Nicolas Eckert
IGE, Grenoble Alpes University, INRAE, France

Grenoble, February 18th 2023

Submission of a revised version of our article to *NHESS* entitled “Development and evaluation of a method to identify potential release areas of snow avalanches based on watershed delineation

Dear Yves Bühler, NHESS scientific editor.

We deeply thank the referee for his/her feedback on our article and their insightful comments. We also thank you for your editorial revue and your suggestions. All points have been addressed in the revised version attached to this submission. In what follows, we further provide a point-by-point answer to all comments and questions and detail the changes made in the revised manuscript. We hope this revised version will be found suitable for publication in NHESS. Thank you for your consideration of our work.

Sincerely,

Nicolas Eckert, on behalf of the authors

Response to referee 1

Dear authors:
I appreciate the opportunity to review the revised version of your manuscript. It is obvious that you have put a lot of effort into the revisions, and the quality of the manuscript has improved substantially. I really appreciate the addition of the more in-depth discussion of the strengths and weaknesses of the various datasets and the parametric analyses. Despite these substantial improvements, I feel that there are several remaining issues that should be addressed before the manuscript can be published.
Authors' Response (A.R.): We deeply thank the referee for his/her positive judgement about our work and his/her meaningful additional suggestions and feedback. We hope our revised version will be found suitable for publication in NHESS.

Major comments
Description of steps of PRA decision method and CLPA procession
I think that the description of the steps of the PRA detection method could be improved by better aligning the description in the text with the graphic presented in Figure 3. Right now, I find the description rather confusing because it talks about three main steps that do not seem obvious in Fig. 3, and while I understand the reason for the split over the two columns, they do not obviously line up with the description in the text. Furthermore, the presentation of the CLPA processing steps in Fig. 4 is visually very different even though some of the steps are the same as in the PRA detection method. Given these similarities and the fact that the PRA detection and CLPA procession steps are closely tied (as explained in the text several times), I think that a more consistent graphic presentation that highlight these connections more obviously (either in a single or two figures) would allow the reader to understand the approach of the analysis more easily.
A.R.: Regarding the description of the method, we tried to rework it once more to make it more explicit. Notably, we removed the reference to "three main steps", which was indeed confusing, as these do not appear on Figure 3.
Regarding Figure 4, it includes both i) a very brief presentation of the data included within the CLPA, ii) how this data is processed to generate a validation sample for our PRA detection method. We reworked the caption of the figure to better underline this (previous caption that mentioned only the processing of the CLPA data was indeed confusing).

Confusion matrix
I am still concerned about the fact that you use the accuracy rate in your study. In reality, you are only looking at the precision/positive predictive value (= true positives/(true positives + false positive)), and your assumption that the true negative rate is 100% artificially produces an accuracy rate value that is halfway between the precision value and 100% without adding any value to the analysis. Similarly, this assumption also creates error rate values that are halfway between the false discovery rate (= false positives/(true positives + false positive)) and 0%. Given your new description of the strengths and weaknesses of the CLPA dataset (which is much appreciated), the assumption of a 100% negative predictive value and 0% false omission rate seems somewhat bold.
In my opinion, it would be more accurate and more transparent to base your evaluation on precision/positive predictive value instead of the accuracy rate. This will not affect the results of your analysis at all, but it will describe the focus of your evaluation more honestly and prevents possible confusion with accuracy rates presented in other studies that actually work with the full confusion matrix. Note that you explicitly point out this limitation yourself on L633. I think it would be very useful for you to highlight in the conclusion section that future studies should aim to assess PRA algorithms with the full confusion matrix.
A.R.: We agree with this comment and have reworked the full paper (including tables and supplements) to provide all results in terms of true positive rates (also known as recall) instead of accuracy rates. Only exception is the description of the confusion matrix for the area of Chamonix,

which is used to introduce the different terms of the matrix and the different scores in a pedagogic way.
By contrast, we did not mention that further research should focus on the full confusion matrix as we think that this is not feasible. As explained in text, we indeed believe that, even with the "best" data set of observed avalanche release areas at hand, one will never be sure that a false positive is simply not a release area or a fraction of a release area that has never been observed so far but could be triggered one day under very specific conditions, see our discussion section for further details.

Comparisons in parametric studies

I appreciate that you now explore the robustness of your approach with a parametric studies. However, I am a bit confused about the fact that parameter values and ranges were only changed in the PRA algorithm and not the validation dataset even though most of them are used the same way in both. It seems obvious that the PRA algorithm that uses the same parameter values as the CLPA processing will naturally perform the best! Applying a different slope or elevation filter in the PRA algorithm but keeping the default one for the CLPA processing obvious decreases the performance. I understand that this relates to the challenging task of defining the "ground truth" (which requires some assumptions), but it seems to me that potential insight from the current approach is limited.

Would it make more sense to also change the parameter values in the CLPA procession like you did for the DEM resolution analysis (L505). I have not completely thought this through, but it would keep the assumptions consistent and allow you to compare apples with apples and not apples with oranges.

A.R.: This is a tough question that we had in mind during the whole work, and we either do not have a definite answer to it. We chose not to recompute the validation sample at each time as, indeed, ground truth should be fixed, but we agree that this favours our "default setting" in the parametric study. For the DEM resolution, we performed both computations as it was a particularly critical point of the analysis for which our findings slightly differ from the state of the art. We could have done this also for all other analyses, but this would have largely increased the number of tables and scores to be analysed (which are already quite numerous). In addition, we do not think that this would have had a large benefit. Indeed, as discussed, our parametric search should not be seen as a way to determine a truly optimal combination of parameters. We do not think we have the data that would allow this. More modestly, our parametric study, as it is conducted, shows that our PRA detection method is, to a certain extent, rather robust over a certain range of parameters which is consistent with the state of the art. Also the DEM analysis example shows that, even when the validation sample is recomputed, the default setting may still be favoured (Table 7). These elements were somehow already present in the discussion of the previous version of the paper but we reworked it slightly to try to be even clearer.

Minor comments

L153: If the optimum DEM resolution is examined in the study, shouldn't all DEM datasets be described in the data section and not just the 25 m one?

A.R.: Indeed, we added in text the precision that the DEMs of finer resolution were also provided by IGN.

L192: I think it would be useful to explicitly explain why you trust the CLPA dataset so much instead of just stating it as a fact.

A.R.: As stated in our discussion, the trust comes from the CLPA long history, with regular updates by skilled and devoted technicians, continuous support by the French ministry of the environment and the inclusion of a large amount of different data sources, so as to be as close as possible to reality. The paragraph has been reformulated as: "Due to its long history, its regular update by devoted technicians, the continuous financial support of the French ministry of the environment and the consideration in the determination of avalanche terrain of a large amount of different data sources, CLPA is very reliable, meaning that an avalanche extent which is within the CLPA is almost surely a true avalanche extent".

L237: It is still a bit unclear how the watersheds are actually delineated. Figure S2 shows how the flow direction and accumulation are calculated but does not show how the actual boundaries are drawn. A slightly bigger example with the actual boundaries drawn would be more informative.
A.R.: To delineate watersheds, we used a standard algorithm well described in the literature. However, the procedure it is not very easy to explain in a few words and to represent within a single figure. As this is not the heart of our paper, we prefer providing the idea only and referring to the source papers. We have added the reference to Djokic and Ye (2000) for a seminal description of the watershed delineation procedure (which includes several illustrations).

L312: The fact that only one pixel of a validation PRA must be identified for a successful match seems a very low bar and a critical assumption of the analysis. It might be worthwhile to justify this choice in more detail and/or explore the effect of different thresholds.
A.R.: We agree that one pixel for a successful match can actually be seen as a "lower bound" (related accuracy is measured by our recall defined on PRA numbers). This is exactly why we also provide a kind of "upper bound" with the recall measured on PRA areas. Our discussion already mentioned that one single metric is certainly not enough to truly assess the efficiency of a detection method, so that we proposed two metrics that cover the most critical dimensions of the problem (PRA numbers and areas). We added to the discussion that, in the future, additional metrics should be considered, notably metrics that combine both information (e.g. successful match for different thresholds defined as minimal matching areas), and/or metrics related to various other characteristics of the detected PRAs (shape, elevation, etc.). This may help assessing even more precisely the strengths and weaknesses of our (or another) PRA detection method.

L540: I appreciate the honest discussion of the limitations of the performance measure here, but I think this could be addressed/avoided by using a more appropriate performance measure that takes the limitations of the dataset into account more honestly earlier (see earlier comment).
A.R.: See our response to the main comment about the choice of the performance measure. The whole text has been reworked accordingly.

L585: It would be better to include the suggestion for a full comparison of different PRA algorithms in the conclusion section where you make other suggestions about future research.
A.R.: This has been done.

L641: I did not read the paper by Giffard-Roisin et al. (2020) in detail, but I think it would be important to briefly mention that while there are benefits to increasing detection power, increasing false positives also has its cost/challenges.
A.R.: We added a note saying that doing so may indeed increase the number of false positives.

L651: It is not completely correct that you validated your PRA algorithm over entire massifs, because your performance measures are only based on the areas where CLPA data is available, which are fractions of the entire massifs.
A.R.: We reformulated as "covering significant proportions of three entire massifs with diverse characteristics"

L656: Are these suggestions meaningful/realistic given the inherent limitations of the CLPA dataset?
A.R.: Probably not all for the CLPA data, the reason why we wrote "and/or with different validation data". Indeed different validation data with different strengths and weaknesses (we doubt that any "perfect' data set may exist) may allow investigating these different issues. We precised as "A similar approach could be further used for comparing different PRA detection methods and/or in other contexts with different validation data having strengths and weaknesses different from those of the CLPA."

Response to editorial comments

The manuscript will require detailed editing as the English is still of limited quality. Below are some comments for improving the writing, but there are likely more issues. I assume that the NHESS editorial team will take care of this before the manuscript is published.
A.R.: We agree that the English of the paper was still improvable. In addition to suggested changes, we did our best to proofread the paper once more.

Abstract
L17-20: I think the performance measures and values used in this study need to be described more accurately in the abstract. See earlier comment on the performance measures.
A.R.: We have reformulated the sentence as: "Comparison to an extensive cadastre of past avalanche limits from different massifs of the French Alps used as ground truth leads to true positive rates (recall) between 80-87% in PRA numbers and 92.4% and 94% in PRA areas,…". See also our response to referee one about the choice of the performance measure.

Introduction
L 62: "Eventually" is not the right term here. You could say "finally" instead. There are many incorrect uses of "eventually" throughout the manuscript. Please replace throughout.
A.R.: This has been done.

L63: The last sentence in the paragraph (As a consequence, …), does not seem properly connected to the rest of the paragraph. Please expand and explain in more detail.
A.R.: We have reformulated the sentence as "Finally, PRA detection methods are primarily oriented towards large avalanches, which are of interest to assess long-term risk for people and settlements downslope, so that a minimal size is generally considered (e.g. Maggioni et al. 2002).

L70: Missing "and" before ii).
A.R.: This has been done.

L 75: Replace "is very dependent" with "depends".
A.R.: This has been done.

L 93: "confront" should be "compare".
A.R.: This has been done.

L101: Replace "summed-up as" with "summarized in". "Summed-up" is used in several locations of the manuscript and should be replaced everywhere.
A.R.: This has been done.

L103: Replace "remain little used so far" with "have only seem limited use so far."
A.R.: This has been done.

L115: Replace "ground" with "build".
A.R.: This has been done.

Data
L150 – Table 1: First, this table seems to include results already. This is rather unusual for a table in the methods/data section. Second, the areas are not explicitly introduced in the text. Their purpose is mentioned on L 139 in general, but the actual areas are not described.
A.R.: For us a Table (or a figure) does not belong to a specific section, it is just located at the place where it is called first in text. Actually, this table is called at several places in text, including in the data,

methods and results sections, so it is logical that it includes information relevant for the case study presentation and for the application of our method. Another solution would have been to split the table in several tables possibly located closer to their use in text, but this would have enlarged the number of tables, which is already high, so that we think it that our solution is sensible.
Regarding the small study areas, we have added the following sentences to introduce them in text (with reference to Figure S1 in the SM where they are mapped): "The Chamonix area is a 34.3 km2 area, which is part of the Mont Blanc massif and includes the municipality of Chamonix Mont Blanc. The Chartreuse / Dent de Crolles area is an even smaller area (7.6 km2) located within the Chartreuse massif and with the Dent de Crolles (2,062 m a.s.l.) in its center (Figure S1 in the SM).

L197: "Avalanche extensions", which is used extensively throughout the manuscript is not the right term. In this particular case, "avalanche records" would work, but most often it refers to the "accuracy of the recorded extent of observed avalanches." Please correct this throughout the manuscript.
A.R.: CLPA really represents avalanche maximal extents and not avalanche records. Notably, it does not include, e.g., the dates and the characteristics of single avalanche events, and even not the contours/extents of individual avalanche events. This is written in the paper, but we understand that it may be confusing for people from countries where habits regarding avalanche data are different. Also, the term "avalanche extension" is the one officially used in the official CLPA caption, and it is very important for us to be precise and consistent from this perspective. Yet, we understand that our formulation was not fully correct from the point of view of the English language. We reworked the paper in order to reach an acceptable compromise, namely we now use "avalanche extent" throughout the text but we kept "avalanche extension" in the captions within the figures when it is necessary (Figs. 4-8), with a note of explanation in the expanded Figure description (the expanded caption below the same figures).

PRA detection
L 236: Delete "(where flow accumulation is equal to zero)" as it is repetitive.
A.R.: This has been done.

L 249: Replace "few" with "too little".
A.R.: This has been done.

L 260: "e.g.," should probably be "i.e.,"
A.R.: This has been done.

Results
L 340 – Caption of Fig. 5: Replace "concordance" with "agreement".
A.R.: This has been done.

Discussion
L646: It seems inaccurate to mention the confusion matrix here since you did not use the full confusion matrix. Instead, you should more strongly highlight that you examined the performance with respect to area and number of" PRAs, which is more novel.
A.R.: The sentence "Finally, confusion matrices and performance criteria were seldom used so far to evaluate PRA detection methods" does not refer to our work and is for us fine. Instead, we specified more clearly what we did in the next sentence as "As a first step towards improved evaluation schemes for PRA detection methods, we proposed to evaluate efficiency with true positive rates (recall) computed both for PRA numbers and areas, which may cover the two most critical dimensions of the problem".

Conclusion
L664: It is unclear to me what you mean with "and close contexts (see below)."

A.R.: We reformulated as “mountain environments with similar characteristics”.

L668: You should explicitly explain how your results contribute to the field and not leave this up to the reader. They might not see it themselves.
A.R.: Our sentence was devoted to refer to the findings that where listed just before. We reformulated as “outcomes of the work include i) the determination of individual PRAs using a watershed delineation algorithm, ii) an approach to define a validation sample from a cadaster of avalanche extents, iii) an evaluation procedure based on two metrics, PRA numbers and area, and iv) a better definition of accuracy scores that should be interpreted in the context of PRA identification. These methodological developments should help progressing towards more efficient approaches for PRA detection and evaluation”. We hope it is clear like this.

L680: Replace “confronted” with “compared” or “contrasted”.
A.R.: We replaced by “compared”.