Reviewer 1

Summary

This paper presents a model chain to reconstruct and project avalanche hazard in northern Norway. The approach builds on the authors' earlier work using Random Forest models for avalanche danger prediction. The model chain combines dynamically downscaled climate model output with the snow cover model SNOWPACK, which provides physically based snow stratigraphy and stability variables. These outputs, together with meteorological variables from the downscaled climate models, are used as inputs to Random Forest classifiers that predict avalanche days for multiple avalanche problem types (e.g., wet snow, storm snow, wind slab, and persistent weak layers). The results show distinct historical trends in the frequency of avalanche days depending on avalanche problem type, as well as statistically significant relationships with large-scale climate drivers such as the Arctic Oscillation (AO). Finally, future projections based on climate scenarios (RCP4.5 and RCP8.5) indicate systematic shifts in avalanche problem regimes, consistent with trends previously reported for Alpine regions in Switzerland and France.

The paper is generally well written, well thought out, and is worthy of publication in The Cryosphere. I would like to thank the authors for a comprehensive review response and changes. I have only technical corrections before publication.

We thank the reviewer for considering our manuscript again and for the favourable decision. The remaining technical comments are briefly addressed below.

Technical corrections:

Line 586: Remove "with" for "Several issues and limitations"

The grammar is correct here since it is "issues with" and not "issues of". That is why we wrote "issues with and limitations of". However, for clarity and brevity we will just change the sentence to "Several limitations of our study should be pointed out."

Line 602-607: Is this somewhere in the supplement information? If available, you can maybe point to the figure/section.

Line 610-611: Same comment as above, add the figure/supplement section if available.

Regarding both above comments, we did not include anything in the supplementary information because the results are very similar to the results obtained in the paper.

Line 613-614: Not very a correction here but mostly a comment on why the performance is not improved with SNOWPACK. The Pwl slab problem would be describe by SNOWPACK

(development of weak layers), but the triggers itself, especially if using natural release, would mostly be weather related like a snowfall. This might be an explanation of why it's not improving your model accuracy as anticipated, in addition to biases you mentioned above in section 6.4.

We thank the reviewer for this insightful comment.

Reviewer 2

This paper combines different tools to study past and future avalanche activity in northern Norway. It uses a Random Forest (RF) model for inferring between avalanche days and non-avalanche days based on meteorological input data from meteorological and climate models and snow information from the snowpack snow cover model. Both the past and future trends on partition between avalanche and non-avalanche days are presented. The main novelty of the paper is to presents trends on northern Norway, a region that benefit from less studies than for instance the European Alps to which the results are compared. The overall methodology is similar to the methodology presented in previous work on European Alps with some adaptations to data available.

The paper is quite long, with a lot of appendices and supplements. It shows the amount of adaptation work needed for Arctic regions but may also make reading more difficult. A proofreading to remove typos and highlight the main conclusions would be valuable. The scientific challenge is interesting and the overall question is relevant for publication in TC. However, my main concern is about the methodology for running snowpack simulations and validation of the results on historical period.

We thank the reviewer for considering in detail our revised manuscript. We appreciate the comments and respond point for point below.

As a general response we note that several of the reviewer's comments were quite fundamental or revealed oversights on our part, prompting us to re-do the model optimisation. Thanks for providing these. The reasons include:

- the inclusion of the inappropriately defined feature lwc_sum,
- the inconsistency of not using both _emin and _emax for all parameters, and
- the inconsistency in the model optimisation procedure regarding the target metric.

Accordingly, we updated all figures (except Figs. 1-3) in the manuscript. However, this led only to small changes in the details of our results and our analysis and conclusions remain the same. The changes of the details of our results may be seen in the tracked changes version of our new manuscript.

Major comments

1. In I understand well, the snowpack model is forced by precipitations and surface temperature of snow. This configuration is generally used when there are snow surface temperature measurements available. Here, you use an emulated definition of surface

temperature coming from a relation between ERA5 data and NORA3 data. However, both data do not have the same representation of snow and in your Figure S1 it is clear that you mix soil with and without snow. You apply this relation to a third model (SNOWPACK) independent in terms of snow coverage. I cannot imagine that there is no discrepancies in the result with high surface temperature (coming from a non-snow situation in the atmospheric model) applied to snow-covered soil in SNOWPACK. The reported RMSE of about 4K seems quite high for me, especially in northern Norway where you state that the air temperature is generally not far from 0°C. There is no sensitivity associated to this quite important parameterization of the input of SNOWPACK. Additionally there is in the paper nothing to judge the relevance of the snowpack represented by SNOWPACK model. Hence, with the presented data, I cannot conclude on the relevance of the variables derived from SNOWPACK model and the conclusions that come from the correlation or absence of correlation with variables from SNOWPACK.

We appreciate the reviewer's concerns, but unfortunately these are the limitations imposed by the data we were working with. Hence, we think that the concerns of the reviewer, while interesting and important, cannot be alleviated in our current work. However, please note that our input to SNOWPACK is an average over several hundreds to thousands of square kilometres. This strong spatial aggregation induces imprecision when it comes to snow-covered and snow-free surface anyway. Given this imprecision we believe our linear model determining the surface temperature is appropriate here.

Note that we discuss the large-scale spatial aggregation and the resulting limitations already quite extensively in our section 6.4 on the limitations of our study. However, please also note that the SNOWPACK data in some aspects revealed valuable insights, especially associated with the presence or absence of snow, and when it comes to the most important features for the PWL slab problem. This is discussed in the paper in section 6. We consider this sufficient to include the SNOWPACK data in our study. An improved implementation of SNOWPACK provided recently by the Norwegian Water Resources and Energy Directorate (including an update of the wind parametrisation; see the brief discussion in section 6.4) is planned to be used for future work.

2. There is some points that would benefit from clarification in the input data and optimisation procedure of the RF model. In table C1 and C2, it was not clear for me on which variables the suffixes are applied. Is it to all features, only some, how they are combined when several suffixes are possible? With the current state of the manuscript, it was very difficult for me to figure out what are the exact inputs of the model.
For the optimisation, you "consider" the average of listed metrics, so F1 score, TSS, FAR,

accuracy. This is a quite suspiring approach. Usually, coherent indicators mixing the different part of the confusion matrix are used but I do not know examples of optimisation on such a mix of scores that, for some of them, already combine the different parts of the confusion matrix (e.g. TSS, F1). This needs at least to be discussed and justified. I fully agree that using only FAR or accuracy may lead to incorrect results but some existing tools already exist to combine indicators coherently (e.g. ROC for combining recall and FAR).

To increase clarity and prevent too long abbreviation names for the features we will change the _emin suffix to _n and the _emax suffix to _x. We recognise that our convention was confusing since we did not add the _emin and _emax suffixes to all features to avoid too long abbreviations. To make it clear, we now add the _n and _x suffixes to every feature and we have amended the table captions to give more information about this.

Please note that we did not consider the "average of listed metrics"—instead, we considered these metrics individually and determined the best hyperparameter values based on a subjective assessment of all of the metrics. In the manuscript we recognise that this "may introduce some inconsistency into the procedure" (lines 291-292) but we wanted to avoid maximising or minimising a single metric. As noted in our response to the general comment above, we have conducted the optimisation procedure again using the more common approach of focusing only on the F1-macro score (as in, e.g., Pérez-Guillén et al., 2022; Eiselt and Graversen, 2025). However, we still consider different metrics to make sure that these do not obtain unacceptable values. We will adjust our description in section 3.3 accordingly.

2.1 there is an inconsistency in the presentation of features when they are several on the same line. e.g. "w1, w3, w7" but "wdrift_2, 3". This does not ease the reading of the table.

We recognise that some of these abbreviations are not easy to understand and can be confusing, especially given that there are so many. Our reasoning here is that wdrift and wdrift3, as explained in Table C1, represent the drift index and the cubed drift index. Thus, following the normal terminology the 3-day averaged cubed drift index would be wdrift33 which appears even more confusing and we instead chose wdrift3_3. The terminology for the 3-day average of the non-cubed drift index is hence wdrift_3. We would like to keep this convention, also to be consistent with our earlier work (Eiselt and Graversen, 2025).

2.2 For the sum of LWC by volume, a clear definition would be valuable as summing percentages on layers of different thicknesses is a nonsense.

Indeed, due to an oversight this parameter is meaningless. We thank the reviewer for recognising this and we have removed the parameter form our analysis when conducting the new optimisation procedure.

2.3 : For SSI, Sk38 and Sn38 variation of the index on 1 to 3 days, I would like to be sure that it is the variation on the same layer even though three days before, an other weak layer have been identified as the weakest layer. Can you explain more in detail this variation?

Our convention is that the variation of these indices corresponds to the variation of the weakest layer within the first 100 cm of the snowpack (_100 parameters) and below (_2 parameters). Thus, the layer can vary of during the 2-to-3-day period. Computing the variation in the way the reviewer is suggesting is slightly different and we will add this as a new index in future work. We thank the reviewer for bringing this to our attention.

3. The presentation of the evaluation of the RF model is quite confusing and surprising. The percentages of figure 4 are not related to the whole sample but subsamples consisting of the different classes. This is not quite clear in the legend (the “instances” are not defined clearly) and difficult to interpret. Moreover the legend and the text line 318 state that two of the four percentages presented are recall score while only one correspond to the common definition of recall (also presented in appendix D). The recall is the part of positive avalanche conditions that are correctly predicted (lower right cell only) and not the part of negative conditions that are correctly predicted (which is usually called a true negative rate or specificity). I suggest to present the scores in a table independent from the confusion matrices presented in Fig 4 and to present all the scores used for optimisation.

We have removed the sentence regarding the recall score from the figure caption. Otherwise, we think our presentation of the evaluation is rather standard (including the row-wise depiction of percentages) and very similar to earlier work (Pérez-Guillén et al., 2022; Hendrick et al., 2023; Eiselt and Graversen, 2025). We have attempted to be clearer on this in the figure caption and also changed the colour-bar label to “Row-wise fraction of days”, similar to the way this is described in Hendrick et al. (2023).

4. Past results are presented with NORA3 data while future (climate) projections are presented with NorCP. However, I have not seen a comparison of the data from NorCP on the historical period with observational data and/or NORA3 data that have been validated in section 4. Usually, climate models are compared on an historical period to observations or other data that give confidence in their past representation of the studied object and

validate their application in a far future. Figure 9 could be enhaced by having for the historical period both the data from NorCP and AvD/non-AvD repartition and/or NORA3 data.

We agree with the reviewer that it is typical to compare models during the historical period with observations/reanalyses. We had done this as a sort of sanity check but not included it in our manuscript since, as the reviewer also notes in the main comment, our paper is already quite long and the assessment of the performance of the NorCP data was not part of our aim with this study. For such an assessment we refer to the respective articles as referenced in section 2.3 in the manuscript. Please also note that the investigation of past and future changes in our article is separate and the future values from NorCP are not compared with the historical values from NORA3. However, based on the reviewer‘s

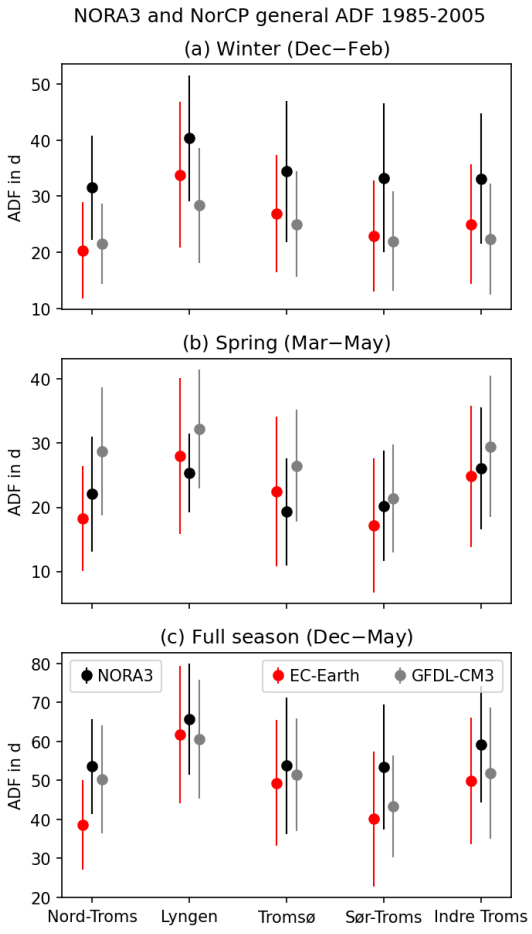

Figure 1: Historical (1985-2005) general avalanche-day frequency (ADF) for NORA3 (black), EC-Earth (red), and GFDL-CM3 (grey). Shown are (a) winter, (b) spring, and (c) the full avalanche season. The dots show the means over the 1985-2005 period for the individual avalanche regions and the errorbars indicate one standard deviation.

comment we decided to include a figure in the Supplement (reproduced here as Fig. 1) showing the comparison of NORA3 and NorCP avalanche-day frequency (ADF) for the historical period 1985-2005 (our "sanity check") and add a brief comment about this in section 6.4 (Limitations). The figure shows that NORA3 and both NorCP models exhibit reasonably similar ADF during 1985-2005. However, it must be noted that because of the influence of the Arctic Oscillation (AO) on the ADF, this 20-year period is likely too short for a robust comparison. It is not clear whether the AO is sufficiently represented in EC-Earth and GFDL-CM3, and even if it is, the AO in the free developing models is unlikely to be in phase with the AO in NORA3. Thus, considerable differences in the ADF between NORA3 and NorCP are to be expected for such a 20-year period. In fact, the ADF in NORA3 is consistently higher in winter and slightly higher in the full season than in both EC-Earth and GFDL-CM3. This may point to the influence of the AO, since the historical period (1985-2005) includes and is likely strongly influenced by the peak of the ADF (especially winter but also in the full season) in NORA3 in the 1990s that we trace to the AO in our manuscript (see e.g. Fig. 6 in the manuscript).

Minor comments

1. The last sentence of the abstract is not easy to understand. A rewriting may allow to deliver the same message more easily.

We have reformulated this sentence to:

"Such a shift can be challenging for avalanche-prone populations as the current knowledge of the local avalanche conditions may become less relevant and increasingly fail to provide protection from the avalanche hazard."

2. Line 41-46, the sentence is very long and some explained acronyms are not (or nearly not) used in the manuscript and may be removed (e.g. SSP, SRES).

We recognise that this sentence is rather long, but this is mostly because of the written-out names. We also recognise that the abbreviations are little used in the remainder of the manuscript, but we think it is appropriate to mention them to give reference and acknowledgement to the work that has been done and on which our and the previous work in this field relies. Moreover, the abbreviations SSP and SRES have essentially become synonymous with these terms, and they are very well-known in this form. That is why we would prefer to keep them.

3. Fig 1 typo in the legend.

Corrected.

4. Figure 3 : Does the line between points have any significance? If not, please remove it.

The line has no significance and was just meant to increase legibility. We have changed this figure to a bar plot.

5.Line 271 : Balancing of classes for RF training. You choose to oversample the minority class. It would also be possible to downsample the majority class or a combination of both. RF models cal also adjust the probability of drawing an observation based on the unbalancing between classes to ensure that the probability of using an observation of minority or majority class is equal. Can you briefly explain and/or justify this choice?

We have tested different ways of class balancing and found only minor differences between the different methods. Thus, we decided to proceed with the minority oversampling as we did in our earlier work (Eiselt and Graversen, 2025). Please note that we mention this in the manuscript in lines 426-428.

6. The abbreviation TSS is used both for snow surface temperature and true skill score, this does not help with readability...

We thank the reviewer for pointing this out. We have changed the abbreviation for surface temperature to TS.

7. Table D1 and Figure 4 does not have prediction/True value at the same place.

Table D1 and Figure 4 were not meant to represent the same things. However, we recognise that the Table D1 does not add much and hence decided to remove this table from our manuscript. We are grateful to the reviewer for raising our attention to this.

8. On Table D1, it is strange to have a, c and d expressed in terms of AvD/non-AvD but not b. Otherwise, a can be called a true positive or a hit, c a miss and d a true negative.

We thank the reviewer for pointing this out. As mentioned above, we decided to remove Table D1.

9. Line 302, point 2 may be rephrased to be easier to read.

Rephrased.

10. I do not clearly see the interest of the “General” model (e.g. Fig 5). I understand all conditions are gathered into this model. However, the climatology of Fig3 show that wind-related problem is largely predominant. Hence, it seems quite logical that the “General” model is quite close to the one trained only on wind slab problem.

Several points lead us to include the general problem in our study: On the one hand, it is not clear that the general problem corresponds mostly to the wind slab. The PWL slab is also rather important and especially towards the end of the 21st century, the wet-snow problem becomes more important. Hence, as the wind slab declines and the wet-snow problem increases, the general problem captures some of the increase of the wet-snow problem and declines less than the wind slab problem. Furthermore, the general problem is the main metric that is (currently) announced to the public on Varsom.no, while the danger levels for the individual problems are not published there. Thus, the development of the general problem has relevance as this is the metric the mountain recreationists are familiar with. To make this clearer, we add a sentence stating that only the general ADL and not the DLs of the individual avalanche problems are published on Varsom. Finally, reporting the general problem facilitates comparison with our earlier work (Eiselt and Graversen, 2025).

11. Figure 7 : There is no uncertainty on this graph while there is on Figure 6. Would it be possible to transfer the uncertainty computed on Fig 6 into Fig 7 because the uncertainty on Fig 6 is interesting for visualisation but not for reading a quantitative value while Fig 7 is designed for that purpose.

While we recognise that adding the uncertainty may increase the informative content of this figure, we think that the figure is already quite crowded and would become considerably less intelligible if the uncertainties were added. Furthermore, the point with this figure was mostly to show that most of the trends are small and not statistically significant (filled markers indicate statistical significance). Thus, we believe that adding the uncertainty does not add enough information to justify the reduced intelligibility of the figure.

12. Figure 7, typo “significance”

Corrected.

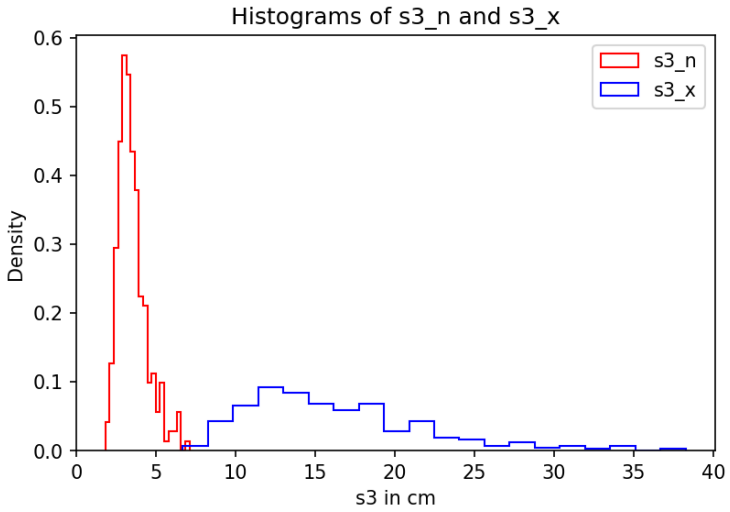

Figure 2: Histograms of s3_n (red) and s3_x (blue) in NORA3 (1970-2024).

13. Mo,e 480-481 : what are the histograms of s3_emin and s3_emax? I wonder if the better correlation with s3_emax is not linked to the fact that at the maximum altitude, you mainly have snow while at lower altitude, you have frequent rain (so s3_emin is zero and could therefore not be well correlated to anything which is not constant).

We show a histogram of s3_emin (now s3_n) and s3_emax (now s3_x) here in Fig. 2. While it is true that the variability of s3_emin is much smaller than s3_emax it is not zero (i.e., s3_emin is not constant) and there is still considerable snow even at lower elevation.

14. Line 569-570 : I think there is a misunderstanding of Castebrunet et al., 2014 as northern French Alps are generally considered of higher elevation than southern French Alps... They mainly state that "results on small scales may be more uncertain, for instance those concerning the southern French Alps".

In section 3.2.1 *Projections of CI values* Castebrunet et al. (2014) state the following (p. 1688-1689): "Figure 9 shows the CI reference distributions and projections for both sub-regions and the different temporal scales considered. It suggests that the overall decrease in avalanche activity forecasted in terms of the CI for the mid 21st century is mostly driven by a strong decrease in the northern French Alps during spring, where the decrease is the strongest (−63 %, Table 6), whereas a slight decrease is also predicted in the winter season (−21 %), contrary to what is expected on the scale of the entire French Alps. By contrast, for the southern Alps, the spring distribution is thinner than for the reference period but with a decrease less marked than on the scale of the entire Alps (−29 %). More dramatically, the winter increase in mean and variance is rather spectacular.

Since the snow and meteorological variable analysis has shown that, at constant altitude, latitudinal gradients (north-south location within the Alps) have little effect on projected changes, these distinct north-south pictures may be attributable to altitudinal effects."

We interpreted this in the way stated in the lines given by the reviewer. However, we will weaken our statement and write that Castebrunet et al. (2014) find "indications" of an altitudinal dependence of avalanche activity changes, while these results remained uncertain.

15. Appendix D : Some metrics are redundant (e.g. FAR and PR).
PR is not redundant as it is used in the definition of the F1-macro score and the FAR is not redundant as it is used in the RF model optimisation procedure (see section 3.3).