Response to Reviewer 2
Responses in blue.

Thank you for the time and effort to review our study and the generally supportive conclusions. We have addressed your criticisms and suggestions below.

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.
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.

Many of the SWE data applied in this study are based on snow accumulation/melt algorithms embedded in different reanalysis models. The derived SWE is therefore a result of the forcing data, primarily temperature and precipitation. When comparing SWE data from these different sources, I miss a discussion on the performance (evaluation scores) of the forcing data used for the various approaches since the SWE estimates will inherit some of their characteristics.
-This would be an interesting analysis but one we are unable to explore fully in this paper. Such an analysis would add quite a bit to an already complex paper and it can only help explain the performance of reanalysis data sets, not the Earth-observation products such as SnowCCI, GlobSnow, and JAXA. Even for the reanalysis datasets, it is clear it will not be able to explain some of the major elements of the performance. For example, ERA5, ERA5-Snow, BTIM-ERA5, and Crocus-ERA5 all use the same meteorology, but span a range of final rankings from 2 through 16 out of 23.

Another issue I feel is almost neglected in the discussion is the role of the native resolution of the gridded datasets. Snow is a property that shows large spatial and temporal variability. Even though the comparison is performed on a joint 0.5°x0.5° grid, the original resolution should have an impact on the estimates. The native resolution of the data sets should be added in table 1.

We answer this question in two steps. First, we note that only the climatological tests in the original analysis used regridded data. For the time-varying results (9 of the 14 tests), the

gridded products were analyzed at their native resolution (the sequence of reference-product pairs used to calculate bias, correlation and standard deviation are selected based on proximity to the reference data using the native resolution of each gridded product). The text in Section 2.4 has been revised to explain this more clearly. Based on this comment and to help simplify our methods we have also altered the NAm climatological test to be more similar to the regional SWE variability tests (and to therefore use the native resolution of each gridded product climatology). Secondly as illustrated in the figure below, we note that resolution is a surprisingly poor explanatory variable for predicting product performance. This is true even after removing products with spurious trends (diagnosed in our manuscript Fig 7 and shown in the figure below in a lighter color).

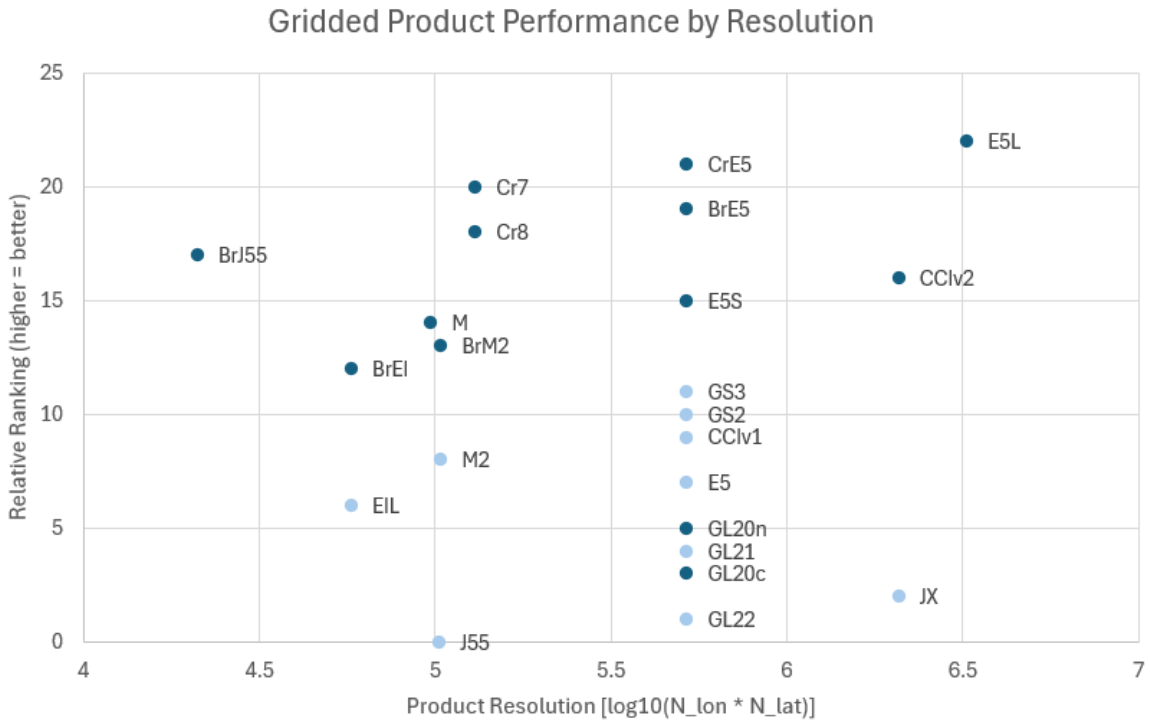

Figure: Relative performance of products according to total number of horizontal grid cells in the NH (log scale). Dark markers denote products with no diagnosed trend issues/discontinuities; pale markers denote products with diagnosed issues. Even removing products with diagnosed trend issues, and removing the remaining two GLDAS products as outliers, the connection to resolution is not strong with multiple products with the same/similar resolutions spanning a range of performance.

A related issue is regarding the benchmark data. It is quite obvious the in-situ data is unevenly distributed in space and also between the regions discussed in the paper. How will the unevenly (and sparsely) distributed snow course data impact the reference data set. And to which extent will that influence the evaluation scores? Further considerations about that would strengthen the paper.

Issues related to sampling bias, data distribution, and SWE measurement type (gamma attenuation measurements or snow courses) are addressed more fully in a companion

paper by Mortimer et al. We originally intended to better coordinate these submissions so that reviewers could be aware of their complementarity in focus, so our apologies for that omission. We have also revised the text to more clearly articulate how our aggregation of the reference data helps to more fairly sample the available distribution of reference data. Finally we note that the scale to which the reference data is aggregated and the regridding of the products does not substantially alter the *relative* assessment of the products. While these decisions do alter the specific values calculated for uRMSE, bias, correlation somewhat, products are still ranked similarly and end up with the same assessed performance overall. This is a reason our overall evaluation strategy focuses on product rankings rather than exact performance measures. In fact our choice to aggregate data within a search radius of 100km tends to improve both the assessed bias and pattern skill for most products. This is demonstrated in the figure below (for mountain regions where this makes a larger difference)[LM1].

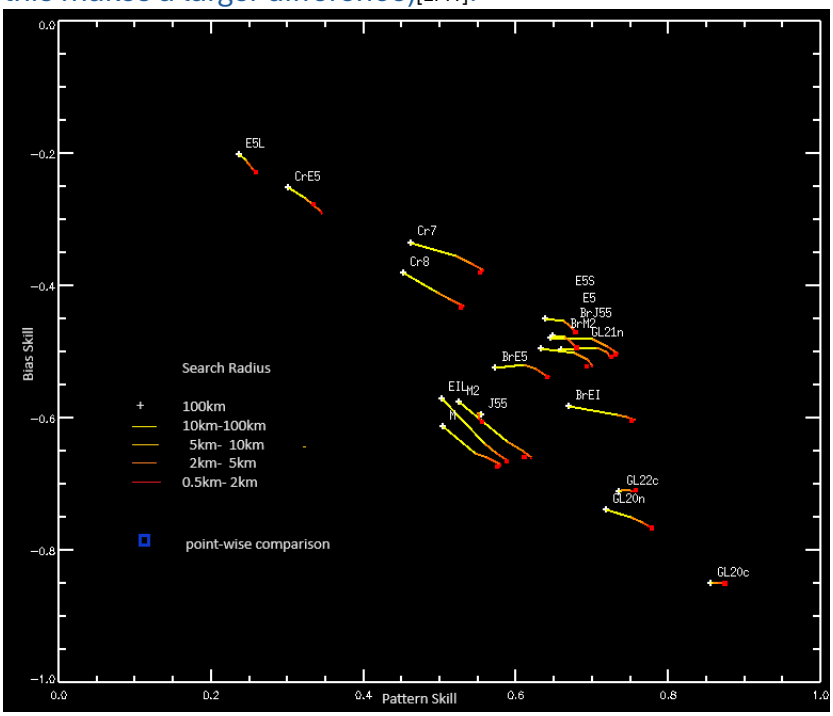

Figure: Dependence of 2-component skill statistics on aggregation radius (including point-wise comparison=no aggregation). Movement of statistics up and to the left indicates improved performance in both statistics.

Further I miss a reflection on the spatial scale of snow cover in mountainous regions. Performing a comparison on 0.5°x0.5° grids will smooth out the natural variability in complex terrain. I think that should be more thoroughly analysed and discussed.
-As mentioned now both the time varying and climatological tests over NAm use the native product resolutions. While we still aggregate the reference data over 200km search windows (100km aggregation radius) in mountain regions, we argue that this is helping to decrease the influence of sampling bias from the insitu data. From the above figure it is clear that while our choices aggregation scale may affect the absolute statistics to a degree, we are still obtaining well-sorted product rankings which are the key output metric of our analysis.

In the beginning of Chapter 2.3 the paper would benefit from a brief introduction in order to prepare and give the readers an idea of the information presented in the next sections.
-We have revised the text in this section in line with comments from reviewer 1 as well.

Line 49: Mortimer et al. 2024 is missing in the reference list (see also comment to line 290)
Line 290: Mortimer et. al, 2023 is missing in the reference list. (see also comment to line 49. If this refers to the paper referred to as submitted in the references I would recommend to be consistent with the references…)
-This is the companion paper that is also under review. We didn't have a link/DOI to cite originally but have now added a temporary citation to the draft manuscript.

Line 57: The term "authoritative" is pretty ambiguous. I would recommend using a more moderate term ;-)
-This word has been changed to comprehensive.

Table 1: Add a column with original resolutions.
We propose to add the requested column if it can be fit by the journal production while maintaining the same portrait orientation for the table (it not immediately apparent to us if there is sufficient width). If there is insufficient width we would prefer to omit the product resolution since it is a poor explanatory variable. We believe the other information provided is more important and the table is more easily read in its current orientation.

Line 109: Explain IMS.
-Added. Thank you.

Line 177. The expression "….method tends to sample" is vague. Please be more specific.
-We have reworded the sentence.

Line 244 (and further lines 487, 489). For consistency, please upcase CCI (to SnowCCI)
-Changed throughout the paper.

Figure 3 contains a lot of information. For easier interpretation I would recommend to add axis titles in all panels.
-Added.

Figure 6: Please explain the term FM.
-We have now written these out as months.

Line 428 (Chapter 3.3). Please be more consistent with the use of "mountainous" and/or "alpine". Maybe stick to one of them?
-The goal was to stick to mountainous, but we were unsuccessful. Now changed.

Line 438 - 440: Why is that causing these anomalous trends? Please add an explanation.

-It's not fully apparent. Our analysis suggests possible explanations are fluctuations in the availability in situ snow depth data or seasonal/regional variability in the detection of snow presence via passive microwave brightness temperature since both of these are assimilated within the JRA-55 analysis. We have added an additional reference to the JRA-55 paper describing how this information is incorporated in the reanalysis and moderated our claimed attribution to this process slightly.

Line 483-485: Is that really the case in all regions? Justification in a graph similar to fig 1a separated into domains would be a nice supplement.
In this content, the in situ data referred to was not the combined snow course/gamma reference data but rather the weather station snow depths assimilated as part of the GS/CCI retrieval algorithms. We have reworded the text to make this clear.

Line 532- . Here I feel the authors are speculating instead of pointing at real properties of the input data. Lines 532-538 need a second look, and maybe rephrasing in order to be more concrete.
-We have reworked this text to make it more explicit and less speculative and brought in references to relevant prior work.

Line 546: I like it!
-no change needed!