Review Fontrodona-Bach et al., (2025)

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.

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 (T_a, α, T_m) 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.

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

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

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.
- Would it be appropriate to standardize the use of the symbol $\hat{(\cdot)}$ for empirically estimated value and $(\cdot)_e$ for a multiple linear regression-based estimate? Also clarify that they are both based on training data only.
- 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 $\hat{(\cdot)}$ 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 $(\cdot)_e$ estimates)?
- In equation (7), α_e refers to a seasonal melt factor estimated through the linear regression; perhaps make it clear that α_d is the corresponding seasonal melt factor derived from observations by representing it as $\hat{\alpha}_d$.
- L196-197: point to the section where this is discussed (S4.2)

- 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.

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.).

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)?

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.

Figure 7: I don't think it makes sense to apply a linear regression to these. Please clarify which melt factor this is (α_e or $\hat{\alpha_d}$).

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.

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

Figure 11: is there a reason why some of the 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.

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?

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…"

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…".

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.

- 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 $\overline{T}$).
- 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).
- Question: what other snow datasets exist for validation and do they overlap with your SWE dataset used here? What are any limitations or opportunities?

Miscellaneous

- Check that temperature-index hyphenation is consistent throughout.