Response to reviewer's comments

Manuscript EGUSPHERE-2025-1572

Assessing the predictive capability of several machine learning algorithms to forecast snow avalanches using numerical weather prediction model in eastern Canada

by Gauthier, Laliberté and Meloche

In the following, we provide (in blue) detailed point-by-point answers to the comments raised by the reviewers (in black, *italic*). In addition, modifications made to the manuscript are highlighted in a separate file with tracked-changes.

Response to Referee #3– Dr. Cristina Pérez-Guillén

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.

1 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.* The following lines were added to the section 3.1 to describe how the observations were collected :*'The avalanche observations are collected by MTMQ patrollers to assess the road conditions for rock and ice falls, snow avalanches, as well as storm surge at all time. The observations are recorded at the hour of discovery of the debris, within a few hours of the actual avalanche. Even when the road is closed to the public, patrollers still moved along the road to assess road conditions and make observations. Then, the hourly events are sum to a daily dataset that contains the number of snow removal interventions on the road, the avalanche corridor, whether or not the road was reached, and the distance travelled (ditch, shoulder, 1 lane, 2 lanes). The type and avalanche size are not recorded by patrollers, and therefore, canoot be used in the analysis.'* The weather station is at the east of the avalanche release areas in Cap-Madeleine, indicated in Figure 1. We added these sentences : *"The weather station is located to the east of the study areas (Figure 1), at the sea level (0 m a.s.l). The altitude of the release areas in in average 50 m a.s.l, and goes up to 200 m in Mont-Saint-Pierre sector (Figure 1).* For the open-source comment, unfortunately the dataset belongs to the MTMQ who to not wan to make it open-source, but could be accessible upon request by researchers.

- *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.* The sentences in section 3.1 were restructured to describe more clearly the event variable :*A binary event day variable (E_{AVA} = 'avalanche day' or 'no avalanche day') was created for days where an avalanche has*

reached the road, regardless of the size or path length because this information was not recorded (Table 1). If more than one avalanche reached the road, duplicate event days were added to the dataset with the following weight of 1, 2, 3, or 4 duplicate days with respect to one intervention, two to five interventions, 6 to 9 interventions, and 10 or more interventions (Table 1)." In addition, Figure 4b was added to justify the choice of the bin weight based on the frequency of avalanches per day.

- *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.* We added a sentence at the beginning of Data and Method section to state that the model will predict both dry and wet avalanches: " Our forecasting goal is to predict both dry and wet avalanches in a unique and simple daily model, which will predict the probability of an avalanche occurring". We also add a sentence in the method section stating that the type was not recorded so it couldn't be used in the analysis: "The type (dry or wet) and avalanche size are not recorded by the road patrol, and therefore, cannot be used in the analysis."

- *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.* The data is merged daily at midnight for each dataset and model. We added "(midnight)" after "daily" in the avalanche event section (3.1) and the meteorological data section (3.3). Unfortunately, this introduces some bias and uncertainty to our method, but it still shows efficient results with few misses, almost the same as traditional forecasting. We also believe that using weather variables derived from 48-hour and 72-hour periods helps reduce the error associated with using a midnight cutoff.

- *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?* We added these sentences to describe the forecast: " The forecasts are issued everyday by Avalanche Québec's forecaster teams specifically for the avalanche paths along road 132. They based their prediction on a specific danger level scale relative to the occurrence of avalanches on these specific coastal paths along the road 132 (Figure 2)."

- *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.* We decided to remove entirely this section and only to present visually the model probability and the avalanche forecast. The danger scale is specific to the path along the coast and was decided by the forecaster and Quebec Ministry of Transportation.

- *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?* We have now provided more details about our feature selection procedure for the LR and NN algorithms in subsection 3.4 'Learning Procedure'. Appendix A includes Table A1, which gives more details on the feature importance of each variable during the RFE-CV (feature selection procedure). We acknowledge that RFE is sensitive to the required number of input variables for the model, or how many features to keep. This is why we chose to perform RFE-CV, where the number of features to keep is defined by the maximum F_1 score. Then, we conduct a second RFE within each group to retain only one feature per group (best feature importance) to avoid collinearity in

the final variables. Thus, the final input variables will always be a maximum of 8 features, which corresponds to the number of correlated groups. Fewer than 8 will be the case if no feature within a correlated group was kept after the first RFE-CV.

- *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.*We added Figure 4-b to show the distribution of both datasets with each weight bins (number of avalanche per day).*The confusion matrices in Tables 6 and 7 do not reflect a balanced dataset. Please check my specific comments on these tables below.* The confusion matrices are balanced only for the *Train* dataset in Table 6. The *Test* dataset, as well as the hindcast for winter 2019 in Table 7, are not balanced to evaluate the predictive capability. Additionally, the confusion matrices have been slightly modified to correct typos in Tables 6 and 7.

- *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.* We added sentences to each machine learning description detailing the packages used, the hyperparameter values, and the specific architecture for each algorithm. However, we did not move this section to the appendix because we consider it an essential part of the methodology. The code will be available online, but the dataset belongs to Québec's Ministry of Transport, and we cannot publish it on their behalf as they are the data owners.

- *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.* We are sorry that original manuscript was not enough clear on that matter. The ground truth in Table 7 is the avalanche event dataset, which was used for the computation of F1 and the AUC score. This information is now present in the caption. *Instead, presenting a correlation analysis between the model outputs and the avalanche forecasts may be more suitable.* The Avalanche forecast is now present as a visual comparison is presented in Figure 5, and the F_1 score of the AvQc forecast is computed with their *Considerable* or *High* danger level (occurence of avalanches on the road) with the avalanche event dataset.

2 Specific comments

- *Figure 1: It is difficult to see the location of the Cap-Madeleine weather station (the legend should also be 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.* The visibility of Cap-Madeline Weather station has been improved and the legend translated.

- *Lines 61–62: The references to Figure 1 are incorrect; they should refer to Figure 2.* Done

- *Tables 6 and 7:*

  - *The captions of these tables could be improved by including more detailed information.* The captions were improved and added more information for clarity.

  - *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?* We improved the caption to state that Table 6 is

for the model selected by each ML models, and Table 7 for the expert models. We also add that, in Table 7, the CT algorithm did not selected the Rain_24h as significant when provided the 4 variables selected by the "expert".

- *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 train non double was removed completely for clarity. The Table now included only the train dataset (Balance) and the test (5 years of Hindcast).
- *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.* The models were constructed 50 times per ML algorithm, thus creating 200 models for the train dataset. The best F_1 score is shown.
- *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.* The typos in the count have been changed, now all the models count 32 events. LR and NN had 38 TP instead of 28 (Table 6). NN in Table 7 was also changed to get 32 events.

- *Table 8:*
  - *The captions of these tables could be improved by including more detailed information.* The captions has been improved for clarity : "The performance metrics are computed with the avalanche event dataset where avalanches reached the road. An event day is considered when the danger level is Considerable or High (avalanche expected on the road), according to Table 2."
  - *Which model is used here?* No model is used here but we compared the danger level (Considerable or High) of AvQc with avalanches on the road.

References