We thank the reviewers and the editor for their constructive second round of reviews. We have revised the manuscript to further improve its clarity and readability, as most suggested changes were textual. We provide a point-by-point response to the reviewers' comments. We are confident the manuscript is now ready for publication and we hope our second round of revisions will satisfy the reviewers and the editor.

Thank you.

The authors.

In black: Referee comments.
In blue: Authors' response.

Line numbers of our responses refer to the TRACK CHANGES MANUSCRIPT

REVISION ROUND 2

Anonymous Referee 1:

Thanks to the authors for considering my suggestions. My comments have been addressed substantially by dividing the data into training and validation sets, adding a flow diagram, and conducting a sensitivity analysis on the temperature threshold.

One additional minor suggestion is that the sensitivity analysis could be presented more clearly. Currently, the description of the method and results for this analysis is overly brief, and only by looking at Figure S7 can a reader get the specific findings. Adding a few sentences describing how the analysis was conducted would be helpful.

We thank the referee once again for reviewing our revised manuscript. We are happy to see the reviewer is satisfied with our revisions and we have gladly addressed the last minor suggestion. We have therefore included a small section with an independent subheader named "Sensitivity analysis" in both the methods section and the results section (lines 246-263 and 413-424), to show the reader clearly and transparently the results of our sensitivity tests.

Anonymous Referee 2:

In its present form, Fontrodona-Bach and co-authors present a manuscript which incorporates the first round of reviewer comments satisfactorily. However, at this point I find that the paper is overly lengthy rather than being a curated summary of the key, novel results. Therefore, I believe another round of revisions is required. My comments below respond to changes made and offer suggested changes that can help streamline the message.

We thank the referee for taking time to review our manuscript and constructively suggesting further improvements. The length of the paper is a product of the original analyses and the additional requested improvements to this during the revisions, including: greatly expanded sensitivity analysis, more detailed threshold justifications, additional validation approaches, and a regional clustering analysis. We believe these improvements provide a clear and impactful manuscript, and note that the overall length is not unusual by common HESS standards, though as noted below, we have taken specific suggestions to shorten or streamline the manuscript.

Abstract:
As it stands, there are too many details in the abstract. In my view, the main value of this study is that it offers an observation-based assessment of the adequacy of a classic temperature index model (Ta , α , Tm) to model snow on a hemispheric scale and that it offers an assessment of regional variability in these parameters for optimized melt-season metrics. After giving some background on the topic, application, and observational dataset used, the main focus could be a qualitative summary such as in L26-31 -- summarizing regional performance -- and some quantitative results related to the melt factor and its spatial variability.

The reviewer is right that the abstract became long and we have therefore taken the reviewer's suggestions to shorten it. Besides the two specific comments below, we have removed the first 4 lines, which provided a too broad background. We have included a more quantitative summary of the mean melt factors and their variability in the abstract.

- L15: I suggest "precipitation phase threshold" as a replacement for "snowfall-air temperature threshold" because it is clearer.

We have rephrased it to "precipitation-phase threshold", as suggested. We have also reworded the sentences around line 15 to make them shorter, while providing a more quantitative summary of temperature thresholds capturing snow accumulation and snowmelt.

- L22-23: This is an example of an overly-detailed sentence. At this point, the authors' nomenclature of "mean melt rates" vs" melt factor" is unclear/uninterpretable by a new reader.

We have removed this sentence, as we agree it may be confusing and the performance of melt rates is already stated a couple sentences after this.

Introduction:

- I suggest omitting snow modelling intricacies that are outside the scope of the present study/TIMs and instead focusing on the main points in paragraphs 1 (there is a need to assess snow model performance), 3/6 (simple models are an important class of snow model, i.e. TIMs), and 5 (we have physically-based interpretations of melt factor values in TIMs).

We are not sure which snow modelling intricacies in the introduction are outside the scope of the study. There are minor comparisons to physics-based snow modelling that were added in response to requests by the reviewers during the previous round of revisions, so we do not think it would be appropriate to remove them now. Regarding the main message for each of the paragraphs, we believe they are also all in scope, and we note which paragraphs were extended in response to reviewer requests in the previous round of revisions:

- P1: Broad background, understanding snow cover changes.
- P2: Snow accumulation modelling (was extended).
- P3: Snowmelt modelling: physics based (was extended).
- P4: Snowmelt modelling: temperature-index based.
- P5: Meaning of the melt factor.
- P6: Defining the knowledge gap.
- P7: Defining the need for this study (was extended).
- P8: Short description of what we do in this study.
- P9: Outlining the novelty of this study (was extended).

We believe all paragraphs are essential to define the background and novelty of this study.

- The last paragraph reflects the contributions of this study well.

We appreciate the comment of the reviewer, especially as this paragraph was added as a result of the first round of revisions, and we believe it improved the clarity of the contribution of our study.

Methods:

- Table 1 is valuable for interpreting the indices. Suggest to shorten some names (e.g. "accumulation onset", "peak swe", and "melt onset") and recommend carefully checking that they are used consistently through the paper.

Thank you. We have shortened the names as suggested and checked they are consistently used through the text and figures (e.g. Figure 9, 10, 11).

- Would it be appropriate to standardize the use of the symbol ^ (·) for empirically estimated value and (·)e for a multiple linear regression-based estimate? Also clarify that they are both based on training data only.

We have standardized the notations “$\cdot_d$” for empirically and “$\cdot_e$” for regression-based estimates. The “^” was only used for the formula of daily empirically derived melt factors, which are then used to compute the median melt factor per station (α_d). We have revised their consistent use across figures (e.g. Figure 6, 7, 8), tables (e.g. Table 2) and text. There is already a paragraph clearly stating that these are based on training data only, as well as other instances where we repeat this information (e.g. line 231).

- Relatedly, I suggest to move L239-242 earlier in the section. Please clarify how the split was done. In the authors' responses it appears that the split in time (for ^ (·) estimates) is taken sequentially as the first half and last half of years. Would the results be affected if instead it was a random selection of years in the two groups? How were the stations split for the 2/3, 1/3 split (for (·)e estimates)? αd .

As stated in the reply to the previous comment, there is already a full paragraph in Lines 221-224 clearly stating that the empirically derived and estimated parameter sets are computed with one part of the data only, while the other part is used for model evaluation. This is repeated in lines 243-245 to remind the reader about this. We do not think this information can be moved to earlier in the section, as there are only 20 lines before the previous mentioning of this aspect, and this reminder.

However, we have included a small section with an independent subheader named "Sensitivity analysis" in both the methods section and the results section, as reviewer 1 had a similar suggestion (lines 246-263 and 413-424). These sections show the reader clearly and transparently how the data splits were performed and replicated, and shows the negligible effect of these replications on the results.

- In equation (7), αe refers to a seasonal melt factor estimated through the linear regression; perhaps make it clear that αd is the corresponding seasonal melt factor derived from observations by representing it as ^

We have standardized the notations according to the suggestions of the reviewer on another comment above. We therefore think the use of α_d and α_e for empirically **d**erived and **e**stimated melt factors is appropriate and clear.

- L196-197: point to the section where this is discussed (S4.2)

We now point the reader to Section 4.1.2 at the end of this sentence (line 200).

- Minor suggestion to use z to represent elevation above sea level and θ to represent latitude in equation 7. These are more typical symbols in the field.

We have changed elevation to z and latitude to θ in equation 7.

Results:

Section 4.1.1 should be refined and shortened to a minimum. I believe Table 3 has the important points of this section -- i.e. a contextualization of the assumptions in the methodology (fixed 0C snow accumulation temperature threshold, the meaning of positive/negative ΔSWE , postiive/negative ΔP , etc.).

We agree Table 3 is important but it needs the contextualization that Figure 3 provides. We note that there are two figures and one table in this subsection, and that 6 full lines were added in response to suggestions in the previous round of revisions. After reading the section again, we do not see how the section can be reduced without removing important information which is part of the messages of our paper.

Replacing the colormap in Figure 4 with a diverging colormap (i.e. white at zero) would help this be clearer. Please include whether this figure is based on the training data (testing data excluded)?

We have replaced the colormap with a diverging colormap. We have specified in the caption that this is only training data, as suggested by the reviewer.

Figure 6b may explain why the linear regression (equation 7) doesn't work well. As I understand it, the interannual variability in the melt factor can be very high and therefore melt should not be expected to merely be a function of the climatological characteristics of a site.

This is a good point and we have added a sentence at the end of lines 351-352: " The high interannual variability of the melt factor on some climates could partly explain the low predictive skill of the linear regression model of Eq. 7."

Figure 7: I don't think it makes sense to apply a linear regression to these. Please clarify which melt factor this is (αe or ^ αd).

We have removed the values of R^2, as we agree the regression did not add any information here. We have clarified in the figure labels and caption that this is α_d.

Suggest moving Figure 9 and 10 to the supplementary and simply concluding no large variations in performance are observed across the three parameter sets (e.g. keep L385-388). Then, move on to Figure 11. The clustering work is a nice addition. I also suggest moving Table 4 to the supplement but including the description of the three clusters: "deep and alpine snowpacks, shallow warm snowpacks, and shallow cold snowpacks" -- this is already embedded in the existing text.

In the previous round of revisions we were asked by a reviewer (we do not know if the same reviewer, or another one) to discuss the reasons for the small differences in performance between the three sets of model simulations. We did that extensively in the revised manuscript, linking to the boxplots of Figure 9. We therefore would not consider it appropriate to remove this part of the results and reduce it to "no large variations in performance are observed across the three parameter sets", as this would contradict the requests from the previous round of revisions.

Figure 10 is a key figure in the results, showing the spatial variation of model performance across the Northern Hemisphere for all variables analysed. Figure 11 shows the performance per climate cluster and therefore moving Figure 10 to the Supplement and not showing performance errors per station displayed on a map would leave a key gap in the manuscript. We have, however, moved Table 4 to the Supplement (now Table S1), as we agree it contains similar information as Figure 11.

For clarity, section 4.2 should repeat that the results shown in this section include the testing data, not the training data.

This is now reminded in Line 395.

Figure 11: is there a reason why some of the colorbars are flipped relative to the others?

Assuming the reviewer means Figure 10, we associate red with any error that is linked to "too warm" or "too little precipitation" (e.g. too low peak SWE, too late accumulation onset). As that differs per variable, some colorbars are flipped relative to the others.

Could the authors include a % of how many data points are outliers in panels f-k? I can interpret the IQR as containing 50% of data points, but it would be instructive to know how many points fall outside the 1.5*IQR range.

This is an interesting indicator to provide and we have therefore included this sentence in lines 451-452 "Outliers in Fig. 11 are a minority and range from 2.4% to 21.7%, therefore never exceeding 10.9% of points on each side of the zero error line."

Discussion:

This section would benefit from highlighting the tradeoffs observed when slight changes to threshold values are introduced. I think another key point is that temperature-index modelling is robust, as we generally expect from past work. That is, if you keep parameters within reasonable bounds, the overall performance remains solid (i.e. current lines 510-512, 588-602). Does this result limit the usefulness of TIMs, especially looking to certain types of climates/regions?

There is an independent paragraph in the discussion (lines 502-504), and one in the conclusions (lines 611-615) highlighting the tradeoffs observed when applying slight changes to threshold values. We do not think it needs further highlighting, and this would also expand the length of the paper.

This is a good point about the usefulness of TIMs, but the results and the narrative of our paper does not point to a limitation of the TIM for certain climates/regions, despite some better or worse performances. We have introduced a sentence pointing this out after the contextualization of results among other modelling exercises (comparison with energy balance models): "Even for the worst performing climate clusters, most stations showed a reasonable performance, showing that simple approaches can be applied across climates, with known limitations described in this paper." (lines 566-567)

Conclusions:

I believe that this study's key strength is the hemispheric coverage of the snow depth data, which allows the simple TIM method to be assessed in a distributed sense, rather than at a single site/watershed. Therefore, I suggest altering L604: "Using a new NH SWE dataset, this paper investigates…"

Thank you, we agree and we have added this important nuance at the opening sentence of the conclusions.

Conclusions should make reference to commonly accepted values, for example, “Our results are consistent with well established choices/physically based values such as 0C for a melt threshold…”.

We have added “Our results are in line with the widely accepted physically based threshold of 0°C separating accumulation from melt processes.” in the second paragraph of the conclusions (line 600-601)

L621-624: Nice point. I agree that a new contribution of this work is showing that tuning the thresholds of the TIM can influence snow metrics but that we cannot uniformly improve them by tuning the threshold. Perhaps the authors can mention other methods of optimal parameter search like Markov Chain Monte Carlo (MCMC); these methods could simultaneously adjust the parameters of the TIM to get the best performance with an appropriately chosen cost function taking into account various metrics.

This is a good point. In our study we discuss large-scale performance of the TIM, but this does not mean that applying optimal parameter search methods could not improve model performance if done on a station-by-station basis instead of in a lumped manner. We have included the following sentence in the conclusions to make this point: “In other words, the temperature-threshold approach yields overall good results with temperature thresholds (accumulation and melt) set to 0°C, but a further model performance improvement would require a joint optimisation of both threshold parameters to individual station data or for different climates (which was not the objective of our work).” (lines 613-616)

- L628-630: This is unclear. Could be aided by changing “melt rate” to “snowmelt rate” as in Table 1 and adding the symbols for the other two variables (like T).

We have reworded it to “snowmelt rate”. However, mean melt season temperature did not have a specific symbol in the manuscript and we do not think it is needed to add one. Regarding the melt factor, as these are the conclusions, we do not think it is necessary to add the specific symbol either.

- I would suggest to move L631-637 to be before L614; I find that this highlights the value of a TIM (it can give a robust estimate of snow variables) and this study (using observations to show how TIM threshold variables can affect these snow variables).

We appreciate the suggestion but we are unconvinced that moving these lines will improve the narrative flow of the conclusions, which follow the logical order of the results and discussion of the manuscript.

- Question: what other snow datasets exist for validation and do they overlap with your SWE dataset used here? What are any limitations or opportunities?

To the best of our knowledge, the only other Northern Hemisphere point-scale SWE dataset is NorSWE (Mortimer and Vionnet, 2025), which was published after the submission of this paper (https://doi.org/10.5194/essd-17-3619-2025), and contains actual SWE observations across the Northern Hemisphere. However, both the spatial and temporal coverage of the SWE observations are more limited than in our dataset. In fact, the strength of the NH-SWE dataset is its spatial and temporal coverage thanks to the far larger availability of snow depth observations across the NH, but of course at the expense of not being direct SWE observations. We also do not know if the data in NorSWE has co-located precipitation and temperature time series, as we had for our study. The limitations of our study have already been extensively described in the manuscript, especially in response to the previous round of revisions. We have included a sentence at the end of the conclusions to outline research opportunities arising from more SWE data becoming available: "Further investigations could also be performed as more SWE data becomes available across the Northern Hemisphere (e.g. Mortimer and Vionnet (2025))." (lines 634-635)

Miscellaneous: Check that temperature-index hyphenation is consistent throughout. Thank you. We have added hyphenation to two instances of temperature-index that were missing it.