Response to Reviewer #1

In this work, the authors utilized snow course and airborne gamma data from North America to comprehensively evaluate the performance of grid snow water equivalent products in both mountainous and non-mountainous areas at various spatial and temporal scales. Additionally, a combined reference SWE dataset for North America was produced. It's a challenging but promising endeavor. Here, I would provide some comments and suggestions for authors' consideration when revising the paper.

Comments:

Thank you for your comments and suggestions. We have responded to them each as listed below. We also wish to point out that this first manuscript was submitted along with a second companion study (Mudryk et al. also currently under review, https://doi.org/10.5194/egusphere-2023-3014). The focus of this first manuscript is to evaluate the extent to which the snow course and gamma reference data can be combined in order to increase total coverage across the North American continent. The focus of the second study is to use the combined reference information along with snow course information over Eurasia to provide a more pan-NH assessment of gridded product performance. This information may help clarify some of the choices we made with regards to aggregation distances as described below. More explicit reference linking these two manuscripts will be added to the introduction.

It is easy for the snow water equivalent in mountainous areas to exceed 1000 mm. However, the 1000 mm SWE was excluded from the validation in the manuscript, did the authors calculate the amount of data for these exclusions, and did they affect the accuracy assessment of SWE in mountainous areas, where snow depth tends to be very large. I suggest the authors add a discussion of the relevant chapters in Uncertainty.

While mountain SWE can indeed exceed 1000mm, observations above this limit represent only a very small proportion of the available reference data - 4% of mountain snow course and <2% of the total snow course observations. It does not apply to the airborne gamma data since its detection limit is ~1000mm. We added the proportion of data removed by applying this threshold.

L124-125: *"This threshold removed <2% of the snow course data (4% of the mountain data) and <0.2% of product SWE with coincident snow course observations."*

As the UA SWE with the highest accuracy, its scatter plot does not show obvious scatter aggregation in the low-value area. So I would like to know if the number of verification points in Figure 1 is the same for each type of SWE product, please give the total number of verification points.

The total number of data pairs after aggregation have been added to Figure 2 and the product names have been revised for consistency with Table 1.

The number of data pairs is smaller for UA SWE and Snow CCI because they do not cover the full spatial domain, and for JAXA AMSR2 and GLDAS 2.2 which are limited temporally.

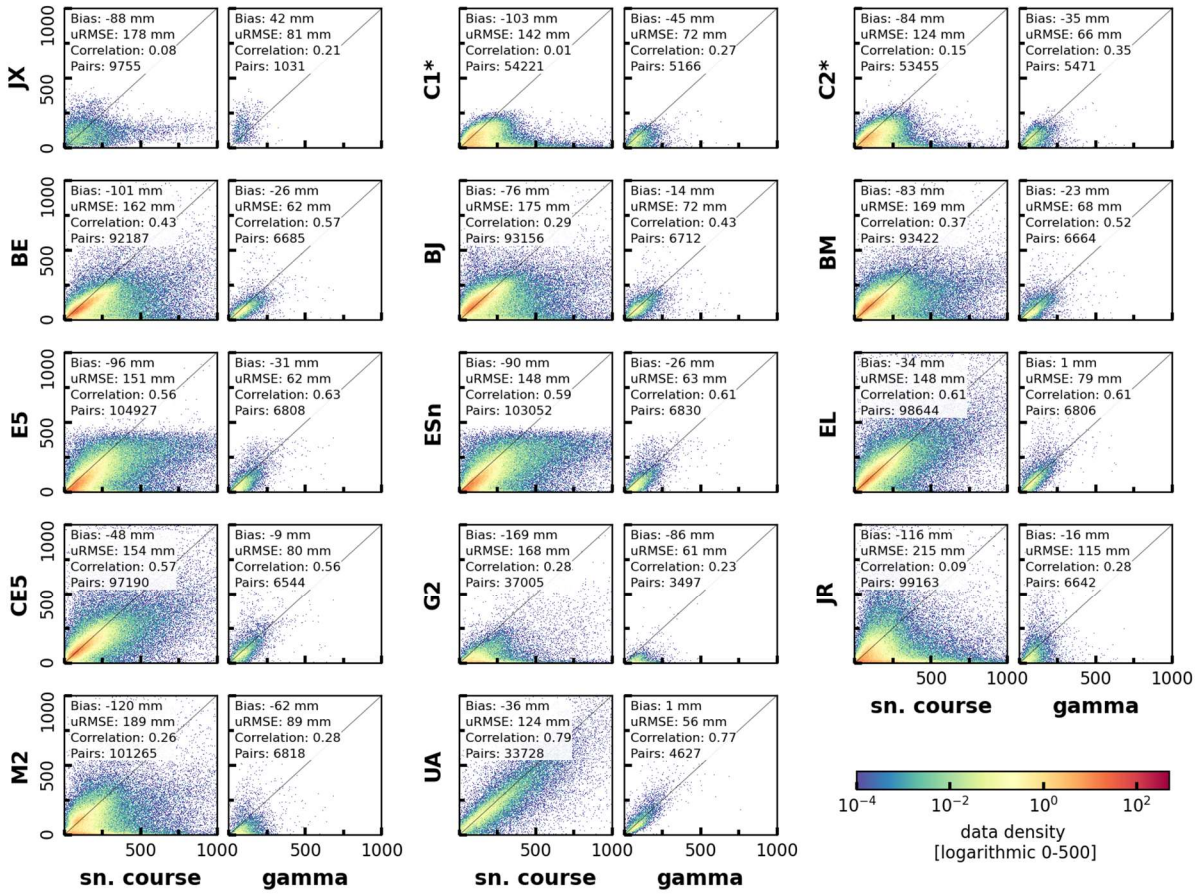

Figure 2: Product vs reference SWE density scatter for measurements > 0 and ≤ 1000 mm during February – April. See Table 1 for product names, descriptions, and time periods. Note that Snow CCI excludes areas of complex terrain, U Arizona is limited to CONUS, and JAXA-AMSR2 (2014-) and GLDASv2.2 (2003-) are limited temporally.

Line 125, How do authors retain two-thirds of these sites, and what are the retained principles?

Text revised for clarity.

L126-128: *"We retained reference sites that have SWE estimates from two-thirds of the products listed in Table 1; this number is roughly equivalent to the number of products covering the full spatial and temporal domain."*

Line 133, the authors highlighted the importance of preventing oversampling in spatially dense areas by limiting the sampling of snow course and gamma SWE to 100 km. However, considering that the resampled grid surpasses the dimensions of certain SWE datasets, could this potentially introduce additional sampling errors that might impact the validation results?

Thank you for your comment. Both reviewers noted the lack of clarity in our description of the spatial aggregation method. We have reworked the description in the revised manuscript (~L134-146) to better describe our approach (see revised text in red further below). We also provide additional clarification immediately below on the rationale for aggregating and on the sensitivity of product statistics to the spatial aggregation distance.

To clarify your question regarding resampling, we did not 'resample' the data as our original text implied. Instead, we averaged the in situ data at the resolution of each product (obtained paired reference-product SWE values) and then aggregated product-reference pairs within a given search radius (100km radius, equivalent to a 200km search window). The spatial aggregation applied in our study is primarily an attempt to minimize biases in the resulting statistics due to the uneven spatial distribution of the reference data. As we show in the figure below (S1), while the choice of aggregation distance impacts the value of the statistics somewhat, it has little-to-no impact on product rankings, and increasing the aggregation distance generally improves product performance up to 200km or so (the increasing amount of aggregation makes the spatial scale of the reference data more consistent with that of most of the gridded products). Our choice of 100-200km for the aggregation scale was intended to obtain a relatively even spatial distribution of the reference data over North America. In addition, it keeps the spatial density of reference data over North American roughly proportional to that over Eurasia, a characteristic that was useful for our companion study noted at the beginning of these responses.

Figure **S1**, presents the validation metrics for spatial aggregation windows between 4km and 500km. The smaller windows (4,10,20,50km) loosely correspond to the resolutions of products tested and allows for better comparison between the aggregated data and native product grid. In general, product ranking remains similar across the range of spatial aggregation distances tested (the horizontal lines rarely cross each other); inter-product differences are minimally impacted by the aggregation. With some exceptions, product metrics improve with spatial aggregation distance as data are smoothed, although the redistribution of data is likely to also have an impact (i.e. less weight given to regions with spatially dense reference data). Unsurprisingly, the change in product metrics with spatial aggregation distance tends to be larger in mountain areas where SWE varies over shorter distances.

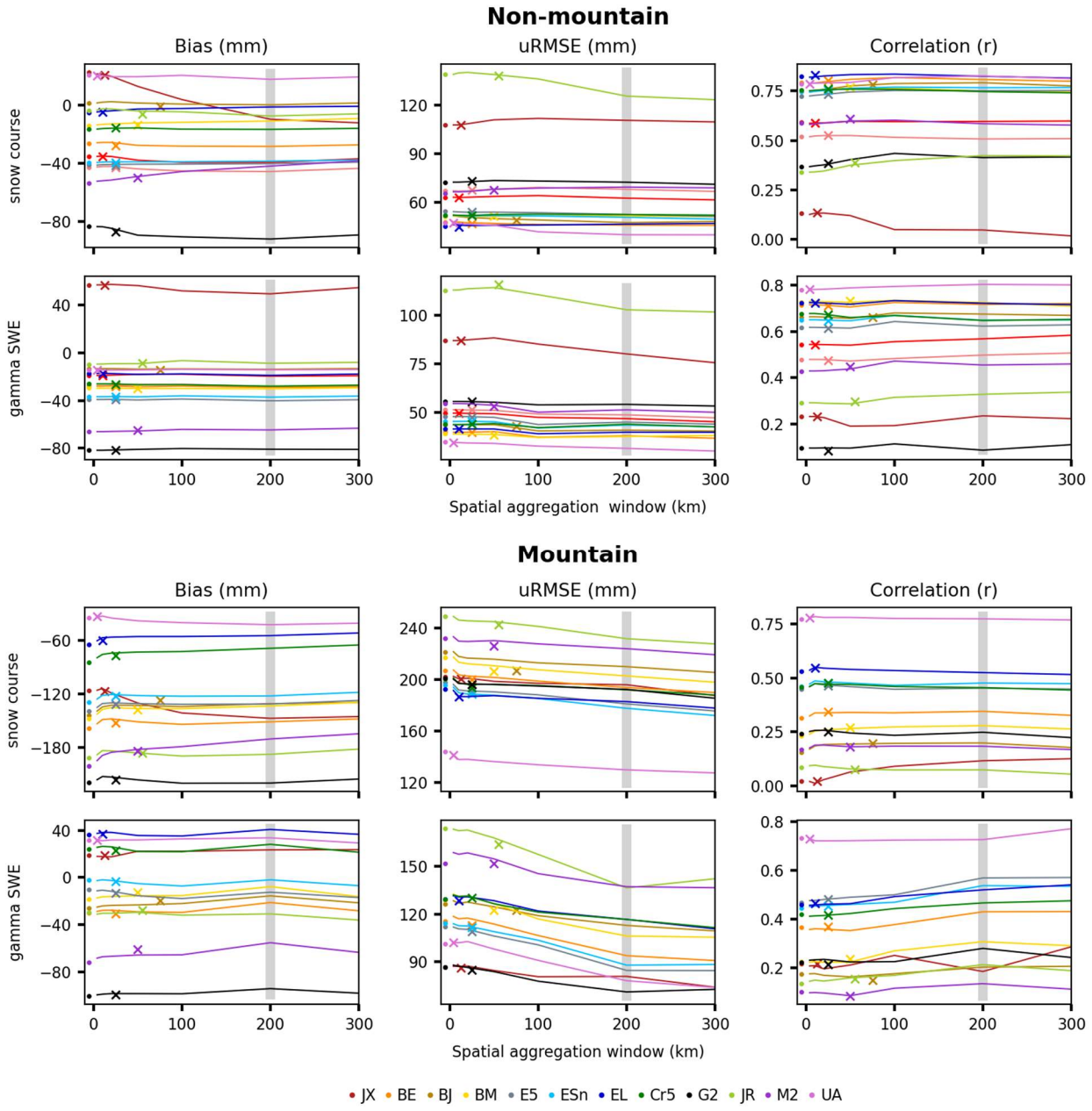

Figure S1: *Product metrics calculated for various aggregation windows (see Sections 3.1 and S0). Crosses show the product metrics calculated at each products' native grid (i.e. all in situ observations on a given date within a product grid cell are averaged together); the circles to the left of zero show the product metrics calculated for all reference-product pairs (no averaging or aggregation). The grey vertical shading at 200km highlights the metrics presented in the manuscript.*

We also propose to add the following to the main body of the text to better explain the purpose of the spatial aggregation step and the general application:

L126-146: *"Reference SWE was matched up in space and time with gridded SWE at the native product resolution. To reduce errors from mismatched water and ice masks, we retained reference sites that have SWE estimates from two-thirds of the products listed in Table 1; this number is roughly equivalent to the number of products covering the full spatial and temporal domain less one to allow for minor differences in product masks. For gamma SWE, we used the midpoint of each flight line for geolocation, which differs slightly from Cho et al. (2019; 2020) and Tuttle et al. (2018) who weighted the average of the gamma SWE footprint (using a fixed diameter of 330 meters assigned to each flight line) contained within each product grid cell. We found that both methods produced similar results, so we used the flight line midpoint for simplicity.*

The reference data were averaged to the resolution of each product. Next, to reduce oversampling of areas with spatially dense networks, all product-reference pairs within sequential 200km windows were averaged (see Supplement Sect. S0). This averaging window corresponds to the range of non-mountain SWE variability (~150-250 km, Pulliainen et al. 2020). Snow course and gamma SWE were considered separately, and mountain measurements were separated from non-mountain. This aggregation approach aims to provide a more even distribution of product errors across landcover types and snow classes. Sensitivity analysis of various spatial aggregation windows between 4 km and 500 km showed little impact of window size on product ranking (Figure S1, limited to 300 km for display purposes). In general, product metrics improve with aggregation window size up to ~100 km but inter-product differences remain fairly consistent. We selected a 200 km aggregation window, as a compromise between sample size and spatial distribution. This approach, which effectively averages the reference data at the scale of the native product grid and then averages product errors within a larger area, is sufficiently flexible to enable the tests of covariates applied in Sections 4.3 through 4.5."

Pulliainen, J., Luojus K., Derksen C., Mudryk L., Lemmetyinen J., Salminen M., Ikonen J., Takala M., Cohen J., Smolander T., Norberg J. 2020: Patterns and trends of Northern Hemisphere snow mass from 1980 to 2018, Nature, 581, https://doi.org/10.1038/s41586-020-2258-0, 2020.

We will also add supplemental text to detail the precise approach. We found that adding these details to the main text induced confusion for the reader.

"As outlined in Section 3.1, we aggregated the reference data at the scale of the native product grid and then averaged the reference-product pairs within a larger window. Because the product grids do not overlay perfectly we did the following:

Sites within 100km of a base site were identified. If, within a given pool of matched reference sites, there were multiple reference-product data pairs within the same native product grid, these pairs were averaged. The mean product and reference SWE within each pool of data were then calculated. This process was repeated sequentially, starting with site ALE-05AA805 and ALE-05FA802 for snow course mountain and non-mountain respectively and AK101 and AB101 for gamma mountain and non-mountain respectively. Sites included in a search pool were dropped from the list and the window moved to the next site on the list. Snow course and gamma SWE were considered separately, and mountain measurements were separated from non-mountain."

Is Figure 2 a scatter plot obtained by sampling the snow course and gamma SWE to 100 km? Why not choose a smaller scale? Will this affect the accuracy of verification results?

Yes, Figure 2 was produced using the aggregated data. As described above aggregating to 200km (100km search radius) minimally impacts the validation. Please also note that during revision we realized what we were describing as a 100km aggregation distance was actually the value used for the reference-product pair search radius and thus equates to a 200km search window. This error has been corrected.

The abbreviations of ESn and CE5 in Figure 2 do not correspond to the abbreviations in Table 1. Please review the product abbreviations throughout the document and make sure they are aligned.

Thank you for noticing these inconsistencies. We will change the abbreviations in Figures 2 and 8 to match those in Table 1 and will modify the abbreviation for ERA5-Snow in the table to ESn to avoid confusion with ERA5 (E5).

Line 133-135: "Sensitivity analysis of various spatial aggregation distances between 50 and 200 km showed little impact of aggregation distance. We selected 100 km as a compromise between sample size and spatial distribution". How can we see "little impact of aggregation distance"? Will the sensitivity analysis results be considered for addition, and further explanation is needed for the selection basis of 50-200 km aggregation distance.

As evidence of this claim, we now include the figure shown above in the supplementary material and refer readers to it.

Since the North American reference SWE dataset was finally constructed in the article, please consider whether further additions are needed using this dataset to verify the SWE grid products mentioned in the article.

A reference dataset very similar to what was used in our analysis is provided here: https://zenodo.org/records/10287093. There are additional data used in our analysis from the Ministère de l'evnrionment, la lute contres les changements climatiques, et de la Faune et des Parcs de Québec that we are unable to share publicly. These data were published following our initial manuscript submission. We have added reference to this dataset in the manuscript where appropriate.

We have reworked the description in the revised manuscript (~L134-146 and Supplement) to better describe our approach (see response and revised text to previous comment).