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

This paper presents several machine learning models for predicting avalanche events in eastern Canada. The study addresses an emergent topic in avalanche forecasting, with increasing demand for developing data-driven models that can be applied in practice. The authors used typical machine learning models, such as Logistic Regression (LC), Tree classification (TC), Random Forests (RF) model, and an Artificial Neural Network (NN), and compared the variation of the performance using variables extracted from meteorological data, local avalanche observations, and avalanche forecasts of Québec region. While the content, structure, and results of the paper are good and relevant to the topic, improvements are needed in the data used to develop the models, the definition of the target variable, and the detailed explanation of the methods and strategies employed. I recommend publishing this paper after the authors have addressed the following comments and suggestions. Please see my detailed feedback below.

General comments:

1. It would be very useful to add more detailed information on the types of avalanches typically released in the area, their locations (including typical paths in the map of Figure 1), and how these data and observations are collected. Are avalanches observed daily? Are the observations human-based? How accurately is the avalanche release time known? What happens during snowstorms? Is the road closed during periods of high avalanche danger? How far is the meteorological station from the avalanche release areas? Specifically, I recommend adding much more information about the type, size, and release time of the avalanche events in the dataset described in Section 3.1. Additionally, making this dataset open source would be beneficial for other studies.

2. I suggest defining the target variable more clearly in Table 1 and the rest of the manuscript. For example, Eava = 'avalanche day' / 'no avalanche day'. This definition should be used consistently in all the confusion matrices of the tables. Also, to qualify as an avalanche day, what are the requirements? For example: at least one avalanche of size X (what is the minimum size or path length to reach the road?) Additionally, the selection of the weights in Table 1 is not justified or supported by previous studies.

3. Given that the meteorological drivers for dry and wet snow avalanches differ, are the models developed to predict dry avalanches, wet avalanches, or both? This is only mentioned in the Discussion section and could be earlier in the manuscript.

4. The explanation of how the data is merged and the target variables used to test the hindcast and forecast models can be explained in more detail. What is the time window of the avalanche forecasts from Avalanche Québec? How is this merged with the variables extracted from the time series of meteorological data and avalanche activity (point data)? The release time is very important when correlating with the meteorological variables used to develop the models. For instance, if an avalanche is released at 01:00, daily meteorological variables computed from 00:00 to 00:00 of the following day would not accurately represent that event.

5. Could you please provide more information about the avalanche forecast issued in Québec? When is this forecast released, and what is the valid time window of the bulletin?

6. Why were these probability thresholds assigned to each danger level? It would be helpful to show the distribution of all the avalanche and non-avalanche events in relation to the danger level forecasts in Québec. This type of analysis could be used to justify or derive the chosen probability thresholds. Additionally, according to the European Avalanche Danger Scale, danger levels 1 and 2 do not imply the absence of avalanche activity and are dependent on avalanche size (1; 2). For instance, the time series results in Figure 4 show that 4 out of a total of 9 avalanches that reached the road occurred when the forecast danger level was 1 or 2.

7. Feature selection: It would be interesting to provide more information about the feature selection approach. For example, including a correlation matrix in the Appendix. Since Recursive Feature Elimination requires as input the number of input variables, it would be good to show how varying this number affects the performance of the LR and NN models. Does the final selection correspond to the number of features that yields the maximum F1-score?

8. Data splitting and balancing: You mention that splitting the data into training and test sets with a 70/30 ratio is optimal, and that the avalanche and no-avalanche events were balanced to train the models. It would be helpful to include, for example, a bar plot showing the total distribution of events for each class and their distribution in the training and test sets. The confusion matrices in Tables 6 and 7 do not reflect a balanced dataset. Please check my specific comments on these tables below.

9. I suggest moving the description of the machine learning models and the definition of the evaluation metrics to the Appendix. Instead, the main text could focus more on the final model selected, including details such as the architecture of the neural network and the hyperparameters used for the other models. Additionally, providing the code and data used to develop the models would improve the reproducibility of the study and be valuable for future research.

10. For evaluating the performance of the hindcast and forecast models (GEMLAM 24h and 48h) shown in Table 7, what "ground truth" is used to evaluate the performance? Are the danger level forecasts merged by combining levels 1 and 2 as "non-avalanche" events and levels 3 and 4 as "avalanche" events? This should be clearly stated in the caption of the tables and the main text. I am not convinced that merging danger levels 1 and 2 as non-avalanche and levels 3 and 4 as avalanche events is appropriate for defining the ground truth, especially since the actual ground truth in this case consists of observed avalanche events. Instead, presenting a correlation analysis between the model outputs and the avalanche forecasts may be more suitable.

Specific comments:

- Figure 1: It is difficult to see the location of the Cap-Madeleine weather station (the legend should be also translated to English). It would be very helpful to show the outlines or locations of the typical avalanche paths used in this study on the map.
- Lines 61–62: The references to Figure 1 are incorrect; they should refer to Figure 2.

Tables 6 and 7:

- The captions of these tables could be improved by including more detailed information.
- The differences between the results shown in the two tables are not clearly explained. Table 6 presents the performance of the models trained with a different set of variables after the feature selection process, right? Table 7 shows the performance of each model using only four variables. Why does the CT model use only three variables?
- The terms "train" and "train non double" are not defined in the main text or in the captions of the tables. Although it is mentioned that the models have been balanced (Line 158), the "train" dataset is not balanced, as shown in the confusion matrices of both Tables. Additionally, both "train" and "train non double" appear to contain significantly fewer samples than the test set. It is unclear whether the results for the test set were obtained using the "train" or "train double" datasets.
- The reason for showing the performance on these two training sets is also unclear, as their results are not discussed in the text. It is assumed that the results are based on 50 repetitions of the models (Line 160). If that is the case, are the reported scores the averages over those 50 repetitions? If so, a different confusion matrix should be obtained for each repetition.
- There is an inconsistency in the count of avalanche events reported in Table 6: the test set contains 42 events for the LR and NN models, while only 32 events for the other models. Also, the event counts in Table 7 for the NN model differ from the rest.

Table 8:

- The captions of these tables could be improved by including more detailed information.
- Which model is used here?

References

1. EAWS. European Avalanche Danger Scale. https://www.avalanches.org/standards/avalanche-danger-scale/ (2021). [Online; last access 22-May-2025].

2. 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).