Response to reviewer RC2 (R. Engeset)

Frank Techel et al.

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

We greatly thank Rune Engeset for his very detailed review and helpful comments.

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

The manuscript https://doi.org/10.5194/nhess-2018-74 provides a significant contribution to the understanding of avalanche danger forecasting. It provides a thorough analysis of a large data set from the European Alps, and provides a method to compare the forecasted danger level within and between Avalanche Warning Services (AWS'). It is well written. The results and conclusions are relevant to the science community as well as the forecasting services. The results are presented and discussed in an adequate manner and the authors use a valid scientific approaches and methods. The manuscript is acceptable for publication in NHESS, when the recommends improvements are carried out.

1 General comments

The way the forecasting regions are reference to varies throughout the text. It will be easier for users to read the text if region names were used followed by the abbreviation in brackets. I recommend to use the following naming of region throughout the manuscript, example: “the regions Tirol (TIR) and Vorarlberg (VOR)”.

Please spell out names of countries, rather than use abbreviation in the text (e.g. P24L10).

The four research questions are formulated on page 3 and 4. However, it is hard to follow how the analysis, discussion and conclusions address these four questions. I recommend that Chapter 2-7 explicitly states how each of the research questions are addressed. This could be done by adding text, such as «In order to test/answer/address research question 1, we analysed the following: » or it could be restructuring the Sub-Chapters or adding Section headings. The conclusions should definitively address each of the four research questions in turn.

The authors should check the manuscript for consistency wrt. the use of "AWS" versus "forecasting centre". The authors should also check for consistency in the use of “warning region” versus “forecasting region”. For example, “warning regions” are used throughout the text, but not in Figs 6, 7 and 8, nor P3L23 or P8L11.

The spelling should be consistent. British spelling is used for some words, such as “neighbour”, while US spelling is used for others, such as “center” and “color”. Please check and correct the spelling.

There is no reference to the material in the Appendix in the text, this should be added or the Appendix skippet.

We will take up all of the before mentioned recommendations and edit the paper accordingly as described in our other two *responses (where the reviewers made similarly useful points which will improve the understandability of our paper..*

2 Specific comments

P1L7. Specify what is meant by "goodness". Is spatial homogeneity equivalent to a good forecast? Probably not, as regions are defined based on, among other characteristics, spatial differences in avalanche conditions. Thus, danger levels are expected to be different from one region to another from time to time. Furthermore, country or AWS-specific user may have developed strategies which account for potential differences between AWS' (in other words are calibrated), and a bias may be only a problem for users who are not familiar with the different products. This could be discussed.

No, with forecast *goodness*, we do not mean spatial homogeneity (P_{agree} = 1 (or 100%))), except in some very special situations. For instance, P_{agree} = 100% could theoretically be possible in cases where a mountain range with similar snow climate is cut by political borders. On the other hand, values of P_{agree} = 100% within the domain of a forecast center means that there is no need to have two warning regions instead of one. However, values of P_{agree} should be relatively similar regardless whether we explore within forecast center boundaries, or across those.

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 (40%) and for $D_{morning}$ 0.36 (36%) (using the distribution shown in Fig. 1a, we simulated 10'000 pairs).*
- *We will emphasize, that in most cases P_{agree} = 100% is not expected.*
- *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 by P_{agree} values of 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 its 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*

P2L26. Add a sentence about avalanche problems, such as "In 2017, EAWS introduced a set of five typical avalanche problems in order to both describe the avalanche hazard in more details and to provide better advice to the end users on how to manage these hazards.".

we will try to incorporate this information in the introduction.

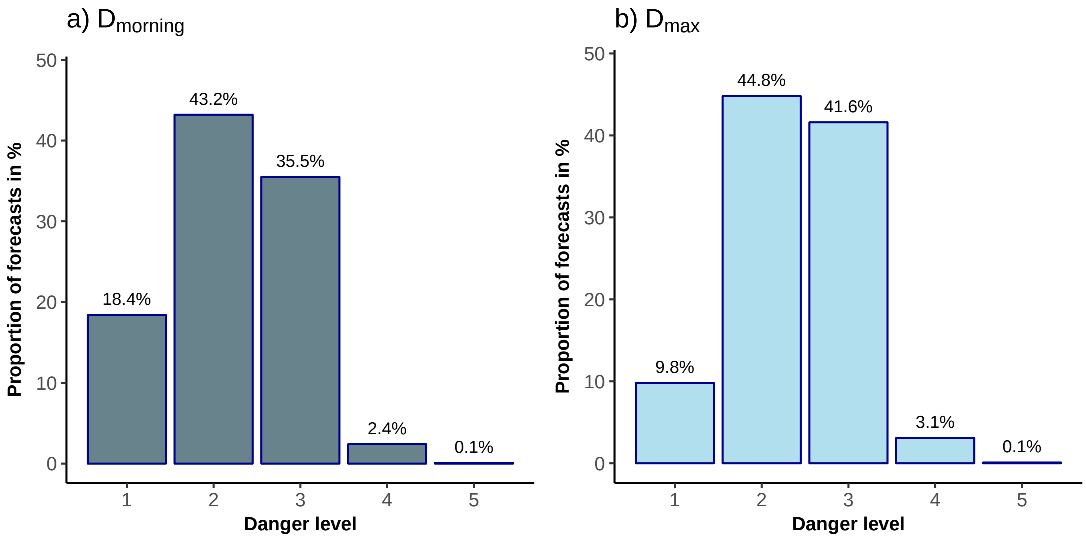

Figure 1. Distribution of forecast danger levels in our dataset, for (a) D_{morning} (danger level valid during first time-step) and (b) D_{max} (highest danger level).

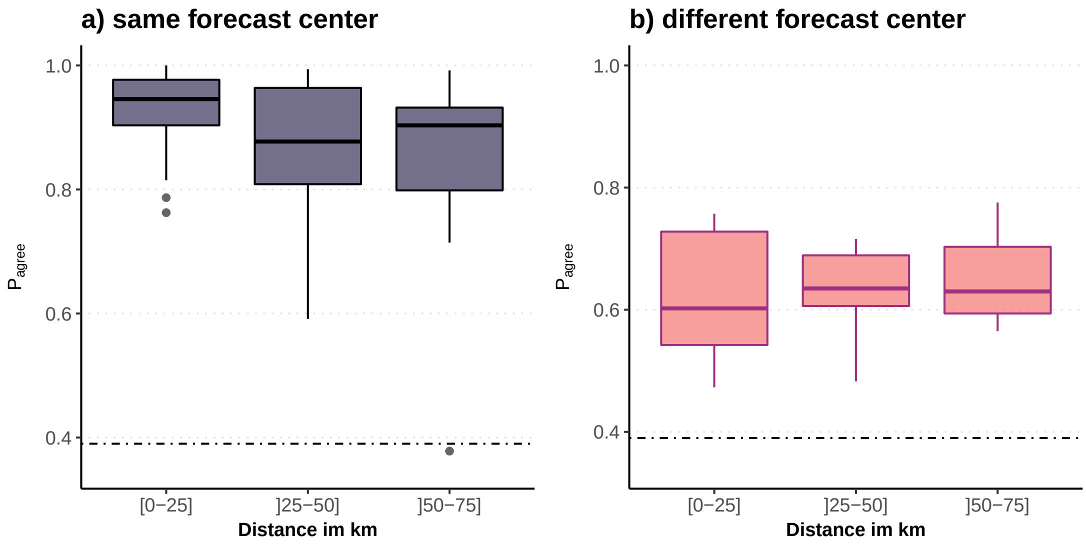

Figure 2. Boxplot showing the agreement rate for neighboring warning region pairs, with similar elevation (Δ elevation$<$ 250 m) and with similar size of the warning regions (the size of the larger warning region is less than 1.5 times the size of the smaller warning region). ($N_{\text{same forecast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

P2L27-28. Add a description of how the danger level is determined, and which factors are used to determine the danger level. Also specify how the level is determined, when the level varies with the spatial and temporal domain of the forecast (e.g., the forecast avalanche danger is the highest expect level in the forecasting time period and geographical region). Furthermore, the authors should provide a short description of possible or actual differences in procedures or practices. For example, avalanche size is an important input factor when the AWS decided which avalanche danger to forecast for a region. The avalanche size may be set differently according to differences in terrain, snow cover, training, culture, etc, and the current definitions of size categories may allow differences in interpretation. These factors should be briefly mentioned in the introduction, and be further elaborate on in Chapters 2 and 5 or 6.

P8L11. Add a description of how the danger level is derived/determined by the different AWS' and what are the contributing factors. For example, if one AWS systematically rate the avalanche size as 3 in cases where the neighbouring centre rates the size as 2, it will also systematically issue danger levels that are higher than its neighbour. Add this as a paragraph in Sub-Chapter 2.2 or as a Sub-Chapter on its own. This is important in order to understand why the danger levels may vary between regions or AWS'.

We would like to emphasize that we wanted to show whether differences in the use of the EADS exist, and whether operational constraints (like the size of the warning regions) or elevation may be able to explain these differences.

We have no information whether avalanche forecasters in different warning services weigh the contributing factors differently. Concerning the weight of contributing factors, we would also like to point out that there is very little found in the literature in that respect: one example is the case-study of four avalanche forecasters in Colorado (Armstrong et al., 1974; LaChapelle, 1980). Despite all of them weighing the input data differently, their overall forecast accuracy at the end of the season was similar. Lazar et al. (2016), who explored the consistency in the use of the North American Danger scale using a questionnaire and a set of ten scenarios, noted differences between countries. Lazar et al. indicated that they also asked forecasters regarding the main contributing factors. Unfortunately, they did not show these results in their paper. Following a personal communication with Brian Lazar, Brian allowed me to quote the following «We did ask forecasters to assign the top 3 environmental factors that most influenced their danger rating assignments. We conducted some initial explorations of the data, and they suggest that forecasters can and do: 1) weigh contributing factors differently while assigning the same danger rating and 2) weigh the factors the same but assign different danger ratings» (Lazar, 2018). In the Alps, we suspect some differences, but we cannot prove these. For instance, in spring-time situations some AWS may weigh the occurrence of natural avalanches more, regardless of size, while others may weigh the expected size of avalanches more. We believe this might be one explanation for the considerable differences noted in the proportion of forecasts with increasing danger levels (Table 7 in the manuscript: France and Switzerland about 15% of forecasts, Piemont (PIE) or Vorarlberg (VOR) about 25%).

Concerning the comment on the rating of avalanche sizes and their impact on the forecast danger level: Using a questionnaire and pictures of avalanches, Moner et al. (2013) explored the consistency of assigning size classes to avalanches. Moner et al. noted considerable variation in the size estimates, but also regarding which of the defining parameters was considered the most important for the size classification. However, only minor systematic differences by country could be observed.

We cannot show/discuss why differences exist. However, we will try to point out where the results indicate that conceptual differences may exist.

P2L22-P3L4. This part of the text should be improved, in order to explain specifically how this is interpreted and addressed in this study.

We will try to be more clear of how we address these points in the study.

P3L7-10. These statements should be explained and substantiated in a better way.

We will add an explanation in that respect.

P3L24-30. The main purpose or goal of the study should be more clearly stated. The current text ("This concept of consistency has in turn important implications for quality and ergo value. In our work, we assume that the quality of forecasting is consistent across all forecast centres, and rather consider the implications for the value of the forecast, as consumed by its users, as a result of potential differences in consistency. We do so by quantifying bias between neighbouring forecast centres and regions in time and space.") is complicated and somewhat hard to follow. What about something along these lines? "Biases in danger level between neighbouring warning regions and centres will decrease the value to users, unless biases are due to difference in avalanche conditions only. The main goal of this study is investigate if such spatial inconsistencies and biases exist, in order to improve the value provided by the European AWS"'.

We will take up this recommendation and rephrase this sentence to be more to the point.

P10Fig3. This map shows region sizes. Region elevation is the other statistics being analysed, I suggest adding a map Fig 3b, showing colour coding according to region elevation, if the elevation differences are possible to display clearly on a map. In this way, the map in Fig 5 may be easier to interpret wrt. elevation as well as size.

We already have such a map, although it was not included in the manuscript as we wanted to keep the manuscript short.

We will include a map as shown in Fig. 3.

P21. Justify why there is no Sub-Chapter on D=2.

There were two reasons: firstly, we wanted to 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 suggest to limit the main part of the paper to the sections on danger level 4 and 5 (currently Sect. 5.2) and danger level 1 (currently Sect. 5.4), as these describe the use of the upper and lower end of the danger scale. As a consequence, we would move the section regarding danger level 3(currently Sect. 5.3) to the appendix or as a supplement.

P23L32. Describe the procedures/practices at the different AWS and discuss if this a factor that causes systematic differences.

Frankly speaking, we don't know whether there are different practices, although we could imagine that both approaches are used. Furthermore, the avalanche danger scale is lacking this information. We are aware, however, that other approaches exist.

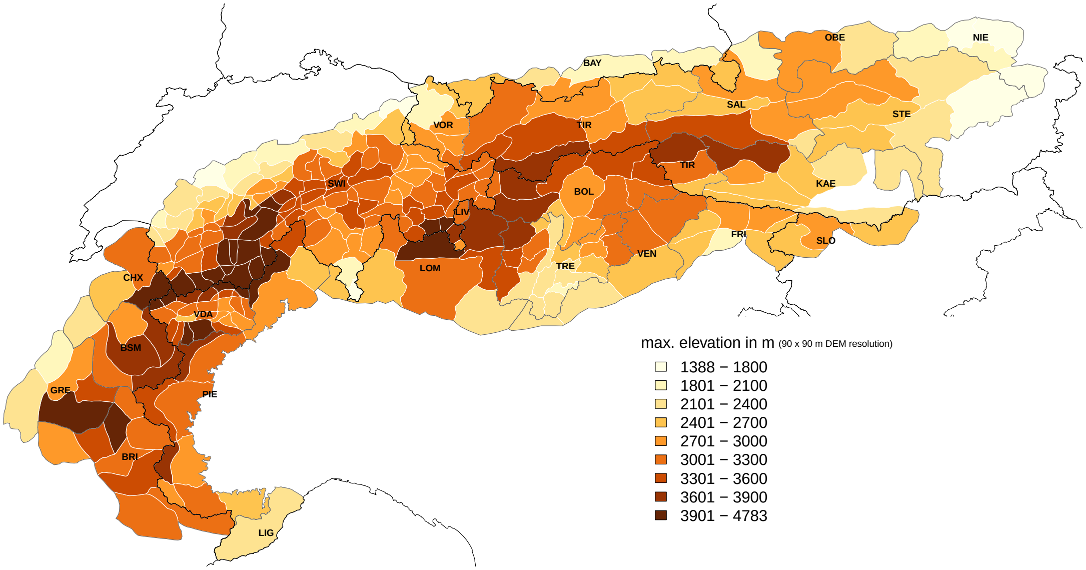

Figure 3. Maximum elevation in each warning region.

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 a so-called "allgemeine Gefahrenstufe" (mean general danger level), which did not always reflect the highest danger level within their forecast domain and forecast time, and which we have not analyzed.

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, 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 $D_{max} \geq 4$ ($P_{v.crit}$, see 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.)

P28L8. Consider to add "EAWS is also in the process of providing clear definitions of the key contributing factors, such as the distribution and likelihood.".

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

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

will be done

P28L26. Discuss what could be the effects of some forecasters or forecasting centres issuing the highest level expected in the forecasting region/period, while others may issue the most probable or general level.

will be done

P29L12. Consider to add “and/or typical avalanche problems” after “regimes”.

will be done

P30L16-19. Consider to specify in more details the why, what, and how of such a study.

We will be more explicit about this.

3 Technical comments

Thank you for pointing out typographical errors, and how we could improve the figures. Wherever possible, we will integrate these suggestions in the revised manuscript.

References

Armstrong, R., LaChapelle, E., Bovis, M., and Ives, J.: Development of methodology for evaluation and prediction of avalanche hazard in the San Juan mountain area of southwestern Colorado, Occasional paper No. 13, university of Colorado, Institute of Arctic and Alpine Research, 1974.

LaChapelle, E.: The fundamental process in conventional avalanche forecasting, Journal of Glaciology, 26, 75–84, 1980.

Lazar, B.: pers. communication, 2018.

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

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

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.