Cold Regions Science and Technology 144 (2017) 52–62

Contents lists available at ScienceDirect

Cold Regions Science and Technology

journal homepage: www.elsevier.com/locate/coldregions

On using local avalanche danger level estimates for regional forecast verification

Frank Techel*, Jürg Schweizer

WSL Institute for Snow and Avalanche Research SLF, Flüelastrasse 11, CH-7260 Davos Dorf, Switzerland

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 hit rate 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.

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^2 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^2, or larger (Zenke, 2013) with a temporal resolution of 6 to 24 h, 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 backcountry 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., 2015). 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^2 (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

[*] Corresponding author.
E-mail address: techel@slf.ch (F. Techel).

http://dx.doi.org/10.1016/j.coldregions.2017.07.012
Received 6 June 2017; Received in revised form 31 July 2017; Accepted 31 July 2017
Available online 02 August 2017
0165-232X/ © 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

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., 2016). 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.

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.

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^2 (Jamieson et al., 2008) to 25 km^2 (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 9553 individual avalanche danger estimates. These estimates were reported either via a website (IFKIS; Bründl et al., 2004; $N = 1774$, 19%) or a mobile app (mAvalanche; Suter et al., 2010; $N = 6531$, 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 ($N = 1248$, 13%). Observations were not distributed evenly across the Swiss Alps, with the most prominent cluster in the region of Davos (Fig. 1) 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 ($N = 1971$, 55 different observers) or by observers based in ski areas ($N = 1423$, 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.

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.

The Alpine warning region comprises an area of 26,400 km^2, 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^2 (Fig. 1). 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. 2), 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

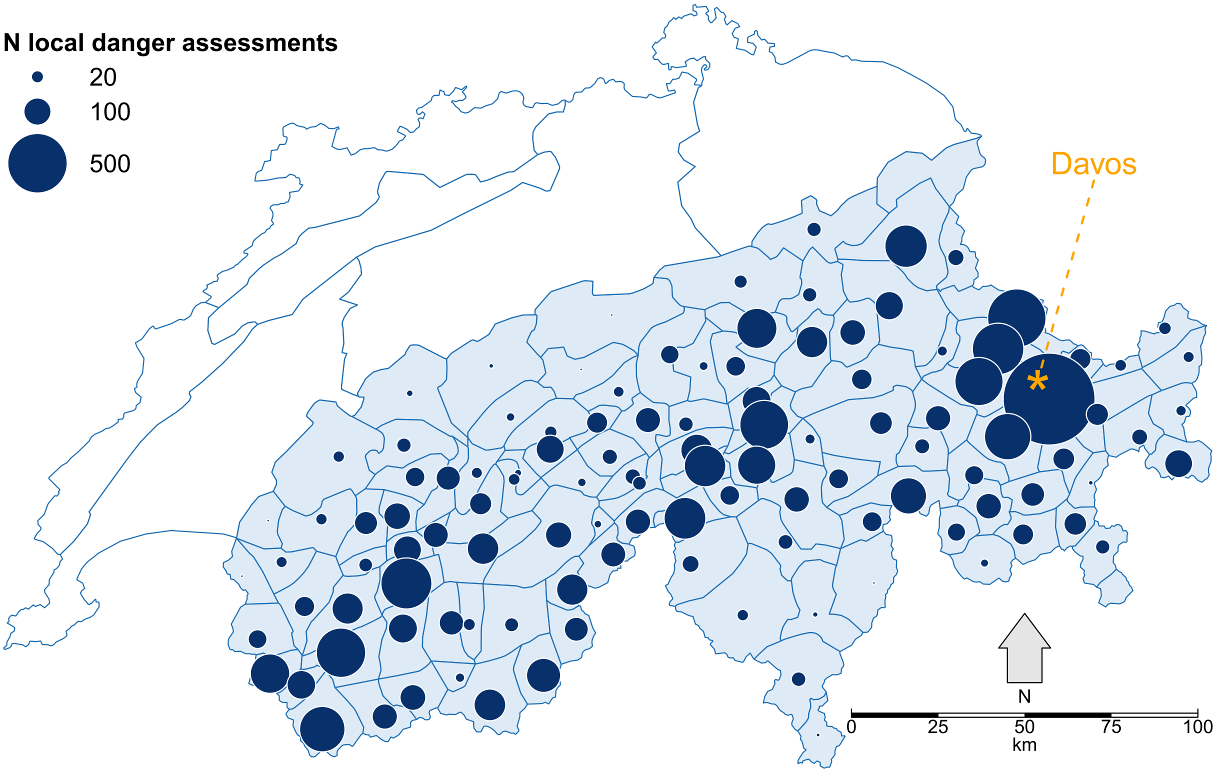
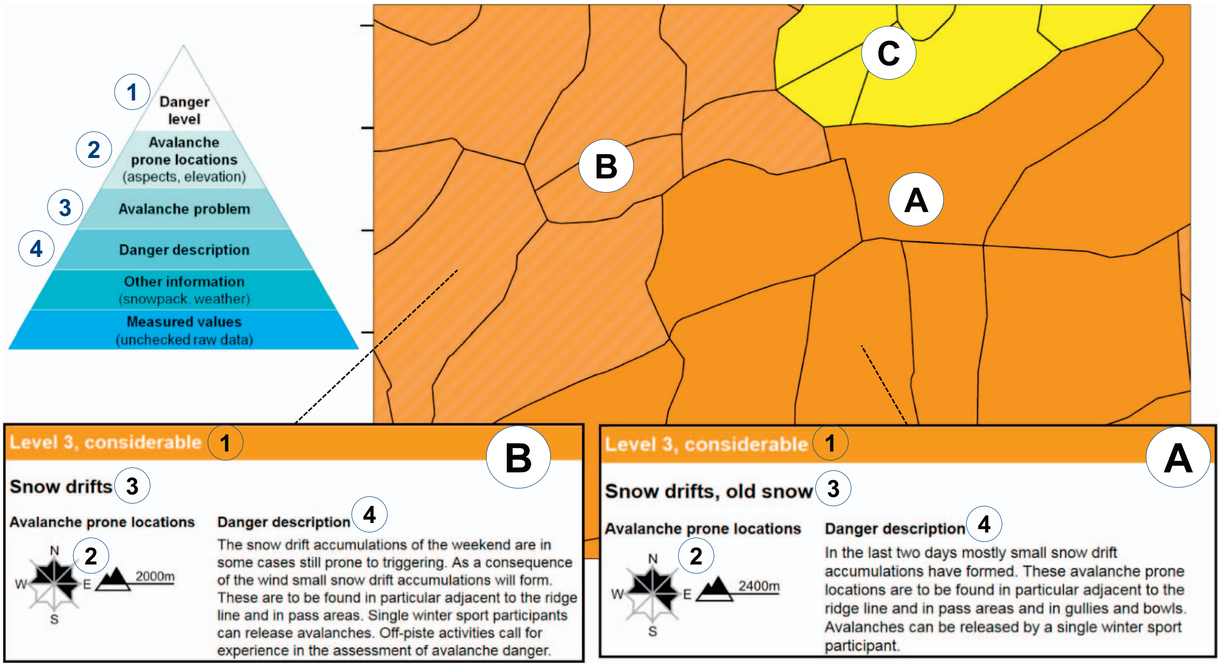


Fig. 1. Map of Switzerland showing number of local danger level estimates per warning region (polygons colored light-blue). The size of the dark-blue circles corresponds to the number of local danger level estimates D_{LN} for each of the 117 warning regions (total forecast area: 26,400 km^2; nine winters, $N = 9553$). The national avalanche warning service is located at SLF in Davos. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 $D_{RF}{}^{evening}$.

The forecast danger level D_{RF} for dry-snow avalanches was level 1-



Level 3, considerable 1
B
Snow drifts 3
Avalanche prone locations 2
Danger description 4
The snow drift accumulations of the weekend are in some cases still prone to triggering. As a consequence of the wind small snow drift accumulations will form. These are to be found in particular adjacent to the ridge line and in pass areas. Single winter sport participants can release avalanches. Off-piste activities call for experience in the assessment of avalanche danger.

Level 3, considerable 1
A
Snow drifts, old snow 3
Avalanche prone locations 2
Danger description 4
In the last two days mostly small snow drift accumulations have formed. These avalanche prone locations are to be found in particular adjacent to the ridge line and in pass areas and in gullies and bowls. Avalanches can be released by a single winter sport participant.

Fig. 2. Map showing an extract of the Swiss avalanche bulletin published on the morning of 7 March 2016 (70 km × 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).

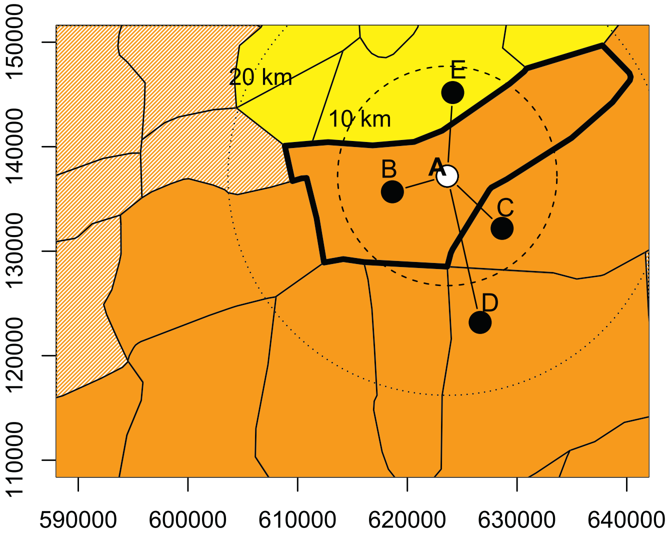


Fig. 3. Map showing an extract of 2000 km^2 of the Swiss Alps (Swiss coordinates in m, 50 km × 40 km) with the color corresponding to the forecast danger level (orange D_{RF} = 3-Considerable and yellow D_{RF} = 2-Moderate, forecast issued in the morning of 7 March 2016, example corresponds to bulletin shown in Fig. 2). 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

3. Methods

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. 3 where the estimates by observers A inside the polygon with outlines in bold are compared to other nearby observers, first of all 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. 3, within region A in Fig. 2). 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. 3), 20 and 30 km, and between more than 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. 3), 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 ΔD_{LN} of an observer X ($D_{LN}{}^{X}$) compared to other observers $D_{LN}{}^{i}$ (i = 1,2,3,…n) is then $\Delta D_{LN} = D_{LN}{}^{X} - D_{LN}{}^{i}$.

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

The rate of disagreement (R_d) for an observer X is therefore the ratio of the number of disagreements between observer X and other observers $N_{\Delta D_{LN}} \neq 0$ to the number of all comparisons between X and other observers $N_{\Delta D_{LN}}$:

$$R_d = \frac{N_{\Delta D_{LN} \neq 0}}{N_{\Delta D_{LN}}}.$$

To explore whether ΔD_{LN} was equally often higher or lower or whether a bias existed, we calculated a bias $\Delta D_{LN}{}^{mean}$ for each observer:

$$\Delta D_{LN}^{mean} = \frac{N_{high} - N_{low}}{N_{high} + N_{equal} + N_{low}}$$

where N_{high}, N_{equal} and N_{low} are the number of higher, equal or lower values of ΔD_{LN}, respectively.

Furthermore, we explored whether the disagreements were randomly distributed or whether a significant bias towards lower or higher disagreements existed. To this end, we calculated the proportion of equally distributed disagreements ($N_{high} = N_{low}$), and the unbalanced disagreements ($\max(N_{high}, N_{low}) - \min(N_{high}, N_{low})$). 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, $|\Delta D_{LN}{}^{mean}| = R_d$ would indicate that all disagreements were either higher or lower. We tested whether N_{high} and N_{low} were significantly different than an equal distribution of N_{high} and N_{low} using the chi-square based non-parametric proportion test (R Core Team, 2016).

We always tested whether the larger number, for instance N_{high}, if N_{high} was greater than N_{low}, deviated significantly from a balanced distribution. 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 $|\Delta D_{LN}{}^{mean}|$, 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.

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 an asterisk, e.g. $R_d{}^{*}$.

3.3. Comparing local nowcasts to regional forecasts

Similar to the procedure described above, we calculated the difference between local nowcast D_{LN} and regional forecast D_{RF} using the integer values of the danger level $\Delta D = D_{LN} - D_{RF}$. If the danger levels agreed ($\Delta D = 0$), we refer to this case as a hit.

The hit rate (HR) 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 $D_{RF.}$

$$HR = \frac{N_{\Delta D=0}}{N_{\Delta D}}.$$

The forecast bias was calculated

$$\Delta D^{mean} = \frac{N_{high} - N_{low}}{N_{high} + N_{equal} + N_{low}},$$

where N_{high}, N_{equal} and N_{low} are the number of higher, equal or lower ΔD, respectively.

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

As we intended to explore whether R_d and HR differed depending on the forecast danger level – the ease of perceiving the hazard –, we divided the D_{RF} data into groups:

First, by the forecast danger level (four groups, 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 ($D_{RF}^{increasing}$), no change in danger rating ($D_{RF}^{no\ change}$), and a decreasing danger ($D_{RF}^{decreasing}$).

Splitting D_{RF} into these groups was motivated by the fact that $D_{RF}^{increasing}$ 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, $D_{RF}^{no\ change}$ and $D_{RF}^{decreasing}$, 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. As for the group $D_{RF}^{no\ change}$, the last group, $D_{RF}^{decreasing}$, 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 ΔD using (a) all individual comparisons, (b) days and regions when observers unanimously 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 R_d. To incorporate R_d, we made the simplifying assumption that the disagreement was always one danger level if $R_d \neq 0$ since deviations of more than one danger level were rare (see below).

To consider the reporting bias and the disagreement rate R_d we proceeded as follows:

- Step 1: As outlined in the bootstrap section before, we randomly selected n data units from the original sample of size n allowing replacement but using the distribution of the forecast D_{RF} and whether D_{RF} had changed from the day before as selection weights for a subset M.
- Step 2: 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 (100–100 $*$ R_d)% of M^2 that the D_{LN} estimate was correct. For example, as will be shown below, R_d was 26% for days with D_{RF} = 3 and $D_{RF}^{increasing}$. 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} was possible, we used the observed distributions of ΔD shown in the Results 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.

Statistical test results were considered significant if $p \leq 0.05$.

All analyses were performed using the statistics software R (R Core Team, 2016).

4. Results

4.1. Local danger level estimates

4.1.1. 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 1).

The disagreement rate R_d 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 1). 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.

As can be noted in Fig. 4 (x-axis), R_d varied considerably between observers. In fact, 8 out of the 40 observers with more than 20 comparisons to other observers had a disagreement rate $R_d \geq 30\%$. 37% of the disagreements were unbalanced for observer pairs within the same warning region (39% for the resampled data, Fig. 4, y-axis). For some observers all the disagreements were unbalanced ($\Delta D_{mean}^{LN} = R_d$, corresponding to the points on the dotted lines in Fig. 4). 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_{LN} estimates (Fig. 4a). Including comparisons with observers in neighboring warning regions with the same danger rating, 9 (or 12%) out of the 75 observers with

Table 1
Disagreement rate R_d 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. 2). Different danger means different danger level. The number of pairs (N) and the median distance is given.

Warning region or distance between observers	Danger	R_d	N	Median 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

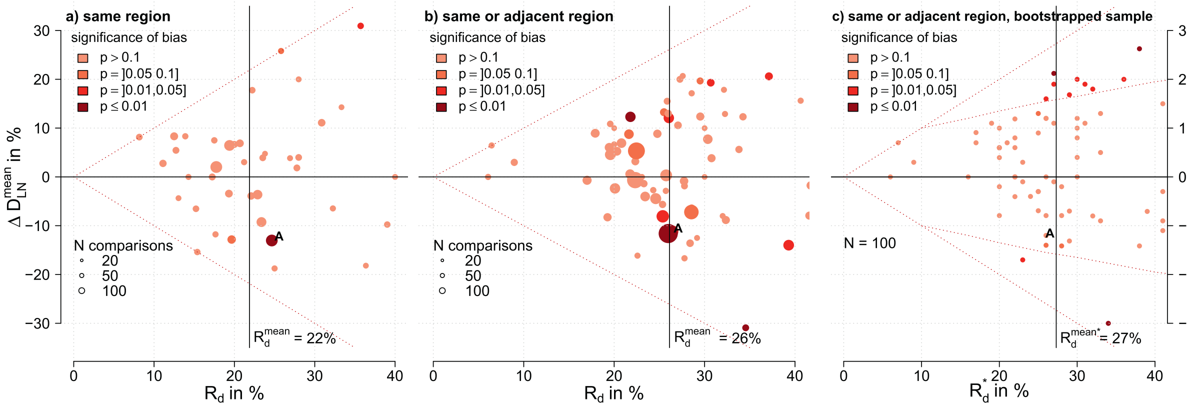


Fig. 4. The disagreement rate R_d and the bias (ΔD_{LN}^{mean}) between local danger level estimates for each observer. (a) within the same warning region, (b) and (c) within the same or an immediately neighboring region with the same danger level and description. In (a) and (b) the size of the circles 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 ($|\Delta D_{LN}^{mean}| = R_d$). Color coding corresponds to significance levels.

Table 2
Disagreement rate R_d^* within the same warning region or in neighboring warning regions with the same danger rating at distances < 10 km, with respect to the forecast regional danger level D_{RF} and whether D_{RF} changed from the previous day. The arrow-symbols indicate whether D_{RF} increased ↗, stayed the same → or decreased ↘. The mean and the standard deviation of the disagreement rate R_d^*, and the number of pairs N are given.

D_{RF}	Mean R_d^*				Standard deviation R_d^*				N comparisons			
	↗	→	↘	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

more than 20 comparisons to others had a significant bias (Fig. 4b).

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 ($R_d = 25\%$) with a ΔD_{LN}^{mean} of 13% (Fig. 4a). While this bias was significant for observer A ($p = 0.002$), a similar or larger ΔD_{LN}^{mean} 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. 4c). Using this latter approach, 13 (or 17%) out of the 75 observers would be considered as being significantly biased ($p \leq 0.05$).

4.1.2. 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 R_d within the group SLF ($R_d^{SLF} = 22\%$, $N = 86$, employees 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 ($R_d^{guides} = 24\%$, $N = 55$) or the whole Swiss Alps ($R_d^{guides} = 22\%$, $N = 516$).

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_{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).

4.1.3. Variations with regard to the forecast regional danger level

In addition to the group-specific variations, we explored whether R_d varied with the forecast danger level D_{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 levels 3-Considerable and 4-High (51% vs. 42%, $p < 0.001$). D_{RF} did not change in 80% of the days and warning regions from one day to the next. D_{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 Table 2, 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, R_d was highest on days when D_{RF} increased (27% ± 3%, mean ± standard deviation) and on days with a D_{RF} 4-High (27% ± 7%), and lowest on days when D_{RF} decreased (14% ± 2%). R_d was particularly low on days when the danger level was lowered from level 4-High to level 3-Considerable (6% ± 1%). R_d was significantly different between days when D_{RF} increased and those when D_{RF} decreased (27% vs. 14%, $p = 0.04$) and when D_{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_{RF}, the disagreement rate was 40% for distances less than 10 km and 49% for distances between 10 and 20 km (Table 1).

4.2. Comparing local nowcasts to regional forecasts

In total, 9543 individual comparisons between local danger level estimates D_{LN} and regional danger level forecasts D_{RF} were analyzed.

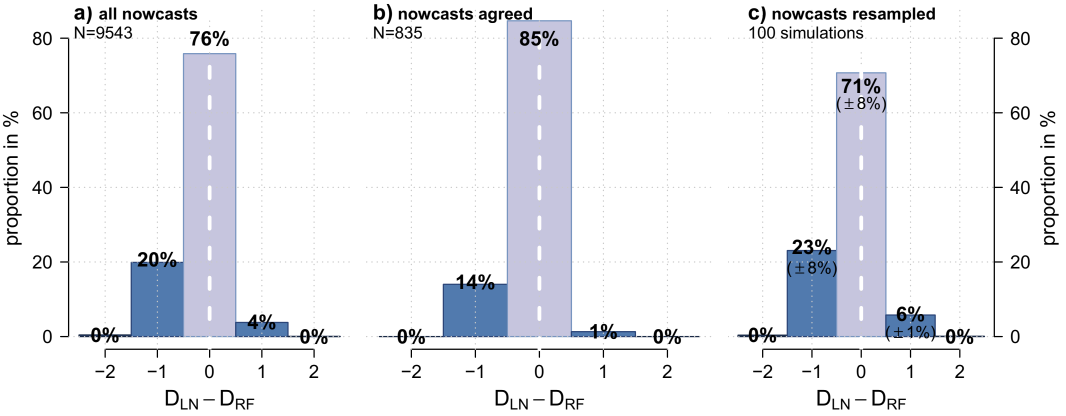


Fig. 5. Distributions of the differences between the local danger level estimate (D_{LN}, nowcast) and the regional danger level forecast (D_{RF}, forecast) for (a) all estimates individually compared to the forecast, (b) for days and regions when the observers in the same region agreed on the danger 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 3) and (2) the disagreement rate R_d between observers (Table 2).

The estimates were provided by 137 different observers on 1076 days and in 115 warning regions.

The hit rate was 76% (Fig. 5a). 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_{RF} was two levels too high or too low. D_{LN} was more often lower than D_{RF} with 20% D_{RF} too high vs. 4% D_{RF} too low. D_{RF} was most frequently considered too low when D_{RF} decreased (11%). In contrast, D_{RF} was most often considered too high when D_{RF} increased (37%). The hit rate was lowest on days when D_{RF} increased (HR = 61%) or when the forecast danger level was 4-High (HR = 28%, Table 3). The latter would indicate that the forecast danger level was perceived mostly as being incorrect. On the opposite side, HR 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 two neighboring warning regions with different danger ratings, observers disagreed often with D_{RF} for the region with the higher danger rating (HR = 51%). In contrast, in the region with the lower rating, observers frequently estimated D_{LN} the same as D_{RF} (HR = 84%).

Considering each observer individually revealed large scatter (Fig. 6). While almost all observers tended to estimate the local danger to be lower than forecast, the frequency on which they considered D_{RF} to be wrong by one danger level varied considerably.

The hit rate was almost identical for those working at SLF (HR = 74%, N = 1047) as for other observers and mountain guides (HR = 76%, N = 8489). However, if just the avalanche forecasters at SLF were considered as a group, a slightly higher hit rate was noted (80%, N = 417). Expanding the comparison to the estimates made during the day in the valley floor (HR = 87%, N = 1971, 55 different observers) or by observers working in ski areas (HR = 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

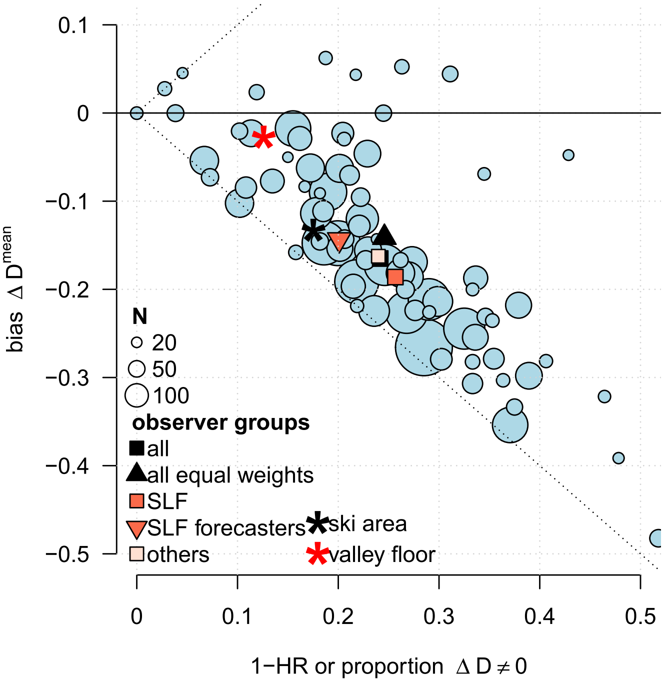


Fig. 6. For each observer, the proportion of days with a local estimate being different than the regional forecast 1 − HR (x-axis) and the bias ΔD^{mean} (y-axis) is shown. The dotted lines correspond to $\Delta D^{mean} = |1 - HR|$ indicating that all differences between D_{LN} and D_{RF} would either 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.

Table 3
Hit rate HR between D_{LN} and D_{RF} ($\Delta D = 0$), 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 ↗, stayed the same → or decreased ↘.

D_{RF}	$\Delta D = 0$				$\Delta D < 0$				$\Delta D > 0$			
	↗	→	↘	All	↗	→	↘	All	↗	→	↘	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%

Table 4
Hit rate HR between D_{LN} and D_{RF} ($\Delta D = 0$) for days and regions, when observers either unanimously agreed on D_{LN} or when a majority opinion existed and depending on the forecast regional danger level and whether D_{RF} changed to the day before in the same region ($N = 835$). The arrow-symbols indicate whether D_{RF} increased ↗, stayed the same → or decreased ↘.

D_{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%

observers estimated D_{LN} significantly often higher than the field observers ($p < 0.01$). Although ski area observers were also more often lower 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_{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 (HR = 75.9%) than with the evening forecast of the previous day $D_{RF}^{evening}$ ($HR^{evening} = 75.3\%$).

Considering only days and regions when two or more observers agreed in their nowcast estimate or when there was a majority opinion on D_{LN}, the agreement with the forecast D_{RF} was higher (HR = 85%, $N = 835$, Fig. 5b). However, as can be seen in Table 4, the values are rather extreme and the hit rate ranges from 0% to 100%. We attribute this to the relatively small sample size in some of the cells in Table 4.

Incorporating the reporting bias (Table 3) and the disagreement rate in the calculation (Table 2) and using the full sample ($N = 9543$), the hit rate was 71% (standard deviation 8%, Fig. 5c, Table 5). Comparing Tables 3 and 5 shows that the hit rate increased for days when the hit rate in Table 3 and $R_{d.}$ in Table 2 were low (for instance for days with D_{RF} = 4-High). In contrast, comparably high HR (e.g. for D_{RF} = 2-Moderate or 1-Low, Table 2) decreased somewhat.

5. Discussion

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^2 (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 wide-spread 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 ≥ 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 in 50% of the time 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

Table 5
Hit rate HR between D_{LN} and D_{RF} ($\Delta D = 0$), proportion of $\Delta D < 0$ and $\Delta D > 0$ incorporating the disagreement matrix (Table 2) and the reporting bias, the frequency of D_{RF} and whether D_{RF} changed to the day before in the same region. The arrow symbols indicate whether D_{RF} increased ↗, stayed the same → or decreased ↘. Cell values represent the mean of 100 repetitions.

D_{RF}	$\Delta D = 0$				$\Delta D < 0$				$\Delta D > 0$			
	↗	→	↘	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%

with the higher danger level (HR = 51%) and a higher accuracy in the region with the lower danger level (HR = 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%.

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^2, 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^2) to forecasts at the mountain range scale (about 1000 to 10,000 km^2) (Schweizer and Kronholm, 2007). In addition, there is a temporal scale mismatch – a forecast valid for a 12 to 24 h period is compared to a local assessment, which is often based on (part of the) day spent in the field (often less than 6 h; 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 hit rate, 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" (R_d = 26% and $R_{d.}$ = 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 traveled in frequently tracked terrain or not, was neither correlated with the disagreement rate nor the hit rate.

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 measureable, 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. 6). In our study, valley-floor and ski area observers as well as SLF forecasters were closer to the forecasted 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.

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 (76%, Fig. 5a).

A higher hit rate (85%, Fig. 5b) 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. (2016). The overall higher hit rate can be expected, as situations with less obvious danger ratings are likely excluded using this sample.

Finally, the third approach, yielding a hit rate of 71% ± 8% (Fig. 5c) incorporated the uncertainty in the D_{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., 1998; 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 hit rate 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 hit rate 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 hit rate 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 7900 km^2 to 6000 km^2. 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 ± 3% to 70.8 ± 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 hit rate was marginally higher (71.8% ± 8%,) than the Swiss average (70.8 ± 9%). In other areas the hit rate was comparable or even higher, for instance in the Lower Valais in the western Swiss Alps (74.8% ± 9%). In contrast, a significantly lower hit rate (66.9% ± 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.

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% ± 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.

Acknowledgements

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.

References

Bakermans, L., Jamieson, B., Schweizer, J., Haegeli, P., 2010. Using stability tests and regional avalanche danger to estimate the local avalanche danger. Ann. Glaciol. 51 (54), 176–186.

Brabec, B., Stucki, T., 1998. Verification of avalanche bulletins by questionnaires. In: Hestnes, E. (Ed.), 25 Years of Snow Avalanche Research, Voss, Norway, 12–16 May 1998. NGI Publication. Norwegian Geotechnical Institute, Oslo, Norway, pp. 79–98.

Bründl, M., Etter, H.J., Steiniger, M., Klingler, C., Rhyner, J., Ammann, W., 2004. IFKIS - a basis for managing avalanche risk in settlements and on roads in Switzerland. Nat. Hazards Earth Syst. Sci. 4 (2), 257–262.

Cagnati, A., Valt, M., Soratori, G., Gavaldà, J., Sellés, C.G., 1998. A field method for avalanche danger-level verification. Ann. Glaciol. 26, 343–346.

Engeset, V.R., 2013. National Avalanche Warning Service for Norway – established 2013. In: Naaim-Bouvet, F., Durand, Y., Lambert, R. (Eds.), Proceedings ISSW 2013. International Snow Science Workshop, Grenoble, France, 7–11 October 2013. ANENA, IRSTEA, Météo-France, Grenoble, France, pp. 301–310.

Föhn, P.M.B., Schweizer, J., 1995. Verification of avalanche danger with respect to avalanche forecasting. In: Sivardière, F. (Ed.), Les Apports de la Recherche Scientifique à la Sécurite Neige, Glace et Avalanche. Actes de Colloque, Chamonix, 30 Mai–3 Juin 1995. ANENA, Grenoble, France, pp. 151–156.

Haegeli, P., 2010. Avaluator: Avalanche Accident Prevention Card. Canadian Avalanche Centre, Revelstoke BC, Canada (30 pp).

Haladuick, S., 2014. Relating Field Observations and Snowpack Tests to Snow Avalanche Danger. University of Calgary, Calgary AB, Canada (M.Sc. Thesis, 178 pp).

Harvey, S., Rhyner, H., Schweizer, J., 2012. Lawinenkunde. Bruckmann Verlag GmbH, München, Germany (192 pp).

Harvey, S., Rhyner, H., Dürr, L., Schweizer, J., Henny, H.M., Nigg, P., 2016. Caution - Avalanches! Avalanche Prevention in Snow Sports. Core team of Instructors, Davos, Switzerland.

Jamieson, J.B., Campbell, C., Jones, A., 2008. Verification of Canadian avalanche bulletins including spatial and temporal scale effects. Cold Reg. Sci. Technol. 51 (2–3), 204–213.

Jamieson, B., Haegeli, P., Schweizer, J., 2009. Field observations for estimating the local avalanche danger in the Columbia Mountains of Canada. Cold Reg. Sci. Technol. 58 (1–2), 84–91.

McClung, D.M., 2000. Predictions in avalanche forecasting. Ann. Glaciol. 31, 377–381.

McClung, D.M., 2002a. The elements of applied avalanche forecasting - part I: the human issues. Nat. Hazards 26 (2), 111–129.

McClung, D.M., 2002b. The elements of applied avalanche forecasting - part II: the physical issues and the rules of applied avalanche forecasting. Nat. Hazards 26 (2), 131–146.

Meister, R., 1995. Country-wide avalanche warning in Switzerland. In: Proceedings International Snow Science Workshop, Snowbird, Utah, U.S.A., 30 October-3 November 1994. ISSW 1994 Organizing Committee, Snowbird UT, USA, pp. 58–71.

Müller, K., Stucki, T., Mitterer, C., Nairz, P., Konetschny, H., Feistl, T., Coléou, C., Berbenni, G., Chiambretti, I., 2016. Towards an improved European auxiliary matrix for assessing avalanche danger levels. In: Greene, E. (Ed.), Proceedings ISSW 2016. International Snow Science Workshop, Breckenridge CO, U.S.A., 3-7 October 2016, pp. 1229–1231.

Page, S.E., 2007. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press, Princeton NJ, U.S.A. (456 pp).

R Core Team, 2016. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ruesch, M., Egloff, A., Gerber, M., Weiss, G., Winkler, K., 2013. The software behind the interactive display of the Swiss avalanche bulletin. In: Naaim-Bouvet, F., Durand, Y., Lambert, R. (Eds.), Proceedings ISSW 2013. International Snow Science Workshop, Grenoble, France, 7–11 October 2013. ANENA, IRSTEA, Météo-France, Grenoble, France, pp. 406–412.

Schweizer, J., 2010. Predicting the avalanche danger level from field observations. In: Proceedings ISSW 2010. International Snow Science Workshop, Lake Tahoe CA, U.S.A., 17-22 October 2010, pp. 162–165.

Schweizer, J., Föhn, P.M.B., 1996. Avalanche forecasting - an expert system approach. J. Glaciol. 42 (141), 318–332.

Schweizer, J., Kronholm, K., 2007. Snow cover spatial variability at multiple scales: characteristics of a layer of buried surface hoar. Cold Reg. Sci. Technol. 47 (3), 207–223.

Schweizer, J., Kronholm, K., Wiesinger, T., 2003. Verification of regional snowpack stability and avalanche danger. Cold Reg. Sci. Technol. 37 (3), 277–288.

Statham, G., Haegeli, P., Birkeland, K.W., Greene, E., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., Kelly, J., 2010. The North American public avalanche danger scale. In: International Snow Science Workshop ISSW, Lake Tahoe CA, U.S.A., 17-22 October 2010, pp. 117–123.

Suter, C., Harvey, S., Dürr, L., 2010. mAvalanche - Smart avalanche forecasting with smartphones. In: Proceedings ISSW 2010. International Snow Science Workshop ISSW, Lake Tahoe CA, U.S.A., 17-22 October 2010, pp. 630–635.

Techel, F., Zweifel, B., Winkler, K., 2015. Analysis of avalanche risk factors in backcountry terrain based on usage frequency and accident data in Switzerland. Nat. Hazards Earth Syst. Sci. 15 (9), 1985–1997.

Techel, F., Dürr, L., Schweizer, J., 2016. Variations in individual danger level estimates within the same forecast region. In: Greene, E. (Ed.), Proceedings ISSW 2016. International Snow Science Workshop, Breckenridge CO, U.S.A., 3-7 October 2016, pp. 466–471.

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences. In: International Geophysics Series, 100. Academic Press, San Diego CA, U.S.A (467 pp).

Winkler, K., Kuhn, T., 2017. Fully automatic multi-language translation with a catalogue of phrases: successful employment for the Swiss avalanche bulletin. Lang. Resour. Eval. 1–23.

Winkler, K., Techel, F., 2014. Users' rating of the Swiss avalanche forecast. In: Haegeli, P. (Ed.), Proceedings ISSW 2014. International Snow Science Workshop, Banff, Alberta, Canada, 29 September - 3 October 2014, pp. 437–444.

Winkler, K., Bächtold, M., Gallorini, S., Niederer, U., Stucki, T., Pielmeier, C., Darms, G., Dürr, L., Techel, F., Zweifel, B., 2013. Swiss avalanche bulletin: automated translation with a catalogue of phrases. In: Naaim-Bouvet, F., Durand, Y., Lambert, R. (Eds.), Proceedings ISSW 2013. International Snow Science Workshop, Grenoble, France, 7–11 October 2013. ANENA, IRSTEA, Météo-France, Grenoble, France, pp. 437–441.

Zenke, B., 2013. Grenzen des Lawinenlageberichtes. bergundsteigen - Zeitschrift für Risikomanagement im Bergsport. Oesterr. Alpenverein. Innsbruck, Austria 22 (4), 30–34.