Review of *Assessing the predictive capability of several machine learning algorithms to forecast snow avalanches using numerical weather prediction model in eastern Canada.* Gauthier et al. (egusphere-2025-1572)

General Comments

In this study, the authors use four different machine learning (ML) techniques for predicting avalanche events impacting a road along the north shore of the Gaspé Peninsula, Quebec, in eastern Canada. They trained the models using 70% of the data from 2003/04 to 2012/13 and tested the data on five winters 2015/16 to 2019/20. In addition to using a suite of 80+ variables, the authors also used a limited number of variables in an 'expert' model. They also assessed the 24 hour predictive capability of these ML techniques using numerical weather products (NWP). The authors detail the accuracy of each specific ML technique using all variables, expert model, and NWP forced regimes. They found the 'expert' models were highly effective and potentially more accessible for forecasting operations.

Overall, the manuscript is generally well written and organized. The dataset appears suitable for this type of analysis, and the methods are appropriate and reasonable. However, there are several major topics the authors should consider. Here, I provide a few general comments and then specific comments below.

1. The first two major issues pertain to data applicability and sample size reporting. The authors use data from ten winter seasons to train the data and then from five winters to test the data. Given non-stationarity of meteorological and climate variables as well as interannual variability, how well do the test data represent the training dataset and vice-versa. In other words, are you testing the model on winters similar enough to the winters on which they were trained? How might using time series data with the potential of non-stationarity influence your results? Can you provide some measure of variability and spread that shows these two winter datasets are reasonable for training and testing? Additionally, when assessing the 2019 season, was the season similar to the test data seasons? Given the decent scores, it appears so, However, when making recommendations for an operation, it is important to note that different variables or interactions of variables different than the training dataset could still result in avalanches resulting in poor model performance (Peitzsch et al., 2012).

2. You state the number of avalanches (n=861) over 153 event days from 1987 to 2020, but not for the period of record you studied. From the confusion matrix it looks like n=209 for the training dataset but n=489 for LR and NN, but n=479 for CT and RF for the test data (Table 6). 1) Why are there over 2x as many event days in the test data compared to the training data when there are 2x as many years in the test data? This seems peculiar. If this is indeed the case, please provide background in the main text. 2) Why are there two different values (n=479 and n=489) for test data in Table 6? It would be helpful to state the sample size overall value of avalanches and event days for the study period, the sample size for $E_{AV\,A}$ during this period, and the sample size for the training and test datasets in the main text.

   Additionally, the weighting for different number of interventions per day seems rather arbitrary. Can you provide more context for using this metric as opposed to using perhaps a modification of the Avalanche Activity Index (Schweizer et al., 2003) that accounts for size. Accounting for size might also allow you to include avalanches impacting the road as well as those that don't.

Forecasters are certainly considering avalanche activity that doesn't reach the object at risk and this might provide operations a continuum of results on which to base their decision.

3. The comparison of ML output probability with Avalanche Québec's danger levels is a bit of a mismatch of scale. As you state, avalanche operations are incorporating more statistical and physically based snowpack models into their workflow. However, they are likely (hopefully) using scale appropriate simulations or datasets. In your study, the models are trained to predict avalanches that reach Road 132, a very specific local area. The danger ratings are a regional hazard forecast, not necessarily a true reflection of avalanche occurrence for that given day. It would be more useful to compare ML predictive capacity to avalanche occurrence (of similar size) throughout the region rather than danger ratings themselves, since the forecasted danger level also contains an unknown amount of uncertainty. Please discuss the rationale for choosing danger ratings and limitation therein.

4. There is currently very little to no mention of limitations and potential sources of uncertainty or error in the manuscript. Given your suggestion to use ML for operational forecasting workflows, it seems necessary to include a section that details the limitations in this study. It would also be useful to include potential considerations or statements of caution if avalanche forecasting operations are contemplating using ML model output for decision-making.

5. As the authors point out, ML techniques have been used to study avalanche occurrence and improve avalanche forecasting. The use of these techniques is not novel. The addition of a comparison with a more parsimonious 'expert' model, the use of a NWP, and the integration of public avalanche forecasts for validation make this paper a worthy contribution to the existing body of literature. However, some of the Discussion reads like a consultant report for a specific client rather than a manuscript for the broader scientific community. For example, there are more suggestions for this specific forecasting operation than broad comparisons with other studies that might mike this manuscript more applicable to other operations. Please expand upon this in the Discussion. Additionally, there is a lot of content in the Discussion devoted to using the model in operational contexts. It would be helpful to mention how incorporating these results/ML output differs from what the operation currently uses. In other words, how would things have improved (or not) if they used this model versus what they currently do? You allude to this in lines 373-375, but providing further evidence of how previous assessments would have been improved could be helpful here. It would also be useful to describe the days the models failed to predict events since the consequences would be more severe.

Specific and Technical Comments

Abstract: please write/define all abbreviations (e.g. MTMQ, LR, and RF)

Line 24: spreading zone? Do you mean runout or deposition zone?

Line 29: 'windy' remove 'y'

Line 84/Figure 3: This is a bit confusing. The 'Data from NWP model' arrow makes it look like this was used to create the expert model variables. I thought the expert model variables were manually selected. Also, where is the expert model in this flow chart?

Line 85/Sec. 3.1: See comment above re: sample size values. It is necessary to know the sample size for training and test datasets in the text for proper evaluation.

Line 91: I don't understand how $E_{AV\,A}$ is a binary variable. It seems there are 4 possible values (1-4) used to weight if the binary value of avalanche hitting the road (1) or not (0) is 1. The variables simply is a weighting factor. Please clarify.

Line 109-110: The current wording is confusing because the probability of an avalanche reaching the road, as determined by the models, can also be less than 50%. Do you mean 'All probabilities above 50% were classified as event days.'? Also, change 'Considerate' to 'Considerable'.

Line 140: BD should be DB, I assume, for database. Either define DB first or use full word.

Line 171: Seems like you need a citation that demonstrates that model output can be daunting.

Line 193: Change TC to CT.

Line 197: CA should be CT.

Line 212: 'data is processed' to 'data are processed'

Line 226: 'unbalanced datasets'. Since you balanced the training datasets (Sec. 3.1), I assume you mean 'unbalanced test datasets' as per the next sentence. Consider clarifying this here.

Table 6,7,8: Please define acronyms for the reader in the caption. Also define Matrice (i.e. confusion matrix).

Line 261: See general comment above re: representativeness of assessing model with just one season.

Line 314-319: It would be helpful to know the proportion of events classified as wet (wet loose, wet slab, or glide) in both the training and test datasets. You are simply including these variables because wet avalanches are likely to become more frequent, but the model is trained on historical data.

Line 321: Use caution when suggesting practitioners can use specific threshold values for forecasting purposes. See general comment above regarding limitations. There is some uncertainty in the models and there should be some confidence intervals surrounding these values if recommending their use. Using specific model-derived values as a threshold for operational decision making can be problematic, particularly in the context of non-stationarity.

References

Peitzsch, E. H., Hendrikx, J., and Fagre, D. B.: Timing of wet snow avalanche activity: an analysis from Glacier National Park, Montana, USA, Proceedings of the 2012 Interational Snow Science Workshop, Anchorage, Alaska, USA, 884-891, 2012.

Schweizer, J., Kronholm, K., and Wiesinger, T.: Verification of regional snowpack stability and avalanche danger, Cold Regions Science and Technology, 37, 277-288, 10.1016/s0165-232x(03)00070-3, 2003.