Response to Reviewer 1
Responses in blue.

Mudryk et al. aim to evaluate 23 different SWE products based on how well they represent SWE climatology, variability, and trends across mountainous and non-mountainous regions in North America and Eurasia. Using existing and newly created reference datasets, the gridded products are scored using skill target diagrams, resulting in a series of Taylor and target plots, eventually leading to an average ranking of the SWE products. The methodology and technical approach to this evaluation is clear. However, the presentation of the datasets, methodology, and results is notably convoluted or disorganized at times, dampening the impact that this thorough analysis could have. In particular, a clear workflow figure could aid in the introduction of the overall evaluation strategy, where related text often references several other sections, causing for much back-and-forth within the manuscript. While repetition in stating the methodology is appreciated, sometimes the methods, results and associated discussion appear in a single section, making the information challenging to process. Thus, I include no major analysis comments and suggest the authors primarily focus on restructuring the manuscript for a clearer portrayal of meaningful results. My more pointed comments below should help with these review items.
Thank you for the time and effort to review our study and the generally supportive conclusions. To respond to the above comments we have rewritten large parts of Section 2 and 3.3 in order to incorporate your specific comments below (and to try and reduce the amount of back and forth that was noted).

We also note the following 3 analysis changes that affect some of the product-specific statistics, but don't alter generalized rankings or conclusions about product performance:

1. There was an error in the skill score components calculated for Figure 3 that has been fixed (the bias and amplitude values used in the figure were incorrectly normalized, but the relative rankings were still correct).
2. We have altered the way we calculated skill scores for climatological SWE in the NAm region to be more similar to the time-varying statistics and to ensure the statistics reflect the native resolution of the individual products. This is related to a comment from reviewer 2.
3. We added an overall scale factor to both S_bias and S_pattern that allows readers to assess how product performance varies across differing regions/terrain throughout the paper (essentially skill scores for a regional/terrain-specific test are scaled by the uRMSE of the worst performing product across all tests). Since this doesn't alter relative rankings on a particular test it only affects the perceived performance on the given test relative to the other tests. Text in the revised manuscript describes this.

Detailed/line-by-line comments:
Suggest writing out each abbreviation first (e.g., ILAMB, AMBER, IMS)
-Will do - thanks for catching that.

Table 1: Suggest placing a reference column and stating, prior to the table, that each "family" of SWE product will be discussed in more detail following the table and/or placing the text prior to the table.
-We have expanded upon the sentence at line 81 to state this more clearly.

Section 2.2
This section is generally challenging to follow, yet it is intended to set up the pertinent evaluation strategy. There is reference to various sections ahead of the current (e.g., full details referenced in section 2.4, reference to 7 products chosen for an ensemble). Perhaps a schematic of the workflow/evaluation scheme would be helpful. As such, it is also unclear when a point system was introduced (line 153-154).
-We have rewritten this section discussing the point system up front and more explicitly and tried to remove some of the back and forth you mention.

Line 156: Unclear what exactly expert judgement is considered in this case
-We have reworded this section and this phrase no longer appears.

Table 2: Suggest including some justification as to how regions were selected. Unclear at what spatial scale these variables are evaluated.
-The rewritten intro to Section 2.2 now states up front that most regions were selected based on the characteristics of the reference data. We have also added more detailed rationales when discussing the reference data in Section 2.3.

Line 172: It would benefit the readership if the text stated the spatial extent in addition to the figure.
-The available coverage over both continents is explicitly stated in the revised text.

Figure 2 c-d: It is unclear what the scale bar is referencing, as it pertains to climatology. If this is "SWE climatology," or peak SWE magnitude, labeling it and stating so in the figure caption, as opposed to "climatology," would be helpful.
-only the nonmountainous climatology is shown now and it has been labelled as "Nonmountainous SWE Climatology (Bias-corrected GlobSnow v3)".

Line 219: It is unclear what "broad a range of meteorological analysis fields as possible" looks like for the 7 selected products, which are spelled out later in this paragraph. Could an example be provided compared to a product that was not selected?
Line 221: Consider removing the following sentence, as it again jumps ahead to several sections from the current and causes confusion: "It will be demonstrated in Sect. 4 that the reanalysis-type products which employ assimilation of surface snow information all have seasonal incongruities with one another." Suggest revising line 124 for similar reasons.
-The rewritten section 2.3 addresses both of these comments (L219 and 221).

Section 2.4: Please number and format equations similarly, as there are many and some build off of one another. Reference to a figure here or in one of the citations may bring additional intuition to this methodology here, prior to seeing the results.
-Equations have been numbered, and we have moved some of the subsequent description on how target diagrams display information to the start of the paragraph.

Line 249: Suggest a ½ to full sentence on why this approach was taken to rank similarity across products (were there other approaches in consideration?).
-The advantages of the two-component skill scores we use compared to uRMSE (which is what is used in Taylor diagrams --- another typical approach) are explained in the following sentences of the paragraph. They are also contrasted in the results presented in Figure 3.

Line 353: Additional annotations on this figure would be helpful. For example, placing notation near Crocus-ERA5 and ERA5-Land on the upper right panel would aid in the necessary scanning between text and figure (especially since the numbers/rankings obviously change between panels). This comment extends to Figure 4 and 5. Suggest also reiterating what is represented on each axis in each figure and/or across panels, particularly in the Taylor plots and in reference to pattern statistics.
-In place of additional notation on an already complex figure, when discussing Figure 3 we have specified the rankings of Crocus-ERA5 and ERA5-Land on the taylor plot to make it easier for readers to identify their positions. We also note the statistics of the two products are better separated (thus easier to read) on the figure using the revised method for assessing product SWE climatologies in mountainous regions.

Section 3.3: Many of these beginning sentences/paragraphs, aside from the sentences explicitly
referring to Figure 7, read as though they belong in the methods or discussion sections, which decreases the impact of the following results. There is a lot of information to unpack in Figure 7. The paragraph structure for each row is appreciated. Perhaps the authors would consider beginning each paragraph with the intended take home point, particularly for the bottom row of results.
-As suggested we have removed much of the methods-related preamble and front-loaded a take-home point in the majority of the paragraphs. We also simplified the messaging regarding the EO-trends (bottom row of results).

Section 4: The final result, presented quite clearly, does seem as though it could live in the results section. Are there comparable results specific to the select regions (NA and Eu mountain and non-mountain)? This question also pertains to my next comment.
-We elected to place the final product rankings in the discussion section because it distinguishes it from the more complex and varied individual results detailed in Section 3 (thereby highlighting it we would argue) and because in discussion of the figure and the overall results leads directly into discussion-appropriate commentary.

-Regarding regionally specific results, to an extent the peak season results presented in Figure 5 partially fulfills this function. But since the full suite of tests does not use the same choice of regions for all tests (for the reasons outlined in Section 2) it's not simple to compare the regions you mention in your comment in a meaningful way. For example, because the performance in mountain regions (which are evaluated over NA only) is a key differentiator of performance, NA-only rankings would be similar to the final rankings. EU-only rankings would still be able to distinguish two distinct product groups (apparent from examining Figures 3-nonNH and 5-EU,): the four GLDAS products, JRA-55, JAXA and ERAint-Land are consistently in the bottom half of the distributions distinct from the remaining products. But the EU-only test would be unable to differentiate as much among the top products because their performance is similar to one another as far as we have available reference data over the region to assess.

Line 535: Can the authors expand and be more explicit about "The relative overall rankings shown in Fig. 8 are meant to function as a guideline only"? The following conclusion also states that "user needs and objectives must guide product selection," however a lot of technical and thorough work went into the culminating Figure 8. Are there thus product recommendations and takeaways for users broadly and by region?
-We have expanded upon this statement in the revised manuscript. We stand by our results to the degree that our coverage of in situ data permits us to generalize. But this statement was meant to acknowledge that some products can have idiosyncratic regional performance. For example, the GLDASv21 performance assessed only over the CONUS (Figure S2) performs much better, especially in CONUS mountainous terrain where it's ranked 4th, compared its overall NAm performance (ranked 14th) and its overall ranking (18th). This is why we have provided the caveat about rankings functioning as a guideline for hemispheric performance, but that for specific regions there may be differences.
Likewise we realize that the absence of reference data from mountainous regions of Europe and western Asia is a clear gap in our ability to assess any deficiencies over these regions that aren't also reflective of the products' performance over North American mountainous regions.

Discussion: There lacks a discussion on limitations to this assessment and consideration of other gridded SWE products
Limitations of the assessment related to the reference data distribution were stated on L555-558. Limitations related to the use of distributions to assign scores were mentioned in L560-562. Beyond this, we are unsure what your comment on consideration of other gridded SWE products means. While we have not evaluated every gridded SWE product that there is, the results provide a general procedure to which additional products could be incorporated. We do state this explicitly now in the new text.