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

Frank Techel
WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland

1 Summary

The study evaluates and compares the predictive performance of four machine learning algorithms for snow avalanche forecasting along a highway corridor in eastern Canada. Models are trained on meteorological data from a local weather station and tested both retrospectively (hindcast) and using 24–48 h numerical weather prediction (NWP) data. The authors also develop simplified "expert models" based on a small set of selected variables and show that these models can match the performance of more complex approaches. They conclude that such models are suitable for operational forecasting and hazard management using NWP data.

While the manuscript is generally well structured and supported by mostly clear figures and tables, several methodological choices reduce the transparency and interpretability of the results. Key concerns include the definition and weighting of the event variable, inconsistent reporting of dataset size and time coverage, the rationale for mapping model probabilities to danger levels, and limited validation based on a single winter season. Additionally, the forecasting goal and performance metrics are not clearly aligned with the operational context. These issues are outlined in detail below, with the intention of strengthening the manuscript's clarity, methodological robustness, and applied relevance.

2 Major concerns

2.1 Size and composition of event data set

The manuscript does not report the total number of days included in the dataset used for model development (though the cumulative number of events since 1987 is mentioned), nor the number of event days used for training and testing, or how many days were covered by avalanche forecasts. This lack of detail makes it difficult to evaluate the size, coverage, and representativeness of the training and validation datasets. Please provide these numbers. Additionally, I recommend including a histogram or table showing the distribution of intervention counts per day. If available, it would be helpful to show, in the same figure, the total number of avalanches per day (including those not classified as events), to help readers understand how

often avalanche conditions were favorable for avalanche release and, among those, how often avalanches reached the road. This would also provide the link to Figure 4, where both interventions and avalanches, which didn't reach the road, are shown.

2.2 Event weighting

The current approach assigns weights to event days based on the number of interventions per day, duplicating rows with weight factors from 1 to 4. While this may aim to emphasize high-impact days in the dataset, the scaling appears arbitrary. The manuscript would benefit from a clear justification for this specific weighting scheme, including a discussion of whether it more reliably captures avalanche activity than alternative approaches — such as duplicating entries based on the actual number of interventions, using no weighting at all, or treating severity as a separate variable. Additionally, the potential for overfitting introduced by this duplication strategy should be addressed.

Potentially, the following comment concerns this weighting scheme: in Tables 6 and 7, the term "train non double" appears as a label for model results, but it is never defined in the text. Presumably, it refers to training on the non-duplicated (unweighted) dataset, but this should be clearly stated. Please define this term explicitly in the methods section and explain how the results under this label differ in training setup from the standard "train" case with duplicated rows.

2.2.1 Mapping forecast probabilities to danger levels

In Section 3.2 and Table 2, the authors introduce a mapping of model-predicted event probabilities to avalanche danger levels. However, the rationale for this mapping remains unclear. It appears that fixed probability ranges (e.g., 50–75% for "Considerable") are assigned to match categorical danger levels, yet it is not explained whether this mapping was derived from empirical data, optimized using forecast performance, or adopted from an existing standard. Shouldn't the probability ranges be obtained following analysis to be in line with the data?

Moreover, it is unclear whether the proposed scale is consistent with the North American Avalanche Danger Scale (Statham et al., 2010), or whether the four-level version used here, and their descriptions which clearly relate to road level avalanche conditions, follow an accepted operational standard, whether this is the definition used by road authorities, or whether it is a custom adaptation. Please clearly state so, justify the choice, and reasoning for this descriptive definition.

Please also provide the respective event and non-event data with the forecast danger levels, i.e., the number of avalanche days to all days for each level, and the average number of events per intervention day.

2.2.2 Forecasting goal, performance metrics, and contextualisation

The manuscript lacks a clearly stated forecasting goal — for example, whether the priority is minimizing false negatives (missed events) or false positives (false alarms) or whether a balanced performance is of interest (Sect. 3.4.6). While F1 score and AUC are widely used, the authors do not critically discuss whether these metrics allow to reflect meaningful operational gains over current forecasting practice. In particular, the F1 score assumes equal cost for false positives and false negatives, which may not be appropriate in avalanche forecasting where missing a high-impact event can have serious consequences.

The claim (L359) that the models are efficient and suitable for supporting road safety management would be stronger if it were grounded in such operational considerations. Ebert and Milne (2022) provides an in-depth discussion of these metrics for evaluating rare-and-severe event forecasts may be useful.

2.2.3 Inconsistent and limited model validation across time periods

It is somewhat confusing — or perhaps just not clearly explained — that results from datasets with differing temporal coverage are directly compared (L261–271): the hindcast includes only a single season (2019), the Test dataset spans five seasons (L156), and the Avalanche Québec forecasts are evaluated over either five or seven seasons (Table 8 refers to 2013–2020, while L156 defines five seasons). Please clarify which time periods are used for each comparison, and ensure consistent reporting of performance metrics across these datasets.

Moreover, while presenting a single season (2019 in Figure 4) as an illustrative example is appropriate, using it as the only hindcast validation case is insufficient to evaluate model robustness across varying snow and weather conditions. As shown in Table 7, only nine event days occurred during this season, which limits the strength of conclusions about model performance (Section 5.2). Please justify the selection of this specific year for hindcast validation and discuss the implications this has for the operational relevance and generalizability of the findings.

3 Minor concerns

- Study area: Please provide the elevation of the Cap-Madeleine weather station and the elevation range of the avalanche release areas. What is the elevation of the GEMLAM grid cell used in the analysis (L186–187)? - This information is essential to assess the representativeness of the input data and the applicability of the forecasts to release zone conditions. In case elevation differs, take this up in the Discussion. Also, in the discussion, address that some road sections lie nearly 100 km from the weather station, which may limit forecast reliability in parts of the forecast area.

- **Section 3.2:** The manuscript uses regional avalanche forecasts issued by Avalanche Québec as a reference, but it is not clear why these were chosen over road-specific hazard assessments. Please clarify their relevance in the study context. In this regard, it may be helpful to include more detail on the operational hazard assessments carried out by road authorities — possibly in the Study Area section. For instance: Are roads preventively closed? Are Avalanche Québec's forecasts intended for, and actively used by, road authorities in their decision-making? When interpreting the results, consider discussing how the avalanche forecasts relate to actual road closures or missed closures (e.g., in terms of precision, false alarm rate, or true positive rate), as this would provide a clearer link to the intended operational application and highlight the challenges in forecasting avalanches that reach the road.

- Figure 1: Please make the location of the Cap-Madeleine weather station more prominent - e.g., using a larger dot or different color.

- **L129:** The notation "1h/24, 4h/24, 8h/24, 12h/24, 24h/24, 1h/48, 4h/48, 8h/48, 12h/48, 24h/48, 48h/48, 72h/72, 96h/96, 120h/120" is not intuitively clear. Additionally, Table 3 lists different values for int_{rain} and $N_{\mathrm{hr,snow}}$. Please add explanatory detail, perhaps moving this notation to Table 3 where similar forms are used for temperature variables.

- Table 3: There are several inconsistencies and typos, such as the use of French ("et") instead of English, and mismatched or unclear time intervals. For example, N_{snow} is described as 24 h on the left but listed as 24 and 48 h on the right. Why are $N_{\mathrm{snow}}(24h)$ and $N_{\mathrm{snow}}(48h)$ grouped on one line, while longer aggregations appear separately? Please revise for consistency and clarity.

- Table 3 and Results section: Is $N_{\mathrm{hr,snow}}$ (Table 3) the same as $Snow_2 4h$ and $Snow_7 2h$ (L304). - Please ascertain that all abbreviations are used consistently throughout.

- Section 3.4.1: This section appears to mix the introduction of NWP data with the process of selecting expert model variables. Please introduce the NWP data in a separate paragraph—either as a dedicated subsection (e.g., Section 3.4) or by restructuring Section 3.3 into subsections for weather station data, NWP data, and derived variables. Additionally, clarify how expert judgment guided variable selection beyond redundancy (step 1), especially since the variables in Table 6 differ significantly from those ultimately selected (L251). Consider dedicating a specific section to this.

- Figure 4: A scale or reference for avalanche activity is missing. It would be helpful to plot the number of interventions and non-intervention avalanche observations above the activity bars. Also, please clarify whether the activity bars represent a histogram or another form of count aggregation.

- Section 4.3: please state whether the "expert models" are analyzed

- **L299–305:** The SnowDriftIndex is discussed as an additional variable of interest, but no results are shown regarding its impact or inclusion in model evaluations. Please report its contribution to performance to ground this discussion in empirical evidence.

References

Ebert, P. A. and Milne, P.: Methodological and conceptual challenges in rare and severe event forecast-verification, Nat. Hazards Earth Syst. Sci., 22, 539–557, https://doi.org/10.5194/nhess-22-539-2022, 2022.

Statham, G., Haegeli, P., Birkeland, K., Greene, E., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J.: The North American public avalanche danger scale, in: Proceedings ISSW 2010. International Snow Science Workshop, 17 - 22 Oct, Lake Tahoe, Ca., pp. 117–123, 2010.