

## Predicting Avalanche Danger in Northern Norway Using Statistical Models (egusphere-2024-2865)

This paper presents several machine learning models for predicting avalanche danger levels in Norway. The study addresses a relevant topic in avalanche forecasting and model development, particularly in regions with limited data, which is significantly challenging for developing models. The authors compare two types of machine learning models, a Random Forest model and an Artificial Neural Network and report some variations in their performance. I recommend publishing this paper after the following comments and suggestions.

Some parts of the manuscript could be shortened or moved to the appendix, while more interesting results could be incorporated into the main text. The model was designed to predict danger levels for both dry- and wet-snow avalanche conditions. Although the dataset is small, I suggest developing and testing models exclusively for dry or wet snow conditions or developing two separate models, as the drivers of each avalanche type differ significantly. Furthermore, justifying the development of a binary model by merging danger levels only based on better performance is insufficient. Even if the performance is lower with four danger levels, this approach is more realistic and a well-defined target in avalanche forecasting. The train/test split and the cross-validation method used for the optimization and development of the models can be better clarified. Below, I provide more detailed comments for each section.

We thank the reviewer for considering our manuscript and for the detailed review. We provide a point-for-point response to the individual comments below. However, we want to respond to some of the comments made here already. The comparison of different machine-learning models was not a main point of our study. We focus mainly on the random forest model and only add the results of the neural network in the supplementary information to show that they are not significantly different from the RF results. Regarding the point about wet and dry snow, we note that at the time of writing the distinction between the different avalanche problems was not available to us, and we still struggle to have this data available. Furthermore, the data availability for individual avalanche problems appears rather small so that it will be highly challenging to generate individual models for the individual problems or even the split between dry and wet avalanches. We would like to turn to this in future work, when we have robust data available.

Regarding the binary model, we explicitly state in lines 189-191 that our reasons for developing this model are not *only* because it performs better, but in fact mainly because we want to use it to generate a hindcast to obtain a measure (rough estimate) of the avalanche activity, linking it to climate modes. We also want to use this model in future work to study the general trend in avalanche activity in future climate projections (see lines 513-521). One could argue that this could be done with the danger levels (i.e., the 4-level model) themselves, but the accuracy of the 4-level model appears too low for obtaining robust results. Moreover, at the ISSW2024 we became aware of two studies that appear to strengthen the argument for using our binary case as a measure for avalanche activity. We reformulate lines 188-191 to make this clearer (note that in response to a later comment we have renamed the “binary-case avalanche activity” to “binary-case frequency”, to distinguish it more clearly from avalanche activity in the sense of actual avalanche occurrence):

“In this study, we consider two types of avalanche-danger scales. First, we employ the full ADL scale (henceforth ‘4-level case’). Second, we generate a binary scale (henceforth ‘binary case’) where the ADLs 1 and 2 are combined to binary-case level (BCL) 0 and the ADLs 3 and 4 are combined to BCL 1. We refer to the number of BCL-1 days per season as the “binary-case frequency” (BCF). Due to its higher accuracy compared to the 4-level case (see section 5), the binary-case model will give a more robust, albeit rougher, estimate of the general tendency of avalanche danger. Furthermore, the BCF appears related to avalanche activity, since, e.g., Perez-Guillen et al. (2024a) in a case study in the Swiss

Alps using an automated seismic avalanche detection system found that on days with no avalanche the mean ADL was  $1.9 \pm 0.8$  while on days with at least one avalanche it was  $3.2 \pm 0.5$ , hence providing a clearly binary appearance. Similarly, in an investigation of Swiss backcountry GPX tracks as a proxy for non-avalanche events, Techel et al. (2024) found that for non-events the median probability of  $ADL \geq 3$  was only 0.14 while for events it was 0.58. Hence, we interpret the BCF as a measure of avalanche activity. We use the binary case in a hindcast for a rough estimate of changing avalanche activity over time and to find potential connections to known climate patterns/modes (see section 6).”

New references:

Peréz-Guillén, C., Simeon, A., Techel, F., Volpi, M., Sovilla, B., and van Herwijnen, A.: Integrating automated avalanche detections for validating and explaining avalanche forecast models, International Snow Science Workshop Proceedings 2024, Tromsø, Norway, pp. 52–57, <https://doi.org/http://arc.lib.montana.edu/snow-science/item/3111>, 2024a.

Techel, F., Helfenstein, A., Mayer, S., Peréz-Guillén, C., Purves, R., Ruesch, M., Schudlach, G., Soland, K., and Winkler, K.: Human vs. machine - Comparing model predictions and human forecasts of avalanche danger and snow stability in the Swiss Alps, International Snow Science Workshop Proceedings 2024, Tromsø, Norway, pp. 31–38, <http://arc.lib.montana.edu/snow-science/item/3108>, 2024.

Please also note that in some cases the line numbers given by the reviewer seem incorrect so that we had to speculate which exact parts of our manuscript were referred to.

Finally, the reviewer asks us to add a Discussion section and to move some of the parts of our article from the sections where they now occur into this section. While we appreciate and understand this suggestion, we are averse to this and prefer the current structure. We prefer to discuss important points where they arise. For example, when we describe in lines 174–185 why we do not have a “tidy” dataset as well as why we do not use remote sensing data, we think it more appropriate to mention this already in the description of the data instead of in an extra section after the results (as suggested by the reviewer below). These lines are rather a justification or extended explanation for using specifically these data and are otherwise unrelated to the results, thus being, in our opinion, unfitting for a discussion.

## Introduction

The introduction should be more concise, re-structured and focused on the main topic: the development of machine learning models for avalanche forecasting in Norway. Overall, the writing should follow a clear structure, avoiding mixing concepts and ensuring it reads like an introduction rather than a discussion or conclusion. Also, it should highlight existing knowledge gaps or limitations of current models, linking them to the relevance of this study.

Specific comments:

- The first paragraph is too general, and part of the content is not very relevant to the paper’s topic.

We agree that this paragraph is quite general, but our point here is that a similar methodology can be used to predict different climate indicators based on meteorological data. We believe it is helpful for a reader not closely familiar with the concepts of avalanche prediction and the capabilities of machine learning in the respect to understand the basic principles on which our work is based. Thus, we would like to retain the introductory paragraph as is.

- Line 43: it is not a number but a scale.

Our point here is that on a given day for a given region there is published *one* avalanche danger level, which is a “single integer”, as we write in the manuscript. We acknowledge that avalanche danger is evaluated on a scale, but the end result, that is, the aggregation of information in specific situations as danger *level*, is a single integer.

- Line 44 and 45: the European and American scales have some differences.

We thank the reviewer for pointing this out and will add the following to our manuscript:

“However, the North American scale was later revised with a focus on risk communication (Statham et al., 2010).”

We have also corrected the year of the adoption of the danger scale in North America, which was 1994 (and not 1997, as we incorrectly stated), according to Statham et al. (2010).

References:

Statham, G., Haegeli, P., Birkeland, K. W., Greene, E., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J.: The North American public avalanche danger scale, Proceedings of the 2010 International Snow Science Workshop, Squaw Valley, CA, pp. 117–123, <http://arc.lib.montana.edu/snow-science/item/353>, 2010.

- Line 52: The danger level model has been operationally integrated into avalanche forecasting in Switzerland since the winter season 2021-2022 (1).

We thank the reviewer for this clarification. We have changed this with reference to the suggested article. We were increasingly confused since there seem different systems in place or in development (see Maissen et al., 2024, and Winkler et al., 2024) and, e.g., van Herwijnen et al. (2023, p. 323) write: “Since the winter season 2020-2021, the Swiss avalanche forecasting service at SLF started using the models described above in operational context.”

References

Maissen, A., Techel, F., and Volpi, M.: A three-stage model pipeline predicting regional avalanche danger in Switzerland (RAvaFcast v1.0.0): a decision-support tool for operational avalanche forecasting, EGUSphere [preprint], <https://doi.org/10.5194/egusphere-2023-2948>, 2024.

van Herwijnen, A., Mayer, S., Perez-Guillen, C., Techel, F., Hendrick, M., and Schweizer, J.: Date-driven models used in operational avalanche forecasting in Switzerland, International Snow Science Workshop Proceedings 2023, Bend, Oregon, pp. 321–326, <http://arc.lib.montana.edu/snow-science/item/2895>, 2023.

Winkler, K., Trachsel, J., Knerr, J., Niederer, U., Weiss, G., Ruesch, M., Techel, F.: SAFE – a layer-based avalanche forecast editor for better integration of model predictions, International Snow Science Workshop Proceedings 2024, Tromsø, Norway, pp. 124–131, <http://arc.lib.montana.edu/snow-science/item/3123>, 2024.

- Lines 52-103: This paragraph is very long and lacks clarity regarding its main topic (e.g., avalanche prediction or avalanche danger level forecasting). Additionally, some explanations fit better for the discussion section.

In our opinion, the discussion of previous similar research should (or at least can) be part of an introduction to a scientific paper and is similarly done in many of the papers we cite in this section

(e.g., Schirmer et al., 2009; Perez-Guillen et al., 2022). We agree that the paragraph in our introduction is somewhat longer than in these earlier studies, but this may mostly reflect the recent increase in research interest in this topic. Also, we are unsure as to the lack of clarity of the paragraph regarding its main topic. The whole point of the paragraph is a review of earlier studies with a very similar aim to ours, i.e., the development of a statistical model to predict avalanche danger level based on meteorological and (potentially) snow-pack data. Moreover, one of the arguments of the editor (Jürg Schweizer) for rejecting a first version of our manuscript was the lack of discussion of previous research. Finally, Reviewer #1 appeared to support our efforts to review previous research and even suggested more references that we should add to this review (this was a major comment of Reviewer #1). We would thus like to keep this discussion of previous research as is, however now including the slight changes we made in response to Reviewer #1.

- Lines 106-110: This should be the motivation of this study and moved above in the introduction.

Since we lead up to this with our more general introduction at the beginning to make the point clear we would like to keep this as is.

- Lines 95: Sections should be referenced in the order in which they appear in the paper.

The structure of the article is given later at the end of the Introduction. Here we were just referencing a specific point (i.e., the selection of the test data), which seem important to us and about which we had more to say later in the article. However, we have removed the cross-reference.

- Lines 106-110: These paragraphs read more like a summary of results mixed with discussion and conclusions rather than an introduction to the content of the paper.

We believe that something went wrong with the given line numbers here as these lines were already addressed above and the reviewer's comment does not fit them. We instead believe the reviewer is referring to lines 122-137. In response to Reviewer #1 we had already shortened these lines to:

“We optimise two different RF models: (1) for the original ‘4-level case’ (ADL 5 has not been forecast in northern Norway) and (2) for a ‘binary case’, where ADLs 1 and 2 and ADLs 3 and 4 are combined. The latter is applied to obtain a hindcast of ‘avalanche activity’ for 1970-2024 and to investigate the linkage between avalanches and regional climate modes, such as the North Atlantic Oscillation (NAO) or the Arctic Oscillation (AO). Our findings have potential implications for the seasonal predictability of avalanche activity and danger, which is a salient point as only a few studies have previously investigated connections between avalanches and regional climate modes.”

We include this very brief summary and the point about the linkage to climate modes as we believe that this topic has received too little research attention and has potential for seasonal avalanche prediction.

## **Data**

Since the danger levels forecast in the public bulletin are used as the target for the machine learning classifiers, the description of the avalanche forecast in Norway should be more detailed, adding the correspondence references. How often is the public avalanche forecast issued in Norway? Is it updated daily? At what time is the bulletin issued, and what is the validity period of the forecasting window? Are the danger levels provided based on the European 5-level Avalanche Danger Scale (2) and the descriptions of the avalanche problems (3)? Given that the meteorological and snowpack factors driving the formation of dry and wet snow avalanches differ, it is important to clarify whether the models have been specifically developed to predict danger levels for dry or wet snow conditions.

Additionally, how are these two types of avalanches forecast in Norway? How are the critical elevation and aspects where the danger level is applied forecast in the Norwegian bulletin? Furthermore, how are the data sets prepared and merged for model development to address these distinctions? I am unsure about the reasons for developing a binary classification model by merging danger levels 1 and 2. The justification that the model performs better should not drive the development of a model with a new target variable, i.e. a new danger level scale. Furthermore, the definitions of danger levels 1 and 2 do not imply an absence of avalanche activity and depend on avalanche size (2; 4). I suggest focusing on the results of the multiclass models.

We agree with the reviewer that our description of the avalanche warnings in Norway was too vague and have now changed the paragraph in our section 2.1 to the following:

“In Norway the ADL assessment is produced under the scope of the Norwegian Avalanche Warning Service (NAWS) which was established in January 2013 (Engeset, 2013; Müller et al., 2013; Engeset et al., 2018b). The NAWS is a member of the EAWS and the ADL assessment follows the EAWS standards (Engeset, 2013). The ADLs are generated and published<sup>2</sup> by a team of experts from the Norwegian Water Resources and Energy Directorate (NVE), the Norwegian Meteorological Institute (MET), and the Norwegian Public Roads Administration (NPRA) aggregating knowledge from snow and weather observations as well as numerical weather prediction modelling. For mainland Norway (i.e., excluding Svalbard), avalanche warnings are published daily from 1st of December to 31st of May<sup>3</sup> for 23 warning regions with an average size of about 9000 km<sup>2</sup>. For 19 further warning regions (average size about 11000 km<sup>2</sup>) avalanche warnings are published on days with ADL 4 or 5. See Table S4 and Fig. S6 in the supplementary material for more detail on the Norwegian warning regions. The avalanche warnings are published before 16:00<sup>4</sup> for three days at a time, with a nowcast for the day of production and a forecast for the next two days (Engeset, 2013). We here use the nowcast data available via NVE’s Regobs platform (Engeset et al., 2018a)<sup>5</sup>, which is conveniently accessible with the Python library Regobslib<sup>6</sup>. Even though the NAWS has published ADLs since 2013, we here use ADL data from the avalanche seasons of only 2016/17 to 2023/24, since the warning-region setup was changed in 2016 (K. Müller, personal communication).

In describing the avalanche danger by a single value per region, the ADL constitutes a large reduction in complexity. In fact, the avalanche forecaster considers several different “avalanche problems” (APs). The NAWS follows the EAWS’s standards, using the following APs<sup>7</sup>: new snow (loose and slab), wind drifted snow (slab), persistent weak layer (slab), wet snow (loose and slab), and gliding snow. Based on the estimated likelihood (based, in turn, on distribution and sensitivity) and size of avalanches the forecaster determines a danger level per AP (Müller et al., 2023). The final ADL in a given region is taken as the highest danger level among the different APs. Hence, the ADL is a result of different APs that are related to different meteorological conditions, complicating the relation between ADL and meteorological data, and thus the modelling of this relation. However, considering only one AP reduces the amount of available data, making a robust training of statistical models more difficult. Also, at least some of the APs may be related to similar meteorological conditions and we thus believe it is still feasible to focus on the general ADL. In future work we will attempt a more detailed decomposition into the different APs.

Footnotes:

2. The Norwegian ADLs are published at <https://varsom.no> (see Johnsen, 2013; Engeset et al., 2018a).
3. See <https://www.varsom.no/en/avalanches/ski-touring-in-norway-important-information/>, last visited 27.11.2024. However, note that in special cases avalanche warnings are sometimes published also in November and June.
4. <https://www.varsom.no/en/avalanches/avalanche-warnings/>, last access 27.11.2024. However, note that on days with ADL 4 or 5, warnings are typically published already before 12:00 (Engeset, 2013).

5. <https://api.nve.no/doc/regobs/>, last access 27.11.2024.
6. <https://pypi.org/project/regobslib/>, last access 27.11.2024.
7. See <https://www.avalanches.org/standards/avalanche-problems/>, last access 27.11.2024.”

#### New references:

- Engeset, R. V.: National Avalanche Warning Service for Norway – established 2013, International Snow Science Workshop Grenoble – Chamonix Mont-Blanc, 2013, pp. 301-310, <http://arc.lib.montana.edu/snow-science/item/1853>, 2013.
- Engeset, R. V., Ekker, R., Humstad, T., and Landrø, M.: Varsom : Regobs – A common real-time picture of the hazard situation shared by mobile information technology, Proceedings, International Snow Science Workshop Grenoble, Innsbruck, Austria, 2018, pp. 1573-1577, <http://arc.lib.montana.edu/snow-science/item/2822>, 2018a.
- Engeset, R. V., Pfuhl, G., Landrø, M., Mannberg, A., and Hetland, A.: Communication public avalanche warnings - what works?, Nat. Hazards Earth Syst. Sci., pp. 2537-2559, <https://doi.org/10.5194/nhess-18-2537-2018>, 2018b.
- Johnsen, E. R.: Modern forms of communication avalanche danger – A Norwegian case, International Snow Science Workshop Grenoble – Chamonix Mont-Blanc, 2013, pp. 423-427, <http://arc.lib.montana.edu/snow-science/item/1829>, 2013.
- Müller, K., Kosberg, S., Landrø, M., and Engeset, R. V.: Report from the first operational winter of the Norwegian Avalanche Centre, International Snow Science Workshop Grenoble – Chamonix Mont-Blanc, 2013, pp. 311-315, <http://arc.lib.montana.edu/snow-science/item/1854>, 2013.

We hope this makes it more clear how and when the danger levels are issued, and specifically that we use the general danger level which combines *all* avalanche problems and does not distinguish between wet and dry. As we have stated several times in the manuscript, a more detailed distinction between avalanche problems is beyond the scope of the current study and left for future work.

At the time of writing, the aggregated information about the critical elevation level and aspects were not available to us and we thus did not consider them.

Regarding the last point about the binary-case levels please see our response to the general comment.

#### Specific comments:

- Lines 149-152: These initial sentences would be more appropriate in the introduction.

The points made in these lines are already included in the introduction (see lines 110-112). They are here used again as an introductory statement to this section, reminding the reader of the reason for our focus on avalanche danger instead of avalanche activity/occurrence.

- Lines 162-165: This explanation should be in the discussion or outlook for future model implementations. Are you also considering a wet snow problem?

We thank the reviewer for pointing out our lack of clarity here, although we note that we discuss this again in the Conclusion section (lines 530-536). Please see our reformulation of this paragraph quoted above.

- Figure 1: I suggest adding here or in a different Figure an example of a danger level map from the public avalanche bulletin issued in the study region.

The danger level maps as published in the bulletin do not lend themselves to be depicted as a figure due their format, but they can be readily accessed by clicking on the link to [varsom.no](https://varsom.no). Moreover, we prefer a figure that gives the topographical information implemented in the NORA3 reanalysis.

- Since the models were developed by merging data from different regions and winter seasons, I suggest modifying this figure to display bar plots showing the distribution of danger levels in the training and test sets. This would provide a clearer visualization of the data volume, the frequency of each danger level, and the proportions used for training and testing.

We suppose the reviewer here refers to Fig. 2. We appreciate the suggestion and have added a bar plot as a third panel to this figure which shows the distribution of the danger levels in the training and test data sets.

- Lines 174-185: This should be moved to the discussion section rather than being included in the description of the data used.

Please see our response to the general comment. We note again that in our opinion this does not fit a discussion section as it does not pertain to the results. It is rather an extended explanation of how and why our data are not as detailed as the data in other studies.

- Lines 202: Please specify the exact present date.

The reason we did not specify the present date is that these data are constantly updated. By the time of writing the most recent available data were for May 2024. Now the most recent data are for August 2024. We have changed these lines to:

“At the time of writing the data availability covers the period January 1970 to August 2024 and is constantly updated with a few months lag.”

- Lines 203-211: this reads more like a discussion than a description of the data used in this study.

It appears natural to us to describe these characteristics of the data in this section rather than in an extra section after the results. It is again somewhat of an extended explanation and justification for using these data.

- Lines 211-214: The values from all the grid points within the elevation band of 400 to 900 meters are averaged per region, is this the elevation band where avalanches usually release? Why are grid points at elevations higher than 900 meters discarded? Are avalanches not released from these higher elevations? When merging the forecast data with the model’s input features, how do you account for the elevation limit of the danger level forecast and the slope aspect? Have you tested averaging grid stations in micro-regions to achieve higher-resolution data and more data for the development of the model? Having more details about the different processes test with the elevation bands would be interesting.

The grid points higher than 900 meters are indeed discarded. Our reasoning for discarding the grid cells outside the given elevation bands is that when averaging over too many grid cells we may average out important variation that the machine-learning model may use to predict ADL. We admit



that our choice in elevation band is somewhat arbitrary, but please note that we have considered different elevation bands with little to changes to the results (see lines 211-214).

Moreover, please consider that the elevations across and within the different regions (note that their average size is about 6800 km<sup>2</sup>; see Fig. S6 and Table S2 in the new supplementary information) are rather different and this is why it is difficult to say at which altitudes avalanches usually release. We have added the point about the large region sizes in Norway to our manuscript. At the time of writing the data regarding the elevation limit and the slope aspect were not available to us, so they are not accounted for. This may be part of future work. We are unsure what the reviewer means by “averaging grid stations in micro-regions to achieve higher-resolution data”. This appears inapplicable to the Norwegian case. Similarly, we are unsure what is meant by: “Having more details about the different processes test with the elevation bands would be interesting.”

- Lines 215-225: More detailed information about the output variables of this model, as well as a visual example of the simulation, would be useful.

We had already rewritten this section in response to Reviewer #1. We have further rewritten the section in response to this comment. Note that we do not include a visual example of the simulation but instead provide a link to the seNorge website where the official Norwegian snow maps (as simulated by seNorge) are published. Rewritten section 2.3:

#### “2.3 seNorge

The NORA3 reanalysis provides no data on the snow conditions at the surface. Thus, in order to obtain information about, e.g., the snow depth and density and snow water equivalent (SWE), we employ the snow model seNorge (Saloranta, 2012) version 1.1.1 (Saloranta, 2014, 2016). Due to a lack of both in-situ and satellite observational data on snow, seNorge is the main tool used to provide information on snow for the avalanche warning system in Norway (Saloranta, 2012; Morin et al., 2020). Daily gridded (1-km resolution) snow maps are generated with seNorge and published on <https://www.senorge.no/>.

The tool seNorge is a simple process-based single-layer snowpack model demanding little computational resources, thus being convenient for application to large high-resolution grids (Saloranta, 2016; Morin et al., 2020). The model consists of two sub-modules for (1) snowpack water balance and (2) snow compaction and density, calculating the snow water equivalent, the melt/refreeze rate, and run-off as well as snow depth and density, respectively. As input data the seNorge only requires daily temperature and precipitation.

To keep our snow and weather data consistent, we rerun the seNorge model using NORA3 daily 2-m air temperature and total precipitation amount as input. To obtain reasonable initialisation data for seNorge, the model was first run for the years 1970 through 1975 with the initial values being zero everywhere. The final simulation outputs from 1975 were then used as model initialisation data for 1970, and the model was run from 1970 through 2024 to produce the snow-cover data.”

- Lines 226-231: This should be in the discussion/outlook of the paper rather than here.

We are thankful for this suggestion and note that we already in response to Reviewer #1 removed this part here and now only discuss these points in the Conclusion section.

- Lines 234-253: Please specify how you are resampling the meteorological time series to match with the forecast window of the danger levels.

This information was supposed to be given in lines 235-237:

“An overview of these potential predictors is presented in Table 2. They include the accumulated new liquid precipitation  $r_1$  on the day of the ADL assessment, as well as the new liquid precipitation accumulated over one to six days before and including the day ( $r_2$ , ...,  $r_7$ ).”



However, given the lack of precision in our description of the publication of the Norwegian avalanche warnings we realise that this may be somewhat unclear as well. In our response to the comment regarding this above we now have added a much more detailed description of the publication of the warnings, specifically the “nowcast”, which we use as our danger level data. We have accordingly rewritten the information about the sampling of the time series as follows:

“An overview of these potential predictors is presented in Table 2. They include the accumulated new liquid precipitation  $r_1$  on the day of publication of the ADL nowcast (see section 2.1), as well as the new liquid precipitation accumulated during one to six days before and including the day of the nowcast ( $r_2$ , ...,  $r_7$ ).”

· Lines 254-253: Please show in a plot or specify in the text the distribution of danger levels and data between the train and test. Ideally, the danger level distributions of test and train should be similar.

The distributions of training and test data are already shown in Fig. 2. As we wrote in line 256, the full seasons of 2020/21 and 2022/23 are chosen as test data while the remainder is used as training data. The distributions of all seasons (including the test data) are shown in Fig. 2 separately. However, we have changed Fig. 2 to now include an extra panel showing the distribution of the training and test data specifically.

## Methods

Since Random Forest is a widely used machine learning algorithm, I recommend moving Sections 3.1 and 3.2, along with Tables 3 and 4, to the appendix to save space. It would be more relevant to specify the programming language and Random Forest implementation used and whether the data and model are open source and available.

Please note that the requested information is to be given in the “Code and data availability” section. However, in the manuscript we mistakenly wrote “decision tree” instead of “random forest” in this section.

We have now added in the Methods section that we use the implementation of random forest in the Python library scikit-learn version 1.5.1.

Furthermore, we have moved section 3.2 as well as Tables 3 and 4 into the appendix as suggested. However, we would like to keep section 3.1 as is. We agree with the reviewer that the random forest method is widely known, but, e.g., the authors themselves were not closely familiar with it before this work. Accordingly, we believe that a brief description of the functionality of the random forest model in the methods section as presented in our manuscript is helpful for the understanding of the optimisation procedure as well as the results. We are furthermore encouraged in our opinion as Reviewer #1 appeared to appreciate our description of the methods.

Specific comments:

· Lines 310-316: The train/test split and the cross-validation method for optimization are unclear. According to Line 257 (Section 2.3), two winter seasons are used as the test set. For model optimization, you used the six available winters, applying cross-validation. Does this mean the test set is also being used for model optimization? Ideally, the optimization and cross-validation process should be applied exclusively to the training data, allowing the model’s performance to be evaluated independently on the test set. Additionally, it would be helpful to provide more details about the SMOTE oversampling technique used.

Please note that these lines are not part of the “Methods” section, but part of the “Random forest optimisation and feature selection” section.

We are thankful to the reviewer for pointing out the lack of clarity here.

We indeed used the whole dataset (six seasons) during the model optimisation (i.e., also in the cross-validation during the grid search) and only later, when testing the accuracy of the final model do we split the data in the training (2018, 2019, 2020, 2022) and test (2021, 2023) set. Essentially, for the section “Random forest optimisation and feature selection” (section 4) we use the whole six-season dataset, while in the section “Model evaluation” (section 5) we use the above-given split into training and test data. We thought this is appropriate since we only had a few seasons of data available. However, since we now also have the 2023/24 season available and, furthermore, the data from 2016/17 became available to us, we have changed the procedure so that the test data (still 2021 and 2023) are no longer included in the optimisation process. We have updated the random forest optimisation section (section 4) accordingly.

Note that this required us to re-do the whole analysis, applying the new models trained on years 2017, 2018, 2019, 2020, 2022, 2024 and tested (as before) on years 2021 and 2023. We have also updated all figures and tables. Importantly however, this only led to minor changes in our results and our discussion and conclusion remain essentially unaltered.

Finally, we have added a brief section (3.2) detailing the SMOTE oversampling technique:

### “3.2 Class balancing – Synthetic minority oversampling

Since our avalanche danger data are highly imbalanced, i.e., the different ADLs have different frequencies (section 2.1, Fig. 2), we employ the widely used (e.g., García et al., 2016; Fernández et al., 2018) synthetic minority over-sampling technique (SMOTE; Chawla et al., 2002; Fernández et al., 2018) to oversample the minority classes. The SMOTE algorithm selects a random instance from the minority class and searches for the  $k$  nearest neighbours ( $k = 10$  in the present study). Then one of these neighbours is randomly chosen and the synthetic instance is generated by interpolating in the feature space between the original instance and the selected neighbour. The new synthetic instance may be visualised as a random point along a “line segment” between the original instance and the selected neighbour (Fernández et al., 2018, see their Fig. 1). We here use the implementation of the SMOTE algorithm in the Python library imbalanced-learn version 0.12.3 (<https://imbalanced-learn.org/>). In this implementation the SMOTE algorithm is applied to each minority class separately, oversampling to the same frequency as the majority class. We note that we have tested several other methods to balance the class frequency (SVMSMOTE, ADASYN; see, e.g., Fernández et al., 2018, for a brief review), but this did not improve the overall accuracy or the distribution of the predicted results.”

New references:

Fernández, A., García, S., Herrera, F., and Chawla, N. V.: SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, *Journal of Artificial Intelligence Research*, 61, 863–905, <https://doi.org/10.1613/jair.1.11192>, 2018.

García, S., Luengo, J., and Herrera, F.: Tutorial on practical tips of the most influential data preprocessing algorithms in data mining, *Knowledge-Based Systems*, 98, 1–29, <https://doi.org/10.1016/j.knosys.2015.12.006>, 2016.

· Lines 334-316: These paragraphs should be included in the discussion section, not the methods section.

We are unsure to which lines the reviewer is exactly referring here, since the order appears to be wrong. We also note again that these lines are not part of the “Methods” section but of “Random forest optimisation and feature selection”. If the reviewer is referring to the lines following 334, we believe that they are appropriate in that section, since we, as mentioned above, prefer a structure with the discussion already in those places where issues arise.

## **Model evaluation**

The first part of this section reads more like a discussion rather than a results or model evaluation section. Even if the performance is lower, I recommend focusing solely on the 4-binary danger level model and comparing the results of the Random Forest model with those of the artificial neural network. The results are difficult to follow with some of them presented in the appendix. I suggest including the most relevant results directly in the manuscript. Also, it would be interesting to evaluate the performance of developing a model exclusively for dry-snow conditions by excluding wet-avalanche days and vice versa.

Regarding the “section read[ing] more like a discussion” point, please see our general response above.

Moreover, please consider our responses above for why the binary case is still important to us. It is a major part of our work, and we would thus like keep the description of the results pertaining to the binary case.

However, we have restructured section 5 and separated the evaluation of the 4-level and binary case into their own sections (5.1 and 5.2, respectively). Note that section 5.1 is much longer than 5.2 because we are unaware of previous work on aggregating the danger levels to binary levels.

The reason we show the results of the ANN only in the supplementary material is that we did not spend any time on optimising the model for our case (as we did for the RF model). We just used the model set-up described by Sharma et al. (2023) and added the results in the supplementary material to show that they are neither significantly worse nor better than those of the RF model. We stated this in lines 120-121 of our manuscript.

As mentioned above, we now have eight instead of six seasons of available data and thus (like for the RF analysis) re-did the ANN training and analysis (see the new submitted supplementary material). We have changed the focus towards the consequences of the inherent randomness of ANNs, because we obtained considerably different overall accuracies when repeatedly training the same model-data set-up (0.54-0.64 and the 4-level case). Accordingly, the text S2 is substantially reworked and most of the figures and tables in the supplementary material pertaining to the ANN are changed. However, the conclusion that the RF and ANN models perform similarly has not changed. As mentioned above, our main focus during our work was on the RF model and we have not spent more time on optimising the ANN and dealing with the issue of its inherent randomness, which is why we would like to keep the text about the ANN in the supplementary material. As a final note, except for some minor changes to some specific values due to the changed training data, text S3 about the hindcast with the ANN remains unchanged.

Regarding the point about training a model exclusively for dry-snow conditions, please see our general response above.

## Hindcasting avalanche danger

The definition of binary-case avalanche activity is based on the model's predictions rather than on avalanche activity data collected in the study area, correct? (Lines 412–413). Therefore, the term activity is not the most appropriate, as it may cause confusion with observed avalanche activity. It would be more relevant to compare the model predictions with actual avalanche observations from the region to validate the model predictions and thus, correlate them with AO.

We agree with the reviewer that the term “avalanche activity” can be confusing and this is why we specifically called it “*binary-case avalanche activity*” (BCA), and not just “avalanche activity”, and refer to it as BCA for the remainder of the text. We specifically mention already in footnote 4 to the Introduction that our measure of avalanche activity is not based on actual avalanche occurrence. We have extensively motivated in several places of the article why we do not view it as feasible to use actual avalanche occurrence as a metric for avalanche activity in northern Norway – that is, that such data are just too sparse (lines 110–115, lines 149–152, lines 174–185) – which is why we hoped it would be clear from our text that, indeed, we do not use avalanche activity from data collected in our region. Using, e.g., avalanche occurrences as derived from remote sensing observations to compare to the AO (e.g., from the data presented in Grahn et al., 2024) appears promising, but it would be a major work in itself and is beyond the aim of this study. Furthermore, the remote sensing data presented in Grahn et al. (2024) are not yet publicly available and may not yet be used in such an analysis (J. Grahn, personal communication).

To reduce the potential confusion, we will rename the “binary-case avalanche activity” (BCA) to “binary case frequency” (BCF). (See also the quoted new passage in our response to the general comment above.)

## References

Grahn, J., Bianchi, F. M., Müller, K., Malnes, E.: Data-driven avalanche forecasting – Using weather and satellite data. International Snow Science Workshop Proceedings 2024, Tromsø, Norway, p. 39–44, <http://arc.lib.montana.edu/snow-science/item/3109>, 2024.

## Summary and conclusions

I suggest adding a discussion section that incorporates some of the content previously presented in the paper and a final conclusions section.

See the general response above for why we would like to refrain from restructuring the manuscript in this way.

## References

1. Perez-Guillen, C., Techel, F., Volpi, M. & van Herwijnen, A. Assessing the performance and explainability of an avalanche danger forecast model. EGU sphere 2024, 1–29, DOI: 10.5194/egusphere-2024-2374 (2024).
2. EAWS. European Avalanche Danger Scale (2018/19). <https://www.avalanches.org/standards/avalanche-danger-scale/> (2021). [Online; last access 26-July-2024].

3. EAWS. Avalanche Problems, Edited, EAWS - European Avalanche Warning Services.  
[https://www.avalanches.org/wp-content/uploads/2019/05/Typical\\_avalanche\\_problems-EAWS.pdf](https://www.avalanches.org/wp-content/uploads/2019/05/Typical_avalanche_problems-EAWS.pdf)  
(2021).
4. Schweizer, J., Mitterer, C., Techel, F., Stoffel, A. & Reuter, B. On the relation between avalanche occurrence and avalanche danger level. *The Cryosphere* 14, 737–750, DOI: 10.5194/tc-14-737-2020 (2020).