Response to reviewer RC1 (K. Müller)

Frank Techel[1,2]

[1]WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland
[2]Department of Geography, University of Zurich, Zurich, Switzerland

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

We greatly thank Karsten Müller for his very detailed review and the helpful comments.

Our response is shown in blue, *intended changes to the manuscript are in italics.*

The presented study analyses the forecasting goodness of avalanche forecasts from 23 different forecasting centers in the European Alps over a period of four years. The authors use the agreement in danger level between neighboring regions (within and between different forecasting centers) as a measure of forecast consistency and bias. They present a method to explore and quantify spatial consistency of forecast regional avalanche danger levels. Bias between neighboring regions could to some extend be attributed to operational constraints of the involved forecast centers. The paper gives a good overview of the different practices and concepts for production and communication of avalanche forecasts in the European Alps. The presented statistics give insight into the different approaches and can provide valuable input for future improvements in avalanche forecasting and communication. The dataset presented is extensive and novel and can certainly help to understand and harmonize avalanche forecasting in the European Alps and worldwide. The text itself is often complicated with long sentences. Simpler and more to the point language throughout the whole paper would be beneficial for the readability and understanding of the paper. Especially for the more technical chapters 3 to 5. Try to avoid repetition. Sometimes terms are defined two or three times throughout the text. Figures and tables are generally good and informative. The study is of value to the avalanche community issuing or using regional avalanche forecasts and suited for publication in NHESS after addressing the following general and specific comments.

1 General comments

The authors follow (Murphy, 1993) to assess forecast goodness based on three factors (quality, consistency and value). While they exclude quality since it is nearly impossible to measure, consistency and value are considered. The authors use P_{agree} as a measure for the consistency of the avalanche forecast. They state that disagreement can be attributed to either climatological or topographical differences or differences in the production of the forecasts between different forecasting centers. I question the value of P_{agree} as a measure of consistency and miss a discussion on the expected agreement rate or consistency. Aside from political borders, the reason for the delineation of individual forecasting regions is that different avalanche conditions are to be expected. An agreement of close to 100% between two neighboring regions indicates that the boundary between them is superfluous? This point is not addressed. On the other hand, there are only five danger levels. A certain agreement is therefore

expected considering that danger levels 2 and 3 are well overrepresented (being issued up 80% of the time) over the course of a forecasting season.

p10 l9: with most of the forecasts during the winter having DL2 or DL3 chances are very high that avalanche danger levels agree between neighboring regions despite differences in size or validity period. Could you present some numbers and discuss this "issue"?

We explored spatial consistency. We consider the agreement rate P_{agree} between neighbouring warning regions as an appropriate statistical indicator of spatial correlation as it provides an easy-to-interpret value ranging from 1 (perfect agreement / correlation) to 0 (total disagreement / no correlation). However, reviewer 1 rightly points out several points, which we now discuss: Firstly, $P_{agree} = 1$ (or 100%), within the domain of a forecast center means that there is no need to have two warning regions instead of one. However, in cases where a mountain range with similar snow climate is cut by political borders, $P_{agree} =$ 100% might theoretically be possible. Furthermore, we do not expect spatial homogeneity. However, we would expect values of P_{agree} to be relatively similar regardless of whether we explore within forecast center boundaries, or across those. Furthermore, as reviewer 1 points out, the frequencies of danger levels 2 and 3 are high. By random chance, P_{agree} will be much larger than 0. So far, we have not addressed these points.

To give the reader some guidance what range of values of P_{agree} would be expected, we will address these issues in the revised manuscript by providing benchmark *values of P_{agree}. The following may help with an interpretation:*

- *We randomly simulate danger levels for two regions using the overall distribution of D_{max} (shown in Fig. 1b), and calculate P_{agree}. This will provide the reader with an approximate lower value. These values are for D_{max} 0.4 (or 40%) and for $D_{morning}$ 0.36 (using distribution shown in Fig. 1a, we simulated 10'000 pairs).*
- *We will emphasize, that in most cases P_{agree} = 100% is not expected (except as mentioned above).*
- *We will compare P_{agree} values within and across forecast center boundaries by stratifying the data by distance (distance of the center points of the warning regions), using only a subset of the immediately neighbouring warning region pairs where the difference in max. elevation is less than 250 m and where the size of larger warning regions is less than 1.5 times the size of the smaller warning region. By using these subsets we compare relatively similar regions. Doing so, our results show agreement rates P_{agree} which differ significantly by more than 0.25 (or 25%; $p<0.001$), confirming the results shown in Table 4 in the manuscript. We propose to show these results graphically (Fig. 2), rather than in their current form as a Table (Table 4 in the manuscript).*
- *we will discuss in more detail, what could be expected and what are the limitations of this approach*

Across political boundaries, avalanche conditions could be expected to be more similar and a disagreement between danger levels could indicate a substantial difference in assessing avalanche danger or interpreting the avalanche danger scale. The study could be strengthened by filtering regions and considering only those that border to regions of different forecasting centers and exclude those that only border with "internal" forecasting regions. Thus, potential conceptual differences between individual forecasting centers might be easier to identify.

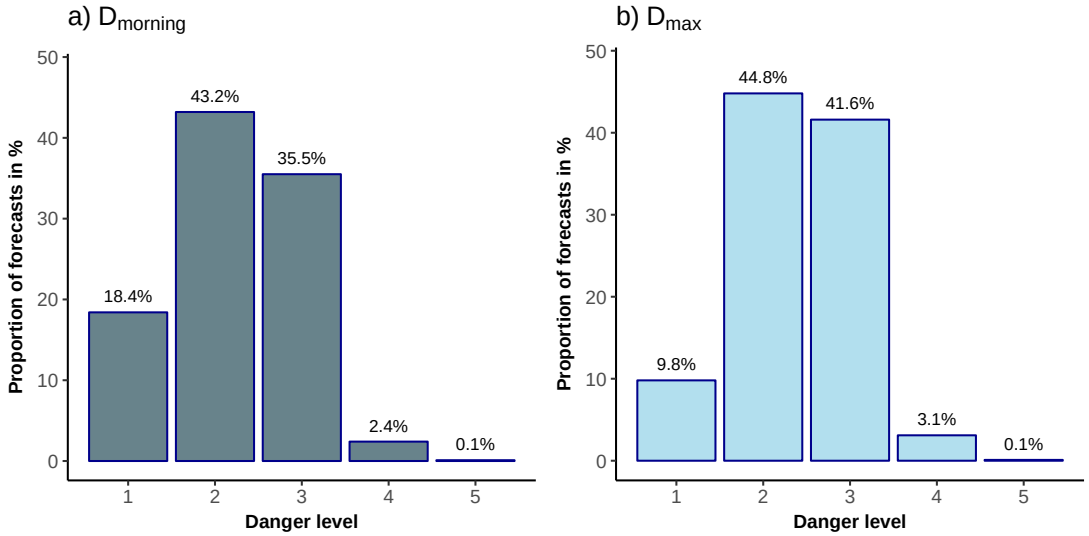

Figure 1. Distribution of forecast danger levels in our dataset, for (a) $D_{morning}$ (danger level valid during first time-step) and (b) D_{max} (highest danger level).

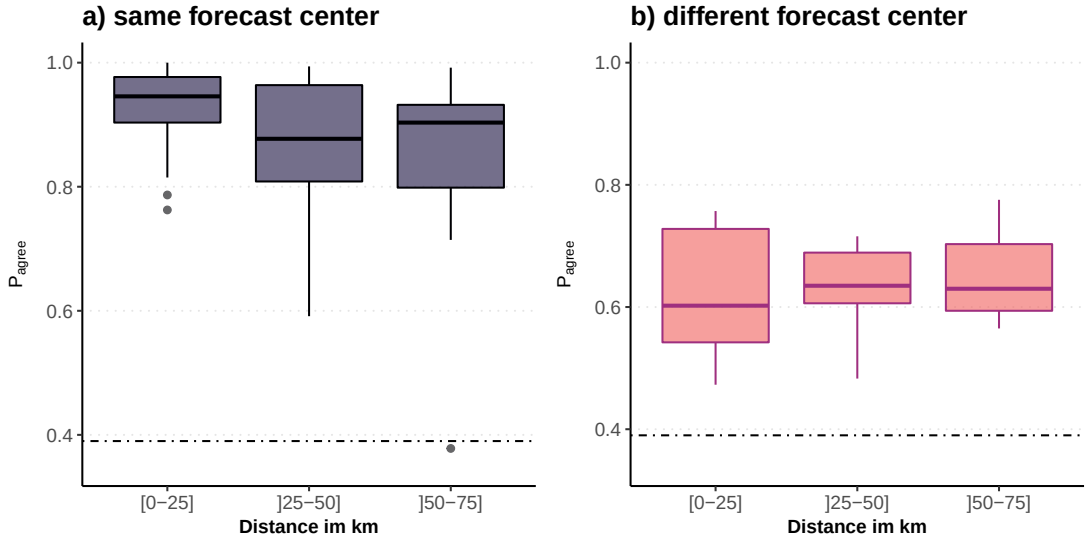

Figure 2. Boxplot showing the agreement rate for neighboring warning region pairs, with similar elevation (Δ elevation$<$ 250 m) and with similar size of the warning regions (the size of the larger warning region is less than 1.5 times the size of the smaller warning region). ($N_{same\ forecast\ center}$ = 108, $N_{different\ forecast\ center}$ = 37).

p27 l11ff: I agree and it would have been interesting to filter the warning regions accordingly and make a separate analysis of regions of neighboring forecasting centers (ideally with an presumably similar snow climate if this information had been

available).

We suggest to show one, possibly two examples of regions which have similar snow climate, but where several warning services / forecast centers provide forecasts. Possibly, this will be a mountain range which lies in Vorarlberg/VOR, Tirol/TIR and Switzerland/SWI (Silvretta mountains), and maybe also a region of Lombardia/LOM which is surrounded on three sides by Switzerland.

"Value" is presented as being both connected to "quality" and "consistency" in the introduction. The authors should be more precise on if and how they evaluate "value". Section 6.4 presents some general reflections around the value of avalanche forecasts to the users, but an assessment of "value" with regard to the presented statistics is lacking in the methods and conclusions.

We intentionally only provided general reflections on *value*. However, we agree with reviewer 1 that we introduced *value*, together with *consistency* as the two *goodness* measures we analyze in detail.

We will state more clearly that we primarily explore spatial consistency *using* P_{agree} *and bias* B_{ij}*. We will emphasize that we will only reflect on* value *in a more general sense, without quantitatively exploring it. Furthermore, we will introduce these terms better.*

The research questions from p3 should be answered in the conclusion. While questions 1 and 2 are addressed answers to questions 3 and 4 should also be given.

this will be done

Based on the author's analysis, region size seems to be an important parameter for the consistency of a forecast. Region size can be adequately analyzed based on the presented data and should be emphasized in the discussion and conclusions.

this will be done

2 Specific comments

We have moved some of the specific comments to the general comments, where we thought it appropriate to answer them together.

p1 l7: Can we actually expect consistency between neighboring regions wrt danger level. In many cases the situation might actually be different and require different danger levels.

p1 l10: Same as for L7 - could be geographical or meteorological reasons for this.

We will revise the abstract introducing better that we explore spatial consistency, what can be expected, and what we found.

p3 l5-7: Can you state that more clearly? I think what you mean is that you compare a single categorical value (given for an area and a certain time span) to a complex and dynamic situation (often over a subset of the valid area and time). This will

even be more pronounced when comparing regions of rather different size.
We will state more clearly what we compare, and take up these points in the discussion as well.

p3 l22: a requirement for this would be that forecasters within each center work consistently, at least with respect to other forecasting centers they are compared to. I assume this is an assumption which is difficult to verify.
Admittedly, we have no information about how consistent forecasters work within each forecast center. We will rephrase this statement.

p3 l24ff: Please be more clear about your use of the terms quality, consistency and especially value. On p3 l19 you state that quality is not measurable. In the abstract and here you state that you focus on consistency which has implications on quality and therefore value. You assume quality to be consistent in your data. On p3 l3 you introduce value as "the benefits or costs incurred by a user as a result of a forecast". Here you state that "implication for the value" are a "result of potential differences in consistency". To me this is somewhat confusing and it is not obvious to me if and how you consider value in your study at all.
we will be more precise about how we use these terms, that we do not explore quality at all, and that we only qualitatively discuss value.

p5 l11: difference between forecast center and AWS not clear.
We made this distinction for the following reason: We expect the greatest consistency between forecasters who work regularly together in the same forecasting office and with the same operational constraints. In contrast, an AWS which has several decentralized forecast centers (as the four forecast centers belonging to AWS MétéoFrance), is characterized by the same operational constraints, but forecasters meet less regularly (compared to them working together). And finally, different AWS have forecasters working in different locations and issue products with different looks / different operational constraints (as for instance forecasters working for the AWS in the federal states in Austria, or when comparing forecast centers in different countries).
We believe this distinction is necessary, to discuss consistency. However, we will introduce these terms more clearly. Furthermore, we will analyze the data primarily by comparing forecast centers. We will reflect on the influence of the AWS in the Discussion section.

p16 l30: in larger regions the distance to the neighboring region can be larger, which makes it more likely to have different danger ratings due to varying parts of each region influencing the danger level. Please discuss.
We will discuss this in greater detail.

p17 l5: the term maximum elevation needs to be introduced and explained earlier; same for the comparison of region sizes. Please explain what you are analyzing and how you calculate $\rho_{elevation}$ and ρ_{area} in the Methods section, e.g. in 4.2.2.

We will introduce these terms and how we calculate them in the Methods section.

p19 l19: what is the reason for remove single years? Please state. Later you argue that the chosen four years are a representative excerpt which would imply no need to remove or filter data by individual years.

p26 l20ff: If you consider your data as sufficiently robust the exercise of removing one of the years does not add value to the study and could be moved to the appendix/supplements.

As the data is limited to four years and to be able to make this statement, we on purpose showed that removing individual years influences results mostly when rather extreme years are removed. However, even removing these years, results remained similar and the interpretation valid.

We suggest to move these to the appendix or provide it as a supplement.

p21 l1: why not an analysis for moderate avalanche danger?

There were two reasons: firstly, we wanted to primarily provide results from the upper and lower end of the danger scale. Secondly, including a section on danger level 2 would make the already rather long manuscript even longer.

We suggest to provide the reader with this information as a supplement. To shorten the main part of the paper, we also suggest to provide the result section regarding danger level 3 as a supplement only.

p23 l32: Is there a difference between forecasting centers? Do some issue the highest while others issue the most representative? If yes, was this considered in the analysis other than for the regions in SWI and VDA?

Frankly speaking, we don't know whether there are different practices, although we could imagine that both approaches are used. Furthermore, the definition of the avalanche danger scale is lacking this information. We are aware, however, that other approaches exist. For instance, in Northern Canada a «hot-zone» approach is used, where the focus is on providing information for the part of (large) regions used most frequently by people (Storm and Helgeson, 2014). Some warning services in Austria, for instance Tirol (TIR) and Vorarlberg (VOR) used to communicate an "allgemeine Gefahrenstufe" (overall danger level), which did not always reflect the highest danger level within their forecast domain and forecast time, and which we have not analysed.

In our paper, it is therefore a hypothetical example, which should highlight how this may influence the forecast of higher danger levels, particularly when warning regions are large.

We will keep this part in the manuscript. However, we will highlight that this is a hypothetical example and that we don't know whether one or the other approach is used. We will also highlight that the EADS lacks this definition. We could present some numbers for Tirol/TIR and Vorarlberg/VOR on how often the "allgemeine Gefahrenstufe" was higher or lower than the highest danger level issued. We could also show this for the proportion of forecasts with $D_{max} \geq 4$ ($P_{v.crit}$, shown in Table 1 below), although the results are similar as shown for Valle d'Aosta/VDA and Switzerland/SWI (Table 5 in the Manuscript). Furthermore, we suggest to discuss potential approaches, including the «allgemeine Gefahrenstufe» , the «hot-zone» approach, or the approach used in Norway, which reviewer Rune Engeset (in a personal communication) made me aware of (generally, the

highest danger level is issued, given that this applies to approx. 100 km^2, or more, of the area within the comparably large Norwegian forecast regions.)

Sec 5.5: Aggregation of smaller regions to larger forecasting regions will necessarily lead to the same danger rating and it is likely that warning regions within the same larger snow-climate region will more often aggregated together. Therefore it is expected that the (rather small) regions in SWI and VDA have a higher agreement rate than in other parts of the Alps where regions are larger and not aggregated. Please discuss.

We agree and will discuss this.

p27 l27: It seems like BRI is somewhat special wrt $P_{v.crit}$. Have you looked into potential reasons for that? Special climate/topography/size/location or conceptual differences in producing or communicating avalanche forecasts?

No, we have not looked into this in detail. As we have no information whether or not conceptual differences exist, we cannot make any definite statements in that respect. However, forecasters at AWS *Météo France* have noted and discussed this at the time. While we cannot exclude that forecasters applied the danger level differently, numerous other explanations are possible. To name just some: possibly atypical avalanche conditions during the four-year study period with a record number of avalanche fatalities in the *Hautes-Alpes* (the northernmost of the three administrative divisions (departments) under the responsibility of BRI), an avalanche climate which favors persistent weak layers, forecast errors due to comparably few field observations in the area and a higher uncertainty concerning the precipitation amounts during Southerly air currents.

In our revised manuscript, we will show whether the variables we explore (max. elevation, size of warning region) are different between the forecast domain in Briançon/BRI and its immediate neighbours in France and Italy. If these are similar, we will briefly discuss potential other explanations.

p28 l23ff: It is expected that the smaller regions will less often have higher danger levels than larger regions since the chance to have a critical situation increases with size. It would have been interesting to see if and/or how large the differences were if equally large regions from different forecasting centers had been compared. E.g. picking or aggregating a 2000 km^2 region from each forecasting center and comparing the frequency of higher danger levels.

We agree that this would be interesting. However, this would also mean that we would have to aggregate warning regions likely not belonging to the same climate region (for Valle d'Aosta/VDA and Switzerland we had some idea about which regions belong to a climate region). We will not take up this recommendation, instead we will show and discuss the examples using the »allgemeine Gefahrenstufe« (overall danger level) which was for instance used in Vorarlberg/VOR and Tirol/TIR. This will allow us to show the influence of either approach using real forecasts for the size of VOR (about 2600 km^2, which is also comparable with one of the aggregation levels shown for VDA and SWI), and TIR (about 12600 km^2). It is of note that the »allgemeine Gefahrenstufe« describes the spatially AND temporally most relevant danger rating.

Table 1. The proportion of forecasts with danger levels *4-High* or *5-Very High* ($P_{v.crit}$): $P_{v.crit}$(max) assumes the communication of the highest danger rating for the entire forecast domain, $P_{v.crit}$(mean) the spatially most relevant danger rating, while $P_{v.crit}$(allgemein) refers to the published *allgemeine Gefahrenstufe* as introduced above.

forecast center	area (km^2)	$P_{v.crit}$(max)	$P_{v.crit}$(mean)	$P_{v.crit}$(allgemein)
Tirol/TIR	12600	7%	2.8%	1.8%
Vorarlberg/VOR	2600	6.1%	5%	3.7%

p30 l20ff: Please try to answer your research question from p3 in your conclusion, especially questions 3 and 4. Emphasize the impact of the size of a forecast region for the consistency.

will be done accordingly

References

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.

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, Banff, Canada, pp. 91–97, 2014.