Reply to Michaela Wenner
June, 2021

Thank You for presenting a detailed review of the manuscript. Your comments will improve the manuscript greatly.

1. Reply of Specific Comments

1.1 Structure of manuscript

In section 4 of manuscript we have included a description of performance measures, cross-validation and model generalisation which should be included in a methods section.
We believe that the data-preprocessing is very specific to the domain of avalanche forecasting and depends on our data-set, this sub-section should remain in Section 4 along with the results.

Following restructuring changes will be made in the revised manuscript:

1. Section 4.2 [performance measures] will be moved into methods section.
2. From Section 4.3 [Model training and hyper parameters training] the details of hyper parameter training will be moved into a sub-section in methods section.
3. From Section 4.4 [Model Generalisation] the details of how the model was trained and the testing scores used will be moved to a sub-section of methods section. These details will only be referred from the section on model generalisation.

1.2 Performance Analysis

1.2.1 Analysis using ROC curves and AUC scores

Table 5 provides the exact FAR/POD trade-off found from the ROC curve i.e a sampling from the curve at uniform FAR intervals of 0.1. It provides additional scores to help readers compare the training and test performance. The AUC scores for testing and training phases are given in the appendix of this reply and will be provided in the caption of Table 5 (in the revision).

1.2.2 Contigency Table Analysis

The contigency tables can be reconstructed from the FAR and POD scores when the number of negetives (Total Negetives in formulas) and positives (Total Positives) in the testing data is known.

$$POD = \frac{True\ Positives}{Total\ Positives} \qquad FAR = \frac{False\ Negetives}{Total\ Negetives}$$

Therefore using FAR,POD, Number of Positives and Number of negetives the contigency table entries are:

$$True\ Positives = POD * Total\ Positives$$

$$False\ Positives = Total\ Positives - True\ Positives$$

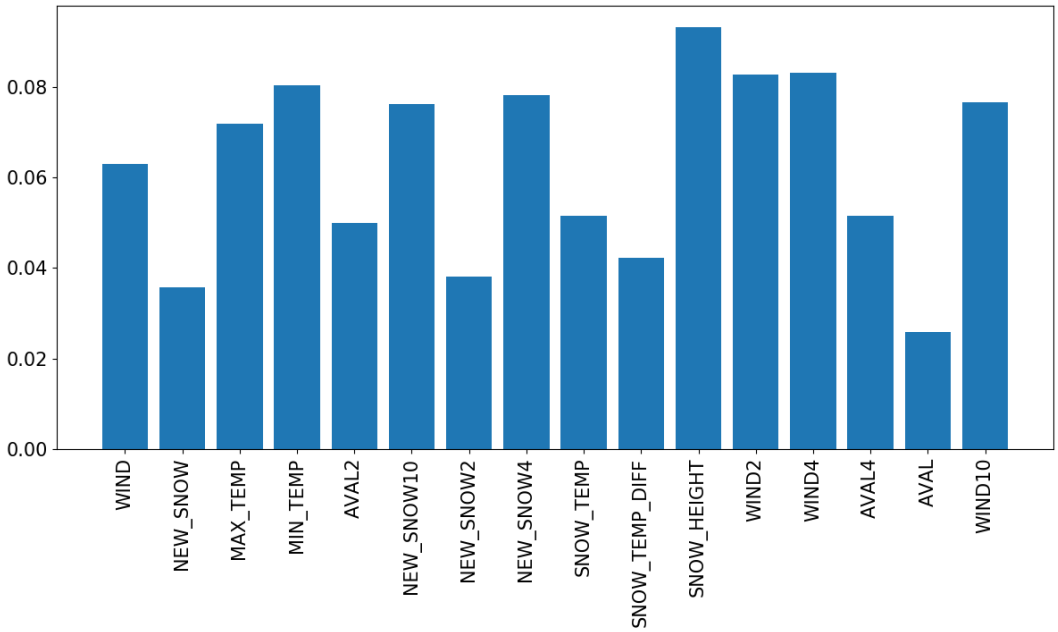
$$False\ Negetives = FAR * Total\ Negetives$$

$$True\ Negetives = Total\ Negetives - False\ Negetives$$

We will provide the contigency tables for each FAR and POD level given in Table 5[in manuscript appendix]. These have been additionaly provided in the appendix of this document.

1.2.3 Feature Importance Analysis

We have computed the importance score for each training feature, presented in the bar graph below.



Some observations:

1. SNOW_HEIGHT has the highest contribution in avalanche formation followed by the variable for cumulative snow-fall of past few days.

2. NEW_SNOW has low contribution, yet the cumulative snow fall is important(NEW_SNOW10, NEW_SNOW4).

3. AVAL has low contribution, yet the avalanche history is important (AVAL4, AVAL2).

From 2 and 3 , we may believe that variables showing past 4 day snow instability (AVAL4) and 4 day snow fall are more valuable for avalanche forecasting than the variables showing immediate instability and snowfall. The weather and snowpack history of past few days contributes in complicated ways to increase hazard. We will analyse this in future work.

1.2.4 Comparison with a simple baseline model to demonstrate data-efficiency

We trained and tested a nearest neighbours model on the training and testing datasets of the RF model, the performance of RF model was found significantly better. The FAR/POD scores obtained and the AUC score of the NN-model are provided in appendix of this file.
These results will be discussed in comparisons section of the revised manuscript. They clearly show the data efficiency of RF model.

2. Technical Corrections

l78 If you mention detection of avalanches using seismic data, you might add our recent paper (Wenner et. al 2021) in which we actually also use RF to detect mass movements (also avalanches).
This reference will be included in the revised version.

l87-88 Is the input not the same as for other ML models? Please clarify what you mean with this statement
Lines 87- 88 all the input parameters used for our model can be collected automatically. This is not true for many avalanche forecasting models, the comparisons section details this.

l108-111 Please rephrase and clarify
To learn a tree from the dataset an algorithm has to find the feature value and its threshold at each tree node. At each iteration the algorithm takes a dataset and gives a threshold (t), feature (f) and two disjoint partitions of the dataset. One partition containes all data points where feature (f) has values <= threshold (t), other contains all the data points where where feature (f) has values > threshold (t). Algorithm starts with the entire datset initially, the split and feature found are recorded in the top-most node, the sub-datasets found are used to define the split and feature values of the right and left sub-childs. This leads to further splitting and a recursive definition of the tree structure.

Table 3 I would find it more informative to see the total number of avalanches days instead of the mean, and accordingly total number of avalanches that happened with a snow height below 0.5m
This information can be derived from information given in Table 3, in revised manuscript we will provide it for completeness.

Table 4: I like the table, it gives a great overview. Maybe you could just add in the caption what i,j are (true label vs classifier label)
This will be done in the revised manuscript.

186 Here, the confusion matrix is described, but then not used in the paper. I strongly suggest to do that though (as mentioned in specific comments).
The confusion matrices for all FAR levels given in Table 5 will be provided in the appendix of revised manuscript.

Figure 4 Add standard deviation from 5-fold-cross-validation. Additionally, make sure that the term ”5-fold-cross-validation” is equally written in the text and the figure title

Please clarify the meaning of standard-deviation here. The results of the 5-fold-cross-validation depend only on the dataset and the classifier used, we are not choosing the 5 – folds randomly so the results dont change when we do it multiple times.

197 Explain quickly what cross-validation is and what the 5-fold means
This will be done in the methods section of revised manuscript.

198 There are more than two hyper-parameters for RF. Explain why you didn't change those or rather why you set it to this value.
Changing them did not result in significant performance difference in our experiments.

l202 Please rephrase and clarify this statement
We get the conditional probability of an avalanche occuring given the input parameters.

l205-209 This should definitely go to the discussion section of the paper
This will be done in the revised manuscript.

l226 Yes, but this is intrinsic with the definition of both. How about the the POD? Also, please address Figure 4a and Figure 4b separately and explain what is shown.
Line 225-226 The sentence "High classification threshold probability means only days when the model is highly confident are classified as positive (avalanche days)" will be rephrased to : "increasing classification threshold gives higher precision i.e the likelihood of an avalanche occurring on a predicted avalanche day increases.". The model gives fewer but more accurrate alarms when the threshold is increased, this leads to lower detection rates and higher precision scores.

l232 What would those complex situations be? Please discuss (in the discussion section)
Complicated situations involve factors which cannot be deduced with a high certainty from the input data alone e.g: burried weak layers, ice layers, depth hoar crystals. To account for the uncertainty involved we use the statistical modelling approach.

l233 Consider using normal words instead of the feature names - new snow instead of new snow (the connection is easy enough)
This will be done in revised manuscript.

l239-241 Consider rephrasing the sentence to: "After hyper-parameter tuning through cross-validation experiments for optimized performance, we constructed (...)"
This will be done in revised manuscript.

l254 How can you find unstable slopes with that? Please clarify
1. By analysing the samples in leaf node.

2. By identifying the important factors associated with the avalanche day, we can identify which slopes they will affect the most: e.g high temperatures affect south aspect slopes most, teperature gradients will affect slopes at a higher altitude most.

3. This can also be done by including terrain features when training the model.

l262 Consider rewriting to "(...) high wind speed as the most indicative feature for avalanche hazard"
This will be done in revised manuscript.

l280-281 Please clarify the sentence
Snow height is an important hazard factor, data analysis shows that avalanche probability is greater when snow height is higher. We found conditions (temperature bound rule) which causes the

hazard of lower snow height days to be greater than days with higher snow height which dont satisfy the conditions.

l283 I suggest to skip the last sentence
This will be done in revised manuscript.

l290-304 I like the comparison to other studies, however I think it could be a bit more comprehensive. Maybe rephrase and make sure to clarify your message. Also, this should be in the discussion section.
We will move this into section 6 [comparisons with other models].

l309 ”The model uses lesser”: what do you mean by lesser? In terms of quality or quantity?
Quantity (The model uses 3 year data, this will also be clarified by giving a comparison with a baseline nearest neighbour approach).

l326-328 Please rephrase and clarify.
Using more information about snow,weather and terrain parameters can improve avalanche forecast. This can be done by including additional features in models e.g: snow wetness index, snow stability index, satellite image features, terrain features etc. To use the new feature, a model must be trained from a dataset containing it. Data efficiency minimises the number of training records required that contain the new feature, this can help if the collection of feature was started recently e.g: by installing new sensor for previously unrecorded parameter.

3. Appendix

Feature Importance From RF model.

WIND	0.06
NEW_SNOW	0.03
MAX_TEMP	0.07
MIN_TEMP	0.08
AVAL2	0.05
NEW_SNOW10	0.08
NEW_SNOW2	0.04
NEW_SNOW4	0.08
SNOW_TEMP	0.05
SNOW_TEMP_DIFF	0.04
SNOW_HEIGHT	0.09
WIND2	0.08
WIND4	0.08
AVAL4	0.05
AVAL	0.03
WIND10	0.08

NN-Classifier Performance: AUC Score NN[0.7], AUC Score RF: [0.82]

FAR	POD [NN]	POD [RF]
0.2	0.6	0.65
0.3	0.68	0.76
0.4	0.75	0.83
0.5	0.77	0.88
0.6	0.85	0.91
0.7	0.87	0.93

Contigency Tables from RF model:

FAR [0.2]	Avalanche	No Avalanche
Avalanche	63	34
No Avalanche	78	310

FAR [0.3]	Avalanche	No Avalanche
Avalanche	74	23
No Avalanche	116	272

FAR [0.4]	Avalanche	No Avalanche
Avalanche	81	16
No Avalanche	155	233

FAR [0.5]	Avalanche	No Avalanche
Avalanche	85	12
No Avalanche	194	194

FAR [0.6]	Avalanche	No Avalanche
Avalanche	88	9
No Avalanche	233	155

FAR [0.7]	Avalanche	No Avalanche
Avalanche	90	7
No Avalanche	272	116