## General Response

We would like extend our gratitude to both reviewers for their valuable feedback and suggestions, which have allowed us to scrutinize our work further and refine it accordingly. We have agreed to most of their feedback and incorporated in the revised manuscript wherever possible. The review, we believe, has helped to improve the manuscript.

Specifically, sensitivity analyses on different thresholds and calibration dates in the study regions have been added to provide new insights regarding our approach in the study regions. Below are our responses to the comments/feedback.

## Response to Juraj Parajka

We thank Juraj Parajka for his continued feedback and constructive critics on our work, which was highly imperative to streamline our manuscript as well as the presentation. Below are our responses to his comments.

1. - *The motivation and context of the study can still be improved. In its current form, the story is not focused and clear. For example, the Introduction states, "Various modeling and measurement techniques are currently in practice . . . " and "Prior studies on the comparison of snow models have highlighted the higher reliability of physically based approaches". It is not clear how these statements support the need to examine simple degree-day approaches in the study? Or, it is not clear how the referenced studies about multiple objective calibrations connect to the study objectives. How are calibrations using scatterometer or evapotranspiration data related to the context of this study? The Introduction can still be more focused and improved.*

   The mentioned statements were put into the introduction section as a problem statement to bolster our methodology which uses readily available data throughout the world in contrast to the data intensive and site specific complex approaches. We agree that multiple objective calibration part was longer than required, but they were put there for an overall picture of how MODIS is being used in hydrological predictions. The latter part was a suggestion in the first round of review. However, the introduction part has been streamlined and made more focused in the revised manuscript.

2. - *Formulation of the novel scientific contribution and research objectives need to be more precise. Is the main objective development of a flexible snow-melt module? I do not think so. Or at least, the results, in their current form, do not refer to a new flexible snowmelt routine.*

   The feedback was well-noted and the novelty and objectives sections are clarified in the Introduction section.

3. - *The formulation of the main novel contribution is also somewhat general and vague. In general, the stepwise calibration of hydrologic models (i.e. calibrating the snow module*

*first, the runoff generation in the next step) is not new. Also, using MODIS for stepwise calibration was already examined and evaluated in previous studies (see, e.g. Szeles et al., 2020, DOI:10.1029/2019WR026153). In my opinion, the novelty is mainly in proposing/testing/examining/evaluation standalone spatial patterns of snow cover (from MODIS) for calibration of the conceptual snow module. The formulation and presentation of the generality of findings can be improved. The study applies numerous assumptions and thresholds and it is not clear how these affect the results and how to set up the thresholds/assumptions in future studies. For example, the study assumes NSDI threshold equals 1 for snow cover classification. The previous versions of MODIS were based on a globally fixed threshold equalled 40. The recent study of Tong et al (2020, https://doi.org/10.1016/j.jhydrol.2020.125548) examined this threshold and found that 40 works well for Austria unless some more detailed information is available. But the threshold=1 likely does not provide the best mapping accuracy compared to observed snow depth. How was this threshold determined? What is the advice for readers in using this threshold in future applications? Other thresholds are, for example, used for snow cover mapping evaluation. Why 2.5mm and 0.5mm? Why 60% of valid pixels? How sensitive are the results to these settings? How should the readers setup these thresholds in future studies? The most important but unclear point is, in my opinion, the definition of which day to use for model calibration? Why selected dates? Again how sensitive are the results to this selection? What to advise to the readers in future applications?*

Following this suggestion, we did sensitivity analyses on different thresholds and sensitivity of dates. The results are available and discussed in the revised manuscript.

4. - *The need for cloud removal techniques is not clear. If only one image is used for model calibration, is this method/step needed? Why this order of steps?*

   The cloud removal techniques were employed to obtain as much data as possible for calibration, using simpler yet effective steps. The model analysis have been changed to sets of images within a season, so the cloud removal technique comes into play in the revised analysis. However, for a single image, the selection can be done so that the image with least amount of clouds can be picked.

5. - *The description of the calibration procedure does not provide sufficient detail to reproduce the experiment. Particularly the ROPE approach and uncertainty analysis can be described in a more detailed way.*

   The ROPE steps were added into the revision. Please refer to Lines 298 - 310.

6. - *The comparison of different degree-day approaches is very brief. In my opinion, it is a missed opportunity to describe and evaluate the differences between the approaches in more detail. It is not clear how the results affect the date selection for model calibration. But are the pixels with the better model performance of model 6 linked to accounting for the radiation input? I will be very interested to see a much more extensive comparison of the approaches. It will be very interesting to see why and where are some approaches more robust than the others. As it is already commented above, the generality of findings based on only one image*

*is rather low and a more detailed assessment will be very interesting to see. For model evaluation, it will be very interesting to see also whether and which models overestimate or underestimate the MODSI snow cover and where?*

We would like to thank you for this suggestion. The model performances in terms of over-, under- and total estimation errors were scrutinized under different elevation zones for both BW and Switzerland and added in the revised manuscript.

7. *- The section about the results of transferability of model parameters does not read well. The method is not introduced in the methods section and it is also not clear how this is linked with the main objectives of the study. Perhaps this part can be a separate study as the forecasting is a quite specific topic and the manuscript in its current form does not provide a clear context for this analysis. I think that more detailed comparisons of the degree-day models and sensitivity analysis of how different thresholds and calibration date selection affect the results will provide already a nice and compact analysis worth to be published.*

   The mentioned section was omitted. This was to show the spatio-temporal flexibility of the calibration which can be presented in a separate study.

8. *- Finally, the Discussion can provide some advice on setting up the analysis (definition of the thresholds, selection of dates) in different regions or time periods. Such lessons will improve the presentation of the generality of findings.*

   Thank you for the suggestion. The discussion and conclusions have been revised.

## Response to the Anonymous reviewer

We would also like to thank the anonymous reviewer for taking the time to carefully read our paper and providing critical remarks and suggestions. We have tried to address all the queries and incorporate his/her valuable suggestions. Below are our responses:

**General comments**

*- The main objective of the study was to develop relatively simple extended degree-day snow models driven by freely available snow-cover images. Authors see the novelty of their research in independent calibration of the snowmelt models on snow cover images which allows standalone estimation of associated parameters and thus a better representation of the snow processes. Output from these snow models were later used as input data in modified HBV model for streamflow simulation in five selected catchments in Germany and Switzerland.*

*First, it should be noted that the paper has been reviewed by two reviewers before and authors created a new version of the manuscript. After reading the reviewers comments and authors replies, it becomes clear that the study has been significantly revised. Nevertheless, I did not base my review on the earlier reviews, and rather tried to comment the revised study without bias.*

*In my opinion, authors did an interesting work. I certainly agree that the focus on testing different variants of degree-day models and their calibration against snow cover area using MODIS data is important, although not fully novel. Similarly, the de-coupling of snow routine from the selected hydrological model and its standalone calibration might bring some new insight on calibration procedures and model equifinality, although many hydrological models are nowadays calibrated using more variables next to streamflow (SWE, snow cover, groundwater levels etc.). Therefore, I found the study important and particularly novel. I thus agree with previous reviews that the study is worth publishing in HESS. However, I have several specific comments and questions regarding the methodology approach and quality of presentation. These comments should be carefully addressed before I can recommend the manuscript for publication. I only partly checked the original manuscript (before the revisions), so I hope I will not be in contradiction with initial reviews.*

- We thank you for your kind response and critical review.

**Specific comments**

1. *- In my opinion, introduction section still needs partial improvement. Especially part within lines 45-66 looks like a list of studies containing just a short description without synthesis of individual information and results. I read the comments of the reviewers in the first round of reviews and their concerns regarding the introduction as well as authors response. Therefore, I do appreciate that authors extended the introduction section, but in my opinion, it resulted only in a partial improvement. Although I understand previous authors argumentation about writing long reviews with citing unrelated studies (and with a deep respect to the second author experience), I still think that it should be possible to write a relatively short and focused introduction section which shows the state of the art of the topic and research gaps which helps the readers to understand what's going on in the topic. Therefore, I would like to encourage the authors to improve the introduction section once again and to better relate*

*individual information to each other.*

As per the review, the introduction section has been modified.

2. - *Two study catchments, Reuss and Aare have some percentage of area covered by glaciers, whereas the glaciation cover for Aare is relatively high (15.5%) and thus the glacier melt considerably influences catchment runoff. Was glacier routine somehow included in the HBV model structure which authors used to simulate streamflow? I did not find this information in the text and thus it seems that glacier routine was not used. If true, I am not sure to what degree the simulations reflect the real observed values (at least for the Aare catchment). Could this somehow influence results interpretation? While I think the missing glacier component is not a problem for snow models and related results interpretation, I think it might be important for interpretation of results related to "standard" HBV and "modified HBV" (although authors assessed NSE values just for cold season months, I assume the simulation itself were done over the whole period 2010-2018). The most straightforward solution would be to include the glacier component for the two glaciated catchments (at least for the Aare catchment), or at least I would like to ask the authors to carefully address this point in discussion section.*

   The objective of this study was to assess the performance of MODIS based calibration on snow accumulation and melt processes. We did not evaluate the Glacial-melt due to the MODIS limitations in identifying the glaciers. Both hydrological models were calibrated without the glacier component for uniformity. Nevertheless, we consider your suggestion as a very pertinent feedback. We added some lines in the discussion section (Lines 570 - 574).

3. - *L 313-317: It is not fully clear to me how exactly authors proceeded when creating the variants of a hydrological (HBV) model. If I understood correctly, authors created six HBV model variants for each catchment (which were named as "modified HBV"). These six HBV variants did not contain snow routines since snowmelt simulations resulting from previously defined six snow routines were directly used as input data to the HBV model. Last variant (seventh) was just a "true" HBV with its snow routine in its original structure (which is partly different than other snow routines, due to, e.g., including water holding capacity and refreezing). Is it right? If yes, then please, consider reformulation of the respective method part to be clearer. The fact that you used HBV snow routine to compare it with other six snow routine variants became clear only from results section to me (mainly from Fig. 7). Therefore, to improve the clarity of methods section, I would suggest modifying it such as you will describe seven variants of the snowmelt model (Model 1 – Model 7); the six you already have and the last (seventh) representing the original HBV snow routine. The seventh snow routine variant should be calibrated in the same way as the other six variant and comparison will be plotted in Fig. 7. In my opinion, using this procedure would make it clearer how well/badly your snow routines perform compared to the original snow routine structure implemented in HBV.*

   Thank you for this suggestion. We have revised these sections accordingly. We presented six snow-melt modules, out of which the radiation based model was selected for further analysis as it reported the lowest Brier-score. This model was used to simulate the melt

using the best parameter vector and this melt was included in a modified HBV (without snow routine) as a standalone input. The other hydrological model was the standard HBV model. These two hydrological models were then calibrated on discharge at catchment level. To differentiate the differences of calibrating on discharge only, we subsetted the HBV's snow routine parameters (1000 best) and used it in HBV's snow routine to simulate the snow-cover distribution and calculate the 1000 Brier-score values, which were compared with the radiation based model's Brier-scores as shown in the mentioned figure (Figure 9 in the revised manuscript). The point here was to show how calibrating on discharge can over-compensate during individual processes simulation. Nevertheless, we have added some clarity in the Model calibration and validation sections.

4. - *Related to comment above, can be the differences between "modified" and original HBV (shown in Fig. 9) attributed to separated calibration of the snow routines or rather to different model structures of the snow routines or both? Can you differentiate between these two influences? Maybe it would be methodologically clearer, if you calibrate the "modified" HBV model against discharge separately for all seven snowmelt inputs (as describe in my comment above) as a first step (this is what you probably did). This way you can better compare which snow routine performs better when implemented in the HBV model. In the second step, you may select just "modified" HBV model with snowmelt inputs from separately calibrated HBV snow routine (snow model variant 7 as suggested in my comment above) and compare it with calibration of the original HBV model. This way, the first step shows the differences between individual snow routine structures (including original HBV snow routine), the second step shows the advantages/disadvantages of separate snow routines calibration compared to "normal" calibration (just against discharge) of the complete HBV model.*

   As mentioned in the discussion section and the response above, our goal was to implement a MODIS based calibration on snow-melt modules, and specifically not to assess the performances of the widely used hydrolgical models like HBV. We just wanted to reiterate our finding that calibrating a hydrological model solely on discharge can have compensating effect on individual sub-processes. We did not use the snow-routine from HBV for snow-distribution calibration. However, it might be good step for future implementation of our approach where we can categorically present the findings in the manner you suggested. For this paper, we have just evaluated the snow-distribution simulation based on a discharge-based and a distribution-based calibration.

5. - *Important question is also whether the model performance should be assessed using NSE only. Current best practise is to use more criteria to make the results more robust. Would results interpretation change in case you will use different objective criteria (logarithmic NSE, volume error, etc.)? With this comment I come a little back to what was mentioned by Reviewer 2 in the first round of reviews, and it is to what degree the values of a single objective function (NSE in this case) could really tell us whether the one model is better than another (especially in case of small differences).*

   This is a valid point. However, as we have pointed out earlier and in the first round of reviews, the objective here was to evaluate the snow-cover based calibration approach and its outputs.

Different objective functions can be added, but the aim was to see if it would add value to the underlying sub-processes. Maybe in the future studies, we can add more criteria including a multi-variable constrained calibration to evaluate the performance of the hydrological model. However, we wanted to show that the parameter set required for calibrating a hydrological model becomes smaller when using a step-wise approach of calibrating the snow module separately. This allows us to identify a more robust set of parameters along with the parameters related to the snow-processes estimated for a 'right reason' with a better representation of underlying snow processes, thereby gaining similar or better performance in terms of discharge simulation. This gain, albeit smaller or even similar, is a gain nonetheless arising from a better representation of the inherent snow accumulation and melt processes.

6. - *In my opinion, the discussion section should be improved since it seems to me that it is not clearly linked with results. It is certainly the matter of personal preferences, but I prefer using the results section just for results description and basic interpretation related to a single figure/table described, and everything which goes beyond a single figure interpretation (it means the results interpretation in a wider context of all your presented results and other literature) should be placed in discussion section. In this respect, the discussion section should be comparable to results in its extend and it may follow (not necessarily) similar structure as the structure of result section.*

   The discussion section has been modified to better connect with the results.

7. - *Overall, the text is often difficult to follow since there are a lot of unclear statements, and it is not often clear how exactly authors proceeded (see also points above). This is also the case of some of figure and tables which are not clearly linked with the text, and they do not provide the reader with all needed information, such as informative caption or correct legend. Please see also my detailed comments in the list below. Maybe my comments and criticism stem just from these unclear issues rather than from real problems in methodological approach and results interpretation. Anyway, I would like to encourage the authors to go carefully through the text and try to make the text clearer and more consistent.*

   Thank you for this comment. We have tried our best to present our findings with more clarity in the revised manuscript.

**Technical corrections**

1. - *L15: Please use "Nash-Sutcliffe efficiency" instead of NSE in abstract.*

   This has been corrected.

2. - *L17: Two full stops at the end of the sentence.*

   This has been corrected.

3. - *L 88-93: I would omit this paragraph since I found it too general. In fact, this is how all scientific papers are organized, thus, the specific description is not necessary here.*

   We have omitted this paragraph in the revision.

4. *Fig. 1: Legend for elevation for the three inset figures (study area) seems not correct to me. As far as I can recognize, the colour scale is continuous in these small inset maps, thus the legend should be displayed accordingly (there aren't only four or five colours in figures, right?). Besides, in case of intervals are used for the colour scale (which is, to my knowledge, the best cartographic practise), the legend should be displayed without spaces between individual coloured rectangles. Additionally, use "Elevation [m a.s.l]" for the respective legend caption and add graphical scale (for all inset maps and the main map).*

   The figure is updated in the revised manuscript.

5. *- L 99, 101: please use "m a.s.l." instead of "masl" (please check also other potential occurrences in the text whenever relevant.*

   This has been done.

6. *- L 101: The highest point of Switzerland is 4634 m a.s.l. (Dufourspitze, Monte Rosa massif). This should be also reflected in legend of Fig. 1 (the last number in the legend). In this context, I would prefer the "real" highest point rather than the highest cell of the DTM raster you used to create the map.*

   This has been modified.

7. *- Please use correct unit conventions (km2, m a.s.l.)*

   This is corrected.

8. *- Section 2.2: Why not to use official Meteoswiss and DWD gridded products (which are available for much finer spatial resolution than your interpolations)? Was it because you needed also Tmax and Tmin while official gridded products were created only for daily Tmean and P? Or was there any other reason? Please clarify shortly.*

   We wanted to align our interpolation to MODIS schema and we needed Tmax and Tmin as well, as interpolated grids. Furthermore, we wanted to test different interpolation techniques especially for precipitation. So for uniformity, we did not use the Swiss or German gridded products.

9. *- L 182: Authors mentioned that their "Basic Degree-day model" (Model 1) is the same model as implemented in HBV. However, this is not fully true since the snow routine implemented in the HBV model accounts also for liquid water holding capacity (which delays the water release from snowpack and thus directly influences daily SWE values) and refreezing (which has usually only a small effect on SWE calculation, at least at seasonal temporal scales). Please also look on my specific comment related to "Model 7").*

   Thank you for the suggestion. We have removed the statement. However, we did find that the refreezing component did not pose significant changes in terms of snow-cover in this approach.

10. *- L 197-198: "falling on the snowpack". While I fully agree that topography (e.g., slope orientation) is important for snowmelt distribution, I would not say it also impacts snowfall temperatures (the shortwave radiation do not much differ between north and south facing*

*slopes during snowfall events). Therefore, I think the Model 4 doesn't make much sense. Nevertheless, I accept authors decision to include it.*

Thank you for the feedback.

11. *- L 257: "grid" instead of "gird".*

    This has been modified.

12. *L 262: I would prefer "seamless" numbering, it means that title "3.1" should follows immediately after title "3". Therefore, I suggest using some title (3.1) for general methodological approach (including Fig. 2 and the list of parameters), continue with title 3.2 named something like "Snow routines variants" (or similar) followed by "3.3 Data requirement of the models" etc.*

    The numbering has been modified in this section.

13. *- Chapter 3.4 would perhaps better fit to discussion.*

    We agree with your statement, but we decided to keep it as it is as we wanted to present before results, how separate calibration helps in reduction of uncertainty. Also regarding the discussion section, we did not want to make it longer, but we have summarized this part in the discussion part.

14. *- L 350 and 362: There are no Figs. 4a and 4b. Or, maybe better put a) to g) labels to individual panels of Fig. 4.*

    This has been revised in the captions.

15. *- Fig. 5: Please add colour scale captions.*

    This figure is omitted from the paper.

16. *- 359: Typo in "efficiency".*

    We wanted to use the term 'efficacy' here.

17. *- L 408: Delete "below" after "Fig. 7" (the figures are placed during post-production and may be placed elsewhere, not necessarily "below").*

    'Below' was omitted.

18. *- Fig. 6: Is the colour scale needed? If I correctly understood, colour scale used here just follows the parameter values, but the parameters are of different physical meaning and different magnitudes thus not comparable to each other. Therefore, I think the colour scale is rather confusing in this context. It would be also good to add units for each parameter. Additionally, please make more informative figure caption. Figures and their captions should be understandable even without the related text. For example, which model variant is shown here? Why the last line represents specific date rather than year as other lines?*

    The figure 6 has been omitted in the revised manuscript. We have tried to make the captions more informative in the revised text.

19. - *Fig. 7: Same as above, please make the Figure caption more informative. Among others, what scores are included within individual plots? Those resulted from 1000 parameter sets? What is represented by the width of individual plots? Please provide clear description in the figure caption. Fig. 8, 9, Table 4 and 5: Same as above, please provide more informative figure caption. For tables, it is not clear what numbers are shown (the fact that it is Brier scores are mentioned only in the text).*

    Figure 7 (now fig. 9) shows the dispersion of Brier scores from the snow-melt model and the HBV snow routine in the different catchments. Figures 8 (now 10) and 9 (now 11) show the parameter ranges and dispersion of NSEs in different catchments. Tables 4 and 5 have been omitted. This information has been added to the respective captions.

20. - *L 425: Typo in "hydrological". Besides, perhaps "Hydrological models validation" would sound better.*

    The typo has been edited.

21. - *L 426-431: This part would fit better to methods section.*

    This is also added in the Methodology section.

22. - *Fig. 10: Why this figure is actually shown here? And why specifically the Horb catchment and the season 2012/13? Please explain it better in the text. I understand that this might be an example to support your conclusion of using separate calibration for snow routines and then for the rest of a hydrological model. However, without any other information it looks like you selected the "best" result to support your conclusion, but without any evidence that also other catchments/years performed similarly well or badly. I would strongly suggest either to put this figure in wider context or remove it. Fig. 10: Y-axis description should contain units.*

    This figure was shown as an illustration of how the approach works in simulating winter flows. However, since other results support the conclusion, we have removed the figure to avoid long manuscript.

23. - *All figures: Besides specific comments above, please check the font size in all figures.*

    We have modified the figure font size.