assumed that local assessors were aware of the forecast, which may introduce a confirmation bias as noted in studies in Canada (e.g. Jamieson et al., 2009; Bakermans et al., 2010).
- Furthermore, the proportion of two  $D_{LN}$  estimates disagreeing increased with increasing  $D_{RF}$  (from 16% at 1–Low to 33% at 4–High). Considering this disagreement rate not only as an indicator that the danger level was in between two danger levels, but also as a measure describing the reliability of the local estimates within the same warning region, less confidence can be placed on  $D_{LN}$  estimates provided on days when 4–High was forecast.

We addressed these uncertainties related to local danger level estimates by relying only on estimates on days and in regions when two observers reported such an estimate. While we believe that cases when two observers reported the same danger level provide a reasonably robust estimate of the avalanche danger level in a warning region, it is not possible to check whether this assumption truly holds and hence, what kind of bias may be present and should be accounted for.

Due to these uncertainties, we suggest to interpret primarily patterns noted in our findings, rather than absolute values.

#### **A.4.5.2 Forecast accuracy and over-forecast bias**

The comparison between local danger level estimates and the forecast danger levels showed an agreement rate of 81% and a rather strong over-forecast bias in case of disagreements between forecast and nowcast. Both the agreement rate between forecast and local nowcasts and the forecast bias are similar to other studies exploring larger data sets (Suter et al., 2010; Techel and Schweizer, 2017) or data from other countries (e.g. in Canada: Jamieson et al., 2008; Sharp, 2014; Statham et al., 2018b). Additionally, we showed that a tendency towards over-forecasting also exists in a spatial and temporal context (Sect. A.4.4.3).

Under-forecasting is of greater concern than over-forecasting, as potentially riskier decisions may be made by forecast users (Jamieson et al., 2009). However, frequent over-forecasting will decrease the credibility of the warning. Hence, forecast accuracy should be improved in general, which will inevitably also reduce the number of days when the forecast danger level is too high.

#### **A.4.5.3 Forecast sub-levels: consistent? And better than random?**

We explored whether anomalies in the use of the sub-levels existed (Sect. A.4.4.1), and whether the forecast sub-levels were better than a random assignment of a sub-level (Sect. A.4.4.3).

##### **Consistent?**

We noted two anomalies in the use of the sub-levels, which may indicate some inconsistency in their use:

- (i) Sometimes, jumps of two or more sub-levels were forecast from one day to the next, or between immediately neighboring warning regions. This anomaly was observed particularly on days when

$D_{RF}$  was lowered from 3–Considerable to 2–Moderate, often during periods when  $D$  decreased rather gradually. On these days,  $D_{RF,sub}$  decreased more often by two levels than on the days immediately before (Sect. A.4.4.1).

- (ii) 2–*plus* was used significantly less often in the forecasts than would be expected, when compared to the frequency of the respective lower and higher sub-levels (Fig. A.17b).

While there are situations, when abrupt changes may be perfectly justified, there are likely also cases, when these are linked to limitations in the availability of relevant observational data, not allowing a more detailed assessment of avalanche danger. In these situations, forecasts are kept more simple reflecting the reduced knowledge the forecasters have. Furthermore, and despite forecasters having full flexibility of aggregating warning regions to a large number of danger regions, allowing in theory to assign more gradual spatial gradients in avalanche danger between warning regions, each of the danger regions must be described with the most critical aspects and elevations, avalanche problem(s) and a danger description. However, in some circumstances, as for instance at 2–Moderate, it may not be possible to make a further distinction in terms of describing avalanche danger.

2–*plus* was used significantly less often in the forecasts than would be expected. We believe this anomaly is linked to both the forecast bias, which was observed in time and space (Sect. A.4.4.3), as well as operational constraints, like the need to provide a danger description for each danger region.

Can this anomaly be addressed in the forecasts?

Some of these cases are likely linked to the forecast bias, observed in time and space. Addressing this bias can only be achieved by actually correcting the forecast danger level  $D_{RF}$ . However, this would be a change in the forecast danger level itself, and not merely a refinement of the danger level, and does not reflect the state of knowledge the forecaster has at the time the forecast is produced. This, clearly, is not a suitable approach as the sub-levels are intended to describe where avalanche danger is situated within a previously assigned danger level.

The danger levels are ordinal values with descriptions for each danger level. Hence, sub-levels cannot be calculated, nor is there a clear definition for them. The width of the sub-levels is therefore up to the subjective assessment of the avalanche forecasters. However, in order to ensure that sub-levels are used more evenly, the Swiss avalanche forecasters should be encouraged to rate, in case of doubt, 2–*plus* rather than 2–*neutral*, and at the same time 2–*neutral* rather than 2–*minus*.

Alternatively, we suggest a more consistent approach, which may reduce both spatial gradients between warning regions (i) and increase the use of 2–*plus* (ii) by automatically refining the sub-level as a function of the danger level in neighboring warning regions:

- sub-level *minus* is assigned, whenever a warning region borders at least one other region with a lower  $D_{RF}$  but no region with a higher  $D_{RF}$
- sub-level *plus* is assigned, whenever a warning region borders at least one other region with a higher  $D_{RF}$  but no region with a lower  $D_{RF}$

In the presented data set, this approach would revise the sub-level of about 13% of the cases. This adjustment would neither affect the danger level communicated to the public, nor the agreement rate between  $D_{RF}$  and  $D_{LN}$ , and it would only marginally and not significantly change the misclassification cost. However, it would reduce the sub-level gradients between neighboring warning regions (i) and would increase the proportion of *2-plus* (ii). However, aggregating the respective regions to form a separate danger region would be difficult since it would require at least some differences in the wording compared to the original description of avalanche danger for users to be able to understand why a separate danger region is given. Thus, with the present format of the avalanche forecast, such a refinement would mainly be useful for internal use or could be a basis for computer-driven models. Furthermore, introducing such a smoothing might be correct on average, but smaller or larger gradients may also be possible.

### Better than random?

Despite these observed anomalies in the use of the sub-levels, the comparison between local estimates and the forecast sub-levels showed that  $D_{RF.sub}$  was better than a random assignment of sub-levels ( $D_{RF.sub.random}$ ):

- $D_{RF.sub}$  was most often *neutral* when  $D_{RF} = D_{LN}$  (Fig. A.19a vs. A.19b). In contrast,  $D_{RF.sub.random}$  was most frequently *minus*.
- In 70% of the cases, when  $D_{RF} > D_{LN}$ ,  $D_{RF.sub}$  was the sub-level closest to the  $D_{LN}$  estimate. Thus,  $D_{RF.sub}$  leaned more strongly towards the local estimate than  $D_{RF.sub.random}$  (Fig. A.19a vs. A.19b).
- The misclassification cost was lower for  $D_{RF.sub}$  compared to  $D_{RF.sub.random}$  (Sect. A.4.4.3).
- *3-plus* was more often associated with natural avalanches (35%) than *3-minus* (4%; Tab. A.16).

This indicates that forecasters, at least when working in a setup as is currently the case at the national warning service in Switzerland, can indeed often refine avalanche danger at a higher resolution, by indicating the trend within the five ordinal danger levels.

#### A.4.5.4 Refining avalanche danger ratings in regional avalanche forecasts - operational prerequisites

The data show that it is possible to determine the regional danger level with greater detail than the five danger levels. Prerequisites for this, which apply to the provision of consistent and reliable forecasts in general, include:

- Relevant and reliable data must be available in a sufficient spatial density and temporal frequency.
- The warning regions, the smallest spatial units in the forecasts, must be sufficiently small, and their aggregation to danger regions must be highly flexible.

If the above requirements are fulfilled, a warning service should refine avalanche danger as detailed as possible, at least for internal assessment. This refinement has the following advantages:

- Expressing the conditions in a level of detail closer to the expected avalanche conditions during the forecast production process will increase consistency. While a categorization into fewer classes is

necessary to reduce the amount of complexity in the forecast product; this should, however, only be done at the end of the forecast process.

- Avalanche forecasters need to be aware of where in the danger level the current situation is located. This facilitates the discussion regarding the conditions and the formulation of consistent danger descriptions.

Such refined avalanche danger ratings may be used, for instance, to train statistical models, or they could be fed into computer-driven models like the Quantitative Risk Reduction method (Schmudlach et al., 2018). Particularly for such modeling approaches, the provision of a refined danger level could be highly relevant, considering that on more than 75% of all the forecasting days, only two of the five danger levels are forecast.

#### **A.4.5.5 Relevance to forecast users?**

In this study, we did not quantitatively explore whether providing sub-levels to the user would actually be beneficial. While we believe that some advanced users could benefit from this information, we suspect that a higher granularity of danger ratings may be primarily useful when integrated into computer models, as for instance those used on web platforms assisting back-country recreationists during the planning phase of a tour (Schmudlach et al., 2018), or to train statistical models assisting avalanche forecasters in their data analysis.

We could imagine that providing more specific information on expected avalanche size, the likelihood of natural avalanches, the additional load required to trigger an avalanche, and the frequency and location of these triggering locations might be of greater value to the user. However, the provision of this information must meet the same quality criteria as we explored for the sub-levels: only when information is of sufficient consistency and quality can it be of value to the user (Murphy, 1993).

#### **A.4.6 Conclusions**

We explored a four-year data-set of avalanche forecasts, which included the indication of three sub-levels refining the forecast regional dry-snow danger level. Comparing forecast danger levels with nowcast estimates, we noted a similar agreement rate of 81% between forecast and nowcast and a similar over-forecast bias as in previous studies. Additionally, we showed that the tendency towards over-forecasting was also present in a spatial and temporal context. Furthermore, we demonstrated that the forecast danger levels refined by sub-levels have skill, that is, they were better than a random assignment of sub-levels. This indicates that forecasters, at least when working in a similar setup as the national warning service in Switzerland, can indeed often refine avalanche danger at a higher resolution, by indicating the trend within the five ordinal danger levels. The results gained from this data analysis may support discussions on optimizing the granularity of avalanche danger ratings, last but not least for the internal assessment process and as a data basis for computer-driven models.

From our perspective, the discussion, whether such sub-level information - or other more specific information - should be provided to the public in avalanche forecast products, must include two aspects: (1) in terms

of consistency and quality, as explored here, and (2) in terms of the benefits from this additional information to the user of avalanche forecasts.

*Data availability:* The data will be made available at [envidat.org](http://envidat.org).

*Author contributions:* FT designed the study, conducted the analysis and wrote the manuscript. CP and KW repeatedly provided in-depth feedback on methodology and subsequent versions of the manuscript.

*Competing interests:* None.

*Acknowledgments:* We greatly appreciate the constructive feedback provided by two anonymous reviewers and Jürg Schweizer, which helped to improve this manuscript.

## A.5 On the importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level

Frank Techel, Karsten Müller, Jürg Schweizer: On the importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level. *The Cryosphere*, 2020. doi: 10.5194/tc-2020-42

### 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 light of uncertainties. Furthermore, the definitions of the danger levels, an ordinal variable, are vague and leave room for interpretation. Decision tools developed to assist in assigning a danger level 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 ( $N = 9,310$ ) and avalanche observations ( $N = 39,017$ ) from two countries related to the three key factors that characterize avalanche danger: snowpack stability, 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 (*very poor*, *poor*, *fair* and *good*), and define frequency classes describing the frequency of the most unstable locations (*none or nearly none*, *a few*, *several* and *many*). Combining snowpack stability, 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 stability distributions together with the largest avalanche size in a step-wise approach, we present a data-driven lookup table for avalanche danger assessment. Our findings may aid in refining the definitions of the avalanche danger scale and in fostering its consistent usage.

### A.5.1 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 (e.g. Statham et al., 2010) with brief definitions of the key factors. The key factors that characterize avalanche danger are (Meister, 1995; EAWS, 2020d, 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 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 frequency and location of the triggering spots is typically unknown. So far, it can only be assessed with laborious extensive sampling (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 or more generally the locations where 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 and/or slope aspects and elevation bands. The frequency distribution describes the question «How often do spots with a certain 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 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, 2017b). In contrast, the location of triggering spots or of snowpack stability refers to «Where in the terrain is avalanche release most likely?» It indicates where in the terrain the frequency is slightly higher (e.g. *where the snowpack is shallow, close to ridgelines, in bowls, . . .*). 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*.

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 (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 the European Avalanche Warning Services (EAWS) in 2005. The Bavarian Matrix, a lookup 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, 2020d). 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 the current version of the EAWS-Matrix and ADAM 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, the frequency distribution of snowpack stability and avalanche size. The data originate from different snow climates, and also from different avalanche warning services (Norway, Switzerland). The key questions are: (1) How do the three factors relate to the danger levels? and (2) Which combination of the actual value of the three factors 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 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.

## **A.5.2 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.

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.

### **A.5.2.1 Avalanche danger level**

The avalanche danger level 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 next (Sect.s A.5.2.2 and A.5.2.3).

Table A.18: Data overview.

parameter		country	N	data from*
avalanches	natural	SWI	29,511	2001-2019
	human-triggered	SWI	3,751	2001-2019
	natural	NOR	4,555	2014-2019
	human-triggered	NOR	1,200	2014-2019
RB		SWI	4,439	2001-2019
ECT		SWI	2,745	2007-2019
		NOR	2,126	2014-2019

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

### A.5.2.2 Snowpack stability

Operationally available information directly related to snow instability includes simple field observations as well as 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. 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  $30^\circ$ , where a 1.5 m  $\times$  2 m block of snow is isolated from the surrounding snowpack and loaded by a person (e.g. Föhn, 1987; Schweizer, 2002). An observer performing a RB records which of the 6 loading steps, referred to as the *score*, caused failure, and what portion of the block slid (the *release type*: whole block, most of block, edge only). If no failure occurs, RB7 is recorded. *Score* and *release type* provide information on failure initiation and crack propagation, essential components of 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 relatively small (30 cm  $\times$  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 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. A.5). 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 4,439 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.s.l. (IQR 2,134–2,547 m) and at 34° (IQR 31–36°). The Norwegian ECT data set consisted of 2,126 ECTs, observed at a median elevation of 760 m a.s.l. (IQR 730–1,067 m). Consistent information on the slope aspect was not available for Norwegian stability data.

### A.5.2.3 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 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:  $\geq 11$ ). In 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 of observed avalanches is described categorically, we assigned numbers as follows: one = 1, few (2-5) = 3, several (6-10) = 8, numerous ( $\geq 11$ ) = 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 33,262 in Switzerland, observed on 6,610 days and regions, and 5,755 in Norway, observed on 1,618 different days and regions (Table A.21).

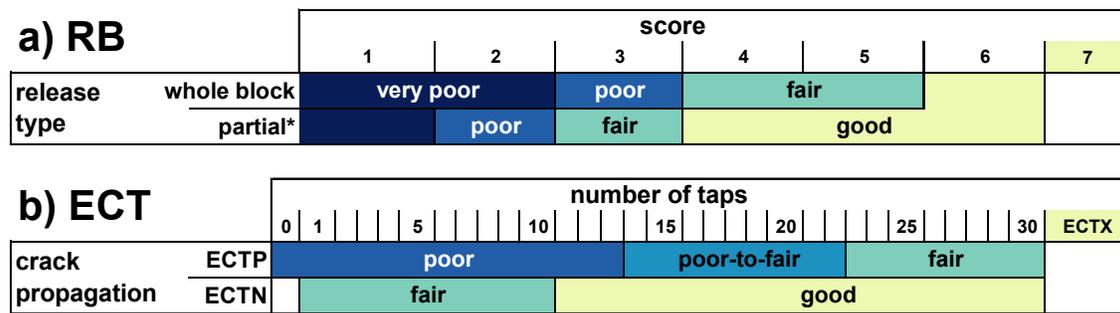


Figure A.20: Stability classification of (a) Rutschblock test results (based on Schweizer (2007a); Techel and Pielmeier (2014)) and (b) Extended Column Test results (based on Techel et al. (2020b)). \* - part of block includes release types most of block and edge only

### A.5.3 Methods

#### A.5.3.1 Classification of snow 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 (*very poor*, *poor*, *fair* and *good* - stability class names are in italics throughout this manuscript).

**Rutschblock test (RB)** results were classified in the four stability classes according to Figure A.20a 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. (2020b), we combined the two classes *very good* and *good* into one class called *good*.

**Extended Column Test (ECT)** results were classified relying on the classification recently suggested by Techel et al. (2020b). Using a combination of crack propagation and the number of taps until failure initiation, four stability classes were defined (Fig. A.20b). As the four stability classes for RB and ECT do not exactly line up, we assigned the following four class labels to the four ECT classes: *poor*, *poor-to-fair*, *fair* and *good* (as in Techel et al., 2020b).

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 *good*), if a failure layer was less than 10 cm below the snow surface (as in Techel et al., 2020b). The lowest stability class was retained for further analysis.

#### A.5.3.2 Simulation of snowpack stability distributions

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

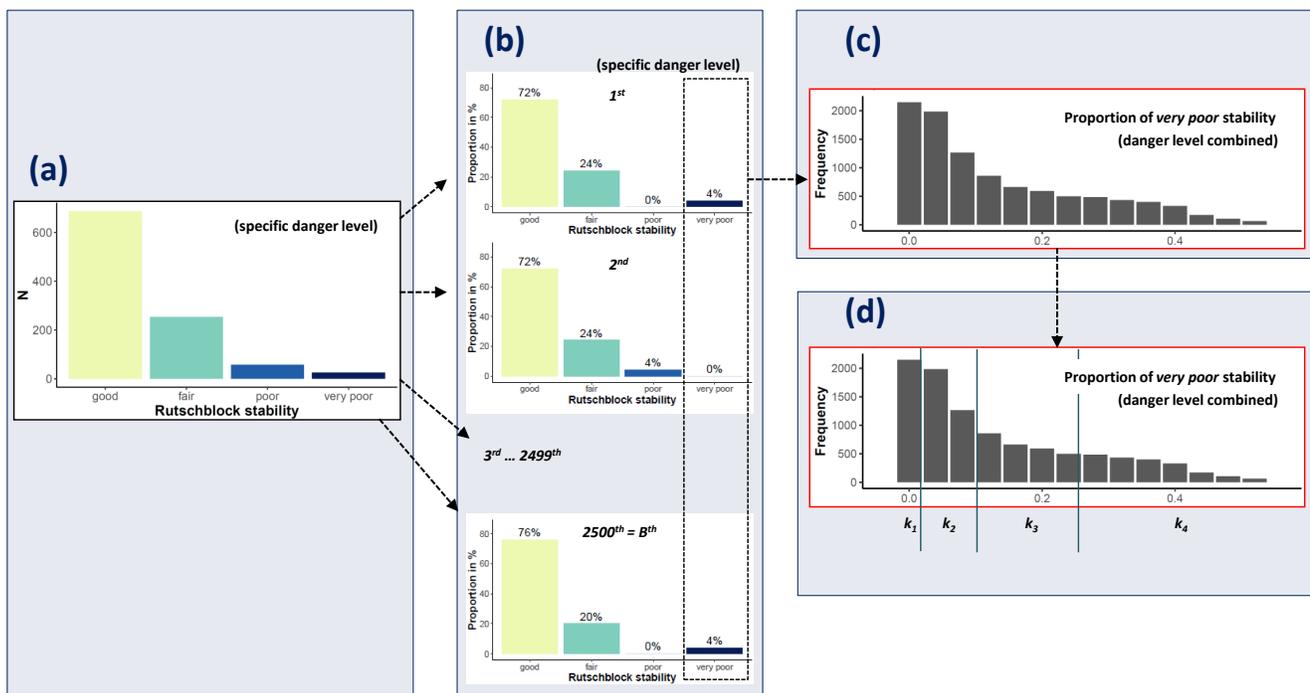


Figure A.21: Schematic representation of the workflow for bootstrap sampling and frequency class definition. a - For each danger level, all stability ratings are combined. b - From the observed stability distributions of a specific danger level (a),  $n$  tests are randomly sampled. This is repeated  $B = 2,500$  times to obtain typical stability distributions for each of the four danger levels. c - The  $4 \times 2,500$  bootstrap samples are merged and the proportion of *very poor* rated stability tests per sample is plotted as a histogram, irrespective of danger level. d - The statistics required for frequency class definitions are calculated and the  $k$  frequency classes defined. For details refer to the description in the Sections A.5.3.2 and A.5.3.3.

To determine the frequency 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 distributions, 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. A.21a and b):

- (i) We randomly selected  $n$  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  $B$  times for each danger level.
- (ii) For each of the  $B$  bootstrap samples, we calculated the proportions of *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  $B$  to be drawn. We used  $B = 2,500$

for each danger level, resulting in 10,000 stability distributions in total.

The second important parameter when bootstrap sampling is the number  $n$  of stability tests drawn in each sample. Small values of  $n$  increase variance, and hence overlap between samples drawn from different danger levels, and reduce the resolution of the desired statistic (e.g. for  $n = 10$ , the resolution is 0.1, for  $n = 100$  it is 0.01). 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 our statistic. Unfortunately, there are no studies we can refer to concerning the amount of overlap that would be appropriate. 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 surroundings of Davos (Switzerland).

### A.5.3.3 Snowpack stability and the frequency distribution of snowpack stability - approach to define frequency classes

Currently, 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 snowpack stability class. We considered the following points:

- Classes should be defined based on the snowpack stability class most relevant with regard to avalanche release, hence the frequency of the class *very poor*. Even though the focus is on the proportion of *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 *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 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 *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  $k$  classes with class interval limits being  $\{0, a, ab, ab^2, \dots, ab^{k-1}, 100\}$ , 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 100 can be given:

$$b = \left( \frac{100 - VP_{med}}{VP_{med}} \right)^{\frac{2}{k}}, \quad (\text{A.12})$$

where  $VP_{med}$  (0 - 100) is the median proportion of *very poor* stability, and  $k$  the number of classes preferred. This approach requires a suitable value of the number of classes  $k$  to be defined. Given  $k$  and  $b$ , the initial class width  $a$  is (Evans, 1977):

$$a = \frac{VP_{med}(100 - b)}{100 - b^{\frac{k}{2}}} \quad (\text{A.13})$$

To derive  $a$  and  $b$ , we generated snowpack stability distributions, as outlined in the previous section (see also Fig. A.21c and d).

#### **A.5.3.4 Combining snowpack stability and the frequency of snowpack stability with avalanche size: a simulation experiment**

When assigning a danger level, the information relating to snowpack stability and the frequency distribution of snowpack stability needs to be combined with avalanche size. 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. 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 snowpack stability and their respective frequency class (Sect.s A.5.3.2 and A.5.3.3). For each danger level, we 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. A.5.42) .

### **A.5.4 Results**

We first present the findings relating to the three contributing factors and their combination making use of Swiss Rutschblock and avalanche data (Sections A.5.4.1 - A.5.4.4). In a second step (Sect. A.5.4.5), the findings regarding snowpack stability and 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. A.5.4.6).

#### **A.5.4.1 Snowpack stability**

##### **Observed Rutschblock test stability distributions**

We analyzed the stability distributions obtained with the RB test at danger levels 1-Low to 4-High (Fig.

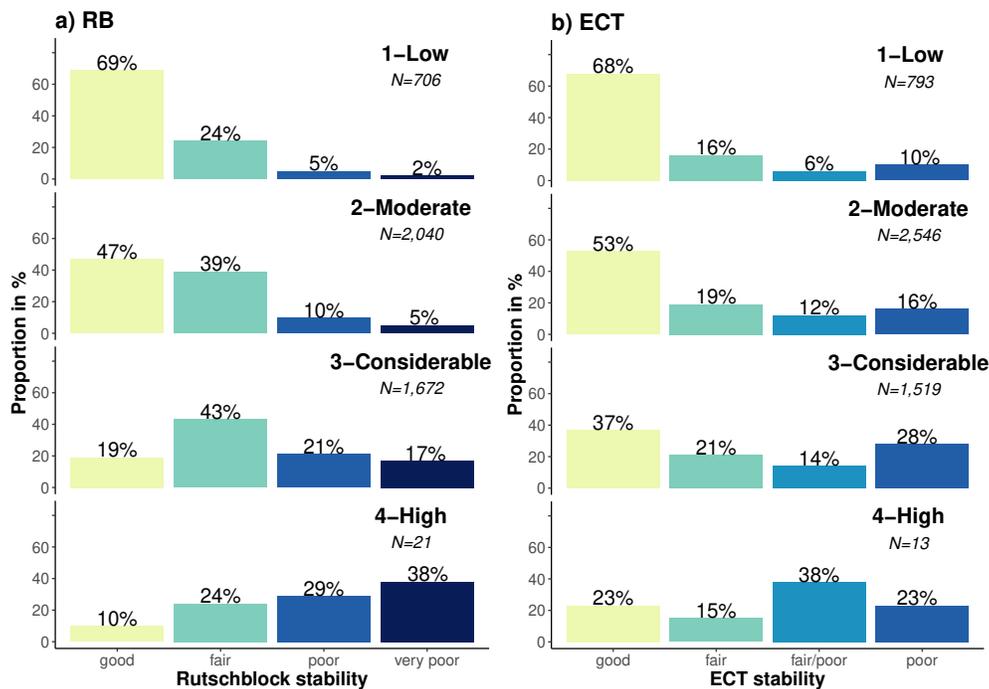


Figure A.22: Distribution of stability ratings for the stability tests (a) Rutschblock (RB) and (b) ECT for danger levels 1-Low to 4-High. For the definition of the stability classes refer to Fig. A.20 and Sect. A.5. Note the small  $N$  for 4-High for both tests.

A.22a). At 4-High, very few RB were observed. The proportion of *very poor* rated RB tests increased monotonically with increasing danger level from 2% at 1-Low to 38% at 4-High (Fig. A.22a). As a consequence, the combined proportion of *very poor* and *poor* rated tests also increased strongly from 7% to 67%, while the proportion of tests rated as *good* decreased accordingly (69% to 10%, Fig. A.22a). These patterns were also confirmed when exploring the correlation between the RB stability class and danger level (Spearman rank-order correlation;  $\rho = 0.4$ ,  $p < 0.001$ ).

### Frequency classes of *very poor* snowpack stability

Here, we describe the four frequency classes based on the frequency of *very poor* stability as sampled from the stability distributions shown in Fig. A.22a (bootstrap sample size  $n = 25$ ). Regarding the sampling and the class definition procedure refer to Sect.s A.5.3.2 and A.5.3.3, regarding the sensitivity of these settings on the results, refer to Sect. A.5.4.6.

Using four frequency classes, and labeling them *none or nearly none*, *a few*, *several* and *many*, the thresholds in the proportion *very poor* stability between frequency class labels were 0%, 4% and 20%, respectively (Tab. A.19). This corresponded to a median proportion *very poor* stability observed in each frequency class of 0%, 4%, 12%, 32%, or, if expressed in the number of *very poor* Rutschblock test results, in 0, 1, 3 or 8 RB out of 25 drawn.

Large proportions of *very poor* stability (e.g.  $\geq 50\%$ ) 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. A.22a), and using a low  $n$  in each of the bootstrap samples, which increases the

Table A.19: Frequency classification derived from the proportion of *very poor* stability ratings, using four frequency classes. The intervals for the frequency of *very poor* stability are shown.  $D(1^{st})$  and  $D(2^{nd})$  indicate the most and second most frequent danger level the samples were drawn from, respectively. 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 Figs A.25 and A.26. For class *none or nearly none* no letter is assigned, as the next higher stability class should be considered.

stability class	frequency class	interval* ( $n = 25$ )	danger level		letter in stability matrix
			$D(1^{st})$	$D(2^{nd})$	
<i>very poor</i>	<i>many</i>	>20% - 100%	4	3	A
	<i>several</i>	>4% - 20%	3	2	B
	<i>a few</i>	>0% - 4%	2	1	D
	<i>none or nearly none</i>	0% - 0%	1	2	
<i>poor</i>	<i>many</i>		2	3	C
	<i>several</i>		2	1	D
	<i>a few</i>		1	2	E
	<i>none or nearly none</i>		1	2	
<i>fair</i>	<i>many</i>		1	2	E
	<i>several</i>		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 one decimal space, the interval thresholds for the frequency of *very poor* stability for sampling with  $n = 25$  were: 0%, 1.8%, 6.2%, 21%, 100%.

variation in the sampled proportions.

The correlation between the frequency class describing the frequency of *very poor* stability and the danger level was strong ( $\rho = 0.81$ ,  $p < 0.001$ ; Fig. A.23). 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. A.23). 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.

#### A.5.4.2 Avalanche size

Most avalanches in the Swiss data set were size 1 (Fig. A.24a), 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 to 3-Considerable shows that the most frequent avalanche size has little discriminating power to differentiate between danger levels. The median avalanche size was

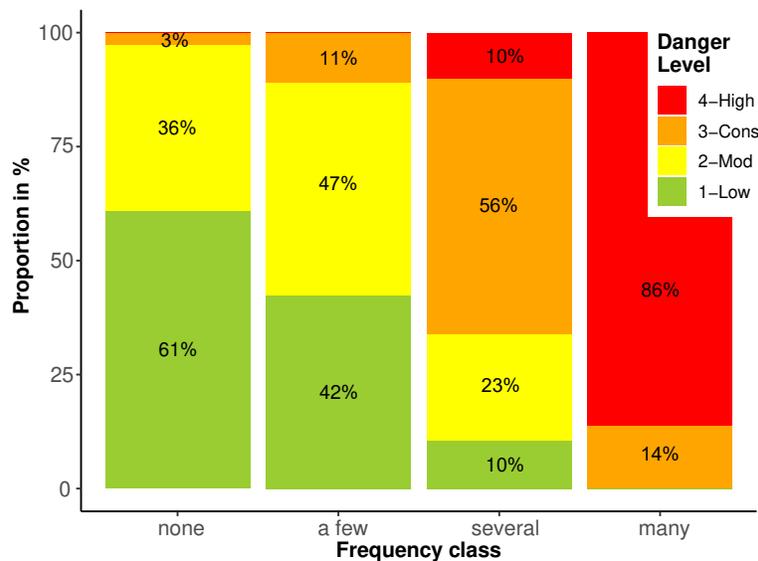


Figure A.23: Distribution of the danger levels for the four frequency classes describing the proportion of *very poor* snowpack stability, derived from sampling 25 Rutschblock tests (as described in Sect. A.5). The respective proportions are indicated for each of the four danger levels.

size 1 at 1-Low and 2-Moderate, size 1 to size 2 at 3-Considerable, and size 2 at 4-High (Fig. A.24a).

Considering the size of the largest reported avalanche per day and warning region showed that the largest avalanche per day and region was most frequently size 2 for 1-Low and 2-Moderate, a mix of size 2 and size 3 at 3-Considerable, and size 3 at 4-High (Fig. A.24b). The proportion of days when size 1 avalanches were the largest observed avalanche decreased significantly with increasing danger level (from 33% to 1%,  $p < 0.001$ ), while the proportion of days with at least one size 3 or size 4 avalanche increased significantly (from 20% to 78%,  $p < 0.001$ ). At 4-High, almost 80% of the days had at least one avalanche of size 3 or 4 recorded.

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

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 *very poor* stability as a proxy.

#### A.5.4.3 Combining the frequency of *very poor* stability and avalanche size

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

Sample 1: 1-Low, a few, largest avalanche size 1

Sample 2: 1-Low, none or nearly none, largest avalanche size 2

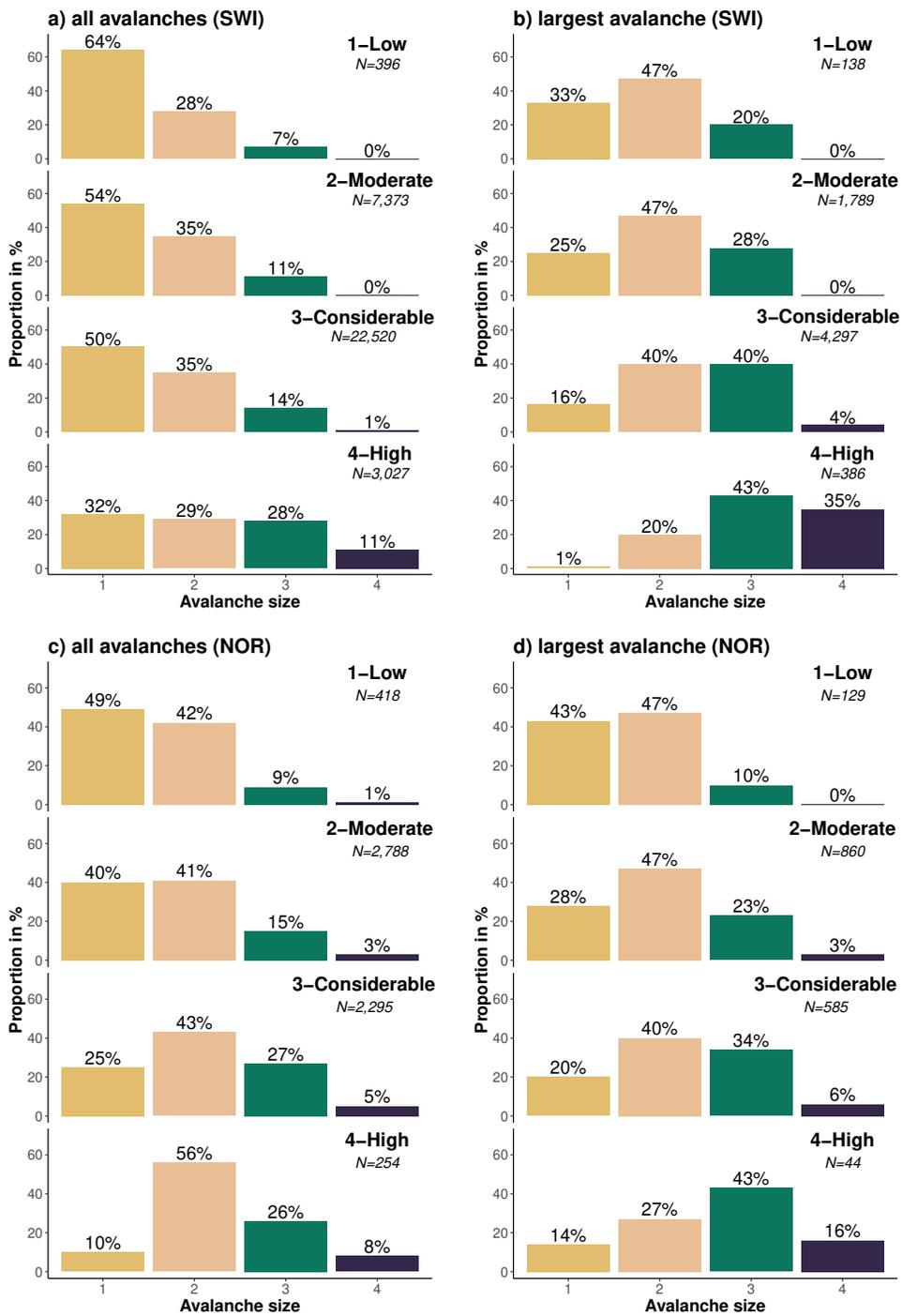


Figure A.24: Size distribution of dry-snow avalanches, which released naturally or were human-triggered for danger levels 1-Low to 4-High, showing all avalanches (a, c) and the largest reported avalanche per day and warning region (b, d) in Switzerland (SWI, upper row) and Norway (NOR, lower row).

Sample 3: 1-Low, a few, largest avalanche size 1

...

Sample B: 1-Low - none or nearly none - largest avalanche size 1

Tab. A.20 summarizes the simulated data set. The most frequent combinations of the frequency class

Table A.20: Table showing the combination of the frequency class of *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 highlight the most frequent combination, "-" indicates that these combinations did not exist.

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

and avalanche size for each danger level were:

- 1-Low: *None or nearly none* locations with *very poor* stability (53% of samples) existed. The largest avalanches were size 2 (48%).
- 2-Moderate: *A few* locations with *very poor* stability (37%) were present. The typical largest avalanche was of size 2 (50%).
- 3-Considerable: *Several* locations with *very poor* stability (75%) existed. The typical largest avalanches were sizes 2 or 3 (79%).
- 4-High: *Many* locations with *very poor* stability (86%) existed. The typical largest avalanche was of size 3 (43%).

#### A.5.4.4 Data-driven lookup table for danger level assessment

Finally, we present a data-driven lookup table to assess avalanche danger (Fig. A.25) 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).

The first matrix (Fig. A.25a), which we refer to as *stability matrix*, combines snowpack stability and the frequency class of the most unstable stability class observed. Cell labels (letters A to E) in this matrix were assigned based on similar danger level distributions behind the respective stability class - frequency class combination (Tab. A.19). 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. A.19, 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 these cells are shown in Figure A.26a.

The second matrix (Fig. A.25b), which we refer to as *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 characterizing this combination. If the second most frequent

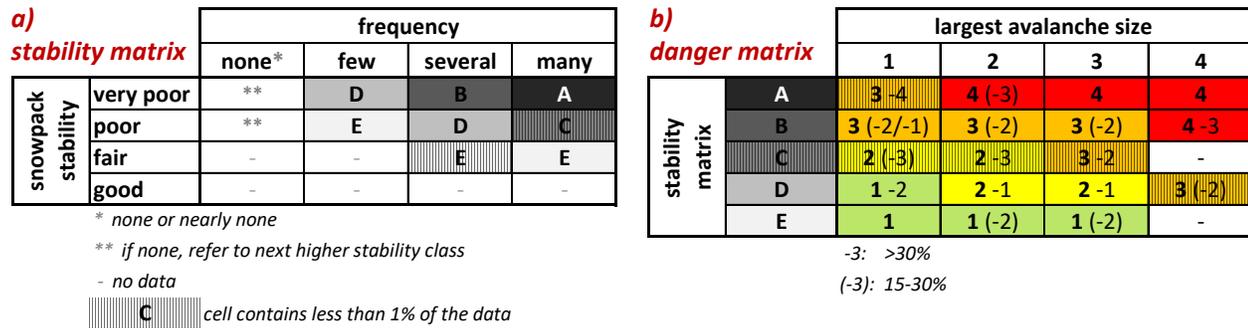


Figure A.25: Data-driven lookup table for avalanche danger assessment (similar to the structure proposed by Müller et al. (2016)). The (a) *stability matrix* combines the frequency class of the most unfavorable snowpack stability class (columns) and the snowpack stability class (rows) to obtain a letter describing specific stability situations, The (b) *danger matrix* combines the largest avalanche size (columns) and the specific stability situations (letter) obtained in the stability matrix (rows) to assess the danger level. In (b): The most frequent danger level is shown in bold. If the second most frequent danger level was present more than 30% of the cases, the value is shown with no brackets, if present between 15 and 30% it is placed in brackets. In (a) and (b): Cells containing less than 1% of the data are marked.

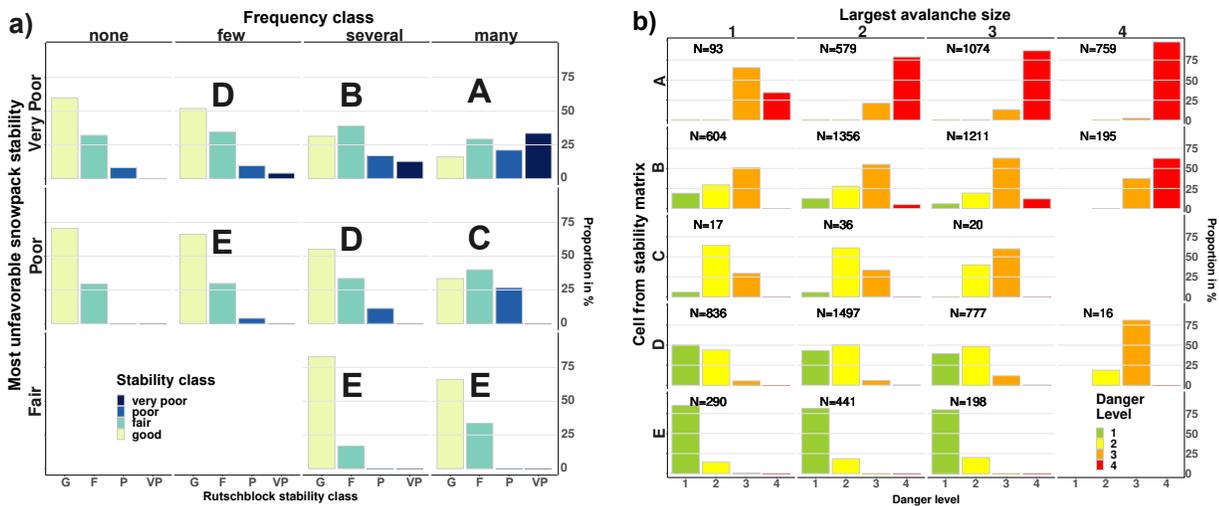


Figure A.26: Data behind the matrices shown in Figure A.25. The layout of the columns and rows is identical to Fig. A.25. The left figure (a) shows the mean simulated stability distributions behind the *stability matrix* (Fig. A.25a). Letters describe cells with the corresponding most frequent and second most frequent danger level. In the right figure (b), the distribution of danger levels for combinations of the typical largest avalanche size and the letters obtained before in the *stability matrix* (A-E, Fig. A.25a) are shown. The most frequent and second most frequent danger levels in each cell - avalanche size combination are shown in the *danger matrix* in the right part of Fig. A.25b.

danger level was present more than 30% of the cases, the value is shown with no brackets, if present between 15 and 30% it is placed in brackets. To illustrate the actual danger level distributions behind this matrix, Figure A.26b summarizes the simulated data.

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

1. In the *stability matrix* (Fig. A.25a), 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. A.25b), where it is combined with the largest avalanche size (Fig. A.25b).
3. The most frequent danger levels that were typical for this combination, are shown.

#### **A.5.4.5 Comparison with other data sets**

For the main results, presented in Sections A.5.4.1 to A.5.4.4, 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.

##### **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. A.22b).

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. A.22b). Again, a positive though weak correlation between stability rating and danger level was noted ( $\rho = 0.22$ ,  $p < 0.001$ ).

In comparison to the RB (Fig. A.22a, Sect. A.5), 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:  $\rho = 0.22$  vs. RB:  $\rho = 0.4$ ).

##### **Avalanche size: comparing Swiss and Norwegian avalanche size distributions**

The avalanche size distributions in Sect. A.5, based on observations made in Switzerland (SWI; Fig. A.24a, b), were compared to observations in Norway (NOR; Fig. A.24c, d).

In Norway, size 1 was the most frequently reported size at 1-Low, while size 2 avalanches were the most frequent size at 3-Considerable and 4-High (Fig. A.24c). 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. A.24d) showed similar trends in the size distributions as in Switzerland (Fig. A.24b). 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.

#### **A.5.4.6 Bootstrap sampling and frequency class definitions - sensitivity analysis**

##### **Bootstrap sampling**

To obtain a variety of frequency distributions of point snow instability, we sampled stability ratings as described in Sect. A.5.3.3. As outlined there, one important parameter affecting such a sampling approach is the number of stability ratings  $n$  drawn in each sample (sample size). In the following, we illustrate the effect of the bootstrap sample size  $n$ .

##### **Influence of sample size**

The results shown in Sections A.5.4.1, A.5.4.3 and A.5.4.4 were based on a sample size  $n = 25$ . To explore the effect of sample size, we in addition sampled using  $n = \{10, 25, 50, 100, 200, 1000\}$ . The histograms displaying the simulated proportion of *very poor* stability for various  $n$  irrespective of danger level showed that the distribution of the proportion of *very poor* stability was skewed towards lower proportions being more frequent than higher proportions, regardless of  $n$  (two examples for  $n = 25$  and  $n = 200$  are shown in the Fig. A.27a and c, respectively). Checking for multi-modality in the histograms by visual inspection and by applying the *modetest* (Ameijeiras-Alonso et al., 2018) showed that increasing the sample size  $n$  impacted the number of modes detected in the histograms, with two or more modes being present when  $n$  reached values of about 50. In the examples showing the proportions of *very poor* stability (Fig. A.27), the number of modes increased from one for  $n = 25$  (Fig. A.27a) to three for  $n = 200$  (Fig. A.27c). 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. A.27b and d). Again, it can be noted that the sampled distributions not only become visually more and more clustered with increasing  $n$ , but the overlap between danger levels decreases. In the examples shown, no obvious clustering can be noted for  $n = 25$  while four distinct clusters exist when sampling  $n = 200$  tests in each bootstrap. 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.

##### **Plausibility of sampled distributions: comparison with observations**

When introducing the bootstrap-sampling approach to create a range of plausible stability distributions (Sect. A.5.3.2), 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. A.28, it can be noted that indeed a range of typical distributions was obtained for the four danger levels. For instance, for  $n = 25$  and 3-Considerable, the interquartile range of the simulated propor-

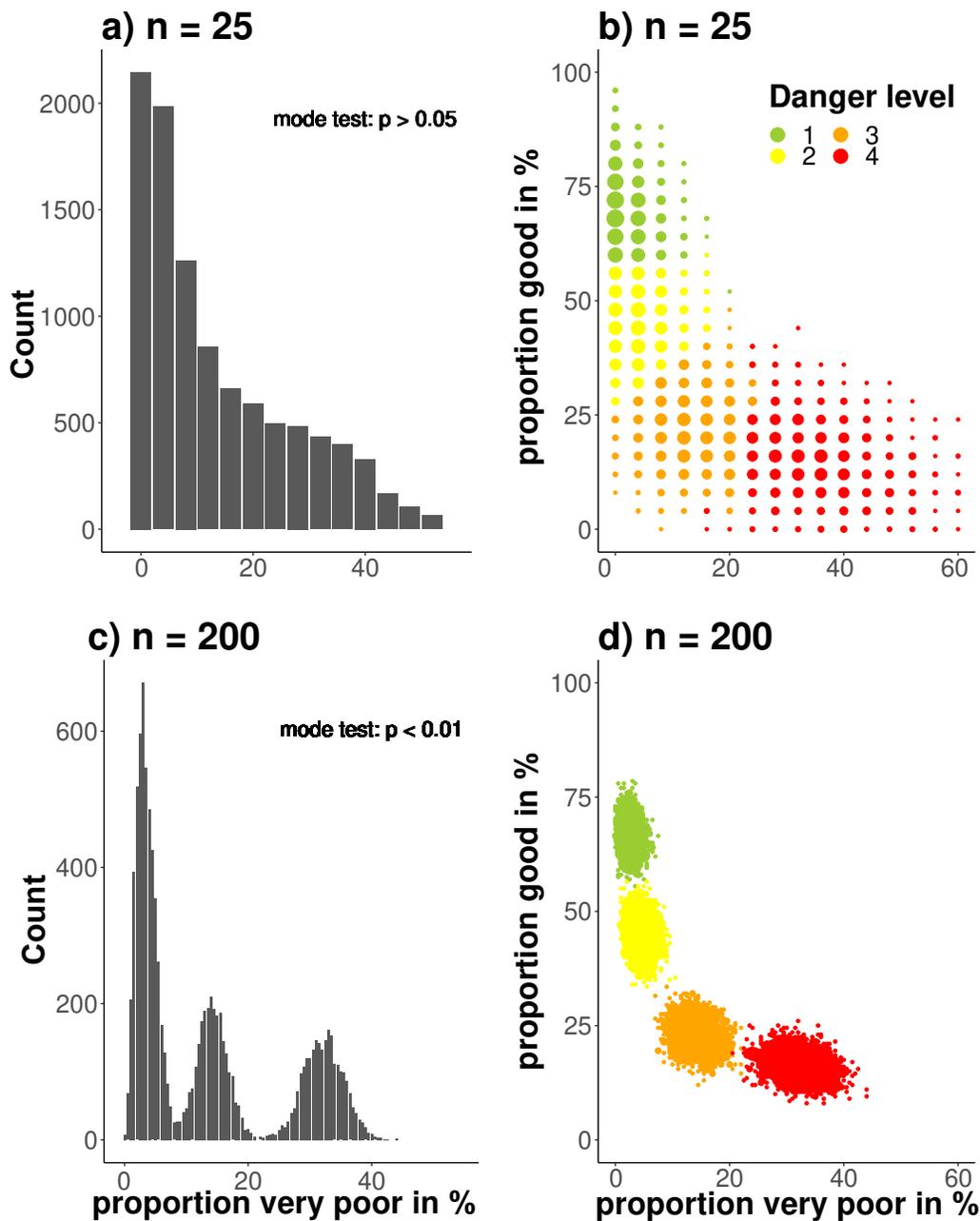


Figure A.27: Simulated proportions of *very poor* and *good* snowpack stability derived from RB tests for different number of samples  $n$  drawn in each of the bootstrap (upper row a and b:  $n = 25$ , lower row c and d:  $n = 200$ ). In the histograms (a, c) the proportion of *very poor* stability is shown, in the scatterplots (b, d) the most frequent danger level for a combination of *very poor* and *good* stability is shown. Note, the histogram in (a) is identical to Fig. A.21c. - The larger the sample size  $n$ , 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.

tions of *very poor* stability was between about 10% and 20% (Fig. A.28d), and for *good* stability between about 15% and 30% (Fig. A.28f).

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 distributions obtained using bootstrap-sampling reflected the variation in the observed distributions reasonably well (Fig. A.28). The exception was  $n = 10$

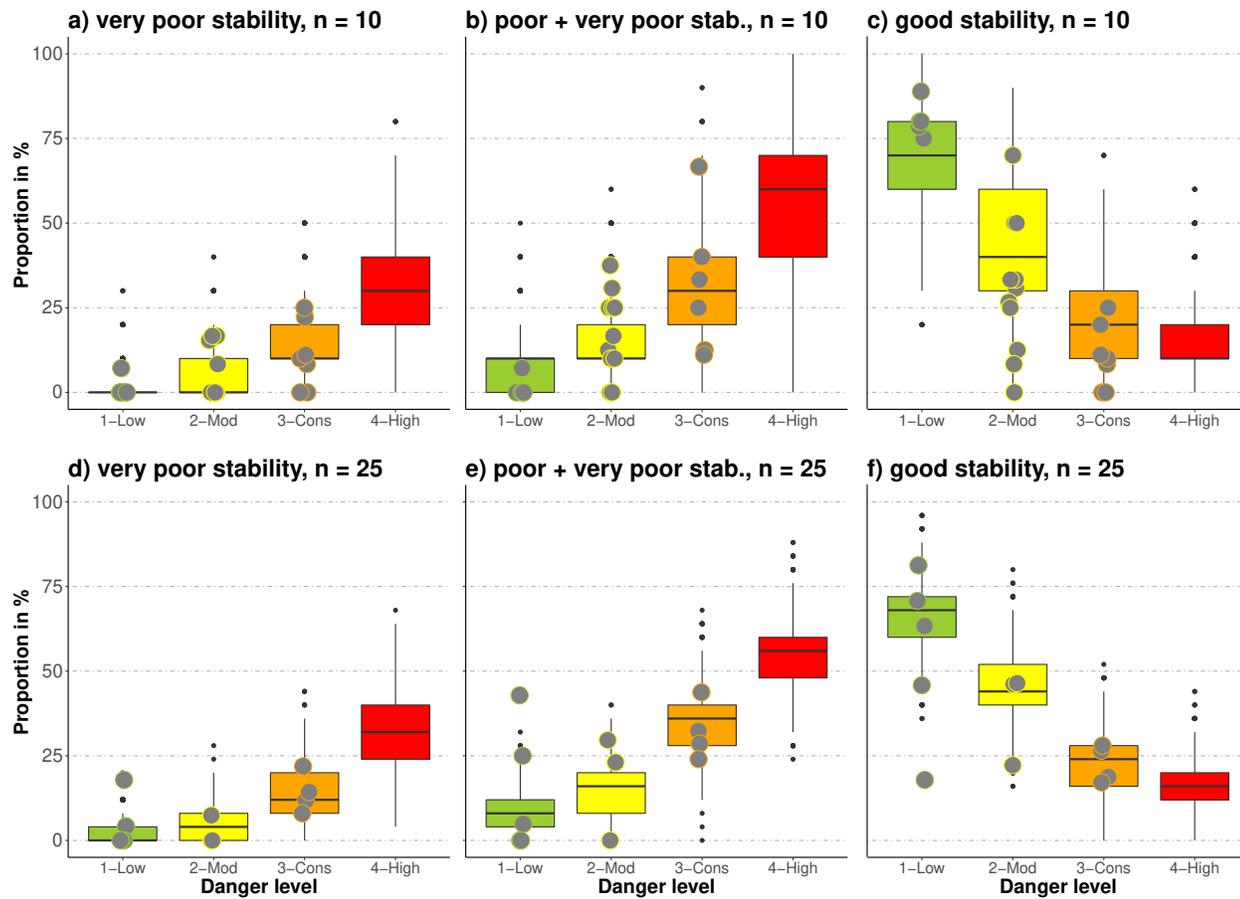


Figure A.28: Comparison of observed (points,  $N = 41$ ) and bootstrap-sampled distributions (boxes) for the proportion of *very poor* (a, d), *very poor* and *poor* combined (b, e) and *good* stability tests (c, f), for two settings of the number  $n$  of tests drawn. When 7 to 15 RB tests were observed on the same day and within the same region, these are shown together with sampled distributions using  $n = 10$  (a-c). When more than 16 tests were collected, these are shown together with sampled distributions using  $n = 25$  (d-f).

and *good* stability (Fig. A.28c): in this case, the observed proportions of *good* stability were significantly different than the sampled distributions at 2-Moderate and 3-Considerable ( $p < 0.05$ , Wilcoxon rank sum test).

In all examples shown in Fig. A.28, 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 in the proportions of a specific stability rating between danger levels, but also variation within.

### Frequency class definition

Relevant parameters for the definition of class intervals, as introduced in Sect. A.5.3.3, 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 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 ( $\rho > 0.7$ ,  $p < 0.001$ ).

## A.5.5 Discussion

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

### A.5.5.1 Data

#### 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. A.22), 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.

#### 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. A.24) 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 the Norwegian data set (Fig. A.24c). 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. 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.

To address potential bias in observations linked to Swiss observational standards (e.g. Techel et al., 2018), we 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). However, the largest avalanche size per day and warning region (Fig. A.24b and d) showed similar overall patterns across countries, with increasing frequencies of *very poor* stability and increasing avalanche size with increasing danger level.

Finally, stability test results, avalanche observations and local danger level 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 observations of natural or artificially triggered avalanches are unambiguous evidence of instability and may thus raise the quality of the local assessment.

#### **A.5.5.2 Methods**

##### **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. (2020b), Fig. A.20). 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. (2020b). 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. (2020b), 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 ECT stability classes and is supported by our findings: The RB stability class distributions changed more prominently from 1-Low (69% *good* stability, 2% *very poor*) to 4-High (10% *good*, 38% *very poor*) than the most favorable and unfavorable ECT stability classes (1-Low: 68% *good* stability, 10% *poor*, 4-High: 23% *good*, 23% *poor*).

### Simulation of stability distributions

We could not rely on a large number of stability 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. A.5.3.2) and by selecting sampling settings which lead to considerably overlapping distributions (Fig. A.28). We argue that some overlap in stability distributions would characterize the large variability of avalanche conditions. However, we do not know which number  $n$  of stability tests drawn captures the variation best. 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 (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 2000 RB stability ratings per danger level, at 4-High the number of observations was very small ( $N = 21$ ). Hence, both the data shown in Fig. A.22 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 *very poor* and *poor* stability increased, and those with *good* stability decreased at 4-High (Fig. A.22), judging whether the observed tests reflect the population well is difficult. Unfortunately, we are not aware of other studies, which have explored the 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/Switzerland (Schweizer et al., 2003), the forecast danger level 4-High was «verified» to be between 3-Considerable and 4-High (Schweizer, 2007b). On this day, 14 Rutschblock tests were observed. 36% of these were either *very poor* or *poor*, thus being close to the average values noted for 3-Considerable (Fig. A.22a). We did not consider these data, as we did not analyze data when for intermediate danger levels.

Comparing the distributions of our 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). For instance, the proportion of *very poor* and *poor* combined was at 2-Moderate about 15% and at 3-Considerable about 40%, which is lower than findings by Schweizer et al. (2003) (20-25% and about 50%, respectively). At 1-Low, about 70% of the RB tests were classified as *good*, while Schweizer et al. (2003) noted about 90% of the profiles to have *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).

### Classification of snowpack stability frequency distributions

In addition to simulating snowpack stability distributions using a re-sampling approach, we developed a data-driven classification of the proportion of *very poor* stability tests. Our approach shows that the number  $n$  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 *very poor* stability tests  $VP_{med}$ , and by the number of classes preferred  $k$ .

Assigning a class to the proportion of *very poor* stability, however, was affected by  $n$  due to the fact that  $n$  influences both the resolution of the statistic and the variance. 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  $k$  may depend on a number of factors. We suggest that defining  $k$  should be guided by keeping classes as distinguishable as possible - for instance by addressing the frequently occurring low proportions of *very poor* stability on one side and the rarely observed large proportions of *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.

#### A.5.5.3 Data interpretation

**Snowpack stability and frequency distribution of snowpack stability** We showed an increasing frequency (or number of locations) of *very poor* snowpack stability with increasing danger level, in line with previous studies exploring point snowpack stability within a region or small basin (Schweizer et al., 2003; Reuter et al., 2016) or the number of natural and human-triggered avalanches within a region (e.g. Schweizer et al., 2020). Furthermore, we showed that high frequencies of *very poor* stability ( $\geq 30\%$ ) were comparably rare (15% of the simulated distributions). Even at 4-High, less than 4% of the distributions had frequencies of *very poor* stability  $\geq 50\%$ .

We explored snowpack stability using RB and ECT, which describe the stability at a specific point. However, within a slope or a region, point snowpack stability is variable (e.g. Birkeland, 2001; Schweizer et al., 2008a). In avalanche forecasts this can be expressed by the frequency a certain stability class exists and by additionally describing the locations more specifically. When describing the avalanche danger level in a region, snowpack stability and the frequency distribution of snowpack stability must therefore be considered. We suggest that primarily the frequency of the lowest stability class is relevant for assigning a danger level, as this stability class combined with 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. These two factors must therefore be assessed in combination for all aspects and elevations. Furthermore, the specific description of triggering locations, for instance *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 (EAWS, 2017b) and the CMAH (Statham et al., 2018a) first combine snowpack stability and the frequency distribution of snowpack stability, 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 *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., 2020b). Thus, depending on the data used to define *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 to capture the frequency of locations where natural avalanches may initiate or where human-triggered avalanches are possible.

### **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 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 were observed per 100 km<sup>2</sup> (Bühler et al., 2019; Zweifel et al., 2019).

### **Combining snowpack stability, the frequency distribution of snowpack stability and avalanche size**

We presented a data-driven lookup table to assess avalanche danger (Fig. A.25). As can be seen in this table, the combination of snowpack stability and the frequency that best matches an avalanche situation (A to E), is highly relevant for danger level assessment. In general, avalanche size 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 snowpack stability, should lead to an increase in danger level by one level. Clearly, the presented data-driven lookup table (Fig. A.25) highlights that a greater focus must be placed on snowpack stability and the frequency distribution of snowpack stability, compared to avalanche size, when assessing avalanche hazard. This was also shown by Clark (2019), who 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 *stability matrix* (Fig. A.25), also had a greater impact on the resulting danger level than avalanche size, even though avalanche size  $\leq 1.5$  (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 lookup table presented in Fig. A.25 should therefore be seen as (a) a tool aiding the discussion of specific situations, and (b) to improve the definitions underlying the categorical descriptions of the danger levels.

### A.5.6 Conclusions

We explored observational data from two different countries relating to the three key factors describing avalanche hazard, snowpack stability, the frequency distribution of snowpack stability and avalanche size. We simulated stability distributions and defined four classes describing the frequency of potential avalanche triggering locations, 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. A.22a, Fig. A.27a-d).

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

- The proportion of *very poor* or *poor* stability test results increased from one danger level to the next higher one (Figures A.22 and A.28). Considering *very poor* snowpack stability and the frequency of this stability class alone, already distinguished well between danger levels (Tab. A.19, Fig. A.23).
- Considering the largest observed avalanche size per day and warning region was most relevant to distinguish between 3-Considerable and 4-High (Fig. A.24 and Tab. A.20). 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. A.24).

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 a significant additional increase in the 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 lookup tables (Fig. A.25), 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 such as the avalanche danger scale.

*Data availability:* The data will become freely available at [www.envidat.org](http://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.

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

## A.6 On snow stability interpretation of Extended Column Test results

Techel, F., Winkler, K., Walcher, M., van Herwijnen, A. and Schweizer, J.: On snow stability interpretation of extended column test results. *Nat. Hazards Earth Syst. Sci.*, 2020, 1941-1953, doi: 10.5194/nhess-2020-50

### Abstract

Snow instability tests provide valuable information regarding the stability of the snowpack. Test results are key data used to prepare public avalanche forecasts. However, to include them into operational procedures, a quantitative interpretation scheme is needed. Whereas the interpretation of the Rutschblock test (RB) is well established, a similar detailed classification for the Extended Column Test (ECT) is lacking. Therefore, we develop a 4-class stability interpretation scheme. Exploring a large data set of 1719 ECTs observed at 1226 sites, often performed together with a RB in the same snow pit, and corresponding slope stability information, we revisit the existing stability interpretations, and suggest a more detailed classification. In addition, we consider the interpretation of cases when two ECTs were performed in the same snow pit. Our findings confirm previous research, namely that the crack propagation propensity is the most relevant ECT result and that the loading step required to initiate a crack is of secondary importance for stability assessment. The comparison with the RB showed that the ECT classifies slope stability less reliably than the RB. In some situations, performing a second ECT may be helpful, when the first test did neither indicate rather unstable nor stable conditions. Finally, the data clearly show that false-unstable predictions of stability tests outnumber the correct-unstable predictions in an environment where overall unstable locations are rare.

### A.6.1 Introduction

Gathering information about current snow instability is crucial when evaluating the avalanche situation. However, direct evidence of instability - as recent avalanches, shooting cracks or whumpf sounds - is often lacking. When such clear indications of instability are absent, snow instability tests are widely used to obtain information on the stability of the snowpack. Such tests provide information on failure initiation and subsequent crack propagation - essential components for slab avalanche release (Schweizer et al., 2008b; van Herwijnen and Jamieson, 2007). However, performing snow instability tests is time-consuming, as they require to dig a snow pit. Furthermore, considerable experience in the selection of a representative and safe site is needed, and the interpretation of test results is challenging (Schweizer and Jamieson, 2010). Alternative approaches such as interpreting snow micro-penetrometer signals (Reuter et al., 2015), are promising, but not sufficiently established yet.

Two commonly used tests to assess snow instability are the Rutschblock test (RB, Föhn, 1987) and the Extended Column Test (ECT; Simenhois and Birkeland, 2006, 2009). For both tests blocks of snow are isolated from the surrounding snowpack. According to test specifications, the block is then loaded in sev-

eral steps. The loading step leading to a crack in a weak layer (failure initiation) is recorded, and whether crack propagation across the entire block of snow occurs (crack propagation). For the RB, the interpretation of the test result is well established and involves combining failure initiation (score) and crack propagation (release type) (e.g. Schweizer, 2002; Winkler and Schweizer, 2009). In contrast, the original interpretation of ECT results considers crack propagation propensity only (Simenhois and Birkeland, 2006, 2009; Ross and Jamieson, 2008): if a loading step leads to a crack propagating across the entire column, the result is considered as *unstable*, else as *stable*. However, Winkler and Schweizer (2009) suggested improving this binary classification by additionally considering the loading step required to initiate a crack and by considering a minimal failure layer depth leading to interpretations of ECT results as *unstable*, *intermediate* and *stable*. Moreover, they hypothesized that performing two tests, and considering differences in test results, may help to establish an intermediate stability class.

As the properties of the slab as well as the weak layer may vary on a slope (Schweizer et al., 2008a), reliably estimating slope stability requires many samples (Reuter et al., 2016) and a single test result may not be indicative. Hence, it was suggested to perform more than one test, either in the same snow pit or in a distance beyond the correlation length, which is often on the order of  $\leq 10$  m (Kronholm et al., 2004). For instance, Schweizer and Bellaire (2010) analysed whether performing two pairs of Compression Tests (CT) about 10 m apart improves slope stability evaluation. They suggested a sampling strategy that essentially suggests that in case the first test does not indicate instability, additional tests can reduce the number of false-stable predictions. Moreover, they reported that in 61–75% of the cases the two tests in the same pit provided consistent results, in the remaining cases either the CT score or the fracture type varied. For the ECT, several authors also noted that two tests performed adjacent to each other in the same snow pit or at several meters distance within the same small slope showed different results (Winkler and Schweizer, 2009; Hendrikx et al., 2009; Techel et al., 2016c). For instance, Techel et al. (2016c) reported that in 21% of the cases the ECT fracture propagation result differed between two tests in the same snow pit. Moreover, they explored differences in the performance between the ECT and the RB with regard to slope stability evaluation and found that the RB detected more stable and unstable slopes correctly than a single ECT or two adjacent ECTs.

Both ECT and RB provide information relating to slab avalanche release. While the Rutschblock provides reliable results, the ECT is quicker to perform in the field, which probably explains why it has quickly become the most widely used instability test in North America (Birkeland and Chabot, 2012). Given the popularity of the ECT as a test to obtain snow instability information and the lack of a quantitative interpretation scheme that includes more than just two classes, our objective is to revisit the originally suggested stability interpretations and to specifically consider cases when two ECTs were performed in the same snow pit. Building on our findings, we propose a new stability classification differentiating between cases when just a single ECT and when two adjacent ECTs were performed in the same snow pit with the goal to minimize false-stable and false-unstable predictions. Additionally, we empirically explore the influence of the base rate frequency of unstable locations on stability test interpretation, which - if neglected - may lead to false interpretations (Ebert, 2019). We address this topic by exploring a large set of ECTs with observations of slope stability collected in Switzerland. Furthermore, ECT results are compared with concurrent RB test results.

## A.6.2 Data

Data were collected in 13 winters from 2006-2007 to 2018-2019 in the Swiss Alps. We explored a data set of stability test results in combination with information on slope stability and avalanche hazard.

At 1226 sites, where slope stability information was available, 1719 ECT were performed (Tab. A.21). At 487 out of the 1226 sites either one (279) or two ECTs (208) were performed (695 ECTs in total). At the other 739 sites, a RB test was conducted in addition to either one (484) or two ECTs (285) in the same snow pit (1024 ECTs in total).

### A.6.2.1 Extended Column Test (ECT) and Rutschblock test (RB)

At sites where ECT and RB were realized in the same snow pit, one or two ECTs were generally performed directly down-slope from the RB (e.g. as described in detail in Winkler and Schweizer (2009)). If no RB was performed but two ECTs were performed, it is not known whether the ECTs were performed side-by-side, or whether the second ECT was located directly up-slope from the first ECT.

Test procedure followed observational guidelines (Greene et al., 2016). For the ECT, loading is by tapping on the shovel blade positioned on the snow surface on one side of the column of snow isolated from the surrounding snowpack (30 loading steps, Fig. A.29a). For the RB, a person on skis stands or jumps on the block (6 loading steps, Fig. A.29b). When a crack initiates and propagates within the same weak layer across the entire column within one tap of crack initiation, it is called *ECTP* for the ECT; for the RB this corresponds to the release type *whole block*. If the crack does not propagate within the same layer across the entire column or within one tap of crack initiation, *ECTN* is recorded for the ECT. Similarly, if the fracture does not propagate through the entire block, *part of block* or *edge only* are recorded as RB release type. If no failure can be initiated including loading step 30 (ECT) or 6 (RB), these are recorded as *ECTX* or *RB7*, respectively.

### A.6.2.2 Stability classification of ECT and RB

To facilitate the distinction between the result of an instability test and the stability of a slope, we refer to test stability using four classes 1 to 4, with class 1 being the lowest stability (*poor* or less) and class 4 the highest stability (*good* or better). In contrast, for slope stability, we use the terms *unstable* and *stable*. We chose four classes as a similar number of classes has been used for RB stability interpretation, as outlined below.

Table A.21: Data overview with the number (N) and proportion of *unstable* rated slopes.

stability tests	N	<i>unstable</i>
single ECT	279	15%
two ECT	208	30%
single ECT and a RB	454	20%
two ECT and a RB	285	20%

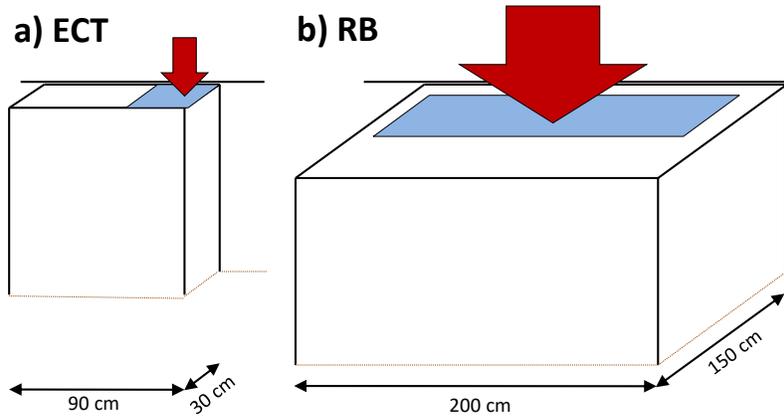


Figure A.29: ECT and RB according to observational guidelines. At the back, the block of snow is isolated by either cutting with a cord or a snow saw. The lightblue area indicates the approximate area, where the skis or the shovel blade is placed. This area corresponds to the area loaded for the ECT, while the main load under the skis is exerted over a length of about 1 m (Schweizer and Camponovo, 2001). Loading is from above (arrows).

**Extended Column Test (ECT):** The stability classification originally introduced by Simenhois and Birkeland (2009) ( $ECT_{orig}$ ) suggested two stability classes:  $ECTN$  or  $ECTX$  are considered to indicate high stability (class 4), while  $ECTP$  indicates low stability (class 1).

The classification suggested by Winkler and Schweizer (2009) ( $ECT_{w09}$ ) uses three classes:

- $ECTP \leq 21$ : low stability (class 1)
- $ECTP > 21$ : intermediate stability (class 2-3)
- $ECTN$  or  $ECTX$ : high stability (class 4)

**Rutschblock test:** We classified the RB in four classes (classes 1 to 4; Fig. A.30). We followed largely the RB stability classification by Techel and Pielmeier (2014), who used a simplified version of the classification used operationally by the Swiss avalanche warning service (Schweizer and Wiesinger, 2001; Schweizer, 2007a). Schweizer (2007a) defined five stability classes for the RB, based on the score and the release type in combination with snowpack structure, while Techel and Pielmeier (2014) relied exclusively on RB score and release type. In contrast to both these approaches, we combined the two highest classes (*good* or *very good*) to one class (class 4).

Shallow weak layers ( $\leq 15$  cm) are rarely associated with skier-triggered avalanches (Schweizer and Lütschg, 2001; van Herwijnen and Jamieson, 2007), which is, for instance, reflected in the threshold sum approach

RB		score							
		1	2	3	4	5	6	7	
release type	whole block	1	2	2	3	4	5	6	7
	partial release*			3	4				

Figure A.30: Classification of RB into four stability classes. \*combines release type *part of block* and *edge only*.

(Schweizer and Jamieson, 2007), a method to detect structural weaknesses in the snowpack. Schweizer and Jamieson (2007) reported the critical range for weak layers particularly susceptible to human triggering as 18-94 cm below the snow surface. Minimal depth criteria were also taken into account by Winkler and Schweizer (2009) in their comparison of different instability tests or by Techel and Pielmeier (2014), when classifying snow profiles according to snowpack structure. We addressed this by assigning stability class 4 if the failure layer was less than 10 cm below the snow surface. If there were several failures in the same test, we searched for the ECT and RB failure layer with the lowest stability class.

### **A.6.2.3 Slope stability classification**

We classified stability tests according to observations relating to snow instability in similar slopes as the test on the day of observation, such as recent avalanche activity or signs of instability (whumpfs or shooting cracks). This information was manually extracted from the text accompanying a snow profile and/or stability test. This text contains - among other information - details regarding recent avalanche activity or signs of instability.

A slope was called *unstable* if any signs of instability or recent avalanche activity - natural or skier-triggered avalanches from the day of observation or the previous day - were noted on the slope where the test was carried out or on neighboring slopes (Simenhois and Birkeland, 2006, 2009; Moner et al., 2008; Winkler and Schweizer, 2009; Techel et al., 2016c).

We called a slope only as *stable* if it was clearly stated that on the day of observation none of the before-mentioned signs were observed in the surroundings. In most cases, surroundings relates to observations made in the terrain covered or observed during a day of back-country touring (estimated to be approximately 10 to 25 km<sup>2</sup>, Meister, 1995; Jamieson et al., 2008).

In the following, we denote slope stability simply as *stable* or *unstable*, although this strict binary classification is not adequate. For instance, many tests were performed on slopes that were actually rated as *unstable*, though did not fail. In other words, *unstable* has to be understood as a slope where the triggering probability is relatively high compared to *stable* where it is low.

If it was not clearly indicated, when and where signs of instabilities or fresh avalanches were observed, or if this information was lacking entirely, these data were not included in our dataset.

### **A.6.2.4 Forecast avalanche danger level**

For each day and location of the snow instability test, we extracted the forecast avalanche danger level related to dry-snow conditions from the public bulletin issued at 17.00 CET, and valid for the following 24 hours.

## **A.6.3 Methods**

### **A.6.3.1 Criteria to define ECT stability classes**

We consider the following criteria as relevant when testing existing or defining new ECT stability classes:

- (i) Stability classes should be distinctly different from each other. The criteria we rely on is the proportion of *unstable* slopes. Therefore, a higher stability class should have a significantly lower proportion of *unstable* slopes than the neighboring lower stability class.
- (ii) The lowest and highest stability classes should be defined such that the rate of correctly detecting *unstable* and *stable* conditions is high, respectively; hence, the rate of *false-stable* and *false-unstable* predictions should be low, respectively. Stability classes in-between these two classes may represent *intermediate* conditions, or lean towards more frequently *unstable* and *stable* conditions, permitting a higher *false-stable* and *false-unstable* rate than the rates of the two extreme stability classes.
- (iii) The extreme classes should occur as often as possible, as the test should discriminate well between *stable* and *unstable* conditions in most cases.

To define classes based on crack propagation propensity and crack initiation (number of taps), we proceeded as follows:

1. We calculated the mean proportion of *unstable* slopes for moving windows of 3, 5 and 7 consecutive number of taps, for *ECTP* and *ECTN* separately. *ECTX* was included in *ECTN*, treating *ECTX* as *ECTN31*.
2. We obtained thresholds for class intervals by applying unsupervised k-means-clustering (R-function *kmeans* with settings `max.iter = 100`, `nstart = 100`; R Core Team (2017); Hastie et al. (2009)) on the proportion of *unstable* slopes of the three running means (step 1). The numbers of clusters *k* tested were 3, 4 and 5.
3. We repeated clustering 100 times using 90% of the data, which were randomly selected without replacement. For each of these repetitions, the cluster boundaries were noted. Based on the 100 repetitions, we report the respective most frequently observed *k*-1 boundaries, together with the second most frequent boundary.
4. To verify whether the classes found by the clustering algorithm were distinctly different (criterion *i*), we compared the proportion of *unstable* slopes between clusters using a two-proportions z-test (*prop.test*, R Core Team (2017)). We considered p-values  $\leq 0.05$  as significant.  
In almost all cases, we used a one-sided test with the null hypothesis  $H_0$  being either  $H_0: \text{prop}(A) \leq \text{prop}(B)$  (or its inverse), where *prop* is the proportion of *unstable* slopes in the respective cluster A or B. The alternative hypothesis  $H_a$  would then be  $H_a: \text{prop}(A) > \text{prop}(B)$  (or its inverse).
5. For clusters not leading to a significant reduction in the proportion of *unstable* slopes, we tested a range of thresholds ( $\pm 3$  taps within the threshold indicated by the clustering algorithm) to find a threshold maximizing the difference between cluster centers and leading to significant differences ( $p \leq 0.05$ ) in the proportion of *unstable* slopes (criterion *ii*). If no such threshold could be found, clusters were merged.

Throughout this manuscript, we report p-values in four classes ( $p > 0.05$ ,  $p \leq 0.05$  when  $p = [0.05, 0.01[$ ,  $p \leq 0.01$  when  $p = [0.01, 0.001[$  and  $p \leq 0.001$ ).

### A.6.3.2 Assessing the performance of stability tests and their classification

When the predictive power or predictive validity of a test is assessed, it is compared to a reference standard, here the slope stability classified as either *unstable* or *stable*. The usefulness of instability test results is generally assessed by considering only two categories related to *unstable* and *stable* conditions (Schweizer and Jamieson, 2010). We refer to these two outcomes as *low* or *high* stability.

There are two different contexts a test's adequacy is looked at: the first explores whether (a) the foundations of a test are satisfactory, and (b) the test is useful (Trevethan, 2017):

(a) Most often the performance of a snow stability test is assessed from the perspective of the reference group (Schweizer and Jamieson, 2010), i.e. what proportion of *unstable* slopes are detected by the stability test. The two relevant measures addressing this context are the sensitivity and specificity, which are considered as the benchmark for the performance:

- The sensitivity of a test is the probability of correctly identifying an *unstable* slope from the slopes that are known to be *unstable*. Considering a frequency table (Tab. A.22) the sensitivity, or probability of detection (POD), is calculated as (Trevethan, 2017):

$$\text{Sensitivity (POD)} = \frac{a}{a + c} \quad (\text{A.14})$$

- The specificity of a test is the probability of correctly identifying a *stable* slope from the slopes that are known to be *stable*. It is also referred to as the probability of non-detection (PON).

$$\text{Specificity (PON)} = \frac{d}{b + d} \quad (\text{A.15})$$

Ideally, both sensitivity and specificity are high, which means that most *unstable* and most *stable* slopes are detected. However, missing *unstable* situations can have more severe consequences and therefore it is assumed that first of all the sensitivity should be high. Nonetheless, a comparably low specificity will decrease a test's credibility.

(b) The second context focuses on the ability of a test to correctly indicate slope stability, i.e. if the test result indicates low stability, how often is the slope in fact *unstable*. This aspect has only rarely been explored for snow instability tests (e.g by Ebert (2019) from a Bayesian viewpoint), and is generally assessed using two metrics:

- The positive predictive value (PPV) is the proportion of *unstable* slopes, given that a test result indicates instability (a low stability class).

Table A.22: 2×2 frequency table cross-tabulating slope stability and test results. A positive test result indicates *low* stability, a negative test result *high* stability.

		slope stability	
		<i>unstable</i>	<i>stable</i>
test result ( <i>stability</i> )	positive ( <i>low</i> )	a	b
	negative ( <i>high</i> )	c	d

$$PPV = \frac{a}{a + b}$$

- The negative predictive value (NPV) is the proportion of *stable* slopes, given that a test result indicates stability (a high stability class).

$$NPV = \frac{d}{c + d}$$

In the following, we will use PPV and 1-NPV in the sense that it reflects the proportion of *unstable* slopes given a specific test result in a setting with up to four test outcomes (classes 1 to 4), which we term the proportion of *unstable* slopes.

PPV and NPV depend strongly on to the frequency of *unstable* and *stable* slopes in the data set (Brenner and Gefeller, 1997). Thus keeping the base rate the same when making comparisons across tests and stability classifications is essential.

To demonstrate the effect variations in the frequency of *unstable* and *stable* slopes have on predictive values like PPV or 1-NPV, we additionally explored this effect for tests observed when either danger level 1-Low, 2-Moderate or 3-Considerable were forecast.

### A.6.3.3 Base rate for proportion of *unstable* and *stable* slopes

As outlined before, the proportion of *unstable* slopes varied within our data set: We noted a bias towards more frequently observing two ECTs when slope stability was considered *unstable* (30%). For single ECT, only 15% of the tests were observed in *unstable* slopes (Tab. A.21). To balance out this mismatch when comparing two ECT results to a single ECT or RB (20% *unstable*), we created equivalent data sets for single ECT and RB containing the same proportion of tests collected on *unstable* and *stable* slopes as found for the data set of two ECTs. For this, we randomly sampled an appropriate number of single ECT and RB observed on *stable* slopes (i.e. we reduced the number of *stable* cases), and combined these with all the tests observed on *unstable* slopes. We repeated this procedure 100 times. We report only the mean values of these 100 repetitions and calculated p-values (*prop.test*) for these mean proportions and the original number of cases in the data set.

The base rate proportion with 30% tests on *unstable* and 70% on *stable* slopes was used throughout this manuscript, except in the section where we evaluate the effect of different base rates.

### A.6.3.4 Selecting ECT from snow pits with two ECT

For snow pits with two adjacent ECTs, we randomly selected one ECT, when exploring single ECT data or the relationship between the number of taps and slope stability. As before, this procedure was repeated 100

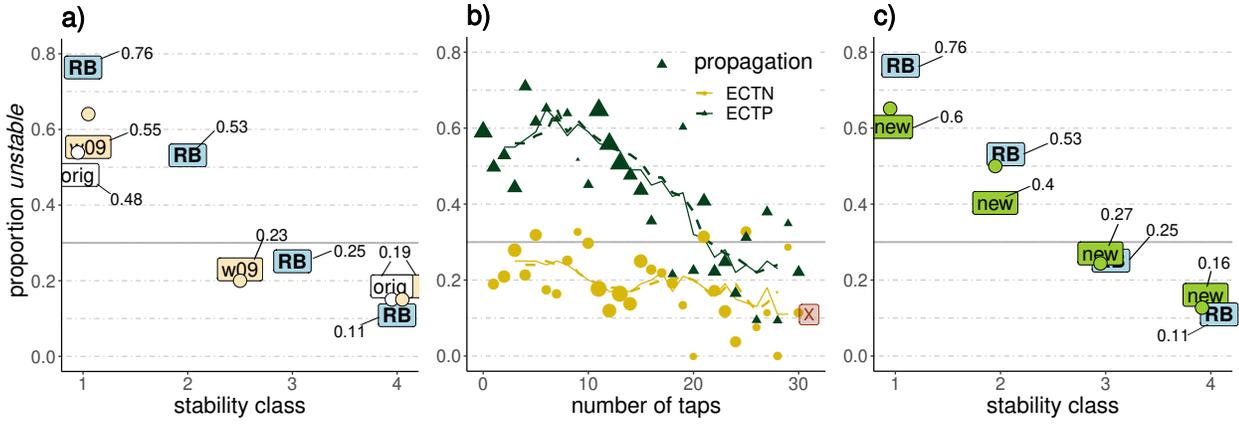


Figure A.31: Proportion of *unstable* slopes (y-axes) for a) the two existing ECT stability classifications ( $ECT_{orig}$ ,  $ECT_{w09}$ ) and the RB, b) the number of taps stratified by propagation, and c) the classification using the  $ECT_{new}$  together with the RB as in a). In a) and c): single ECT results are indicated by the respective text labels, two ECTs resulting in the same stability class by circles. For single ECT and RB, additionally the actual values for the proportion of *unstable* slopes are indicated. In b): The lines represent the mean proportion of *unstable* slopes calculated for moving windows including five or seven consecutive numbers of taps. a) to c) 30% *unstable* and 70% *stable* slopes were used (i.e. the grey line shows the the base rate proportion of *unstable* slopes).

times. The respective statistic, generally the mean proportion of *unstable* slopes, was calculated based on the 100 repetitions.

## A.6.4 Results

### A.6.4.1 Comparing existing stability classifications

We first consider the results for a single ECT. The original stability classification  $ECT_{orig}$  led to significantly different proportions of *unstable* slopes for the two stability classes (0.48 vs. 0.19,  $p < 0.001$ , Fig. A.31a). The  $ECT_{w09}$  classification, with three different classes, showed significantly different proportions of *unstable* slopes between the lowest and the intermediate class (0.55 vs. 0.23,  $p \leq 0.001$ ), but not between the intermediate and the highest class (0.23 and 0.19,  $p > 0.05$ ). Although  $ECT_{w09}$ -class 1 had a larger proportion *unstable* slopes than  $ECT_{orig}$ -class 1, the difference was not significant ( $p > 0.05$ ).

Considering the results obtained from two adjacent ECTs resulting in the same stability class 1, between 0.54 ( $ECT_{orig}$ ) and 0.64 ( $ECT_{w09}$ ) of the slopes were *unstable*. Although the proportion of *unstable* slopes was higher by 0.06 to 0.09 than for a single ECT, this difference was not significant ( $p > 0.05$ ). When both ECTs indicated the highest stability class, the proportion of *unstable* slopes was 0.15, not significantly different than for a single ECT resulting in this stability class (0.19,  $p > 0.05$ ). When one test resulted in the lowest and the other in the intermediate  $ECT_{w09}$ -class, 0.21 of the slopes were *unstable*. While this was clearly less than when both resulted in  $ECT_{w09}$ -class 1 ( $p < 0.05$ ), it was not significantly different than two ECT with  $ECT_{w09}$ -class 4 (0.15,  $p > 0.05$ ).

Regardless whether a single ECT or two ECTs were considered, the  $ECT_{w09}$ -classification had a 0.07-0.08 larger proportion of *unstable* slopes for stability class 1 than the  $ECT_{orig}$ -classification. For stability class 4

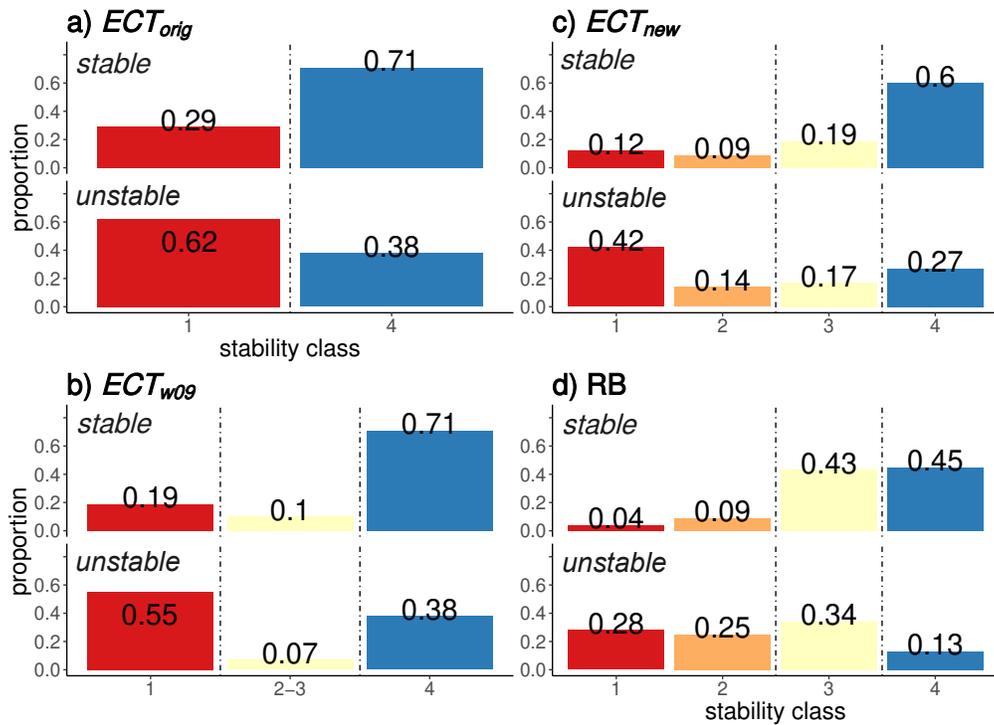


Figure A.32: Distribution of stability classes by slope stability for the different stability test and classification approaches: a) with two classes ( $ECT_{orig}$ ); b) with three classes ( $ECT_{w09}$ ); and c) and d) with four classes ( $ECT_{new}$  and RB, respectively). The vertical dashed lines indicate the thresholds when the primary slope stability associated with a test result changed from one slope stability to the other. Reading subfigures row-wise provides an indication of POD and PON. Comparing proportions column-wise corresponds to a base rate of 0.5. If no clear prevalence was observed, the stability class is considered as intermediate (light yellow colour). Stability classes were considered as having no clear prevalence, when the ratio of the proportion of *unstable* cases to the combined proportions of *unstable* and *stable* was between 0.4 and 0.6. As an example, for RB stability class 3 this ratio would be  $0.34/(0.34+0.43)$ .

there was no difference, as the definition for this class is identical.

The sensitivity was higher for  $ECT_{orig}$  (0.62) than for  $ECT_{w09}$  (class 1: 0.55, Fig. A.32a and b). However, this comes at the cost of a high false alarm rate (1-specificity) for  $ECT_{orig}$  (0.29), considerably higher than for  $ECT_{w09}$  (0.19).

The optimal balance between achieving a high sensitivity and a low false alarm rate was found to be at  $ECTP_{\leq 21}$  (R-library *pROC* (Robin et al., 2011)), exactly the threshold suggested by Winkler and Schweizer (2009).

#### A.6.4.2 Clustering ECT results by accounting for failure initiation and crack propagation

So far, we explored existing classifications. Now, we focus on the respective lowest number of taps stratified by propagating ( $ECTP$ ) and non-propagating ( $ECTN$ ) results. If in the same test for different weak layers  $ECTN$  and  $ECTP$  were observed, only  $ECTP$  with the lowest number of taps was considered.

As can be seen in Fig. A.31b, the proportion of *unstable* slopes was higher for  $ECTP$  compared to  $ECTN$ , regardless of the number of taps and in line with the original stability classification  $ECT_{orig}$ . However, a

notable drop in the proportion of *unstable* slopes between about 10 and 25 taps is obvious ( $ECTP$ , from about 0.6 to almost 0.25).

Clustering the ECT results shown in Figure A.31b with the number of clusters  $k$  set to 3, 4 and 5, and repeating the clustering 100 times, each time with 90% of the data, split the data at similar thresholds. In the following, we show the results for the two most frequent cluster thresholds obtained for  $k = 4$ . The frequency, the respective cluster threshold was selected in the 100 repetitions, is shown in brackets:

- $ECTP \leq 14$  (48%),  $ECTP \leq 13$  (36%)
- $ECTP \leq 20$  (37%),  $ECTP \leq 18$  (36%)
- $ECTN \leq 10$  (29%),  $ECTN \leq 9$  (22%)

Setting  $k$  to 3 resulted in clusters being divided at  $ECTP \leq 14$  and at  $ECTP \leq 21$ ,  $k = 5$  resulted in cluster thresholds  $ECTP \leq 9$ ,  $ECTP \leq 14$ ,  $ECTP \leq 20$  and  $ECTN \leq 10$ . The second most frequent threshold was almost always within  $\pm 1$  tap of those indicated before. Applying the same approach with 80% of the data (rather than with 90%) resulted in very similar class thresholds.

To maximize the difference in the proportion of *unstable* slopes between classes, we varied the thresholds defining clusters by testing  $\pm 3$  taps. The following four stability classes for single ECT ( $ECT_{new}$ ) in combination with the depth of the failure plane criterion were obtained ( $p$ -values indicate whether the proportion of *unstable* slopes differed in relation to the previously described group):

1.  $ECTP \leq 13$  - capturing test results with the largest proportion of *unstable* slopes. The proportion of *unstable* slopes (0.6) was double the base rate (0.3).
2.  $ECTP > 13$  and  $ECTP \leq 22$  (proportion of *unstable* slopes = 0.4,  $p \leq 0.05$ ) - transitioning from a high (0.6, for  $ECTP \leq 13$ ) to a lower proportion of *unstable* slopes (0.27, for  $ECTP > 22$ ). However, the mean proportion of *unstable* slopes was still higher than the base rate.
3.  $ECTP > 22$  or  $ECTN \leq 10$  (0.27,  $p \leq 0.01$ ) - the proportion of *unstable* slopes was lower than the base rate.
4.  $ECTN > 10$  or  $ECTX$  (0.16,  $p \leq 0.05$ ) - capturing test results corresponding to the lowest proportions of *unstable* slopes (about half the base rate).

#### A.6.4.3 Evaluating the new ECT stability classification

##### Stability classification for single ECT

The  $ECT_{new}$  classification showed continually and significantly decreasing proportions of *unstable* slopes with increasing stability class (0.6, 0.4, 0.27, 0.16 for classes 1 to 4, respectively,  $p \leq 0.01$ , Fig. A.31c). The lowest  $ECT_{new}$ -class had a larger proportion of *unstable* slopes (0.6) than the lowest classes for  $ECT_{w09}$  (0.55) or  $ECT_{orig}$  (0.48), though this was only significant compared to  $ECT_{orig}$  ( $p \leq 0.05$ ). In contrast, only marginal differences were noted when comparing the proportion of *unstable* slopes for stability class 4 ( $ECT_{new}$  0.16,  $ECT_{orig}$  0.19). Considering  $ECT_{new}$  class 1 as an indicator of instability, the sensitivity was

Table A.23: Proportion *unstable* slopes when randomly selecting one of two ECTs as the first test ( $ECT_{new}(1^{st})$ ) (prop *unstable* 1<sup>st</sup>) and the number of cases (N) , and the respective proportion *unstable* slopes 2<sup>nd</sup> following the outcome of the second ECT ( $ECT_{new}(2^{nd})$ ).

$ECT_{new}(1^{st})$	prop <i>unstable</i> 1 <sup>st</sup>	N	$ECT_{new}(2^{nd})$	N	prop <i>unstable</i> 2 <sup>nd</sup>
1	0.58	114	1 or 2	98	0.64
			3 or 4	16	0.19
2	0.47	52	1 or 2	38	0.53
			3 or 4	14	0.32
3	0.23	78	1 or 2	17	0.27
			3 or 4	61	0.21
4	0.13	209	1 or 2	14	0.22
			3 or 4	195	0.13

0.42. When considering classes 1 and 2 together, the sensitivity increased to 0.56 (Fig. A.32c).

#### Stability classification for two adjacent ECTs

70% of the time two ECTs indicated the same  $ECT_{new}$  class, in 19% they differed by one class and in 11% by two (or more) classes. Two ECTs resulting in the same  $ECT_{new}$  class resulted in pronounced differences in the proportion of *unstable* slopes for classes 1 to 4 (0.65, 0.5, 0.24 and 0.13, respectively; Fig. A.31c).

Randomly picking one of the two ECTs as the first ECT yielded the proportion of *unstable* slopes as shown in Table A.23. Additionally considering the outcome of a second ECT increased or decreased the proportion of *unstable* slopes for some combinations. For instance, if a first ECT resulted in either  $ECT_{new}$  class 1 or 4, the second test would often indicate a similar result: class  $\leq 2$  in 86% of the cases, when the first ECT was class 1, and class  $\geq 3$  in 93% of the cases, when the first ECT was class 4. However, if the first ECT was either  $ECT_{new}$  class 2 or 3, a large range of proportion of *unstable* slopes resulted depending on the second test result (0.21 - 0.53, Tab. A.23), including some combinations resulting in the proportion of *unstable* slopes being close to the base rate.

#### A.6.4.4 Comparison to Rutschblock test results

The proportion of *unstable* slopes decreased significantly with each increase in RB stability class (0.76, 0.53, 0.25 and 0.11 for classes 1 to 4, respectively;  $p < 0.01$ ; Fig. A.31c). If a binary classification were desired, classes 1 and 2 would be considered as indicators of instability, classes 3 and 4 as relating to *stable* conditions. Employing this threshold, the sensitivity was 0.53 and the specificity 0.88 (Fig. A.32d). Considering RB class 3, also termed «fair» stability (Schweizer, 2007a), as an indicator of stability is, however, not truly supported by the data. This class had a proportion *unstable* slopes of 0.25, not significantly lower than the base rate.

Comparing RB with the ECT showed that the proportion of *unstable* slopes for RB stability class 1 was

Table A.24: Proportion *unstable* for  $ECT_{new}$  and RB class 1, classes 1 and 2 combined, and class 4, stratified by regional forecast danger level ( $D_{RF}$ ).

test	$D_{RF}$	all classes		class 1		classes 1 or 2		class 4	
		N	prop. <i>unstable</i>	N	prop. <i>unstable</i>	N	prop. <i>unstable</i>	N	prop. <i>unstable</i>
ECT	1-Low	134	0.02	10	0.1	15	0.07	102	0.02
	2-Moderate	523	0.1	73	0.33	128	0.23	302	0.05
	3-Considerable	451	0.38	103	0.7	153	0.65	202	0.22
	all	1108	0.21	186	0.52	296	0.44	606	0.1
RB	1-Low	78	0.01	2	0.5	3	0.33	54	0
	2-Moderate	334	0.1	21	0.48	52	0.31	145	0.05
	3-Considerable	315	0.36	42	0.74	98	0.61	81	0.16
	all	727	0.2	66	0.64	153	0.57	280	0.07

significantly higher ( $p < 0.01$ ) and for class 4 by about 0.05 lower ( $p > 0.05$ ) than for the respective ECT classifications (Fig. A.31a, c). This indicates that the RB stability classes at either end of the scale captured slope stability better than the ECT results, regardless which of the ECT classification was applied, and whether a second test was performed. Fig. A.31a and c also highlight that RB class 2 and ECT class 1 ( $ECT_{w09}$ ,  $ECT_{new}$ ) had similar proportions of *unstable* slopes.  $ECT_{new}$  stability class 2 had a lower proportion of *unstable* slopes than RB class 2 ( $p < 0.05$ ), but a higher proportion than RB class 3 ( $p < 0.05$ ). The proportions of *unstable* slopes for the two highest  $ECT_{new}$  classes were not significantly different than for the two highest RB classes ( $p > 0.05$ ).

The false alarm rate of the RB (classes 1 and 2) was lower than for any of the ECT classifications (Fig. A.32). However, in our data set a comparably large proportion of RB tests (0.34) indicated stability class 3 in slopes rated as *unstable*. This ratio is higher than for single  $ECT_{new}$  class 3. However, the frequency that stability class 4 (false *stable*) was observed in *unstable* slopes was lower than for  $ECT_{new}$  class 4 (0.13 vs. 0.23, respectively).

The  $ECT_{new}$  stability class correlated significantly with the RB stability class (Spearman rank-order correlation  $\rho = 0.43$ ,  $p < 0.001$ ), a correlation which was stronger for ECT pairs resulting twice in the same ECT stability class ( $\rho = 0.64$ ,  $p < 0.001$ ).

For both tests, stability class 3 was neither truly related to *unstable* nor *stable* conditions, and may therefore be considered to represent something like «fair» stability.

#### A.6.4.5 The predictive value of stability tests - including base rate information

Now, we explore the predictive value of a stability test result as a function of the base rate proportion of *unstable* slopes. In our data set the base rate proportion of *unstable* slopes increased strongly, and in a non-linear way, with forecast danger level: for the 1108 snow pits with at least one ECT it was 1-Low: 0.02, 2-Moderate: 0.1, 3-Considerable: 0.38 (Tab. A.24).

Considering single  $ECT_{new}$  class 1 and RB class 1 showed that the proportion of *unstable* slopes (PPV) was

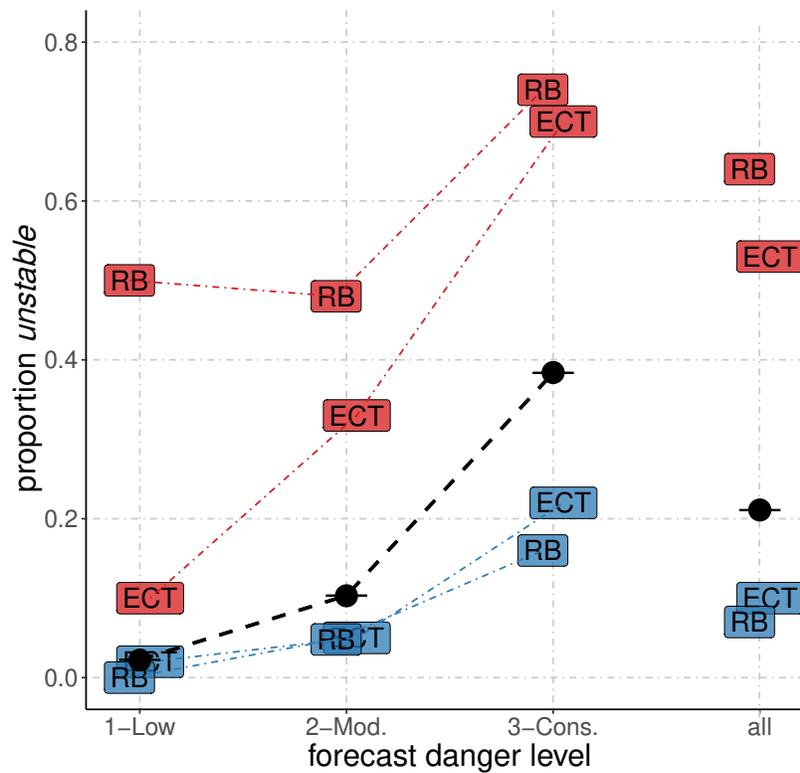


Figure A.33: Proportion of *unstable* slopes (position of labels, RB - Rutschblock, ECT = single  $ECT_{new}$ ) are shown compared to the respective base rate proportion of *unstable* slopes (black dots and black dashed line) for danger levels 1-Low, 2-Moderate (2-Mod) and 3-Considerable (3-Cons), and for the entire data set (all). The proportion *unstable* values are shown for the respective lowest (red colour, labels above base rate line) and highest stability classes (blue, labels below base rate line).

always higher than the base rate proportion (Fig. A.33), indicating that the stability test predicted a higher probability for the slope to be *unstable* than just assuming the base rate. This shift was more pronounced for the Rutschblock than for the ECT, particularly at 1-Low and 2-Moderate. The proportion *unstable* for  $ECT_{new}$  class 1 remained low at 1-Low and 2-Moderate (proportion *unstable*  $\leq 0.33$ , Tab. A.24), indicating that it was still more likely that the slope was *stable* rather than *unstable* given such a test result (Tab. A.24).

Figure A.33 also shows the shift in the proportion *unstable* (1-NPV), when considering  $ECT_{new}$  or RB stability class 4 (high stability). In these slopes, the proportion *unstable* was lower than the base rate, indicating that the probability the specific slope tested to be *unstable* was less than the base rate. The resulting proportion *unstable* was still higher compared to the base rate proportion *unstable* of the neighboring next lower danger level.

Analyzing the entire data set together, regardless of the forecast danger level, the proportion *unstable* slopes was 0.21, and thus somewhat between the values for 2-Moderate and 3-Considerable. Again, the informative value of the test can be noted (Fig. A.33). However, ignoring the specific base rate related to a certain danger level, leads - for instance - to an underestimation of the likelihood that the slope is *unstable* at 3-Considerable (RB or  $ECT_{new}$  class 1), or an overestimation for the presence of instability at 1-Low (RB or  $ECT_{new}$  class 4).

At 1-Low, observations of RB stability class 1 were much less common (3%, or 2 out of 78 tests, Tab. A.24)

compared to  $ECT_{new}$  class 1 (7%). Similar observations were noted for classes 1 or 2: at 1-Low 4% of the RB and 11% of the ECT fell into these categories, increasing to 31% (RB) and 34% (ECT) of the tests at 3-Considerable. This shift from the base rate proportion of *unstable* slopes to the observed proportion was more pronounced for the RB compared to the ECT.

As shown in Figures A.31c, the two extreme RB stability classes correlated better with slope stability than the respective two extreme  $ECT_{new}$  classes. This is also reflected in Fig. A.33 by the stronger shift from the base rate proportion of *unstable* slopes to the observed proportion of *unstable* slopes. It is important to note that a stability test indicating stability class 4 was observed in 10% (ECT) or 7% (RB) of the cases in slopes rated *unstable*. This clearly emphasizes that a single stability test should never be trusted as the single decisive piece of evidence indicating stability.

## A.6.5 Discussion

### A.6.5.1 Performance of ECT classifications

We compared ECT results with concurrent slope stability information, applying existing classifications and testing a new one.

Quite clearly, whether a crack propagates across the entire column or not, is the key discriminator between *unstable* and *stable* slopes (Fig. A.31b). This is in line with previous studies (e.g. Simenhois and Birkeland, 2006; Moner et al., 2008; Simenhois and Birkeland, 2009; Winkler and Schweizer, 2009; Techel et al., 2016c) and with our current understanding of avalanche formation (Schweizer et al., 2008b). Moreover, our results confirm the proposition by Winkler and Schweizer (2009) that the number of taps provides additional information allowing a better distinction between results related to *stable* and *unstable* conditions. The optimal threshold to achieve a balanced performance, i.e. high sensitivity as well as high specificity, was found to be between  $ECTP20$  and  $ECTP22$ , depending on the method (*kmeans*-clustering, *pROC*-cutoff point). This finding agrees well with the threshold proposed by Winkler and Schweizer (2009) who suggested  $ECTP21$ . Using the binary classification, as originally proposed by Simenhois and Birkeland (2009), increased the sensitivity but led to a rather high false alarm rate. Moving away from a binary classification increased PPV and NPV for the lowest and highest stability classes, respectively, but came at the cost (or benefit) of introducing intermediate stability classes.

Only in some situations did pairs of ECTs performed in the same snow pit show an improved correlation with slope stability: when two tests were either  $ECT_{new}$  stability class 1 or 2, or when either both tests were class 4, or one class 3 and one class 4.

### A.6.5.2 Comparing ECT and Rutschblock

To our knowledge, and based on the review by Schweizer and Jamieson (2010), there have only been three previous studies that compared ECT and RB in the same data set.

Moner et al. (2008), in the Spanish Pyrenees, relying on a comparably small data set of 63 RB (base rate 0.44) and 47 single ECT (base rate 0.38) observed a higher unweighted average accuracy for the ECT

(0.93) than the RB (0.88). In contrast, Winkler and Schweizer (2009,  $N = 146$ , base rate 0.25) presented very similar values for RB (0.84) and the ECT (0.81). However, Winkler and Schweizer (2009) partially relied on a slope stability classification which is based strongly on the Rutschblock. Therefore, they emphasized that the RB was favored in their analysis. And finally, the data presented by Techel et al. (2016c) is to a large part incorporated in the study presented here.

In that respect, this study presents the first comparison incorporating a comparably large number of ECT and RB conducted in the same snow pit, where slope stability was defined independently of test results. Seen from the perspective of the proportion of *unstable* slopes, the lowest and highest RB classes correlated better with slope stability than the respective ECT classes. Incorporating the sensitivity, the proportion of *unstable* slopes detected by a test, a mixed picture showed: Single ECT and RB (classes 1 and 2) detected a comparable proportion of *unstable* slopes (0.56 vs. 0.53, respectively, Fig. A.32c, d). Missed *unstable* classifications, however, were comparably rare for the RB (0.13) compared to single ECT (0.21). Similar findings were noted for *stable* cases and stability class 4: RB results indicating instability on *stable* slopes (0.13) were less frequent than ECT indicating instability on *stable* slopes (0.27).

#### A.6.5.3 Predictive value of stability tests

We recall the three lessons drawn by Ebert (2019) in his theoretical investigation of the predictive value of stability tests using Bayesian reasoning in avalanche terrain, as this inspired us to explore these aspects using actual observations and compare them to our results:

(1) «A localised diagnostic test will be more informative the higher the general avalanche warning.» (Ebert, 2019, p. 4). With general «avalanche warning» Ebert (2019) referred to the forecast danger level as a proxy to estimate the base rate. As shown in Fig. A.33, the observed proportion of *unstable* slopes (PPV) increased for both ECT and RB class 1 with increasing danger level, and hence base rate, supporting this statement.

(2) «... Do not 'blame' the stability tests for false positive results: they are to be expected when the avalanche danger is low. In fact, their existence is a consequence of the basic fact that low-probability events are difficult to detect reliably» (Ebert, 2019, p. 4). Fig. A.33 supports this statement: at 1-Low and 2-Moderate an ECT indicating instability (class 1) was much more often observed on a *stable* slope than an *unstable* one. Only once the base rate proportion of *unstable* slopes was sufficiently high, in our case at 3-Considerable, tests indicating instability were observed more often on *unstable* rather than *stable* slopes. When the base rate was low, the predictive value of the RB was higher than of the ECT, suggesting that it may be worthwhile to invest the time required to perform a RB rather than an ECT.

(3) «In avalanche decision-making, there is no certainty, all we can do is to apply tests to reduce the risk of a bad outcome, yet there will always be a residual risk» (Ebert, 2019, p. 5). The proportion of *unstable* slopes (PPV) was greater than the base rate proportion of *unstable* slopes for tests indicating instability, regardless whether we considered an ECT or a RB result and regardless of the danger level, while the proportion of *unstable* slopes (or 1-NPV) was lower for tests indicating stability. From a Bayesian perspective, we can say that a positive test (a low stability class) always increases our belief that the slope is *unstable*, and vice versa when a test is negative (a high stability class). In summary, both instability tests are useful despite the

uncertainty which remains.

#### A.6.5.4 Sources of error and uncertainties

Beside potential misclassifications in slope stability, which we address more specifically in the following section, Schweizer and Jamieson (2010) pointed out two other sources of error. The first of these is linked to the test method, which are relatively crude methods and where, for instance, the loading may vary depending on the observer. The second error source is linked to the spatial variability of the snowpack. The constellation of slab and underlying weak layer properties vary in the terrain and may consequently have an impact on the test result. Furthermore, this data set did not permit to check whether the failure layer of avalanches or whumpfs was linked to the failure layer observed in test results. Such information about the «critical weak layer» was, for instance, incorporated by Simenhois and Birkeland (2009) and Birkeland and Chabot (2006) in their analyses. However, from a stability perspective, considering the actual test result is the more relevant information.

#### A.6.5.5 Influence of the reference class definitions and the base rate

So far we have explored ECT and RB assuming that there are no misclassifications of slope stability. However, as the true slope stability is often not known (particularly in stable cases), errors in slope stability classification will occur. Such errors, however, may potentially influence all the statistics derived to describe the performance of tests (Brenner and Gefeller, 1997). For instance, if there are at least some slopes misclassified, classification performance will drop. However, in such cases, POD and PON will additionally be influenced by the true (though unknown) base rate (Brenner and Gefeller, 1997).

In previous studies exploring ECT (Moner et al., 2008; Simenhois and Birkeland, 2009; Winkler and Schweizer, 2009), slope stability classifications were generally well described and the base rate for the applied slope stability classification given. However, slope stability classification approaches differed somewhat. For instance, a stability criterion used by Moner et al. (2008) was the occurrence of an avalanche on the test slope, while Simenhois and Birkeland (2009) additionally considered explosives testing of the slope as relevant information. Winkler and Schweizer (2009), on the other hand, additionally considered the manual profile classification used operationally in the Swiss avalanche warning service (Schweizer and Wiesinger, 2001; Schweizer, 2007a). They already considered a location as *unstable*, when profiles were rated as «very poor» or «poor». As this classification relies rather strongly on the RB result, the RB would be favored in such an analysis (Winkler and Schweizer, 2009).

We have no knowledge about the uncertainty linked to our classification. However, we can demonstrate the impact of variations in the definition of the reference class on summary statistics like POD and PON, and using different data subsets for analysis: Let us assume we are not interested in comparing ECT and RB, but want to explore only the performance of a binary ECT classification with *ECTP22* as the threshold between two classes. We will, however, use the RB together with the criteria introduced in the Methods-Section to define slope stability:

- Without using the RB as an additional criterion, POD and PON for the ECT was 0.56 and 0.79, re-

spectively (Fig. A.32c).

- If only slopes were considered *unstable*, when the RB stability class was  $\leq 2$ , and those as *stable* with RB stability class 4, the resulting POD was 0.70 and PON was 0.91. The base rate in this data set was 0.32 and  $N = 243$ .
- Being even more restrictive, and considering only slopes *unstable*, when the RB stability class was 1, and those as *stable* with RB stability class 4, the resulting POD was 0.74 and PON was 0.91. The base rate in this data set was 0.2 and  $N = 206$ .

Of course, one could also be interested in exploring the performance of a binary classification of the RB, and define slope stability by using ECT results as additional criterion to those in the Methods-Section. Without relying on ECT results, POD and PON for the RB were 0.53 and 0.88, respectively (Fig. A.32d). Considering only slopes as *unstable*, when additionally  $ECT_{new}$  stability class  $\leq 2$  was observed, and those with  $ECT_{new}$  class 4 as *stable*, POD and PON would increase to 0.66 and 0.94 ( $N = 307$ , base rate 0.29), or 0.71 and 0.94, respectively when considering only  $ECT_{new}$  stability class 1 as *unstable* and class 4 as *stable* ( $N = 285$ , base rate 0.23).

The combination of various error sources, together with varying definitions of slope stability and differences in the base rate make it almost impossible to directly compare results obtained in different studies. Therefore, performance values presented in this study, but also in other studies regarding snow instability tests, must always be seen in light of the specific data set used and allow primarily a comparison within the study.

#### A.6.5.6 Proposing stability class labels

For the purposes of this manuscript, we introduced class numbers to assign a clear order to the classes rather than assigning class labels. However, the introduction of class labels rather than class numbers may ease the communication of results.

We believe suitable terms should follow the established labeling for snow stability, which includes the main classes: poor, fair, and good (e.g. CAA, 2014; Greene et al., 2016; Schweizer and Wiesinger, 2001). Hence, we suggest the following four stability class labels to rate the ECT results (Fig. A.34a):

- *poor*:  $ECTP \leq 13$
- *poor to fair*:  $ECTP > 13$  to  $ECTP \leq 22$
- *fair*:  $ECTP > 22$  or  $ECTN \leq 10$
- *good*:  $ECTN > 10$

Introducing these four labels allows an approximate alignment with the labels used for the RB (Fig. A.34b), and reflects the variations in the proportion of *unstable* slopes observed between classes (Fig. A.31c; proportion of *unstable* slopes for the four RB classes: 0.76, 0.53, 0.25, 0.11, respectively; and the four ECT classes: 0.6, 0.4, 0.27, 0.16, respectively).



*Acknowledgments:* We greatly appreciate the helpful feedback provided by the two referees Bret Shandro and Markus Landro, the questions raised by Eric Knoff and Philip Ebert, which all helped to improve this manuscript.

## A.7 List of publications and conference contributions

### Peer-reviewed publications

- 2020 Techel, F., Müller, K. and Schweizer, J.: On the importance of snowpack stability, the frequency distribution of snowpack stability and avalanche size in assessing the avalanche danger level. *The Cryosphere*, 2020, doi: 10.5194/tc-2020-42
- 2020 Techel, F., Pielmeier, C. and Winkler, K.: Refined dry-snow avalanche danger ratings in regional avalanche forecasts: consistent? And better than random? *Cold Reg. Sci. Technol.*, 2020
- 2020 Techel, F., Winkler, K., Walcher, M., van Herwijnen, A. and Schweizer, J.: On snow stability interpretation of Extended Column Test results. *Nat. Hazards Earth Syst. Sci.*, 2020, 1941-1953, doi: 10.5194/nhess-2020-50
- 2019 Berlin, C., Techel, F., Moor, B., Zwahlen, M. and Hasler, R.: Snow avalanche deaths in Switzerland from 1995 to 2014 - Results of a nation-wide linkage study. *PLOS ONE*, 2019, 14, doi: 10.1371/journal.pone.0225735
- 2019 Morin, S., Horton, S., Techel, F., Bavay, M., Coléou, C., Fierz, C., Gobiet, A., Hagenmuller, P., Lafaysse, M., Ližar, M., Mitterer, C., Monti, F., Müller, K., Olefs, M., Snook, J. S., van Herwijnen, A. and Vionnet, V.: Application of physical snowpack models in support of operational avalanche hazard forecasting: A status report on current implementations and prospects for the future. *Cold Reg. Sci. Technol.*, 2019, 102910, doi: 10.1016/j.coldregions.2019.102910
- 2019 Schweizer, J., Mitterer, C., Techel, F., Stoffel, A. and Reuter, B.: On the relation between avalanche occurrence and avalanche danger level. *The Cryosphere*, 2019, 14, 737-750, doi: 10.5194/tc-2019-218
- 2018 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.*, 2018, 18, 2697-2716, doi: 10.5194/nhess-18-2697-2018
- 2018 Wever, N., Vera Valero, C. and Techel, F.: Coupled snow cover and avalanche dynamics simulations to evaluate wet snow avalanche activity. *J. Geophys. Res. Earth. Surf.*, 2018, 123, 1772-1796, doi 10.1029/2017JF004515
- 2017 Techel, F. and Schweizer, J.: On using local avalanche danger level estimates for regional forecast verification. *Cold Reg. Sci. Technol.*, 2017, 144, 52 - 62, doi: 10.1016/j.coldregions.2017.07.012
- 2017 Pasquier, M.; Hugli, O.; Kottmann, A. and Techel, F.: Avalanche accidents causing fatalities: are they any different in the summer? *High Alt. Med. Biol.*, 2017, 18, 67-72, doi: 10.1089/ham.2016.0065
- 2016 Techel, F., Jarry, F., Kronthaler, G., Mitterer, S., Nairz, P., Pavšek, M., Valt, M. and Darms, G.: Avalanche fatalities in the European Alps: long-term trends and statistics. *Geographica Helvetica*, 2016, 71, 147-159, doi: 10.5194/gh-71-147-2016

**Peer-reviewed publications (continued)**

- 2016           Badoux, A., Andres, N., Techel, F. and Hegg., C.: Natural Hazard fatalities in Switzerland from 1946 to 2015. *Nat. Hazards Earth Syst. Sci.*, 2016, 279-294, doi: 10.5194/nhess-16-2747-2016

**Non peer-reviewed publications and conference proceedings papers**

- 2020           Techel, F., Birkeland, K., Chabot, D., Earl, J., Moner, I. and Simenhois, R. Comparing Extended Column Test results to signs of instability in the surrounding slopes - exploring a large, international data set. *The Avalanche Review*, 2020, 39 (1), 24 - 25
- 2018           Techel, F., Ceaglio, E., Coléou, C., Mitterer, C., Morin, S., Purves, R. S. and Rastelli, F.: Consistency in avalanche forecasts: a look across borders. In: *Proceedings ISSW 2018. International Snow Science Workshop, 7-12 October 2018, Innsbruck, Austria*, 2018, 1496 - 1500
- 2018           Schweizer, J., Mitterer, C., Techel, F., Stoffel, A. and Reuter, B.: Quantifying the obvious: the avalanche danger level. In: *Proceedings ISSW 2018. International Snow Science Workshop, 7-12 October 2018, Innsbruck, Austria*, 2018, 1052 - 1058
- 2018           van Herwijnen, A., Heck, M., Richter, B., Sovilla, B. and Techel, F.: When do avalanches release: investigating time scales in avalanche formation. In: *Proceedings ISSW 2018. International Snow Science Workshop, 7-12 October 2018, Innsbruck, Austria*, 2018, 1030-1034
- 2016           Techel, F., Dürr, L. and Schweizer, J.: Variations in individual danger level estimate within the same forecast region. In: *Proceedings ISSW 2016. International Snow Science Workshop, 3-7 October 2016, Breckenridge, Co.*, 2016, 466-471
- 2016           Techel, F., Walcher, M. and Winkler, K.: Extended Column Test: repeatability and comparison to slope stability and the Rutschblock. In: *Proceedings ISSW 2016. International Snow Science Workshop, 3-7 October 2016, Breckenridge, Co.*, 2016, 1203-1208
- 2016           Winkler, K., Fischer, A. and Techel, F.: Avalanche risk in winter backcountry touring: status and recent trends in Switzerland. In: *Proceedings ISSW 2016. International Snow Science Workshop, 3-7 October 2016, Breckenridge, Co.*, 2016, 270-276
- 2016           Fierz, C., Egger, T., Gerber, M., Techel, F. and Bavay, M.: SnopViz, an interactive visualization tool for both snow-cover model output and observed snow profiles. In: *Proceedings ISSW 2016. International Snow Science Workshop, 3-7 October 2016, Breckenridge, Co.*, 2016, 637-641

**Workshop and conference contributions (without proceedings paper)**

- 2020 Techel, F.: Comparing Extended Column Test results to signs of instability in the surrounding slopes - exploring a large, international data set. *19<sup>th</sup> Colorado Snow and Avalanche Workshop, 14-16 October 2020 (virtual conference), organizers: Forest Service National Avalanche Center, Colorado Avalanche Information Center, Friends of the CAIC.*
- 2019 Techel, F.: Two extraordinary avalanche situations with danger level 5-Very High in the Swiss Alps. *18<sup>th</sup> Colorado Snow and Avalanche Workshop, 4 October 2019, Breckenridge, Co.*
- 2019 Techel, F.: On the verification of regional avalanche forecasts. *18<sup>th</sup> Colorado Snow and Avalanche Workshop, 4 October 2019, Breckenridge, Co.*
- 2019 Techel, F.: On the key elements defining avalanche hazard: «What can data tell us?» *General Assembly of the European Avalanche Warning Services, 12 - 14 June 2019, Oslo, Norway*
- 2018 Techel, F.: Field observations provided by the public - a useful data-source for operational avalanche forecasting? *Innopol workshop: Who is behind your data? A conversation across geographic disciplines, 9 March 2018, University of Zurich*
- 2017 Techel, F.: On the use of the danger levels in avalanche bulletins in the Alps. *General Assembly of the European Avalanche Warning Services, 12 - 15 June 2017, Tutzing, Germany*

# Appendix B

## Supplement to Data section

### B.1 Forecast verification data: Switzerland, Norway, Canada, Colorado

Table B.1 shows the joint distributions of forecast - nowcast pairs used in Section 5.3 (p. 54) for Swiss data. The Swiss data set is described in detail in the publication «On using local avalanche danger level estimates for regional forecast verification» (Appendix A.3, p. 122ff). The respective data used for Norway, Canada and Colorado is shown in Table B.2.

#### Local nowcast estimates - regional forecast pairs - Norway ( $DL_{NOR}$ )

Data was extracted from the operational data base by Ragnar Ekker, at the time avalanche forecaster at the national Norwegian avalanche warning service NVE (Norwegian Water Resources and Energy Directorate). Only  $D_{LN}$  estimates were used, if observers had been in the field, and if they had participated in at least one of the observer training courses.

In 2018, NVE issued avalanche forecasts for 21 A regions (daily forecasts) covering an area of about 165'000 km<sup>2</sup>. The size of the warning regions ranged between 3'000 and 14'000 km<sup>2</sup> (median 7'600 km<sup>2</sup>).

#### Regional nowcast assessment - regional forecast pairs - Canada ( $DL_{CAN}$ )

For the avalanche forecasts in the Banff, Yoho, and Kootenay regions in the Rocky Mountains, published by Parks Canada, Statham et al. (2018b) compared the regional forecast danger level  $D_{RF}$  with regional nowcast assessment. These assessments were made by the forecasters during the forecast process, often after a day in the field (Statham, 2019). The forecast domain has a size of 1950 km<sup>2</sup> and consists of one single warning region (forecast zone).

Table B.1: Joint distributions of forecast danger level  $D_{RF}$  and local nowcast estimates  $D_{LN}$  for Switzerland. wr - warning region

a) <b>Switzerland</b> ( $N = 11,760$ )		<b><math>D_{LN}</math> - individual</b>				
		1-Low	2-Mod	3-Cons	4-High	5-vHigh
$D_{RF}$	1-Low	0.061	0.010	0	0	0
	2-Mod	0.055	0.337	0.027	0	0
	3-Cons	0.004	0.117	0.365	0.002	0
	4-High	0	0	0.013	0.009	0
	5-vHigh	0	0	0	0	0.001
b) <b>Switzerland</b> ( $N = 842$ )		<b><math>D_{LN}</math> - two the same, same warning region</b>				
		1-Low	2-Mod	3-Cons	4-High	5-vHigh
$D_{RF}$	1-Low	0.049	0.001	0	0	0
	2-Mod	0.031	0.369	0.013	0	0
	3-Cons	0	0.076	0.438	0	0
	4-High	0	0	0.018	0.004	0
	5-vHigh	0	0	0	0	0
c) <b>Switzerland</b> ( $N = 1,158$ )		<b><math>D_{LN}</math> - two the same, same/neighbor wr</b>				
		1-Low	2-Mod	3-Cons	4-High	5-vHigh
$D_{RF}$	1-Low	0.049	0.001	0	0	0
	2-Mod	0.033	0.354	0.008	0	0
	3-Cons	0	0.081	0.460	0	0
	4-High	0	0	0.013	0.001	0
	5-vHigh	0	0	0	0	0

### Regional hindcast assessments - regional forecast pairs - Colorado ( $DL_{COL}$ )

In Colorado (United States) five office-based forecasters in the Boulder central office assessed the forecast danger level one or two days after the forecasts were valid. For this assessment, any information considered relevant was integrated and the forecast assessed at the scale of the warning region (in Colorado termed forecast zones) for three elevation bands (below treeline, at treeline, above treeline). According to Logan (2020), the assessor tended to rate  $D$  over the entire warning region, which had a size between 3,900 km<sup>2</sup> and 11,700 km<sup>2</sup> (Logan and Greene, 2018). Spatial variations within these large regions were not expressed in this assessment. These reassessments were made during the 2017-2018 forecast season, for a total of 147 forecast days.

Only summary information in form of a frequency table was available. Not all forecasters reassessed all regions and elevation bands every day.

Table B.2: Joint distributions of forecast danger level  $D_{RF}$  and local nowcast estimates  $D_{LN}$  for Norway, and regional nowcast or hindcast assessments in Canada and Colorado.

a) Norway ( $N = 4,511$ )		$D_{LN}$ - individual				
		1-Low	2-Mod	3-Cons	4-High	5-vHigh
	1-Low	0.116	0.008	0	0	0
	2-Mod	0.105	0.431	0.020	0	0
$D_{RF}$	3-Cons	0.005	0.130	0.170	0.003	0
	4-High	0	0	0.005	0.006	0
	5-vHigh	0	0	0	0	0
b) Norway ( $N = 310$ )		$D_{LN}$ - two the same				
		1-Low	2-Mod	3-Cons	4-High	5-vHigh
	1-Low	0.103	0	0	0	0
	2-Mod	0.035	0.532	0.010	0	0
$D_{RF}$	3-Cons	0	0.132	0.177	0	0
	4-High	0	0.001	0.006	0.003	0
	5-vHigh	0	0	0	0	0
c) Canada ( $N = 2,774$ )		$D_{RN}$				
		1-Low	2-Mod	3-Cons	4-High	5-vHigh
	1-Low	0.278	0.021	0	0	0
	2-Mod	0.032	0.301	0.025	0	0
$D_{RF}$	3-Cons	0.003	0.045	0.222	0.011	0
	4-High	0	0.001	0.019	0.037	0.004
	5-vHigh	0	0	0	0	0
d) Colorado ( $N = 2,018$ )		$D_{RH}$				
		1-Low	2-Mod	3-Cons	4-High	5-vHigh
	1-Low	0.388	0.034	0	0	–
	2-Mod	0.057	0.359	0.027	0	0
$D_{RF}$	3-Cons	0	0.036	0.089	0.004	0
	4-High	0	0	0.005	0.001	0
	5-vHigh	0	0	0	0	0

## B.2 Avalanche recordings - mapped avalanches Davos / Switzerland

Avalanches observed in the region surrounding Davos (Switzerland) have been manually mapped since decades, but in greater detail since about the winter of 2004-2005. This data set of manually mapped avalanches, or subsets of these avalanches, has been used in several publications (e.g. Mitterer et al.,

2009; Wever et al., 2018; Harvey et al., 2018; Schweizer et al., 2020) or Master thesis (Völk, 2020), where the data set is described in greater detail.

The following selection rules were applied to define the data set:

- Avalanches were considered, which were within the area defined as the core mapping area around Davos (180 km<sup>2</sup>), and which released in the months Dec to April. Furthermore, only avalanches mapped in the 15 years 2004-2005 to 2018-19 were considered, as in the years before almost ten times less avalanches were mapped annually.
- The release type of avalanches was sometimes missing. Therefore, the study region was further restricted to the area outside of the ski areas, as defined by Bühler et al. (2018), as often avalanches are released artificially to secure ski runs. This area had a size of 167 km<sup>2</sup>.
- Furthermore, only those avalanches were retained which were not triggered artificially (by humans or explosives).
- The median elevation of the observed avalanche release area (defined below) had to lie above 1800 m above sea level, which is lower than the treeline in the region. An elevation of 1800 m above sea level is nowhere more than 350 m above the valley floor. 96% of the observed avalanche release areas were above this elevation, and in 96% of the avalanche forecast the elevation above which the danger level prevailed was indicated to lie at 1800 m or higher.

With these restrictions, the data set contained 6,729 avalanches, reported on 744 days. The avalanche data set was merged with the forecast danger levels for the region of Davos, resulting in 2,205 days with information about avalanche activity and avalanche hazard. For each of these days, an avalanche activity index (AAI) was calculated. The AAI sums up all avalanches by assigning weights to their size (size 1 to size 4, weights 0.01, 0.1, 1, 10, respectively; Schweizer et al., 1998).

# Bibliography

- AC: Data set: size of warning regions used in public avalanche forecasts in Canada, data provided by J. Floyer, Avalanche Canada, 2019.
- Agresti, A.: An Introduction to Categorical Data Analysis, Wiley, Hoboken, NJ, 2 edn., 2007.
- Ameijeiras-Alonso, J., Crujeiras, R., and Rodríguez-Casal, A.: multimode: An R package for mode assessment, URL <https://arxiv.org/abs/1803.00472>, 2018.
- Baker, J. and McGee, T.: Backcountry snowmobilers' avalanche-related information-seeking and preparedness behaviors, *Society & Natural Resources*, 29, 345 – 356, doi:10.1080/08941920.2015.1103387, 2016.
- 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.
- Ballou, D. and Pazer, H.: Modeling completeness versus consistency tradeoffs in information decision contexts, *IEEE Transactions on Knowledge and Data Engineering*, 15, 240–243, doi:10.1109/TKDE.2003.1161595, 2003.
- Birkeland, K.: Spatial patterns of snow stability through a small mountain range, *Journal of Glaciology*, 47, 176–186, doi:10.3189/172756501781832250, 2001.
- Birkeland, K. and Chabot, D.: Minimizing «false-stable» stability test results: why digging more snowpits is a good idea, in: *Proceedings ISSW 2006. International Snow Science Workshop, 1 - 6 October 2006*, Telluride, Co., 2006.
- Birkeland, K. and Chabot, D.: Changes in stability test usage by Snowpilot users, in: *Proceedings ISSW 2012. International Snow Science Workshop, Anchorage, AK., 2012*.
- Birkeland, K. and Landry, C.: Power-laws and snow avalanches, *Geophysical Research Letters*, 29, doi:10.1029/2001GL014623, 2002.
- Bivand, R.: classInt: Choose univariate class intervals, URL <https://CRAN.R-project.org/package=classInt>, r package version 0.1-24, 2017.
- Bivand, R. and Piras, G.: Comparing implementations of estimation methods for spatial econometrics, *Journal of Statistical Software*, 63, 1–36, URL <http://www.jstatsoft.org/v63/i18/>, 2015.
- Bivand, R., Pebesma, E., and Gómez-Rubio, V.: *Applied spatial data analysis with R*, Springer Science + Business Media New York 2013, 2 edn., doi:10.1007/978-1-4614-7618-4, 2013.
- Blume, J.: Likelihood methods for measuring statistical evidence, *Statistics in Medicine*, 21, 2563—2599, doi:10.1002/sim.1216, 2002.

- Bovee, M., Srivastava, R., and Mak, B.: A conceptual framework and belief-function approach to assessing overall information quality, *International Journal of Intelligent Systems*, 18, 51–74, doi:10.1002/int.10074, 2003.
- Bowler, N. E.: Explicitly accounting for observation error in categorical verification of forecasts, *Monthly Weather Review*, 134, 1600 – 1606, 2006.
- Brabec, B. and Stucki, T.: Verification of avalanche bulletins by questionnaires, in: *Proceedings 25 Years of Snow Avalanche Research at NGI*, pp. 79–83, Norges Geotekniske Institutt NGI, 1998.
- Brenner, H. and Gefeller, O.: Variations of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence, *Statistics in Medicine*, 16, 981–991, doi:10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N, 1997.
- Brun, E., Martin, E., Simon, V., Gendre, C., and Coléou, C.: An energy and mass model of snow cover suitable for operational avalanche forecasting, *Journal of Glaciology*, 35, 1989.
- Brun, E., David, P., Sudul, M., and Brunot, G.: A numerical model to simulate snow-cover stratigraphy for operational avalanche forecasting, *Journal of Glaciology*, 38, 1992.
- Bründl, M., Etter, H., Steiniger, M., Klingler, C., Rhyner, J., and Ammann, W.: IFKIS - a basis for managing avalanche risk in settlements and on roads in Switzerland, *Nat. Hazards Earth Syst. Sci.*, 4, 257–262, doi: 10.5194/nhess-4-257-2004, 2004.
- Bühler, Y., von Rickenbach, D., Stoffel, A., Margreth, S., Stoffel, L., and Christen, M.: Automated snow avalanche release area delineation – validation of existing algorithms and proposition of a new object-based approach for large-scale hazard indication mapping, *Natural Hazards and Earth System Sciences*, 18, 3235–3251, doi: 10.5194/nhess-18-3235-2018, URL <https://www.nat-hazards-earth-syst-sci.net/18/3235/2018/>, 2018.
- 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, URL <https://www.the-cryosphere.net/13/3225/2019/>, 2019.
- Burelli, G., Ceaglio, E., Contri, G., Debernardi, A., Frigo, B., and Pivot, S.: *Rendiconto nivometeorologico. Inverno 2011-2012*, Tech. rep., Regione Autonoma Valle d'Aosta e Fondazione Montagna sicura, 184 p., 2012.
- Burelli, G., Ceaglio, E., Contri, G., Debernardi, A., and Pivot, S.: *Rendiconto nivometeorologico. Inverno 2015-2016*, Tech. rep., Regione Autonoma Valle d'Aosta e Fondazione Montagna sicura, 152 p., 2016.
- CAA: *Observation guidelines and recording standards for weather, snowpack and avalanches*, Canadian Avalanche Association, NRCC Technical Memorandum No. 132, 2014.
- Cagnati, A., Valt, M., Soratroi, G., Gavaldà, J., and Sellés, C.: A field method for avalanche danger level verification, *Annals of Glaciology*, 26, 343–346, 1997.
- 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.
- Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, *Psychological Bulletin*, 70, 213–220, 1968.
- Coléou, C.: *Enneigement dans les massifs français dans l'hiver 2011-2012*, *Neige et Avalanche*, 139, 28 – 29, 2012.

- Dale, M. and Fortin, M.-J.: Spatial analysis: a guide for ecologists, Cambridge University Press, 2 edn., 2014.
- 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.
- Doswell, C.: Weather forecasting by humans - heuristics and decision making, *Weather and Forecasting*, 19, 1115 – 1126, 2004.
- Doswell, H. and Brooks, H.: Probabilistic forecasting - a primer, online (The National Severe Storms Laboratory), URL [https://www.nssl.noaa.gov/users/brooks/public\\_html/prob/Probability.html](https://www.nssl.noaa.gov/users/brooks/public_html/prob/Probability.html), last access: 26/03/2020, 2020.
- Dürr, L. and Darms, G.: SLF-Beobachterhandbuch (Observation guidelines), WSL Institute for Snow and Avalanche Research SLF, Davos, URL [https://www.slf.ch/fileadmin/user\\_upload/WSL/Publikationen/Sonderformate/pdf/SLF-Beobachterhandbuch.pdf](https://www.slf.ch/fileadmin/user_upload/WSL/Publikationen/Sonderformate/pdf/SLF-Beobachterhandbuch.pdf), 2016.
- EAWS: Bavarian matrix, URL [http://www.avalanches.org/eaws/en/main\\_layer.php?layer=basics&id=5](http://www.avalanches.org/eaws/en/main_layer.php?layer=basics&id=5), 2005.
- EAWS: Typical avalanche problems, Tech. rep., URL [https://lawine.tirol.gv.at/data/eaws/typical\\_problems/EAWS\\_avalanche\\_problems\\_EN.pdf](https://lawine.tirol.gv.at/data/eaws/typical_problems/EAWS_avalanche_problems_EN.pdf), approved by General Assembly of EAWS, Munich, 2017; last access: 01/12/2017, 2017a.
- EAWS: EAWS Matrix, Tech. rep., URL <https://www.avalanches.org/standards/eaws-matrix/>, last access: 2020/01/31, 2017b.
- EAWS: Content and structure of public avalanche bulletins, Tech. rep., URL <https://lawine.tirol.gv.at/data/produkte/basics/ContentAndStructureAvalancheBulletin.pdf>, last access: 2017/06/01, 2017c.
- EAWS: Memorandum of understanding for the European Avalanche Warning Services (EAWS), Tech. rep., URL [https://lawine.tirol.gv.at/data/eaws/MoU\\_EAWS.pdf](https://lawine.tirol.gv.at/data/eaws/MoU_EAWS.pdf), last access: 02/09/2018, 2017d.
- EAWS: European Avalanche Danger Scale (2018/19), [https://www.avalanches.org/wp-content/uploads/2019/05/European\\_Avalanche\\_Danger\\_Scale-EAWS.pdf](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, URL <https://www.avalanches.org/standards/avalanche-size/>, last access: 09/09/2019, 2019.
- EAWS: Standards: Avalanche Danger Scale, website, URL <https://www.avalanches.org/education/avalanche-danger-scale/>, last access: 14/02/2020, 2020a.
- EAWS: Standards: Avalanche Problems, website, URL <https://www.avalanches.org/standards/avalanche-problems/>, last access: 14/02/2020, 2020b.
- EAWS: Standards: Information Pyramid, website, URL <https://www.avalanches.org/standards/information-pyramid/>, last access: 14/02/2020, 2020c.
- EAWS: EAWS Matrix, URL [https://www.avalanches.org/wp-content/uploads/2019/05/EAWS\\_Matrix\\_en-EAWS.png](https://www.avalanches.org/wp-content/uploads/2019/05/EAWS_Matrix_en-EAWS.png), last access 31/01/2020, 2020d.
- Ebert, P. A.: Bayesian reasoning in avalanche terrain: a theoretical investigation, *Journal of Adventure Education and Outdoor Learning*, 19, 84–95, doi:10.1080/14729679.2018.1508356, 2019.

- 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.
- Ekker, R.: Information regarding the data set of local danger level estimates provided by Norwegian observers (Norwegian Water Resources and Energy Directorate NVE), personal communication, 2018.
- Elder, K. and Armstrong, B.: A quantitative approach for verifying avalanche hazard ratings, in: Symposium at Davos 1986 on Avalanche Formation, Movement and Effects, vol. 162 of *International Association of Hydrological Sciences Publication*, pp. 593 – 603, 1987.
- Engeset, R.: National Avalanche Warning Service for Norway - established 2013, in: Proceedings ISSW 2013. International Snow Science Workshop, 7 - 11 October 2013, Grenoble - Chamonix Mont-Blanc, France, pp. 301–310, 2013.
- Engeset, R.: The Chairman's Report, website, URL <https://www.avalanches.org/downloads/#publicdocuments>, chairman's report to the 20th General Assembly (GA) of the European Avalanche Warning Services (EAWS), 12-14 June 2019 in Oslo, Norway, 2019.
- Engeset, R. V., Pfuhl, G., Landrø, M., Mannberg, A., and Hetland, A.: Communicating public avalanche warnings – what works?, *Nat Hazards Earth Syst Sci*, 18, 2537–2559, doi:10.5194/nhess-2018-183, 2018.
- ESRI: ArcGIS online basemap: world topo map, 2017.
- Evans, I.: The selection of class intervals, *Transactions of the Institute of British Geographers*, 2, 98–124, 1977.
- Faillietaz, 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.
- Finn, H.: Examining risk literacy in a complex decision-making environment: A study of public avalanche bulletins, Master's thesis, School of Resource and Environmental Management. Simon Fraser University, Burnaby, B.C., m.R.M. research project no. 745, 2020-04, 2020.
- Floyer, J., Klassen, K., Horton, S., and Haegeli, P.: Looking to the 20's: computer-assisted avalanche forecasting in Canada, in: Proceedings ISSW 2016. International Snow Science Workshop, 2–7 October 2016, Breckenridge, Co., pp. 1245–1249, 2016.
- 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.
- Furman, N., Shooter, W., and Schumann, S.: The roles of heuristics, avalanche forecast, and risk propensity in the decision making of backcountry skiers, *Leisure Sciences*, 32, 453–469, doi:10.1080/01490400.2010.510967, 2010.
- Galimberti, G. and Soffritti, G.: Modern analysis of customer surveys: with applications using R, chap. 15: Tree-based methods and decision trees, pp. 283 – 308, *Statistics in practice*, Wiley, 524 p., 2012.

- Giraud, G., Lafeuille, J., and Pahaut, E.: Evaluation de la qualité de la prévision du risque d'avalanche, *Int. Ass. Hydrol. Sci. Publ.*, 162, 583–591, 1987.
- Goetz, D.: Bilan nivo-météorologique de l'hiver 2013-2014, *Neige et Avalanche*, 147, 12 – 14, 2014.
- Gordon, N. and Shaykewich, J.: Guidelines of performance assessment of public weather services, World Meteorological Organization, wMO/TD No. 1023, 2000.
- Greene, E., Wiesinger, T., Birkeland, K., Coléou, C., Jones, A., and Statham, G.: Fatal avalanche accidents and forecasted danger level: patterns in the United States, Canada, Switzerland and France, in: *Proceedings ISSW 2006. International Snow Science Workshop, 1 - 6 October 2006, Telluride, Co.*, pp. 640–649, 2006.
- Greene, E., Birkeland, K., Elder, K., McCammon, I., Staples, M., and Sharaf, D.: *Snow, weather and avalanches: Observational guidelines for avalanche programs in the United States*, American Avalanche Association, Victor, ID., 3 edn., 104 p., 2016.
- Haegeli, P.: *Avaluator V2.0 – avalanche accident prevention card*, Revelstoke, BC: Canadian Avalanche Centre, 2010.
- Haegeli, P. and McClung, D.: Expanding the snow-climate classification with avalanche-relevant information: initial description of avalanche winter regimes for southwestern Canada, *J. Glaciol.*, 53, 266–276, 2007.
- Haladuick, S.: *Relating field observations and snowpack tests to snow avalanche danger*, Master's thesis, Applied Snow and Avalanche Research, University of Calgary, Canada, 178 p., 2014.
- Harvey, S., Meister, R., Leuthold, H., and Allgöwer, B.: Local Verification of the Swiss Avalanche Bulletin, in: *Proceedings ISSW 1998. International Snow Science Workshop 1998, Sun River, Or.*, pp. 339–343, 1998.
- Harvey, S., Rhyner, H., and Schweizer, J.: *Lawinenkunde*, Bruckmann Verlag GmbH, München, 2012.
- Harvey, S., Rhyner, H., Dürri, L., Schweizer, J., Henny, H., and Nigg, P.: *Caution Avalanches!, avalanche prevention in snow sports*. Core team of instructors, Davos, Switzerland, 2016.
- Harvey, S., Schmudlach, G., Bühler, Y., Dürri, L., Stoffel, A., and Christen, M.: Avalanche terrain maps for backcountry skiing in Switzerland, in: *Proceedings ISSW 2018. International Snow Science Workshop Innsbruck, Austria.*, pp. 1625 – 1631, 2018.
- 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.
- Hendrikx, J., Birkeland, K., and Clark, M.: Assessing changes in the spatial variability of the snowpack fracture propagation propensity over time, *Cold Reg. Sci. Technol.*, 56, 152–160, 2009.
- Hijmans, R.: raster: Geographic data analysis and modeling, URL <https://CRAN.R-project.org/package=raster>, r package version 2.5-8, 2016.
- Hilton, R.: The determinants of information value: synthesizing some general results, *Management Science*, doi:10.1287/mnsc.27.1.57, 1981.

- Horton, S., Bellaire, S., and Jamieson, B.: Modelling the formation of surface hoar layers and tracking post-burial changes for avalanche forecasting, *Cold Regions Science and Technology*, 97, 81 – 89, doi:10.1016/j.coldregions.2013.06.012, 2014.
- Horton, S., Nowak, S., and Haegeli, P.: Enhancing the operational value of snowpack models with visualization design principles, *Natural Hazards and Earth System Sciences Discussions*, 2019, 1–20, doi:10.5194/nhess-2019-344, URL <https://www.nat-hazards-earth-syst-sci-discuss.net/nhess-2019-344/>, 2019.
- Horton, S., Nowak, S., and Haegeli, P.: Enhancing the operational value of snowpack models with visualization design principles, *Natural Hazards and Earth System Sciences*, 20, 1557–1572, doi:10.5194/nhess-20-1557-2020, URL <https://www.nat-hazards-earth-syst-sci.net/20/1557/2020/>, 2020.
- Jacob, T., Tennenbaum, D., and Krahn, G.: Family interaction and psychopathology, chap. 8: Factors influencing the reliability and validity of observation data, pp. 297 – 328, Springer Science+Business Media New York, 1987.
- Jamieson, B. and Johnston, C.: Interpreting rutschblocks in avalanche start zones, *Avalanche News*, 46, 2–4, 1995.
- Jamieson, B., Campbell, C., and Jones, A.: Verification of Canadian avalanche bulletins including spatial and temporal scale effects, *Cold Regions Science and Technology*, 51, 204–213, doi:10.1016/j.coldregions.2007.03.012, 2008.
- 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.
- Jarvis, A., Reuter, H., Nelson, A., and Guevara, E.: Hole-filled seamless SRTM data V4, Tech. rep., International Centre for Tropical Agriculture (CIAT), URL <http://srtm.csi.cgiar.org>, 2008.
- 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*, 7 - 11 October 2013, Grenoble - Chamonix Mont-Blanc, France, pp. 316 – 319, 2013.
- Kronholm, K., Schneebeli, M., and Schweizer, J.: Spatial variability of micropenetration resistance in snow layers on a small slope, *Annals of Glaciology*, 38, 202–208, doi:10.3189/172756404781815257, 2004.
- LaChapelle, E.: The fundamental process in conventional avalanche forecasting, *Journal of Glaciology*, 26, 75–84, 1980.
- Landrø, M., Hetland, A., Engeset, R. V., and Pfuhl, G.: Avalanche decision-making frameworks: Factors and methods used by experts, *Cold Regions Science and Technology*, 170, 102897, doi:<https://doi.org/10.1016/j.coldregions.2019.102897>, 2020a.
- Landrø, M., Pfuhl, G., Engeset, R. V., Jackson, M., and Hetland, A.: Avalanche decision-making frameworks: Classification and description of underlying factors, *Cold Regions Science and Technology*, 169, 102903, doi:10.1016/j.coldregions.2019.102903, 2020b.
- 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*, 2–7 October 2016, Breckenridge, Co., pp. 457 – 465, 2016.
- Lehning, M. and Fierz, C.: Assessment of snow transport in avalanche terrain, *Cold Reg. Sci. Technol.*, 51, 240–252, 2008.

- Lehning, M., Bartelt, P., Brown, B., Russi, T., Stöckli, U., and Zimmerli, M.: Snowpack model calculations for avalanche warning based upon a new network of weather and snow stations, *Cold Reg. Sci. Technol.*, 30, 145 – 157, doi: 10.1016/S0165-232X(99)00022-1, 1999.
- Logan, S.: Information regarding forecast verification at the Colorado Avalanche Information Center (CAIC), personal communication, 2020.
- 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, 7 - 12 Oct 2018, Innsbruck, Austria*, pp. 1059–1062, 2018.
- LWD Steiermark: *Ergebnisse der Online-Umfrage des LWD Steiermark 2015*, 2015.
- Malamud, B. and Turcotte, D.: Self-organized criticality applied to natural hazards, *Natural Hazards*, 20, 93–116, 1999.
- Mansiot, O.: Indices de risque 2 et 3: utilise-t-on correctement les informations du bulletin d'estimation du risque d'avalanche?, *Neige et Avalanches*, 155, 8–10, 2016.
- Marazzi, S.: *Atlante orografico delle Alpi: SOIUSA - suddivisione orografica internazionale unificata del Sistema Alpino*, Pavone Canavese: Priuli & Verlucca, 2005.
- Mayer, S., van Herwijnen, A., Olivieri, G., and Schweizer, J.: Evaluating the performance of an operational infrasound avalanche detection system at three locations in the Swiss Alps during two winter seasons, *Cold Regions Science and Technology*, 173, 102 962, doi:10.1016/j.coldregions.2019.102962, 2020.
- McCammon, I.: The role of training in recreational avalanche accidents in the United States, in: *Proceedings ISSW 2000. International Snow Science Workshop, Big Sky, MT.*, 2000.
- McCammon, I. and Haegeli, P.: Comparing avalanche decision frameworks using accident data from the United States, in: *Proceedings ISSW 2004. International Snow Science Workshop, Jackson, WY.*, 2004.
- McCammon, I. and Hägeli, P.: An evaluation of rule-based decision tools for travel in avalanche terrain, *Cold Regions Science and Technology*, 47, 193–206, doi:10.1016/j.coldregions.2006.08.007, 2007.
- McClung, D.: Predictions in avalanche forecasting, *Annals of Glaciology*, 31, 377–381–, 2000.
- McClung, D.: The elements of applied avalanche forecasting, part I: The human issues, *Natural Hazards*, 26, 111–129, doi:10.1023/A:1015665432221, 2002a.
- McClung, D.: The elements of applied avalanche forecasting, part II: The physical issues and the rules of applied avalanche forecasting, *Natural Hazards*, 26, 131–146, doi:10.1023/A:1015604600361, 2002b.
- 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.
- McClung, D. M.: The strength and weight of evidence in backcountry avalanche forecasting, *Natural Hazards*, 59, 1635–1645, doi:10.1007/s11069-011-9856-y, 2011.
- Meister, R.: Country-wide avalanche warning in Switzerland, in: *Proceedings ISSW 1994. International Snow Science Workshop, 30 Oct - 3 Nov 1994, Snowbird, UT*, pp. 58–71, 1995.

- Météo France: Guide avalanche, Météo France, Saint-Mandé Cedex, édition 2012-2013 edn., 2012.
- Mitterer, C. and Schweizer, J.: Analysis of the snow-atmosphere energy balance during wet-snow instabilities and implications for avalanche prediction, *The Cryosphere*, 7, 205–216, doi:10.5194/tc-7-205-2013, URL <http://www.the-cryosphere.net/7/205/2013/>, 2013.
- Mitterer, C., Mott, R., and Schweizer, J.: Observations and analysis of two large wet snow avalanche cycles, *Proceedings ISSW 2009. International Snow Science Workshop, Davos, Switzerland*, pp. 262–266, 2009.
- Mock, C. and Birkeland, K.: Snow avalanche climatology of the Western United States mountain ranges, *Bulletin of the American Meteorological Society*, 81, 2367 – 2392, doi:10.1175/1520-0477(2000)081<2367:SACOTW>2.3.CO;2, 2000.
- 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, 21–27 September 2008, Whistler, Canada*, pp. 386—392, 2008.
- Moner, I., Orgué, S., Gavalda, J., and Bacardit, M.: How big is big: results of the avalanche size classification survey, in: *Proceedings ISSW 2013. International Snow Science Workshop, 7 - 11 October 2013, Grenoble - Chamonix Mont-Blanc, France*, 2013.
- Monti, F., Schweizer, J., and Fierz, C.: Hardness estimation and weak layer detection in simulated snow stratigraphy, *Cold Regions Science and Technology*, 103, 82 – 90, doi:10.1016/j.coldregions.2014.03.009, 2014.
- Morin, S., Horton, S., Techel, F., Bavay, M., Coléou, C., Fierz, C., Gobiet, A., Hagenmuller, P., Lafaysse, M., Ližar, M., Mitterer, C., Monti, F., Müller, K., Olefs, M., Snook, J. S., van Herwijnen, A., and Vionnet, V.: Application of physical snowpack models in support of operational avalanche hazard forecasting: A status report on current implementations and prospects for the future, *Cold Regions Science and Technology*, p. 102910, doi: <https://doi.org/10.1016/j.coldregions.2019.102910>, 2019.
- Moser, S. C.: Communicating climate change: history, challenges, process and future directions, *WIREs Climate Change*, 1, 31–53, doi:10.1002/wcc.11, 2010.
- 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, 2–7 October 2016, Breckenridge, Co., USA*, pp. 472–479, 2016.
- Munter, W.: *3x3 Lawinen*, Agentur Pohl und Schellhammer, Garmisch-Partenkirchen, 1st edn., 1997.
- Murphy, A. H.: What is a good forecast? An essay on the nature of goodness in weather forecasting, *Weather and Forecasting*, 8, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.
- NAC: Size of forecast zones used in the public avalanche forecasts in the United States, data provided by K. Birkeland and S. Trautman, National Avalanche Center (NAC), 2020.
- NVE: Data set: size of warning regions used in public avalanche forecasts in Norway, data provided by R. Ekker, Norges vassdrags- og energidirektorat (NVE), Oslo, Norway, 2018.
- ÖLWD: Saisonbericht der österreichischen Lawinenwarndienste 2011/2012, Arbeitsgemeinschaft der österreichischen Lawinenwarndienste, 180 p., 2012.

- ÖLWD: Saisonbericht der österreichischen Lawinenwarndienste 2013/2014, Arbeitsgemeinschaft der österreichischen Lawinenwarndienste, 236 p., 2014.
- ÖLWD: Saisonbericht der österreichischen Lawinenwarndienste 2014/2015, Arbeitsgemeinschaft der österreichischen Lawinenwarndienste, 272 p., 2015.
- Page, S. E.: *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (New Edition), Princeton University Press, URL <http://www.jstor.org/stable/j.ctt7sp9c>, 2007.
- Pahaut, E. and Bolognesi, R.: Prévisions régionale et locale du risque d'avalanches, in: *Guide Neige et avalanche. Connaissances, Pratiques, & sécurité*, edited by Ancey, C., ISSN 2-85744-797-3, chap. 7, pp. 161–178, Édisud, Aix-en-Provence, France, 3 edn., electronic version, based on second edition, 2003.
- Procter, E., Strapazzon, G., Dal Cappello, T., Castlunger, L., Staffler, H., and Brugger, H.: Adherence of backcountry winter recreationists to avalanche prevention and safety practices in northern Italy, *Scand. J. Med. Sci. Sports*, 24, 823 – 829, doi:10.1111/sms.12094, 2014.
- Purves, R., Morrison, K., Moss, G., and Wright, D.: Nearest neighbours for avalanche forecasting in Scotland: development, verification and optimisation of a model, *Cold Regions Science and Technology*, 37, 343–355, 2003.
- R Core Team: *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>, last updated: June 2017, 2017.
- 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., Schweizer, J., and van Herwijnen, A.: A process-based approach to estimate point snow instability, *The Cryosphere*, 9, 837–847, doi:10.5194/tc-9-837-2015, 2015.
- 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:10.1002/2015JF003700, 2016.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M.: pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC Bioinformatics*, 12, 77, 2011.
- Ross, C. and Jamieson, B.: Comparing fracture propagation tests and relating test results to snowpack characteristics, in: *Proceedings ISSW 2008. International Snow Science Workshop*, 21–27 September 2008, Whistler, Canada, pp. 376—385, 2008.
- Ruesch, M., Egloff, A., Gerber, M., Weiss, G., and Winkler, K.: The software behind the interactive display of the Swiss avalanche bulletin, in: *Proceedings ISSW 2013. International Snow Science Workshop*, 7 - 11 October 2013, Grenoble – Chamonix-Mont Blanc, France, pp. 406–412, 2013.
- Schmudlach, G.: skitourenguru, *Bergundsteigen*, 96, 68 – 78, 2016.
- Schmudlach, G. and Köhler, J.: Automated avalanche risk rating of backcountry ski routes, in: *Proceedings ISSW 2016. International Snow Science Workshop*, 2–7 October 2016, Breckenridge, Co., 2016, pp. 450–456, 2016.
- Schmudlach, G., Winkler, K., and Köhler, J.: Quantitative risk reduction method (QRM), a data-driven avalanche risk estimator, in: *Proceedings ISSW 2018. International Snow Science Workshop*, 7 - 12 Oct 2018, Innsbruck, Austria, pp. 1272–1278, 2018.

- Schwarb, M., Frei, C., Schär, C., and Daly, C.: Hydrologischer Atlas der Schweiz, chap. 2.7 - Mean Seasonal Precipitation throughout the European Alps 1971–1990, Geographisches Institut der Universität Bern – Hydrologie, 2001.
- 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.: Predicting the avalanche danger level from field observations, in: *Proceedings ISSW 2010. International Snow Science Workshop*, 17 - 22 Oct, Lake Tahoe, CA, pp. 162–165, 2010.
- Schweizer, J. and Bellaire, S.: On stability sampling strategy at the slope scale, *Cold Regions Science and Technology*, 64, 104–109, doi:10.1016/j.coldregions.2010.02.013, 2010.
- Schweizer, J. and Camponovo, C.: The skier's zone of influence in triggering slab avalanches, *Annals of Glaciology*, 32, 314–320, doi:10.3189/172756401781819300, 2001.
- Schweizer, J. and Föhn, P.: Avalanche forecasting - an expert system approach, *Journal of Glaciology*, 42, 318–332, 1996.
- Schweizer, J. and Jamieson, B.: A threshold sum approach to stability evaluation of manual profiles, *Cold Regions Science and Technology*, 47, 50–59, doi:10.1016/j.coldregions.2006.08.011, 2007.
- 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 Kronholm, K.: Snow cover spatial variability at multiple scales: Characteristics of a layer of buried surface hoar, *Cold Regions Science and Technology*, 47, 207–223, doi:10.1016/j.coldregions.2006.09.002, 2007.
- Schweizer, J. and Lütschg, M.: Characteristics of human-triggered avalanches, *Cold Reg. Sci. Technol.*, 33, 147–162, doi:10.1016/S0165-232X(01)00037-4, 2001.
- 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: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.: Quantifying the obvious: the avalanche danger level, in: *Proceedings ISSW 2018. International Snow Science Workshop, 7 - 12 Oct 2018, Innsbruck, Austria.*, pp. 1052 – 1058, 2018.
- 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.
- Shandro, B. and Haegeli, P.: Characterizing the nature and variability of avalanche hazard in western Canada, *Natural Hazards and Earth System Sciences*, 18, 1141–1158, doi:10.5194/nhess-18-1141-2018, 2018.
- Sharp, E.: Avalanche forecast verification through a comparison of local nowcasts with regional forecasts, in: *Proceedings ISSW 2014. International Snow Science Workshop, 29 September - 3 October 2014, Banff, Canada*, pp. 475–480, 2014.
- 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, 1 - 6 October 2006, 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.
- SLF: *Beobachterhandbuch (observation guidelines)*, Eidg. Institut für Schnee- und Lawinenforschung SLF, Davos, Switzerland, 1987.
- SLF: *Interpretationshilfe zum Lawinenbulletin des Eidgenössischen Institutes für Schnee- und Lawinenforschung Weissfluhjoch, Davos*, in: *Mitteilung des Eidgenössischen Institutes für Schnee- und Lawinenforschung*, vol. 49, p. 24, Eidgenössischen Institutes für Schnee- und Lawinenforschung, 1993.
- SLF: *Handbuch für Flachfeldbeobachter (observation guidelines for study plot observers)*, WSL Institute for Snow and Avalanche Research SLF, 70 p., 2002.
- SLF: *Avalanche bulletins and other products. Interpretation guide. Edition 2015*, WSL Institute for Snow and Avalanche Research SLF, Davos, URL [http://www.slf.ch/lawineninfo/zusatzinfos/interpretationshilfe/interpretationshilfe\\_e.pdf](http://www.slf.ch/lawineninfo/zusatzinfos/interpretationshilfe/interpretationshilfe_e.pdf), 16th revised edition, 50p., 2015.
- SLF: *Avalanche bulletin interpretation guide*, WSL Institute for Snow and Avalanche Research SLF, URL [http://www.slf.ch/lawineninfo/zusatzinfos/interpretationshilfe/interpretationshilfe\\_e.pdf](http://www.slf.ch/lawineninfo/zusatzinfos/interpretationshilfe/interpretationshilfe_e.pdf), edition December 2017, 53p., 2017.
- SLF: *Destructive avalanche database, website*, URL <https://www.slf.ch/en/services-and-products/data-and-monitoring/extracts-from-the-destructive-avalanche-database.html>, last access 2018/04/08, 2018.
- SLF: *Avalanche bulletin interpretation guide*, WSL Institute for Snow and Avalanche Research SLF, URL [https://www.slf.ch/files/user\\_upload/SLF/Lawinenbulletin\\_Schneesituation/Wissen\\_zum\\_Lawinenbulletin/Interpretationshilfe/Interpretationshilfe\\_EN.pdf](https://www.slf.ch/files/user_upload/SLF/Lawinenbulletin_Schneesituation/Wissen_zum_Lawinenbulletin/Interpretationshilfe/Interpretationshilfe_EN.pdf), edition December 2019, 52p., 2019.

- 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.
- SRTM: SRTM data V4, Website, URL <http://www.cgiar-csi.org>, download 19/04/2017, 2017.
- St. Clair, A.: Exploring the effectiveness of avalanche risk communication: a qualitative study of avalanche bulletin use among backcountry recreationists, Master's thesis, School of Resource and Environmental Management. Simon Fraser University, Burnaby, B.C., m.R.M. research project no. 738, 2019-10, 2019.
- Statham, G.: Information regarding the nowcast assessments at Parks Canada, personal communication, 2019.
- 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, 17 - 22 Oct, 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, 7 - 12 Oct 2018, Innsbruck, Austria., pp. 1491 – 1496, 2018b.
- Stewart, T. R.: Principles of forecasting: a handbook for researchers and practitioners, chap. Improving reliability in judgemental forecasting, pp. 81–106, Springer Science + Business Media, LLC, 2001.
- Storm, I. and Helgeson, G.: Hot-spots and hot-times: exploring alternatives to public avalanche forecasts in Canada's data sparse Northern Rockies region, in: Proceedings ISSW 2014. International Snow Science Workshop, 29 September - 3 October 2014, Banff, Canada, pp. 91–97, 2014.
- Suter, C., Harvey, S., and Dürr, L.: mAvalanche - smart avalanche forecasting with smartphones, in: Proceedings ISSW 2010. International Snow Science Workshop, 17 - 22 Oct 2010, Squaw Valley, Ca., USA, pp. 630–635, 2010.
- Techel, F.: Field observations provided by the public - a useful data-source for operational avalanche forecasting?, innopool workshop: Who is behind your data? A conversation across geographic disciplines, 9 March 2018, University of Zurich, 2018.
- 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., Pielmeier, C., Darms, G., Teich, M., and Margreth, S.: Schnee und Lawinen in den Schweizer Alpen. Hydrologisches Jahr 2011/12, WSL-Institut für Schnee- und Lawinenforschung SLF Davos: 118 pages (WSL Ber. 5), 2013.
- Techel, F., Stucki, T., Margreth, S., Marty, C., and Winkler, K.: Schnee und Lawinen in den Schweizer Alpen. Hydrologisches Jahr 2013/14, WSL-Institut für Schnee- und Lawinenforschung SLF Davos: 87 pages (WSL Ber. 31), 2015a.

- Techel, F., Zweifel, B., and Winkler, K.: Analysis of avalanche risk factors in backcountry terrain based on usage frequency and accident data in Switzerland, *Nat. Hazards Earth Syst. Sci.*, 15, 1985–1997, doi:10.5194/nhess-15-1985-2015, 2015b.
- Techel, F., Dür, L., and Schweizer, J.: Variations in individual danger level estimate within the same forecast region, in: *Proceedings ISSW 2016. International Snow Science Workshop, 2–7 October 2016, Breckenridge, Co.*, pp. 466–471, 2016a.
- Techel, F., Jarry, F., Kronthaler, G., Mitterer, S., Nairz, P., Pavšek, M., Valt, M., and Darms, G.: Avalanche fatalities in the European Alps: long-term trends and statistics, *Geographica Helvetica*, 71, 147–159, doi:10.5194/gh-71-147-2016, 2016b.
- Techel, F., Walcher, M., and Winkler, K.: Extended Column Test: repeatability and comparison to slope stability and the Rutschblock, in: *Proceedings ISSW 2016. International Snow Science Workshop, 2–7 October 2016, Breckenridge, Co.*, pp. 1203–1208, 2016c.
- 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, 2018.
- Techel, F., Birkeland, K., Chabot, D., Earl, J., Moner, I., and Simenhois, R.: Comparing Extended Column Test results to signs of instability in the surrounding slopes - exploring a large, international data set, *The Avalanche Review*, 39, 24–25, 2020a.
- 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*, 20, 1941–1953, doi:10.5194/nhess-2020-50, 2020b.
- Tremper, B. and Diegel, P.: The wisdom of crowds in avalanche forecasting, in: *Proceedings ISSW 2014. International Snow Science Workshop, 29 September - 3 October 2014, Banff, Canada*, pp. 78–84, 2014.
- Trevethan, R.: Sensitivity, specificity, and predictive values: foundations, pliabilitys, pitfalls in research and practice, *Frontiers in Public Health*, doi:10.3389/fpubh.2017.00307, 2017.
- Valt, M. and Cianfarra, P.: La stagione invernale 2013/2014. Innevamento e attività valanghiva sulle Alpi, *Neve e Valanghe*, 81, 10–19, 2014.
- van Herwijnen, A. and Jamieson, B.: Snowpack properties associated with fracture initiation and propagation resulting in skier-triggered dry snow slab avalanches, *Cold Regions Science and Technology*, 50, 13–22, doi:https://doi.org/10.1016/j.coldregions.2007.02.004, 2007.
- van Herwijnen, A. and Schweizer, J.: Monitoring avalanche activity using a seismic sensor, *Cold Regions Science and Technology*, 69, 165–176, doi:10.1016/j.coldregions.2011.06.008, 2011.
- Vernay, M., Lafaysse, M., Mérindol, L., Giraud, G., and Morin, S.: Ensemble forecasting of snowpack conditions and avalanche hazard, *Cold Regions Science and Technology*, 120, 251 – 262, doi:http://dx.doi.org/10.1016/j.coldregions.2015.04.010, URL <http://www.sciencedirect.com/science/article/pii/S0165232X15000981>, 2015.
- Vionnet, V., Guyomarc'h, G., Lafaysse, M., Naaim-Bouvet, F., Giraud, G., and Deliot, Y.: Operational implementation and evaluation of a blowing snow scheme for avalanche hazard forecasting, *Cold Regions Science and Technology*, 147, 1 – 10, doi:https://doi.org/10.1016/j.coldregions.2017.12.006, 2018.

- Völk, M. S.: Analyse der Beziehung zwischen Lawinenauslösung und prognostizierter Lawinengefahr - Quantitative Darstellung einer regionalen Lawinenaktivität am Beispiel Davos (Schweiz), Master's thesis, Leopold-Franzens-Universität Innsbruck, Austria, Fakultät für Geo- und Atmosphärenwissenschaften, Institut für Geographie, 110 p. Supervisor: R. Sailer, F. Techel, 2020.
- Vul, E., Harris, C., Winkielman, P., and Pashler, H.: Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition, *Perspectives on Psychological Science*, 4, 274–290, 2009.
- Wand, M.: Data-based choice of histogram bin width, *The American Statistician*, 51, 59–64, doi:10.1080/00031305.1997.10473591, 1997.
- Wever, N., Vera Valero, C., and Techel, F.: Coupled snow cover and avalanche dynamics simulations to evaluate wet snow avalanche activity, *Journal of Geophysical Research: Earth Surface*, 123, 1772–1796, doi:10.1029/2017JF004515, 2018.
- Wilks, D.: Statistical methods in the atmospheric sciences, vol. 100 of *International Geophysics Series*, Academic Press, San Diego CA, U.S.A, 3rd edn., 2011.
- Williams, K.: Credibility of avalanche warnings, *Journal of Glaciology*, 26, 93–96, doi:10.1017/S0022143000010625, 1980.
- Winkler, K. and Kuhn, T.: Fully automatic multi-language translation with a catalogue of phrases - successful employment for the Swiss avalanche bulletin, *Language Resources and Evaluation*, 51, 13–35, doi:10.1007/s10579-015-9316-5, 2017.
- Winkler, K. and Schweizer, J.: Comparison of snow stability tests: Extended Column Test, Rutschblock test and Compression Test, *Cold Regions Science and Technology*, 59, 217–226, doi:10.1016/j.coldregions.2009.05.003, 2009.
- Winkler, K. and Techel, F.: Users rating of the Swiss avalanche forecast, in: *Proceedings ISSW 2014. International Snow Science Workshop*, 29 September - 3 October 2014, Banff, Canada, pp. 437–444, 2014.
- Winkler, K., Bächtold, M., Gallorini, S., Niederer, U., Stucki, T., Pielmeier, C., Darms, G., Dürr, L., Techel, F., and Zweifel, B.: Swiss avalanche bulletin: automated translation with a catalogue of phrases, in: *Proceedings ISSW 2013. International Snow Science Workshop*, 7 - 11 October 2013, Grenoble – Chamonix Mont-Blanc, France, pp. 437–441, 2013.
- Winkler, K., Fischer, A., and Techel, F.: Avalanche risk in winter backcountry touring: status and recent trends in Switzerland, in: *Proceedings ISSW 2016. International Snow Science Workshop*, 2–7 October 2016, Breckenridge, Co., pp. 270–276, 2016.
- Zenke, B.: Grenzen des Lawinenlageberichts, *Bergundsteigen*, 4, 30 – 34, 2013.
- Zenkl, G.: 10 Jahre Lawinenwarndienst Niederösterreich, *Saisonbericht der österreichischen Lawinenwarndienste 2015/2016*, pp. 192–193, 2016.
- Zweifel, B., Rätz, A., and Stucki, T.: Avalanche risk for recreationists in backcountry and in off-piste area: surveying methods and pilot study at Davos, Switzerland, in: *Proceedings ISSW 2006. International Snow Science Workshop*, 1 - 6 October 2006, Telluride, Co., pp. 733–741, 2006.

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.