Reply to Referee #2

Dear reviewer

We would like to thank you for your review and valuable comments. We address your questions in detail below (in blue).

1. This study focused on the classification of snow height measurements based on AI techniques. However, the scientific significance was not sufficient enough, or why this work is important?
    - The snow height data are key for many research and operational applications, in particular avalanche forecasting, i.e. warning the public about imminent avalanche danger. Therefore, the classification of the snow height signal is a key part of the signal quality assessment process. QA and therefore also snow height classification, is important to provide high-quality data for research and operational applications. In the operational setting, manual checking is not feasible, and the definition of thresholding rules does not yield good results. We will make the relationship between QA and snow classification more explicit to improve clarity of the manuscript.
    - The key contributions of our work include, among others, the following three points:
        - First, the machine learning based method presented in our manuscript has advantages over traditional approaches which are described below (in the answer to your question number 2). We have chosen an existing model specifically designed to process time-series inputs and thoroughly described how to apply it in a novel setting of snow height classification.
        - The second major contribution of this work is a dataset of manually annotated snow height measurements, which can be used to further improve models for snow height classification and quality assessment.
        - Third, the proposed method strongly reduces the manual work which is required to use snow height signal in other fields, as for example in vegetation science, which is demonstrated by the case study in section 4.9.

2. What's the relationship between quality assessment and AI classification? In my view, this work doesn't aim at improving data quality, just distinguishing possible

anomalies from all station measurements. So how to reflect the advance of AI method in this work?

- We consider distinguishing between possible anomalies and flagging different behaviors using ML as a good step towards QA. We agree, however, that a full QA process might require additional work. This paper covers the snow height classification part and provides a practical solution for operational requirements as well.
- In our opinion, flagging the snow height signal which does not correspond to snow, but to e.g. vegetation, drastically improves the quality of the snow height data.
- Below we list advantages of using machine learning over other existing methods:
    - Compared to thresholding rules, which have been applied in the past, CleanSnow generalizes to most of the stations out-of-the-box, without need of threshold manipulation for every station. The pre-trained CleanSnow model, which is provided together with this submission, is straightforward to run assuming basic Python knowledge and minimum machine learning experience.
    - Machine learning methods are good in learning important relationships in the data which are extremely difficult to model explicitly using e.g. thresholding rules or by manually crafting features. Humans cannot learn and implement all the statistical relationships occurring, but only impose some simple physical laws that are bound to be suboptimal due to noise in sensors, variability in weather, etc.
    - Moreover, the source code provided as a part of this submission contains everything necessary to train new models on different datasets with potentially different sets of input features. Making changes to models based on thresholding or to the SNOWPACK model is nontrivial and much more time consuming.

3. The structure of this article is not clear enough, please improve it and maintain some important research work. Now
    - As per comments of RC1, we will revisit the structure while preparing the revised manuscript.

4. In figure 1, how to determine the training and testing stations?
    - Stations in our dataset have been selected together with domain experts and people responsible for the Intercantonal Measurement and Information System (IMIS), so that they span across the whole

Switzerland and different elevations. Stations in the test set have been selected so that they are interesting and challenging enough test cases (are at different elevations, contain signals with vegetation growth, are known to have some specific problems, etc.). This is described in Section 2.2.1 of the manuscript. We will revisit the description and improve clarity.

5. P3, lines 70-80. These descriptions should be moved to Section introduction.
    - We will resolve this issue in the revised manuscript.

6. P3, lines 80-85. Please give the physic basis.
    - We will be more explicit about the reasons why the sensor measurements listed are key factors in the revised manuscript, as already provided for reflected SW radiation.

7. Please provide a flowchart for this paper.
    - We provide a flow diagram of our model in Figure 3. We believe it covers well how the model works. We will improve the arrows in the diagram so that it better explains the flow of the data from the sensor to the output.

8. How to determine the truth data?
    - Ground truth is determined manually by experts that are involved in the work on IMIS. Domain experts can manually annotate the data while looking at HS measurements in the context of other signals measured at the same station (e.g. air temperature, snow surface temperature, reflected solar radiation, etc.). Section 2.2 of the manuscript describes how we prepared the data. We will revisit this section and improve clarity wherever appropriate.

9. P6, lines 110-135. This paragraph should belong to methodology, thus, the title '3 Machine learning based snow cover classification' is not suitable. This section should be method or methodology.
    - We will adjust the title of Section 3 in the revised manuscript.

10. P8, '4.1 dataset' should be introduced in methodology section, not here.
    - We will make changes accordingly in the revised manuscript.

11. It is difficult for me to understand the logic and structure of this study.

- We will modify the structure and improve the clarity of the text. Allow us to summarize our work in a few points:
    - Our study is motivated by the fact that a snow height sensor measures the height of any underlying object, independently of whether it is snow or not. This poses challenges for researchers interested in the snow height signal, but also gives the possibility of using snow height signal in vegetation science during summer periods. Other use cases, such as calculation of SWE, are included in the Introduction section.
    - Automatically disentangling snow from vegetation or other surfaces is non-trivial task, which has been in the past approached with sets of thresholding rules. These thresholding rules are, however, often station-specific and require therefore still a lot of manual work.
    - Machine learning would be of help, however, datasets to train and evaluate machine learning models are missing.
    - We therefore first created a new dataset which allows us to train and evaluate machine learning models for snow height classification.
    - Subsequently, we provide guidelines and a thorough analysis of how to apply machine learning models to the snow height classification problem. This is followed by presenting results of our model.
    - As an outcome, we provide a new model together with its source code to the community – CleanSnow – which can be either used out-of-the-box, already pre-trained, or trained from scratch with different data or a different set of input features.
    - We also provide the manually annotated dataset, which allows for further development in the field, as well as other analyses, which require snow height signal cleaned of spurious measurements, such as grass or bare ground.

Best regards,

Jan Svoboda, on behalf of all co-authors