Predicting Avalanche Danger in Northern Norway Using Statistical Models (egusphere-2024-2865)

This paper presents several machine learning models for predicting avalanche danger levels in Norway. The study addresses a relevant topic in avalanche forecasting and model development, particularly in regions with limited data, which is significantly challenging for developing models. The authors compare two types of machine learning models, a Random Forest model and an Artificial Neural Network and report some variations in their performance. I recommend publishing this paper after the following comments and suggestions.

Some parts of the manuscript could be shortened or moved to the appendix, while more interesting results could be incorporated into the main text. The model was designed to predict danger levels for both dry- and wet-snow avalanche conditions. Although the dataset is small, I suggest developing and testing models exclusively for dry or wet snow conditions or developing two separate models, as the drivers of each avalanche type differ significantly. Furthermore, justifying the development of a binary model by merging danger levels only based on better performance is insufficient. Even if the performance is lower with four danger levels, this approach is more realistic and a well-defined target in avalanche forecasting. The train/test split and the cross-validation method used for the optimization and development of the models can be better clarified. Below, I provide more detailed comments for each section.

Introduction

The introduction should be more concise, re-structured and focused on the main topic: the development of machine learning models for avalanche forecasting in Norway. Overall, the writing should follow a clear structure, avoiding mixing concepts and ensuring it reads like an introduction rather than a discussion or conclusion. Also, it should highlight existing knowledge gaps or limitations of current models, linking them to the relevance of this study.

Specific comments:

- The first paragraph is too general, and part of the content is not very relevant to the paper's topic.
- Line 43: it is not a number but a scale.
- Line 44 and 45: the European and American scales have some differences.
- Line 52: The danger level model has been operationally integrated into avalanche forecasting in Switzerland since the winter season 2021-2022 [1].
- Lines 52-103: This paragraph is very long and lacks clarity regarding its main topic (e.g., avalanche prediction or avalanche danger level forecasting). Additionally, some explanations fit better for the discussion section.
- Lines 106-110: This should be the motivation of this study and moved above in the introduction.
- Lines 95: Sections should be referenced in the order in which they appear in the paper.
- Lines 106-110: These paragraphs read more like a summary of results mixed with discussion and conclusions rather than an introduction to the content of the paper.

Data

Since the danger levels forecast in the public bulletin are used as the target for the machine learning classifiers, the description of the avalanche forecast in Norway should be more detailed, adding the correspondence references. How often is the public avalanche forecast issued in Norway? Is it updated daily? At what time is the bulletin issued, and what is the validity period of the forecasting window? Are the danger levels provided based on the European 5-level Avalanche Danger Scale [2] and the descriptions of the avalanche problems [3]? Given that the meteorological and snowpack factors driving the formation of dry and wet snow avalanches differ, it is important to clarify whether the models have been specifically developed to predict danger levels for dry or wet snow conditions. Additionally, how are these two types of avalanches forecast in Norway? How are the critical elevation and aspects where the danger level is applied forecast in the Norweigan bulletin? Furthermore, how are the data sets prepared and merged for model development to address these distinctions?

I am unsure about the reasons for developing a binary classification model by merging danger levels 1 and 2. The justification that the model performs better should not drive the development of a model with a new target variable, i.e. a new danger level scale. Furthermore, the definitions of danger levels 1 and 2 do not imply an absence of avalanche activity and depend on avalanche size [2; 4]. I suggest focusing on the results of the multiclass models.

Specific comments:

- Lines 149-152: These initial sentences would be more appropriate in the introduction.
- Lines 162-165: This explanation should be in the discussion or outlook for future model implementations. Are you also considering a wet snow problem?
- Figure 1: I suggest adding here or in a different Figure an example of a danger level map from the public avalanche bulletin issued in the study region.
- Since the models were developed by merging data from different regions and winter seasons, I suggest modifying this figure to display bar plots showing the distribution of danger levels in the training and test sets. This would provide a clearer visualization of the data volume, the frequency of each danger level, and the proportions used for training and testing.
- Lines 174-185: This should be moved to the discussion section rather than being included in the description of the data used.
- Lines 202: Please specify the exact present date.
- Lines 203-211: this reads more like a discussion than a description of the data used in this study.
- Lines 211-214: The values from all the grid points within the elevation band of 400 to 900 meters are averaged per region, is this the elevation band where avalanches usually release? Why are grid points at elevations higher than 900 meters discarded? Are avalanches not released from these higher elevations? When merging the forecast data with the model's input features, how do you account for the elevation limit of the danger level forecast and the slope aspect? Have you tested averaging grid stations in micro-regions to achieve higher-resolution data and more data for the development of the model? Having more details about the different processes test with the elevation bands would be interesting.
- Lines 215-225: More detailed information about the output variables of this model, as well as a visual example of the simulation, would be useful.
- Lines 226-231: This should be in the discussion/outlook of the paper rather than here.
- Lines 234-253: Please specify how you are resampling the meteorological time series to match with the forecast window of the danger levels.
- Lines 254-253: Please show in a plot or specify in the text the distribution of danger levels and data between the train and test. Ideally, the danger level distributions of test and train should be similar.

Methods

Since Random Forest is a widely used machine learning algorithm, I recommend moving Sections 3.1 and 3.2, along with Tables 3 and 4, to the appendix to save space. It would be more relevant to specify the programming language and Random Forest implementation used and whether the data and model are open source and available.

Specific comments:

- Lines 310-316: The train/test split and the cross-validation method for optimization are unclear. According to Line 257 (Section 2.3), two winter seasons are used as the test set. For model optimization, you used the six available winters, applying cross-validation. Does this mean the test set is also being used for model optimization? Ideally, the optimization and cross-validation process should be applied exclusively to the training data, allowing the model's performance to be evaluated independently on the test set. Additionally, it would be helpful to provide more details about the SMOTE oversampling technique used.
- Lines 334-316: These paragraphs should be included in the discussion section, not the methods section.

Model evaluation

The first part of this section reads more like a discussion rather than a results or model evaluation section. Even if the performance is lower, I recommend focusing solely on the 4-binary danger level model and comparing the results of the Random Forest model with those of the artificial neural network. The results are difficult to follow with some of them presented in the appendix. I suggest including the most relevant results directly in the manuscript. Also, It would be interesting to evaluate the performance of developing a model exclusively for dry-snow conditions by excluding wet-avalanche days and vice versa.

Hindcasting avalanche danger

The definition of binary-case avalanche activity is based on the model's predictions rather than on avalanche activity data collected in the study area, correct? (Lines 412–413). Therefore, the term activity is not the most appropriate, as it may cause confusion with observed avalanche activity. It would be more relevant to compare the model predictions with actual avalanche observations from the region to validate the model predictions and thus, correlate them with AO.

Summary and conclusions

I suggest adding a discussion section that incorporates some of the content previously presented in the paper and a final conclusions section.

References

1. Pérez-Guillén, C., Techel, F., Volpi, M. & van Herwijnen, A. Assessing the performance and explainability of an avalanche danger forecast model. *EGUsphere* **2024**, 1–29, DOI: 10.5194/egusphere-2024-2374 (2024).

2. EAWS. European Avalanche Danger Scale (2018/19). https://www.avalanches.org/standards/avalanche-danger-scale/ (2021). [Online; last access 26-July-2024].

3. EAWS. Avalanche Problems, Edited, EAWS - European Avalanche Warning Services. https://www.avalanches.org/wp-content/uploads/2019/05/Typical_avalanche_problems-EAWS.pdf (2021).

4. 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).