Overall Evaluation

The manuscript presents a valuable and well-structured reconstruction of snow cover using a hybrid gap-filling framework that combines decision-tree and machine learning approaches. The long-term dataset and the integration of multiple satellite sources represent a significant contribution to snow monitoring and hydrological applications.

The methodology is generally sound, and the results are relevant and promising. However, several methodological aspects require clarification to improve transparency and reproducibility, particularly regarding model configuration, data processing choices, and evaluation procedures. In addition, some figures and descriptions would benefit from clearer explanations to facilitate interpretation.

Overall, the manuscript is of good quality and suitable for publication after minor revisions addressing the points raised below.

Detailed Comments and Suggestions

Comment on Section 2.2 (snowMapper model overview):

The model overview is clear and well structured, and Figure 2 is informative. However, this section remains largely descriptive and would benefit from additional clarification.

Thank you for your comment. We acknowledge your point regarding Section 2.2 of our manuscript being largely descriptive. This decision was made intentionally, in an effort to deliver a brief overview of a relatively complex model, to accommodate the varied interests of a wider audience. This way, readers who are interested in the results of the climatological analysis rather than the model, would be able to get a quick understanding about how the model operates before proceeding to the following sections. On the other hand, readers interested in the technical aspect of snowMapper, will be able to get high detail from the following sections/subsection of the methodology.

Specifically:

- The novelty of snowMapper relative to existing approaches is not clearly articulated. Please highlight the key contributions and what differentiates this framework from previous methods.

  We will articulate more clearly the novelty of snowMapper relative to existing approaches presented in the Introduction.

- The term “physics-informed” should be better defined (e.g., are physical constraints explicitly enforced, or are only physically meaningful variables used?).

  We will add more information regarding the term “physics-informed” machine learning.

- The “assimilation” step appears to rely on direct replacement of observations rather than a formal data assimilation approach; this should be clarified.

  While this is described in section “2.4.4 Assimilation” (lines 272–275), we appreciate your comment. Our approach does not constitute formal data assimilation, as it involves direct replacement of simulated pixel values with available clear-sky satellite observations without statistical weighting. We will revise the manuscript to replace the term “assimilation” with “direct insertion of observations” (or similar wording) to more accurately reflect the method.

- The workflow could be described more explicitly (e.g., step-by-step process or simplified schematic), as Figure 2 is relatively complex and not always easy to follow.

  We will modify section “2.2 snowMapper model overview” in order to link it better with the different steps displayed in the schematic of Figure 2, thus helping the reader to follow.

Comment on Section 2.3.1 (Satellite imagery and MODIS processing):

The satellite data processing is generally well described; however, several methodological choices require further justification:

- The use of a fixed NDSI threshold (NDSI > 0.4) is not justified. Since optimal thresholds can vary depending on region, illumination conditions, and land cover, please explain how this value was validated for the Greek mountains. A sensitivity analysis or regional calibration would strengthen the approach.

  While optimized thresholds have been shown to increase accuracy in the classification of binary snow cover from optical satellite sensors, this calibration requires a large amount of *in situ* observations (Notarnicola, 2020; Poussin et al., 2023). Unfortunately, Greece has no such network of snow depth stations, and therefore such an optimization is not yet possible. The use of a steady NDSI > 0.4 threshold was selected due to its ability to still offer high classification accuracy and maintain consistency across time and space. Furthermore, while testing different thresholding schemes falls outside of the scope of this present study, the model does allow users to (a) choose from a list of five threshold-based snow binarization schemes, (b) configure their own thresholds, or (c) add a new custom snow binarization scheme that best fits their research needs.

  Notarnicola, C.: Hotspots of snow cover changes in global mountain regions over 2000–2018, Remote Sensing of Environment, 243, 111781, https://doi.org/10.1016/j.rse.2020.111781, 2020.

  Poussin, C., Timoner, P., Chatenoux, B., Giuliani, G., and Peduzzi, P.: Improved Landsat-based snow cover mapping accuracy using a spatiotemporal

NDSI and generalized linear mixed model, Science of Remote Sensing, 7, 100078, https://doi.org/10.1016/j.srs.2023.100078, 2023.

- The 50% FSC threshold used to binarize MODIS data also appears empirical. Please clarify whether this threshold was calibrated or evaluated against alternative values.

  The FSC threshold used to binarize MODIS was derived from the literature (lines 151-152), where it has been widely used (Notarnicola, 2020; Shen et al., 2025). While we did not perform any further calibration or evaluation, as this would fall outside of the scope of this study, as we mentioned above, the model does allow users to configure their own custom thresholds. Furthermore, in the case of MODIS data, we would like to highlight that it is only used in an auxiliary capacity. This means that MODIS data will only be used in pixels that (a) do not have satellite observations, and (b) do not satisfy the first temperature-based decision-tree gap-filling criterion (lines 259-261).

  Notarnicola, C.: Hotspots of snow cover changes in global mountain regions over 2000–2018, Remote Sensing of Environment, 243, 111781, https://doi.org/10.1016/j.rse.2020.111781, 2020.

  Shen, Y., Wang, X., Zhu, R., Che, T., and Hao, X.: A Downscaling Algorithm for Snow Cover Extent Over the Tibetan Plateau Based on a Similar Conditional Probability and Otsu's Method, IEEE Transactions on Geoscience and Remote Sensing, 63, 1–14, https://doi.org/10.1109/TGRS.2025.3543433, 2025.

- MODIS data were resampled from 500 m to 100 m using bicubic interpolation. Please justify this choice, as such resampling does not introduce new spatial information and may lead to smoothing artifacts. Why was this approach preferred over simpler methods (e.g., nearest neighbor) or dedicated downscaling techniques?

  We chose bicubic resampling over nearest neighbor not despite its smoothing effect, but because of it. We believe that at the 100 m scale, smoothing the hard edges of 500 m MODIS grid cells before extracting binary snow cover values from them, allows for a more realistic representation of snow cover at our final spatial resolution. At the same time, although, as you correctly point out, new spatial information and a more dedicated downscaling technique would offer a more accurate representation, we have seen bicubic resampling used in a similar context with good results in a much less computationally demanding approach (Kollert et al., 2024; lines 152-154). Although a more dedicated MODIS downscaling module might be developed as part of future versions of the snowMapper, we would like to emphasize that MODIS is currently only used at an auxiliary capacity.

  Kollert, A., Mayr, A., Dullinger, S., Hülber, K., Moser, D., Lhermitte, S., Gascoin, S., and Rutzinger, M.: Downscaling MODIS NDSI to Sentinel-2 fractional snow cover by random forest regression, Remote Sensing Letters, 15, 363–372, https://doi.org/10.1080/2150704X.2024.2327084, 2024.

- The use of MODIS Terra only is not justified. Combining Terra (MOD10A1) and Aqua (MYD10A1) products is commonly used to reduce cloud contamination and improve temporal coverage. Please explain why Aqua data were not included, particularly given the importance of gap-filling in this study.

  While MODIS Aqua may have provided additional snow cover information, we decided against its integration for three reasons:

  1. Overpass time in our study region is ~13:30. This is inconsistent with MODIS Terra as well as the higher-resolution missions (Landsat, Sentinel-2), which all pass around 10:00, allowing us to calculate daily aggregates of the meteorological conditions around that time, in order to obtain the most up-to-date meteorological information for each pixel (lines 164-166).
  2. In a comparison of MODIS Terra's and Aqua's snow detecting capabilities over the Pyrenees, the latter was found to be less accurate (Gascoin et al., 2015).
  3. As mentioned earlier, MODIS data are used only in an auxiliary capacity, during gap filling, and therefore we believe that the Terra collections alone are able to satisfy the needs of the model at that stage.

  Gascoin, S., Hagolle, O., Huc, M., Jarlan, L., Dejoux, J.-F., Szczypta, C., Marti, R., and Sánchez, R.: A snow cover climatology for the Pyrenees from MODIS snow products, Hydrology and Earth System Sciences, 19, 2337–2351, https://doi.org/10.5194/hess-19-2337-2015, 2015.

Comment on Section 2.3.4 (In situ data):

The training data are derived from stations in the Alps and Pyrenees rather than from Greece. Please justify the transferability of the model to Mediterranean snow conditions, which may differ significantly.

As we point out in section "2.3.4 In situ data" (lines 199-200), no data are available for Greece. However, this is what inspired us to create our 'physics-informed' machine learning algorithm, which simulates snow cover conditions by performing iterations that take into account the continuously cumulating meteorological conditions, including heating, cooling, precipitation, and continuous snow cover days. To ensure that the machine learning algorithm takes into account only physical variables, we have removed all geographical ones (lines 224-225). We believe that the accuracy assessment of the model provides satisfactory evidence that this approach does indeed work, and can be transferable to other regions. Having said that, users are of course still able and welcome to create their own classifier with data from their own study region, or a region with similar climate conditions to theirs, enabling the model to always provide the most accurate simulation (modules for preprocessing station data & training a machine learning classifier are included in our code repository to facilitate this process).

Comment on Section 2.4.1 (Machine learning classifier):

The Random Forest hyperparameters (e.g., number of trees = 30, minimum leaf size = 1, bag fraction = 0.5) are specified, but their selection is not justified. Please clarify how these values were chosen (e.g., cross-validation, sensitivity analysis, or empirical testing).

In the case of number of trees, we used 30, following an error matrix sensitivity analysis. In the case of minimum leaf size and bag fraction, we used the default settings recommended in Google Earth Engine, due to no information on these appearing in our search in the literature for applications of machine learning methods in snow science.

Comment on Section 2.4.5 (Final output):

The computation of monthly aggregates is not clearly described. Please clarify how daily snow cover is aggregated to monthly values (e.g., mean, maximum, or fraction of snow-covered days). In addition, the method used to convert daily binary snow maps into monthly fractional snow cover (FSC) should be explicitly defined.

In this step, daily binary snow cover values are subject to two consecutive aggregations:

- First, an aggregation by mean is applied in the temporal domain, ultimately describing the fraction of snow-covered days in a given pixel (i.e., temporal FSC at pixel scale), and
- Second, an aggregation by mean is applied in the spatial domain, ultimately resulting in a final value per mountain/study area, per month (i.e. spatio-temporal FSC at study scale).

Indeed, this information was missing, and we thank the reviewer for pointing it out. We will provide further clarification on the updated manuscript.

Comment on Figure 4:

Figure 4 is not easy to interpret. The definition of “fraction of pixels” is unclear, and it is not specified how these monthly proportions are computed. Please provide additional information in the figure caption. In addition, the machine learning contribution appears relatively constant over time; please clarify how this fraction is computed and whether it varies across years.

By fraction of pixels, we simply refer to the percentage of pixels that each month came from clear-sky satellite observations (Landsat/Sentinel-2), or were gap-filled using decision trees, or machine learning. These metrics are derived from flags during the daily snow cover reconstruction, which are then aggregated on a monthly scale. Therefore, 100% = all pixels of the area, across all days of that given month.

Regarding the contribution of the machine learning gap-filling, the graph correctly shows that more weight is given to decision tree gap filling and consequently less to machine learning after 2000, due to MODIS (decision tree gap filling step No.2 - lines 262-263). This variation across years is described in lines 353-355.

We will provide further clarification, as requested.

Comment on Figure 5:

Although Figure 5 describes the temporal aggregation of the metrics, the evaluation methodology is not fully clear. Please clarify what datasets are being compared (e.g., model outputs vs. observations) and whether the evaluation is performed at the pixel level over the study area.

As described in section “2.5 Model evaluation & bias correction”, first a pixel-level true positive/false positive/true negative/false negative classification is given to the model data, by comparing pre-assimilation results with any clear-sky observations (Landsat/Sentinel-2). We do not need to use a subsection of those observations for validation, as they are still independent during the evaluation step – replacing simulated values with the observed clear-sky ones comes after that evaluation. Once the model run is complete, during a postprocessing step, all tp/fp/tn/fn values are aggregated in time, and space (over the study area), and used to calculate the monthly accuracy, underestimation, and overestimation metrics using Eq. 11-13 (lines 290-298).

Thank you very much for supporting the improvement of our paper! All your comments are very valuable.