Dear editor and referees,

Thank you for your constructive and thorough reviews. We have substantially revised the manuscript to improve the clarity, readability and scientific impact of the study.

The revisions do not materially change the main results, but they strengthen the presentation, analyses, and conclusions. A summary of the main changes to the manuscript is provided below:

- We have included a **new sensitivity analysis** of fixed and smoothed temperature thresholds on temperature-index model performance.
- After correcting a small mistake while revising the code in response to reviewer's comments, the number of available stations in the study has **decreased from 5,560 to 4,736, with no impact** on the previously presented results.
- The introduction has a **stronger focus on the value** and contribution of our study. This is also incorporated through minor modifications to the abstract and the conclusions.
- The manuscript now more clearly **acknowledges the superiority of physics-based** models when sufficient data is available.
- We have expanded our **discussion of dataset and TI model assumption limitations**.
- More **detailed discussions on the effectiveness of different parameter sets**, and why there are small differences between them.
- A **new regional analysis** of temperature-index model performance has been added, strengthening the impact of our findings (new Fig. 11, Table 4) .
- Because of the additional and supporting analyses performed, we **now have a Supplement to the manuscript**, which includes also the previous figures from the Appendix.

We are confident the manuscript has improved substantially and will be of wider interest to the HESS community. Thank you.

The authors.

In black: Referee comments
In blue: Authors' response

Line numbers of reviewers' comments refer to the original preprint manuscript. Line numbers of our responses refer to the TRACK CHANGES MANUSCRIPT, unless otherwise stated.

RC1: 'Comment on egusphere-2025-1214', Anonymous Referee #1, 24 Mar 2025

The temperature index method is a convenient and widely used method in snow simulation. This study estimates the important parameters in the temperature index method based on the published SWE and climate datasets, and analyzes the influence factors of these parameters. The results can provide insights into the temperature index method, making this paper worth publishing in HESS. Having said that, I would like to point out some major concerns that should be addressed before publication.

We thank the reviewer for their thorough and constructive review of our paper. We appreciate them finding the paper worth publishing in HESS, and **we have addressed the concerns** highlighted to improve our study.

1. Potential Circular Logic: The temperature threshold and melt factors are estimated based on the SWE dataset, which is subsequently used to evaluate the performance of the temperature index model. This approach risks circular reasoning. A more rigorous method would involve dividing the dataset into two subsets—one for parameter estimation and another for model validation.

We agree this risk needed addressing and have done so for each of the three parameter sets separately:

1) Common parameter set (Run 1)**:** We use a single melt factor (the across-site median). A 1,000-replicate random half-sample test yields medians 3.61–3.66 mm $°C^{-1}$ day^{-1}, indicating the common value is robust and that the results are negligibly affected by which half is used. Therefore, **we have not changed this** in the manuscript
2) Empirically derived per-station set (Run 2): **We now split each station's time series temporally** (non-overlapping halves). Parameters are derived on one half and evaluated on the other. This is indicated in Lines 239-242. We found that the key performance metrics did not materially change (Fig. 9 updated).
3) Estimated per-station set (Run 3): **We now split stations spatially (two-thirds for regression training; one-third for evaluation)**. This is indicated in Lines 239-242. The model fits changed marginally (snowfall threshold R^2: 0.67 to 0.68; melt factor R^2: 0.16 to 0.13), and Equations 5 and 7 have been updated accordingly. Importantly, we find the performance remained stable (Fig. 9)

2. Subjectivity in Determining the Melt Threshold: In Section 4.1.2, the melt threshold appears to be assumed as 0°C, with supporting analyses provided. However, if the threshold were slightly adjusted around 0°C, the conclusions in Section 4.1.2 would

still hold. To minimize subjectivity, a quantitative approach should be employed. While the melting process is expected to occur at 0°C from a physical standpoint, the temperature data used do not precisely reflect the conditions at the exact location where phase changes occur (e.g., the snow surface for melting and the atmosphere for precipitation partitioning).

We agree that there is subjectivity in determining the melt threshold. In the literature, this threshold varies between -1°C and 1°C, so we heuristically chose 0°C as the central value of this range, and we supported it with our analyses in Section 4.1.2 (Figure 5 in the manuscript). While a fully data-driven melt-threshold estimate is not feasible with daily-scale NH-SWE, because short (1–2 day) decreases in SWE can reflect sublimation, redistribution, or noise rather than melt, **we did quantify the sensitivity to reasonable alternatives (Line 253).** We find a slightly higher 0.5 °C melt threshold reduces melt-onset bias but worsens snow-season end timing, consistent with the physics of initiating vs continuing melt (Supplement Fig. S7; Results Lines 433-435; Discussion Lines 505-507). Importantly, we find that no single threshold was able to consistently improve all metrics (Lines 510-512).

3. Clarity in Time Scale: The study computes temperature thresholds and melt factors at the daily scale in some instances, while averaging them in others. This inconsistency makes the methodology difficult to follow. A clearer distinction between different time scales should be provided.

The reviewer is right that the time scales across sections and figures might be confusing. The analyses of the two melt thresholds and the melt factor start at the daily time scale by comparing the daily time series of temperature and precipitation with the daily time series of SWE. This leads to the insights on temperature thresholds and melt factor at the daily time scale from Figure 3, Figure 5, and Figure 6a. In this study we use temperature-index modelling with constant parameters (non-varying in time, only in space), therefore we compute the median of the daily temperature thresholds and melt factor to obtain a seasonal value for each year (e.g. Figure 6b). The calculated seasonal values are further averaged to obtain a mean temperature threshold and melt factor for each station (that is Figure 4, Figure 6c, Figure 7, Figure 8).

To make all these differences clearer to the reader, **we have adapted Figure 2 (flow diagram)** to include the steps involved in deriving and estimating temperature thresholds and melt factors, adding the words daily or seasonal where appropriate in the diagram. **We have also specified** whether temperature thresholds and melt factors are **daily/seasonal/annual** in several instances **in the text and in the figures** (e.g. the x-axis labels of Figure 3 and Figure 5, the panel labels of Figure 6). We have also added an equation for the calculation of the daily melt factors (Equation 6) and have adapted the mathematical notation throughout the methods to improve the clarity in time scale.

4. Effectiveness of a Single Parameter Set: The model employing a common single parameter set outperforms the other two models in certain aspects. This raises the question of whether complex parameter estimation methods are necessary. A

discussion on the added value of these methods compared to a simpler approach would be beneficial.

We agree that a more thorough discussion about this was needed in the manuscript. We find it encouraging that a simple parameter set has good performance and provides a useful justification for studies with no or little physical information on which to base their model. On the other hand, we also want to understand and provide some physical basis for the spatial variability of the melt factors and their performance.

There are a variety of reasons leading to small differences in model performance between the three parameters sets. First, the melt temperature threshold is the same for all sets of simulations, which limits the possible differences in results. Second, the snowfall temperature threshold varies per station but only minimally, as we set a minimum value of 0°C, for which **we have now added a justification (Lines 246-247)**, in response to one of the minor issues from former L257). For the empirically derived and estimated parameter sets, 79% and 86% of stations have a threshold value of 0°C, respectively. For the rest of the stations, values are mostly below 2°C (see Figure R1 below, now **Fig. S10 in the Supplement**). The small variability of the snowfall temperature threshold across parameter sets may also contribute to the small differences in model performance across parameter sets. Note that the figure excludes stations with a snowfall threshold of 0°C.

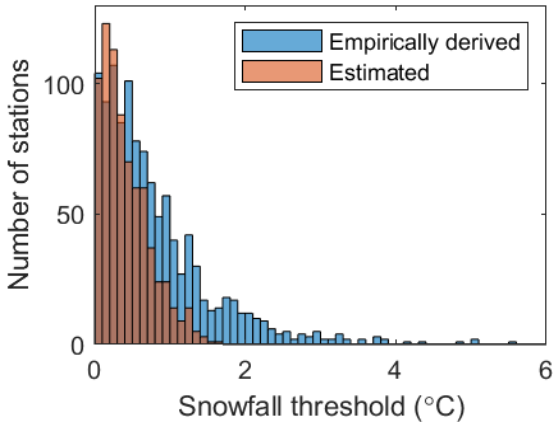

Figure R1. Distribution of the snowfall threshold in the empirically derived and estimated parameter set, excluding stations with a 0°C threshold (which represent 79% and 86% of stations for the empirically derived and estimated parameter set, respectively).

The last factor of model performance variability across parameter sets is the melt factor. This is the parameter that varies most between stations and between parameter sets, as seen in Figure R2 left panel below (**now Fig. S11** in the Supplement). The melt factor differs between the two parameter sets (Figure R2 right panel) because our parameter estimation model based on climate variables has low predictive skill (L275 in original manuscript). The bimodal distribution of melt factor observed in Figure R2 is the result of the bimodal distribution of the station latitudes in the data set (as our estimation model includes latitude as a predictor).

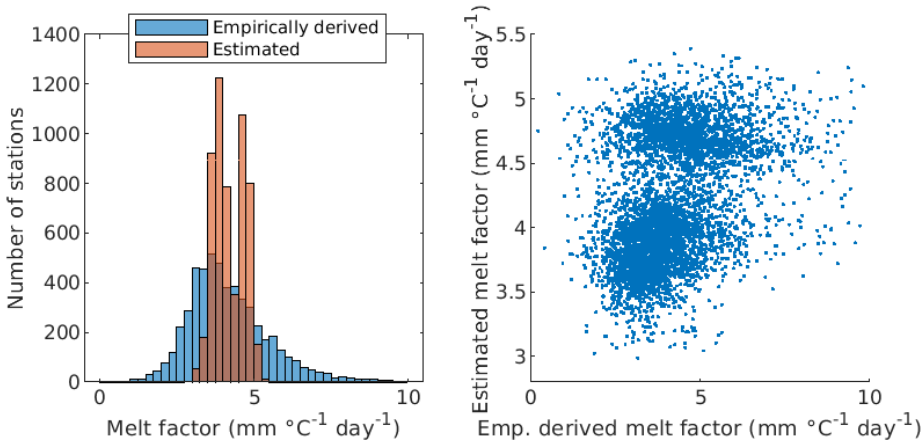

Figure R2. Melt factor distribution in the empirically derived parameter set vs the estimated parameter set based on climate variables.

The model performance variables most sensitive to the melt factor are the melt rate and the time of the end of the snow season. However, there is a strong correlation in model errors between the two sets of model simulations (see Figure R3 below, **now Fig. S12** in the Supplement). This indicates that model performance is not very sensitive to model parameters, at least when evaluated over long-term time series.

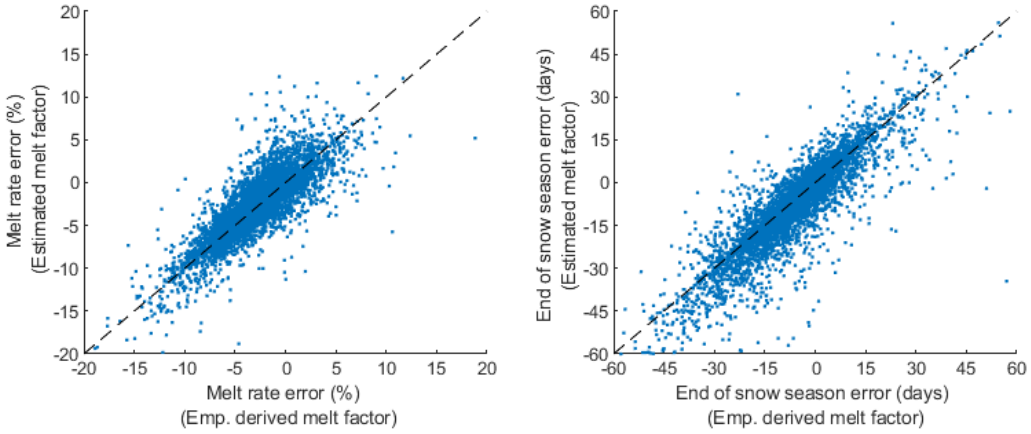

Figure R3. Model errors in the empirically derived parameter set vs the estimate parameter set based on climate variables. Melt rate error (left) and timing of end snow season error (right). Dashed line indicates the 1:1 line.

As the reviewer points out, this suggests that for long-term studies using temperature-index modelling, it is not necessary to use complex parameter estimation methods, as the use of commonly applied values in the literature will lead to similar model performances. **We have added this discussion in Lines 588-602,** and Figures R1-R3 are in the new **Supplement (S10-S12).**

Minor issues:

We thank the reviewer for pointing out these minor issues. We have incorporated their suggestions accordingly.

L106: Provide the full name of SNOTEL.
Done

L123: Ensure consistency in terminology (e.g., CHCN-d, CHCNd, and CHCN-Daily).
Both GHCNd and GHCN-Daily are used by NOAA. We have adapted all to be GHCNd.

3.1: Make it clear what you are specifically referring to here. There are more than two indices in the Table 1.
We have modified Table 1, and the text in 3.1, to specify the difference between climate indices and snow season terms.

L162: “toe” should be “to”.
Corrected.

L163-166: Difficult to understand. Please rephrase and explain it more clearly.
Rephrased to “The minimum temperature at which a decrease in SWE occurs cannot reliably indicate a melt threshold, as the decrease at that temperature may not be due to snowmelt but to other processes.” We hope this is clearer for the reader **(Lines 211-213)**.

3.4: Consider merge sections 3.2 and 3.4, moving the descriptions of the model to the beginning of 3.4 section.
The rationale behind this decision was that it was confusing to start describing the estimation of model parameters, without having initially described the model we are using. We appreciate the reviewer suggestion but we believe merging the sections would also create the problem of describing parameters before describing the model.

Figure 3d: what does the dark orange mean?
This is the overlap of orange and blue bars. We have added this detail in the caption, to clarify.

L257: Why is a threshold lower than 0℃ not allowed?
We assume the reviewer means former line L237. Due to our definition of snowfall threshold in **lines 193-196** in the manuscript, our snowfall threshold estimation model predicts threshold temperatures below 0℃. However, based on Figure 3 in the manuscript, precipitation below 0℃ is very likely to be snow. We therefore only allow the snowfall threshold to be 0℃ or higher, to avoid capturing too much snowfall as rain. A thorough analysis of this choice is done in response to comment 2 of reviewer 2. A justification is added in **lines 246-247**.

In some figures, the text is overlapped. Please check and modify them.
Thank you for pointing this out, we have revised overlapping text in figures.

RC3: 'Comment on egusphere-2025-1214', Anonymous Referee #2, 18 Apr 2025

Temperature index models can provide valuable estimates of snowpack characteristics and hydrological variables. One such model is examined in the present preprint. While there are compelling reasons to use simplified modelling, it must be acknowledged that state of the art snow modelling today can be much more complex, explicitly including multiple snow layers, snowpack energy balance, and more processes (wind redistribution, snowpack temperature gradients, liquid water in snowpack, refreeze). I found interesting

ideas in this study, including the attempt to find a relationship for the spatial variability of the melt factor. However, I am concerned with the authors' claim to "comprehensively assess temperature index model assumptions, parameterizations, and performance across a range of snow climates". The study's scope is narrower than this and there are some issues that I believe need to be addressed.

We thank the reviewer for their thoughtful and constructive feedback. We appreciate that they found our study interesting and believe that their comments have helped improve our manuscript.

We fully agree that state-of-the-art snow models based on energy balance formulations can resolve a much broader range of snowpack processes than the temperature-index approach. Our manuscript does acknowledge this, but we have revised the introduction to more clearly and explicitly explain the limitations of temperature-index modelling relative to more complex physical models. This has been **emphasized in the introduction in Lines 62-63, and 72.**

However, the intent of our study is not to argue for the superiority of the temperature-index model, nor to suggest it as a substitute where full energy balance modelling is feasible. Rather, our goal is to provide a systematic, climate gradient evaluation of temperature-index model behaviour, assumptions, and performance. This is important given its widespread use, due partly to ease of implementation and minimal data requirements, especially where detailed forcing or energy balance data are unavailable. **We have included a statement about the preferability of physics based models** in the abstract (lines 5-6) and introduction (see **lines 92 and 109-113**) to better clarify this scope and intended contribution.

Regarding the concern that our claim to “comprehensively assess” TI model assumptions may overstate the scope, we appreciate the opportunity to clarify our intent. While we do not aim to evaluate every process or modelling approach in snow hydrology, we use the term ‘comprehensively’ in the context of TI modelling and its assumptions.

1. Suitability of the datasets. This SWE dataset, derived from snow depth observations and co-located with precipitation and temperature observations, covers extensive spatial and temporal domains. This is a great strength. However, the temperature index model linking precipitation, temperature, and SWE relies on a balance between accumulation and melt processes. Therefore, the fact that the observational datasets cannot detect when both accumulation and melt have happened within one time step is problematic. This is addressed as a limitation for snow decreases in L161, but the same problem could affect the assumption in L151. Findings presented in L198-203 could also be affected by days with both melt and accumulation occurring, in addition to the explanations offered by the authors. L210-220 suggest that some SWE changes are not detectable in this dataset, but once again this could be confounded by a mixture of processes occurring in one time step. While small SWE changes may indeed be missed in the dataset, this means there is added uncertainty on derived estimates such as onset of snow, peak SWE, onset of melt, etc. L244-246 highlight more errors implicit in using this dataset for this work. These

issues could be examined by using some other dataset to examine the magnitude of errors introduced by some of the provided plausible explanations.

We thank the reviewer for these thoughtful points. As noted, the SWE dataset we use (NH-SWE), based on snow depth observations co-located with precipitation and temperature, is a major strength of this study due to its unprecedented spatial and temporal coverage. At the same time, we acknowledge that, like any observational dataset, it comes with limitations which we discuss in the manuscript.

Regarding the concern about possible co-occurrence of accumulation and melt within a single daily time step, this is indeed a potential source of uncertainty, and we agree that we did not explicitly discuss it in sufficient detail. **We have revised the manuscript to acknowledge this limitation**. However, the available evidence suggests this issue is actually very rare and unlikely to bias our overall findings. More specifically:

- As shown in Figure 5c and discussed in **lines 300-302**, the potential for both melt and accumulation within a single day is mostly confined to temperatures near 0 °C. Days with recorded precipitation and SWE loss (indicative of possible rain-on-snow or mixed processes) represent fewer than 10% of events at 0 °C, and drop to less than 1% at −5 °C.

- Similarly, Figure 3c supports this interpretation as most cases with simultaneous precipitation and SWE loss above 0 °C are attributable to rain-on-snow rather than snowmelt–snowfall overlap.

- While we cannot resolve sub-daily variability with daily data, this limitation applies primarily to small-magnitude melt events during the accumulation season. Larger snowmelt events during the melt season are confidently captured, as demonstrated by Fontrodona-Bach et al. (2023), which shows minimal bias in total melt estimates.

A discussion with the above points has been added in Lines 300-305 in the results section, and 484-486 in the discussion.

We also note that the statement in L151 (now line 189), that increases in SWE can be confidently attributed to snow accumulation, is supported by the physical basis of this measurement. Increases in snow depth correspond to new snow accumulation, except in rare cases of measurement noise, which are negligible at the scale of this study. **We have clarified this** in **line 190**.

Regarding the undetected SWE changes noted in former lines 210–220 and 244–246, we confirm that these correspond to well-known issues such as snowfall undercatch or small precipitation events that fall below the detection threshold. Again, these represent a small fraction of the total dataset. Line 288 and Figure 3d clarify that these are typically very small (< 3 mm) and unlikely to significantly impact seasonal-scale results. **We have also extended the sentence** in **line 325-327** to clarify that undetected melt events are very small events in frequency and magnitude.

Furthermore, the NH-SWE dataset used here has been independently evaluated (Fontrodona-Bach et al., 2023), showing high accuracy in snow onset (forced by observations), and only small median biases in peak SWE (−1.7%) and melt onset (−1 day). These uncertainties are also very small relative to the interannual and spatial variability explored in our study.

In summary, whilst this dataset has some acknowledged limitations, including the inability to explicitly resolve overlapping processes within the daily time step, we are confident these do not materially impact the results and conclusions. **We have endeavoured to clarify these points in the revised manuscript and strengthened our discussion of these specific uncertainties.**

2. Structural uncertainty introduced by binary thresholds, some thresholds not varied. The snow accumulation and snow melt thresholds are both assumed to be strict cutoffs for the respective processes. However, observations (e.g. Dai, 2008) suggest that there is a smooth rain-snow phase transition, though much uncertainty and fundamental difficulties remain (e.g. Jennings et al., 2025) . What is the consequence of this structural assumption for the temperature index model? Furthermore, while 0C is a reasonable guess for a melt threshold, and the data does not contradict it, there is no test provided to show that it is the best choice and there is even some indication that another choice would be more suitable (as described in section 4.2, 5.1, and 5.3). Further tests could examine the melt threshold's effect on the relative performance of the three simulations.

We thank the reviewer for this insightful comment. We agree that the use of fixed temperature thresholds introduces structural simplifications, especially given evidence from higher-resolution studies (e.g., Dai, 2008; Jennings et al., 2025) that the rain–snow phase transition is more gradual. We have included the use of these references in the revised introduction **(lines 57-58)** and **we performed tests to examine the effect of a fixed threshold vs smooth threshold** and the implications on the results (mentioned in the methods section, **lines 196-197**). The results of this examination are also incorporated in the manuscript (see below in detail).

Snow accumulation threshold:

We note that the above-cited studies use sub-daily (3-6 hourly) data, while our analysis is based on daily time steps. At daily resolutions, uncertainty in the timing of precipitation events makes the application of these smoother temperature thresholds less straightforward. Nonetheless, we agree that this simplification deserves testing. We have therefore **conducted additional sensitivity analyses using both fixed and smoothed** snow accumulation temperature thresholds on our daily data.

To reduce confounding and noise from more minor accumulation events that have little impact on SWE, we focus on precipitation days with more than 10 mm recorded. As shown in Figure R4 (**new Figure S4** in the Supplement), no single fixed threshold performs optimally: low thresholds tend to misclassify snow days as rain, while higher thresholds misclassify rain as snow. Smooth thresholds, where snowfall probability transitions linearly between two temperatures, offer only marginal improvements for total snowfall but do not

improve peak SWE or snow season onset (Figure R5 below, **new Figure S5** in the Supplement).

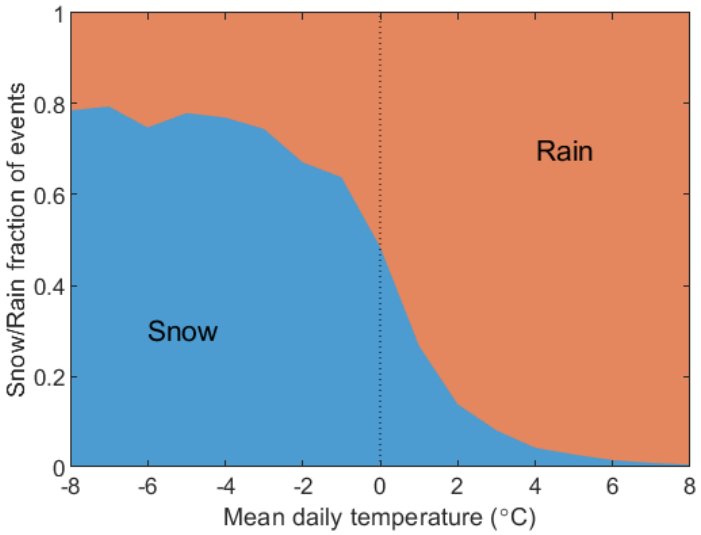

Figure R4. Fraction of events with snow accumulation and no snow accumulation (rain). Based on 1,321,260 days with observed precipitation of more than 10 mm. If the precipitation caused an increase in snow accumulation, the precipitation is considered as snow.

Interestingly, a lower threshold (e.g., 0 °C) improves estimates of total snowfall and snow onset by reducing false accumulation on ephemeral snow days outside the core season (Figure R5 top panel). Conversely, a higher threshold (e.g., 1.5 °C) improves peak SWE by reclassifying marginal rain events as snow (Figure R5 middle panel). However, it is important to emphasise that our data clearly demonstrates that no threshold, fixed or smooth, consistently improves all metrics or reduces interquartile uncertainty ranges.

The results on the examination of different fixed and smooth snowfall thresholds has been added in Lines 408-416 and Figures R4 and R5 added in the Supplement (Figures S4 and S5).

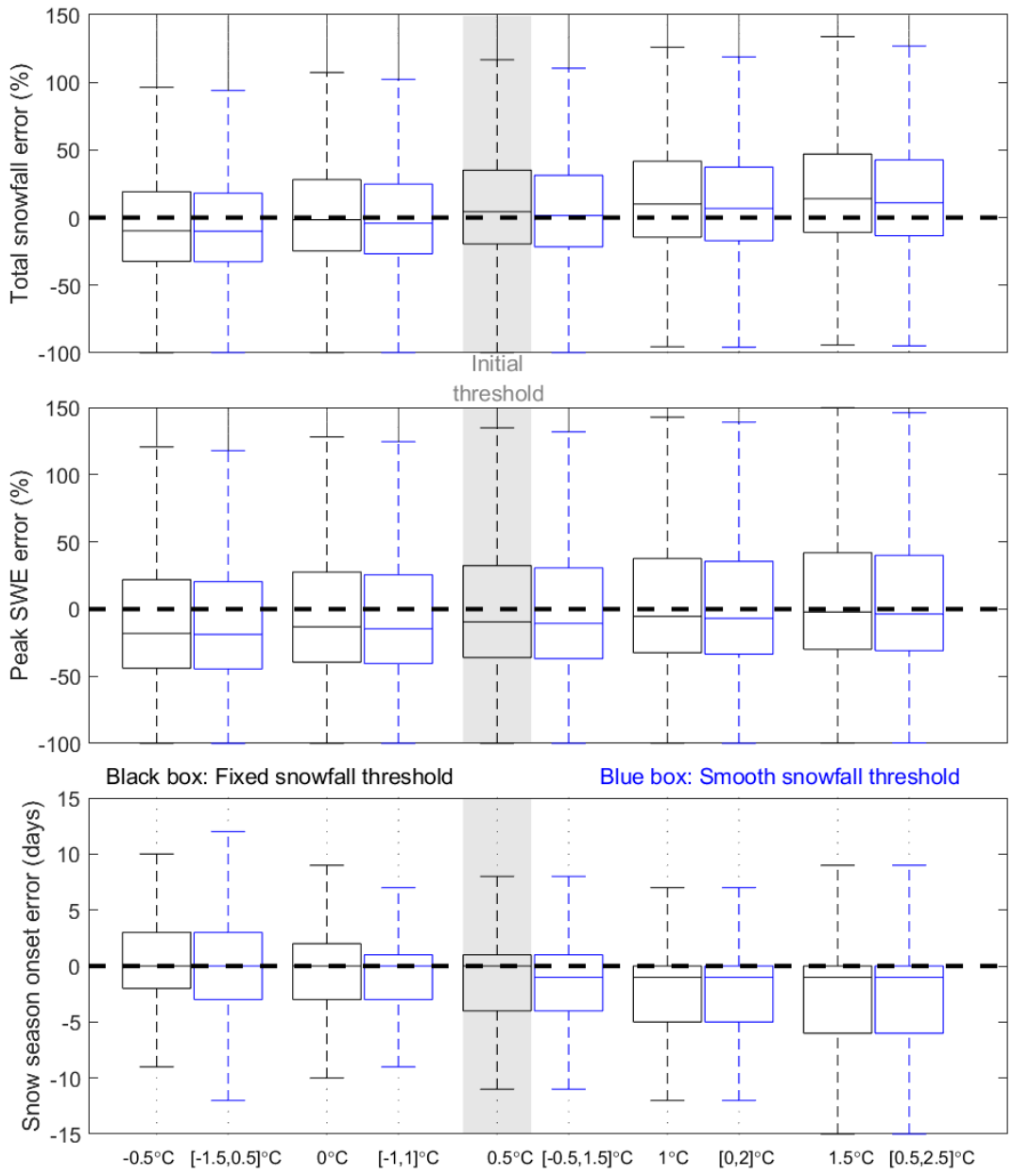

Figure R5. Sensitivity of results to varying fixed and smoothed snowfall temperature threshold. Black boxes indicate the results for fixed thresholds, and their neighbouring blue boxes show the results of a smooth snowfall threshold with a 1°C lower and 1°C higher range around the fixed threshold, where the amount of precipitation falling as snow linearly scales from all snow in the lowest temperature to all rain in the highest temperature. The grey shaded box indicates the threshold used in the original submitted manuscript.

Snowmelt threshold:

We also **tested varying the melt temperature threshold** (Figure R6). A slightly higher threshold (0.5 °C) does improve melt onset estimation, consistent with the physical requirement for additional energy input to initiate melting (Molotch et al., 2009; Jennings et al., 2018a). However, this same adjustment degrades performance in predicting snow season end dates, which are more sensitive to daily temperatures near 0 °C once melt has already begun. We have **added Figure R6 of this reply to the new Supplement (new Fig. S7)**, and referred to this in the results (**lines 505-507**).

As with snowfall thresholds, these results highlight trade-offs, whereby adjusting thresholds improves performance for some variables at the cost of others. Interquartile ranges remain broadly similar across threshold choices, suggesting that model performance is relatively insensitive to the precise value of the threshold within the ranges we tested. **This has been added in the discussion lines 510-512.**

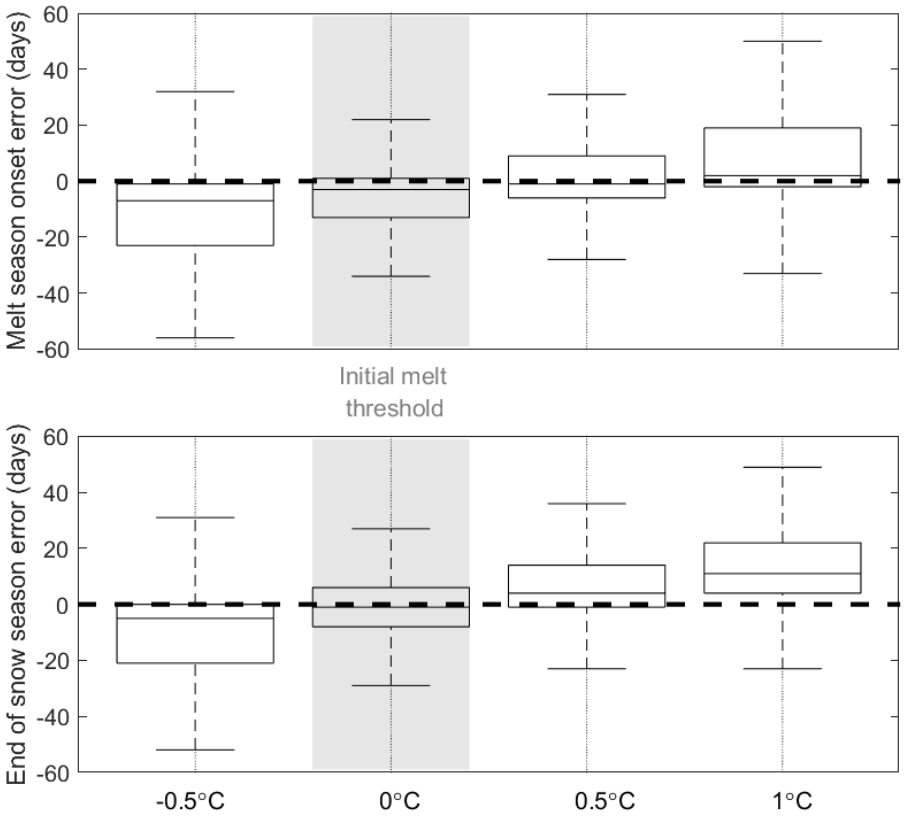

Figure R6. Sensitivity of results to varying melt temperature thresholds. The grey shaded box indicates the threshold used in the original submitted manuscript.

Although the fixed thresholds used in our original analysis are simplifications, the **new sensitivity tests** show that alternatives yield mixed results and no consistent performance gains. We have **included these new analyses as Supplement figures** in the revised manuscript and **expanded our discussion** of threshold-related structural uncertainty. We appreciate the reviewer's suggestion, which has helped strengthen the manuscript's methodological transparency.

3. Needs more focus on value added. There are interesting ideas in this paper, despite some data and methodological challenges.
(We split point number 3 into several subpoints).

Our manuscript investigates the performance, assumptions, and parameter sensitivities of the temperature-index (TI) modelling approach across a wide range of Northern Hemisphere snow climates. While our goal was not to develop or optimize a new TI model, our analysis offers several novel contributions: 1) a systematic cross-climatic evaluation of TI model behaviour across >4,500 sites, 2) a transparent assessment of performance limitations under different snowpack regimes, 3) a sensitivity analysis of model

assumptions (e.g. thresholds, melt factors) and their spatial variability, 4) new spatial analyses that reveal how and where model performance varies systematically with snow climates. We recognize that the discussion and conclusion did not highlight these contributions strongly enough, and **we have revised the abstract, the last paragraph of the introduction (lines 119-125), and the conclusions, to state these contributions.**

I believe that refocusing on the regional results is more critical than summary statistics covering the large regions. We see in Figure 10 that many regions remain poorly represented (description L314-318), but the reasons are left mostly unexplored.

We agree that further exploration of regional performance strengthens the manuscript. While the smaller scale variability is outside the scope of this work, **we have conducted additional analyses to delve into regional patterns** in model error. Because large regions (e.g. North America) can contain very different types of snowpacks, we have used **k-means clustering to classify snow climate clusters** over the Northern Hemisphere. We used five geographic and climate variables to define the clusters, and we obtained the best Silhouette score with 3 snow climate clusters. These are, in short, deep and alpine snowpacks, shallow warm snowpacks, and shallow cold snowpacks. The three snow climate clusters have very clear spatial/regional patterns, and distinct median performances for the different metrics evaluated. These new analyses are presented in **new Table 4 and new Figure 11 in the manuscript**, and described in **Lines 441-458 in the Results and Lines 552-565 in the Discussion**. We believe these additional analyses strengthen the manuscript by clearly and quantitatively showing spatial differences in performance between three main snow climates across the Northern Hemisphere.

Further, while the median performance for most snow metrics is good (Figure 9), the box and whiskers show a large range. Which regions specifically contribute to this wide spread?

We agree that we can better explain the wide range of errors observed in Figure 9. As shown in Figure R7 below (**new Figure S3** in the Supplement), relative errors in peak SWE are largest at stations with very shallow snowpacks (peak SWE < 100 mm). For such sites, small absolute deviations translate to large relative errors (e.g. 100 mm modelled vs 10 mm observed = +1000%). However, for stations with deeper snowpacks, errors are consistently lower. Median underestimation for snowpacks >100 mm is ~25%. **We have included this analysis and complementary figures in lines 397-402** and Figure S3. It helps clarify that while outliers are prominent, they are concentrated in a specific subset of sites and not indicative of general performance. Furthermore, it must be noted that in Figure 9 all station-years are plotted, instead of station medians. In our **new Figure 11** where station medians are plotted per climate cluster, **the outliers and whiskers are highly reduced** compared to all stations-years plotted in Figure 9, showing that the very large errors are not the norm.

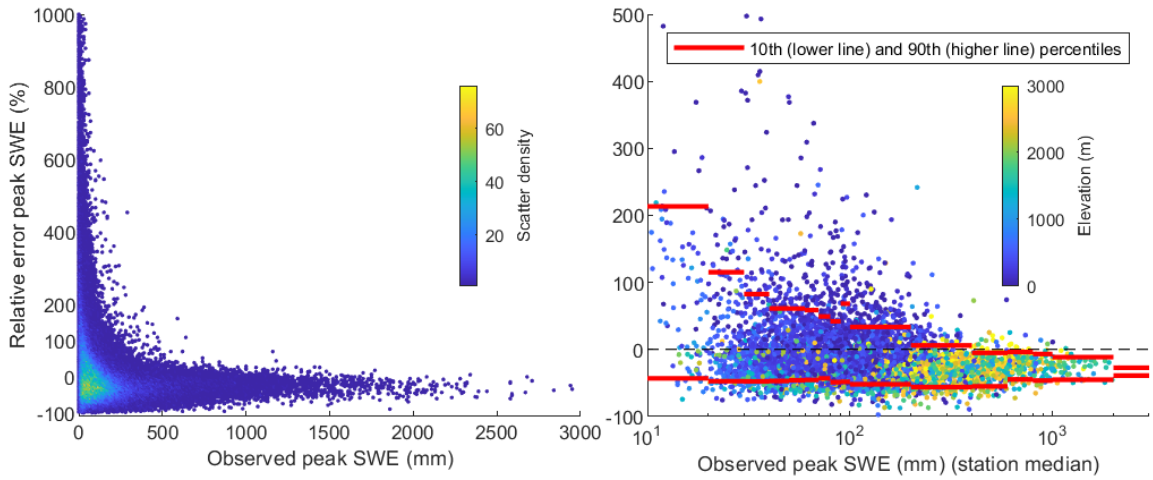

Figure R7 . Analysis of relative errors of peak SWE. All station-years (left) and median per station (right).

How do all of these results compare to observational uncertainty or model uncertainty from state-of-the-art snow models? Re-focusing the study on new findings and placing the results in context is needed to justify the conclusion that the model performs satisfactorily.

This is an important contextual point. While a direct comparison to energy balance models is not possible at this scale, due to the lack of input data and information on appropriate parameter ranges, we can offer a useful benchmark by comparing reported performances from recent model evaluations. Menard et al. (2021) report peak SWE biases from −50% to +35% across ESM-SnowMIP models. Cho et al. (2022) found that land surface models underestimated peak SWE by 268 mm on average at SNOTEL sites. In comparison, our TI-based approach underestimates peak SWE at SNOTEL sites by 114 mm on average, and the timing of peak SWE is 23 days early vs. 36 days early in Cho et al. (2022). **We have incorporated this comparison into the discussion lines 569-576**. In our new regional analysis in Figure 11, we show station-median performance boxplots (previously only mapped in Figure 10), which help in comparing our results to other studies that also report station-medians. While such comparisons are not one-to-one, they demonstrate that TI models, when interpreted cautiously, can yield useful constraints.

Cho, E., Vuyovich, C. M., Kumar, S. V., Wrzesien, M. L., Kim, R. S., & Jacobs, J. M. (2022). Precipitation biases and snow physics limitations drive the uncertainties in macroscale modeled snow water equivalent. Hydrology and Earth System Sciences Discussions, 2022, 1-22.

Menard, C. B., Essery, R., Krinner, G., Arduini, G., Bartlett, P., Boone, A., ... & Yuan, H. (2021). Scientific and human errors in a snow model intercomparison. Bulletin of the American Meteorological Society, 102(1), E61-E79.

While particularly intriguing because it could offer a way to extrapolate melt factors from the point-scale, the multi-linear regression model for melt factor also appears minimally predictive. In addition to the good practice of preserving a random subset of data from the regression in order to use it for testing, other variables could be added (longitude, mean SWE amount, perhaps a typical "snow type" (Sturm & Liston, 2021).

We agree that the predictive power of the regression model is modest, which we now **emphasize in line 232**. As suggested by both reviewers, we have withheld a random test set to validate the model **(specified in lines 239-242)** and assess generalisability more robustly. We have already tested a range of covariates that did not improve model skills, suggesting that estimating melt factors without actual melt data is a complex exercise. We have not additionally explored whether including a snow type classification (e.g. Sturm & Liston, 2021) improves predictive skill, but a discrete variable such as snow classification type would not be easily tested along other continuous climate variables. Nevertheless, **we have included the updated regression analysis (Eq. 5 and 7)** which is built with a random test set only.

Other questions and minor points:

L62 typo "toe" should be "to"
Corrected.

L112-115/L445-453: How much missing data is there that needs gap-filling? Could this affect any of the results (e.g. giving deltaSWE=0)?
Gap-filled data represents only 3% of the final database, so we do not think this could significantly affect any of the results. We have added this number in the gap-filling description **in the Supplement**.

L170: It could be possible for melt to occur on days with negative temperatures (e.g. if some period of the day exceeded 0C). Could it also be explained by the snow depth decreasing due to wind compaction?
The ΔSNOW model (Winkler et al. 2021) is used to convert snow depth to SWE and accounts for this process, but it is possible that in some cases the model reaches the maximum density (which is parameterised) and sees a further decrease in snow depth as melt, when it actually was further compaction. A more comprehensive analysis is provided by Winkler et al. (2021) and Fontrodona-Bach et al. (2023). **We have added the possibility of model inaccuracies** when discussing data limitations in **line 211**, citing the references of Winkler et al. (2021) and Fontrodona-Bach et al. (2023).

L235: Would equation (5) be sensitive to changes in the binary structure of equation (3)?
Equations 3 and 5 are independent and do not interact with each other. While equation 3 describes the snow accumulation routine in the temperature-index approach, equation 5 describes a model that derives the accumulation temperature threshold from the NH-SWE time series and the observations of temperature, but there is no modelling involved.

L242 typo "is only be" should be "is only"
Corrected.

L250 typo "blue and read" should be "blue and red"
Corrected.

L260-261, L375-376: It would be good to really highlight these spatial results which show where temperature index models can or can't work robustly.

We have now highlighted spatial results with our new regional analysis by snow climate cluster in Table 4 and Figure 11, strengthening the analyses of where temperature-index models perform better or worse.

L386-388: It is somewhat misleading to present hemispheric scale results when there are such big regional differences. Could focus the summary text onto the range of results in Figure 10, perhaps then aggregating results (e.g. for Western North America, Arctic, Europe, and Scandinavia) to give median summary statistics with regional grouping.
We agree and find this suggestion very useful. In combination with one of the main comments from the reviewer, **we have included a new spatial and regional analysis** (Table 4, new Figure 11, Lines 441-458 in the Results and Lines 552-565 in the Discussion). We believe a snow climate clustering is more adequate than splitting the Northern Hemisphere into large regions. In some cases (e.g. North Asia) snowpacks may be quite similar across large areas, but in other cases (e.g. alpine areas) the large regions may cover very different types of snowpacks. Our snow climate clustering clearly grouped together roughly similar types of snowpacks across the Northern Hemisphere.

L404-407: I think this is a key aspect for discussion. Should be revisited in the Conclusions section.
We agree it is an interesting finding and have **further delved into the sensitivity of the melt rate to temperature and to the melt factor** (see below the response to comment about L424-427). We now also mention in the conclusions that melt rate model performance is not overly sensitive to the accurate estimation of the melt factor because the mean melt rate is more sensitive to the mean melt season temperature than to the melt factor (Lines 629-630).

L416-418: Could reference other data products that use simple snow modelling and data assimilation of snow observations already (e.g. CMC daily snow depth product; Brown and Brasnett, 2010), or else clarify if "assimilation" used here means something different.
As we do not really test the possibility of snow data assimilation, we thought it best to **remove this statement** from the conclusions.

L420: Could the SWE decreases below freezing be due to some other explanation, including observational errors? If such errors are potentially present, then could they affect your other findings?
They could also be observational errors, similar to the response to the minor comment about line L170. This is **discussed in more detail** in reply to major comment number 1.

L419-423: There were no tests showing 0C to be more valid than another choice of melt threshold, or another method of partitioning precipitation into rain and snow. It was fixed throughout the whole study.
These tests are now provided. This is **discussed in more detail** in reply to major comment number 2.

L424-427: How do these melt factors compare to those found in the literature? With respect to the values of the two rates, this is still a large range. While it has been diagnosed that there are regions where there is high interannual variability in melt factors, what might

cause this variability? Does it disqualify temperature index modelling from being used in those regions?
The range of melt factors found in our study has been further contextualised in the discussion in lines 517-518. We now also discuss potential reasons for the high and low interannual variability of the melt factor in **lines 523-527**, owing likely to a high interannual variability of melt energy sources for shallow snowpacks, and a low interannual variability of dominant melt energy sources for deep snowpacks. We do not think it disqualifies TI modelling in those regions, since the performance evaluation of melt rates is good. **We further delved into the reasons why the performance is good despite the high interannual variability**: **new Figure S9** shows that mean melt rates are strongly linked to mean melt season temperatures but are less sensitive to melt factors. This means that accurate temperature data and reasonable values for melt factors can yield robust melt rate estimates across the Northern Hemisphere snow climates considered. This is outlined in **lines 518-584**.

L429: Modelling peak SWE requires the right balance of accumulation and ablation processes. Which of these is most contributing to these challenges?
We have included a new small supplementary analysis which quantitatively analyses the balance between accumulation and ablation process during the accumulation season to accurately model peak SWE. It shows that precipitation is the main culprit. Where peak SWE is underestimated (e.g. deep and alpine snowpacks), not enough precipitation is captured during the accumulation season, probably linked to undercatch and measurement challenges in alpine areas. In areas where peak SWE is overestimated (e.g. shallow snowpacks in cold, continental climates) too much precipitation is captured during the accumulation season, probably not capturing snow redistribution processes which might be quite dominant for dry low density snowpacks. This new analysis and supporting **Figure S8** guide the discussion in **Lines 552-565**.

Further editor comments:

Dear authors,

Thank you for submitting your manuscript to HESS. Both reviewers provided constructive feedback to enhance the quality of your work, and I concur with their assessment that major revisions are required before the manuscript can be considered for publication. Below, I outline my own concerns, which I hope will further strengthen your study.

We thank the editor for their constructive comments which further help improve our manuscript. **We have addressed all the comments** from the reviewers above, and the comments from the editor below, and we believe our revised manuscript is clearer and much more impactful than before, with new results and interesting insights.

Major Comments:

Organization of Methods and Results:

The distinction between the Methods and Results sections needs clarification. For instance, Equations 5 and 6 (presented in Table 2 within the Methods section) are only fully explained in the Results section. I recommend restructuring these sections to ensure methodological details are self-contained before presenting results.

We agree that the order of the presentation of methods and results may seem confusing. As suggested by the editor, **Equations 5 and 6 (now 5 and 7) are now contained within the methods section**, and we have adapted the text of the methodology and the results to make them fit appropriately and following a logical order.

Consideration of Key Snow Hydrological Processes:

While your study focuses on snow accumulation and melt, other critical snow-related processes—such as sublimation, snowpack water holding capacity, and liquid water refreezing (Seibert, 1997; Pomeroy and Essery, HP 1999)—could significantly influence your findings. Please discuss whether these factors were considered and, if not, justify their exclusion or address their potential impacts on your conclusions.

We have now specifically mentioned in the introduction that the temperature-index approach can not resolve other processes than snow accumulation and melt **(lines 72 and 111-113)**. Our revised results, however, suggest that snow redistribution processes may play a part in the temperature-index model overestimation of peak SWE over cold, continental climates with dry snow (**lines 555-557**). However, since our data can not resolve more processes, it is not straightforward to analyse what is the influence of these unresolved processes on the results. Our study robustly shows accumulation and melt patterns. Investigating what happens with melt water (whether it is held in the snowpack, refreezes, or flows out) is a very interesting question, but we believe it is beyond the scope of this paper.

Study Area Definition:

The manuscript positions this as a Northern Hemisphere study, yet it excludes major snow-dominated regions such as the High Mountain Asia, including Tibetan Plateau (Gao et al., HP 2011), Himalayas (Immerzeel et al., Science, 2010), and high mountains in Central Asia (Gao et al., HP 2017). These high-elevation zones exhibit substantial snow cover and hydrological importance. I recommend either:

Clearly justifying the exclusion of these regions, or

Expanding the analysis to incorporate them for a more representative Northern Hemisphere assessment.

It is true that our study excluded highly glacierised and snow-dominated areas of the Northern Hemisphere. The reason is that data in these regions is either temporally or spatially scarce, or not freely or easily accessible. Therefore, it was not possible to include these regions while complying with FAIR principles and the data requirements of our study: minimum 5 years of gap-free data including snow water equivalent (or snow depth),

and temperature and precipitation data. **We have included this statement** in **lines 152-154** in the Data Section.

I believe addressing these points will further improve the manuscript's rigor and scope. I look forward to receiving your revised work.

Best regards,

Hongkai Gao

Thank you.

The authors.