General comment :

The paper is well-structured, well illustrated, and easy to read. Particular care was taken to provide neat and readable figures and explanations, which I thank and congratulate the authors for. The paper's topic is both interesting and timely, in line with the increasing availability of diverse snow depth data, sometimes produced by non-professional networks or organizations, and that could serve scientific goals provided they can be qualified. The methods proposed and the analysis of the results are sound and provide a balanced evaluation of the proposed automatic quality assessment tool. I have some suggestions that I hope will help clarify some methodological points related to metrics/evaluation, and complement the perspectives. I recommend the publication of this article provided these minor suggestions are taken into account.

We appreciate the reviewer comment and acknowledge that recommended changes will improve the clarity of our work. We think all suggested modification are feasible, thus we will work towards this direction to improve our work.

Detailed comments :

*** "what is the accuracy of a Random Forest classifier algorithm in automatically performing QA/QC of near-surface snow depth observations?" Although the choice of a RF classifiers is well justified in the paper, it seems other AI algorithm could also be used. This could be something to explore in future work. Especially, the consideration of snow height measurements as a time series and not just separate features, could possibly help identify spikes better ; this could be done through the use of AI algorithm incorporating memory features like recurrent networks or LSTM (see last point of the Detailed comments).**

We agree with this suggestion and in general with the recommendation of exploring the performances of LSTM over a Random Forest. Indeed, we believe there could be possible interest in investigating the ability of a LSTM to handle time series. Although we chose a Random Forest because of its easier implementation than LSTM, we will add a reference to this in the Discussion. We will add the paragraph below:

In recent years, Deep learning has proven successful in dealing with many complex task (Camps-Valls et al. 2021). Future research questions may investigate the ability of other algorithms in this classification problem, such as neural networks, which are able to deal with time series and incorporate memory features. One concrete example in this regard a recurrent neural networks or LSTM (Long Short Term Memory). In particular, it would be important to explore the performances of such algorithms in dealing with the recognition of the error class.

*** L149-150 "After the oversampling procedure, a sample of 1.9×10**6 over-sampled measurements was used". It is somewhat not so clear whether 1.9x10**6 is the size of the total training set (including majority classes + oversampled minority class, which I believe it is) or just the oversampled minority class, which the "over-sampled measurement" in the sentence makes think. Could you be a bit more specific in the description ?**

The mentioned sample size is the size of the total training set, including both the majority and the oversampled minority classes. We will modify the description to clarify.

*** L162-165 : it should be stated that the metrics are going to be used to characterize the performance of the RF for each class separately, and then globally to characterize the multi-class performance through use of a macro-average. It may be also useful to explain the term macro-average to enhance readability.**

We thank the reviewer for their suggestion. We will clarify the meaning of macro-average and specify the use of the metrics. We will add this paragraph after L162-165:

The metrics of precision and recall were used to characterize the performance of the Random Forest for each class separately. Then macro-averages of both measures were computed to characterize the multi-class performance. A macro average is the arithmetic mean computed giving equal weight to all classes, and is used to evaluate the overall performance of the classifier.

*** L 179 : which radiation is this ? Incoming longwave, shortwave ; reflected or upcoming ones from the ground ? It should be specified because it matters for the interpretation of the importance of and relationships to this predictor. Typically, reflected shortwave radiation could be a super-help to detect snow vs grass-ground, but I assume this is not the kind of radiation that was used.**

We agree with the reviewer here, thus we will specify the type of radiation used: incoming shortwave radiation.

*** Both the test set and the evaluation set share years with the training set. The effect of different years on the RF performance is assessed in Fig 7 and related text, but actually, it seems to me that the temporal transferability of the algorithm (= transferability to other, completely unknown years) was not thoroughly tested within the split-sample procedure, though this is probably one key application of this algorithm. You very wisely discuss that your results "may point to our Random Forest being robust to different climatic regimes". Is there a particular reason why you did not choose an**

evaluation set enabling an evaluation of the RF model in full spatio-temporal extrapolation mode ? Maybe related to that, how sensitive would be the performance of the RF algorithm to a moderate reduction in size of the train set, for instance to a withdrawal of the 3 complete years 2018, 2020, 2022 that would then enable an evaluation of the model in spatio-temporal extrapolation ?

We thank the reviewer for their comments and acknowledge that more details were needed on this point.

We addressed spatial extrapolation through the validation test performed on the rest-of-Italy sample. This is because this validation test includes areas experiencing a variety of climates that are only weakly correlated with those in Aosta valley (Avanzi et al. 2022). Regarding temporal extrapolation, we preferred not to withdraw specific water years and proceed with a more standard out-of-bag validation in an effort to maximize the number of training points and climate variability in our training sample. This is particularly critical for random errors, which were the least represented class and would have been further penalized by the withdrawal of complete water years. We acknowledge that the above does not represent a full evaluation of the spatio-temporal extrapolation skills of our Random Forest and will add this consideration to our revised manuscript.Here the proposed text that will be add in the Discussion section:

It is worth mentioning that, although the choice of validation dataset allowed for testing the spatial extrapolation abilities, a full evaluation of the spatio-temporal extrapolation skills was not achieved. The algorithm was trained on all the years available, with a standard out-of-bag validation in an effort to maximize the number of training points and climate variability in our training sample. No year was withdraw. It was aimed at reducing the impact of impoverishment of the sample on the least represented class of random errors.

Finally, the rationale behind the very short section 4.3 should be described either ahead of this section in the introduction, or within the section.

We agree with the need for further information as suggested by the reviewer, thus we will add this paragraph in Section 3.2:

For each year in the Valle D'Aosta dataset, a Random Forest algorithm was trained with 80% of the data and then tested on an out-of-bag sample of 20% of the same year data. The aim of this test was to investigate the possible correlation between the performance of the classification by the Random Forest algorithm and annual climate. For each year, the F1 score on the test sample was analyzed against annual mean values of features used for the classification, computing correlation factors.

*** L 320-324 : I am not sure that the addition of more data will distinctively refine the accuracy in the "errors" class, except if you use a super huge amount of new data. I would hypothesize that other**

strategies may pay out with respect to this issue, and may be either explored or at least cited if you find them relevant :

We agree with the reviewer that, despite the use of more data being likely the most straightforward option to detect rare random errors, other options (such as other algorithms) could also be a solution.

- maybe using other, pre-processed features could help, as for instance a Delta-HS = HS(t)-HS(t-1) with HS = Height of Snow. This could help detect unrealistic spikes or drops in snow data like the spikes remaining after RF treatment in Fig A1. This hypothesis is very basic to test.

We believe the solution suggested by the reviewer to be a potentially effective one; moreover, the literature suggests its feasibility in operational application (e.g. meteoIO (Bavay & Egger 2014)). Nevertheless, our aim here was to develop a fully Machine Learning procedure requiring only minimal pre-process. We would like to keep this focus for the present study. Nonetheless, we agree that coupling our Machine Learning procedure with other statistical methods could be beneficial, and we will comment on this in our Discussion section.Here a proposed paragraph:

The use of more data is likely the most straightforward option to detect rare random errors. However, other options may prove to be effective. In light of this, the proposed algorithm may be coupled with classical QA/QC procedures imposing a-priori thresholds, like those already proposed by Bavay & Egger (2014). Such procedures could, e.g., help with the detection of spikes in data using climatological snow-depth thresholds for maximum values.

- alternatively, using AI algorithms suited to dataseries and incorporating some memory, like recurrent network or LSTM, could help if fed with snow height time-series or small extracts of them.

We agree with the reviewer that the use of a deep learning algorithm may help in improving the performances of classification of rare random errors, requiring however further study. We will mention this as a future opportunity in the Discussion as explained in our answer to the first comment above.

- finally, have you considered the use of webcam images from nearby the stations within the same elevation/aspect, that could provide a simple, maybe not completely reliable snow-nosnow information, but with errors maybe not completely correlated with the RF errors ?

We believe the suggestion made by the reviewer could be interesting in specific research settings, especially because coupling different data types and data sources may enrich the algorithm performances (Karpatne et al. 2018). At the same time, webcams are not systematically installed at operational, automatic weather-snow stations in Italy and elsewhere, which significantly limits the applicability of this approach in the real world.

Edits : L5-6 : "with particular reference to differentiate snow cover from grass or bare ground data and to detecting random errors (e.g., spikes)" -¿ to detect ?

L54 : "It is clear then the necessity for a quality checking procedure, that ought to... " it seems there is a syntax issue

Fig 2 : adding the contours of Italy would be nice

L143 : end,askowleding

L143 : " the work of (Ponziani et al., 2023) in which no clear evidence of out-performance of any strategy, " It seems some words are missing

L 162 : "precision(measure of"

L 208 : I guess a "." is missing before "Fig 5".

Caption of Fig 5 : "model.In".

Fig 8 : maybe use the same vertical scale across rows, as the amplitudes are otherwise quite hard to compare esp. in the 3rd column.

References : there is an issue with the Avanzi et al 2020, 2021 and 2022 references that are always stated twice.

We thank the reviewer for their comments. We will modify the text accordingly.

References

Avanzi, F., Gabellani, S., Delogu, F., Silvestro, F., Pignone, F., Bruno, G., Pulvirenti, L., Squicciarino, G., Fiori, E., Rossi, L. et al. (2022), 'It-snow: a snow reanalysis for italy blending modeling, in-situ data, and satellite observations (2010–2021)', *Earth System Science Data Discussions* pp. 1–30.

Bavay, M. & Egger, T. (2014), 'Meteoio 2.4.2: a preprocessing library for meteorological data', *Geoscientific Model Development* **7**(6), 3135–3151.
URL: *https://gmd.copernicus.org/articles/7/3135/2014/*

Camps-Valls, G., Tuia, D., Zhu, X. X. & Reichstein, M. (2021), *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*, John Wiley & Sons.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A. & Kumar, V. (2018), 'Machine learning for the geosciences: Challenges and opportunities', *IEEE Transactions on Knowledge and Data Engineering* **31**(8), 1544–1554.