Second Review of: Assimilation of airborne gamma observations provides utility for snow estimation in forested environments

Overview: The revised manuscript has adequately addressed a majority of my major concerns during the first round of review, including performing some simple sensitivity analysis with Noah-MP to address the poor performance of the open loop simulation, and including some additional details regarding model set up and methods as well as significant steps to improve replicability of the study. Furthermore, the DA of gamma flightline measured SWE in forested regions of the Northeast US is of potential value since this can address a number of issues in the region related to snow characterization on fine scales. In particular, the result quantifying the impacts of localization distance on model performance can inform future data collection strategies and constrain regional SWE estimates from blended model/observational approaches. The DA approach is reasonable, and the gamma-SWE dataset is well validated and widely accepted within the community. Taken altogether, this study is of potential high-value to the community and worthy of publication. However, there are still some lingering larger concerns with the study that should be addressed prior to publication. Additionally, the modifications during the first round of revisions introduced a number of minor technical issues that need to be corrected.

Major comments:

1. While the authors have made substantial strides towards addressing the poor performance of the OL simulations, it is still concerning just how bad the model performance appears to be. While, there are and can be fairly large model errors with Noah-MP, particularly around SWE max and during the melt season these are by far the most egregious that I can recall seeing in the literature. Accordingly, because the results of the OL simulation are so questionable for a widely used and accepted numerical model, I think the bar to publishing these results should be quite high. In the first revision, the authors have made solid efforts to add context to the model performance, however to reach this bar, in addition to what the authors have already done, I recommend adding (or at least investigating, even if these results don't end up in the final manuscript) three specific things:

    a. Find a single example location within the model domain where with an in-situ SWE measurement and compare SWE from the UA dataset, the OL simulation, and different model-based product for snow (e.g., NLDAS). Snow depth could be used as a backup if there are no in-situ SWE measurements in the model domain.

    b. Perform a simple "reality check" analysis of the MERRA-2 forcing against a handful (or even a single) in-situ weather station to explore possible biases associated with the reanalysis forcing in a more direct way than comparing LSM results from different versions.

    c. Include one more Noah-MP simulation with the rain/snow partitioning threshold set to 0.0 instead of using the BATS = 2.5C threshold. I suspect that the lack of sensitivity to the precipitation phase partitioning is tied to the fact that both Jordan (1991) and BATS have that 2.5 threshold for rain/snow within them.

Addressing these three items would do the following: 1) directly contextualize and ground truth the OL model results and the UA dataset with an on-the-ground SWE observation within the region, 2) illustrate, directly, possible biases in the model forcing, and 3) provide a full sensitivity analysis to phase-partitioning. Once these have been done, I think that frees up the authors to be more speculative regarding the model performance, and perhaps even make more general statements on how to improve the model in this region.

2. To me there appears to be some weirdness going on in figure 4 that should be explained.
    a. How does the AMSR-E DA run end up with a later melt out date than all other simulations despite consistently lower SWE throughout the season, and a number of "zero" SWE assimilations in the spring? That is, why is the AMSR-E snowmelt rate dramatically more gradual than the OL or the Gamma-DA simulations?
    b. What is going on at the Rumney time-series that allows for a late-season spikes in SWE time-series in the Gamma-DA simulation that is not reflected in the OL simulation? My first assumption was that there was a late-season snow event each year, but then wouldn't that also show up in the OL simulation? I'm confused how there is this spike in SWE in the DA simulation that appears without a flightline gamma observation to explain it.

3. Additional model details would be helpful here in section 4. For example, if the model is run over a gridded domain, what is the grid-spacing? How does this grid-spacing compare to the MERRA forcing? Would it be helpful to show an outline of the model domain in figure 1? Is figure 1 already *showing* the model domain? Finally, while the authors indicated in their responses that the MERRA-2 data was interpolated to the LIS grid, was it *downscaled* to the terrain at all? (e.g., was temperature adjusted using a lapse-rate?). There is a lot of terrain in NH and ME that the flightlines sample, and the MERRA forcing almost certainly doesn't capture it adequately.

Minor comments:

Line 137: What does UA stand for, I think this is the first time this acronym shows up, please define it.

Line 139: What is the snow density parameterization, reference?

Line 191: The BATS and CLASS albedo schemes are specifically for the snow albedo, not the total ground albedo, please correct this.

Line 197: Would it be helpful to elaborate on the purpose of the ensemble here, e.g., to generate model uncertainty metrics for the DA?

Line 224: “of limited used” should be “of limited use”

Line 231: Consider replacing “the degree of the SWE updates” with “the magnitude of the SWE DA adjustment” or something along those lines.

Line 244 – 249: I suggest rewording “However, this was a consequence of the fact that the overestimated SWE during the accumulation season and early in the melt season was offset by the underestimated SWE during the snowmelt season (i.e., April and May). When the gamma SWE observations exist during the accumulation period (which is a typical case), DA corrected the overestimated SWE, whereas it further underestimated SWE in the snowmelt season (**Figures 3** and **4**), resulting in the increased (negative) bias, as presented in **Figure 2**.”

As

However, this was a consequence of the fact that in correcting the overestimated SWE during the accumulation season, the DA introduced a greater underestimate during the melt season (Figures 3 and 4).

Line 251: I suspect that the r-value decrease is due almost entirely to the increase in the number of zero – to – not-zero comparisons involved in the analysis. The physical lower-boundary of snow as a variable really complicates a lot of these statistical metrics. Perhaps a BETTER comparison would be to only compare data-points where both the simulated and observed SWE are greater than zero.

Line 261: Again, I have trouble with the statement “limited model physics” here. Pending changes made to address major comment 1 above, I think this can be addressed with a simple change of wording to something along the lines of (changes in ***bold/italics***):

“However, the assimilation of the airborne gamma SWE measurements was not able to improve the snow ablation timing ***due to sparse gamma data during the spring in combination with the overall poor model performance during the melt season***”

Essentially, any rewording that more generally acknowledges the poor performance in this specific instance, rather than speculating that there is something intrinsically wrong with the model physics.

Line 264: “single gamma SWE” should be “single gamma SWE flight?”

Line 273: I suggest that you remove “In the figure” from the beginning of the sentence.

Line 275: “has low bias and RMSD than OL SWE”. This is an incomplete though, the model has a low bias, and a ??? RMSD compared to OL SWE? Higher? Lower? Equivalent? Please fix.

Line 275: Suggested replace: “DA performances show relatively” to “DA led to”

Line 279 (and more generally throughout the manuscript): Consider replacing the word “updated” with “improved.” Using the word updated is ambiguous regarding whether or not the simulation was improved. I’ll note it here, but I recall seeing the use of “updated” where a better word choice is possible at other places throughout the manuscript as well.
Line 294: consider replacing “surrounding areas, where gamma flights do not exist” with “In areas surrounding the gamma-flights”

Figure 6 and Lines: 297 – 307: This is a nice result. Is there a way to assess statistical significance in this comparison? E.g., at what localization distance are the improvements compared to the OL no longer statistically significant (or are all of these significant)? The results here are somewhat convincing visually, so I leave any decision to include a statistical significance analysis here to the authors.

Line 297: what is the “effective surrounding areas”? is this a fixed number for ALL localization radii? Or only gridcells within the localization radii?

Line 301: Why is the OL RMSD decreasing as localization radii?

Line 316: replace “reduced the model SWE errors” with “improved model SWE”

Line 317: There is an errant open parenthesis here.

Line 332: replace the word “ground” with the word “snow”

Line 334: add, “snow albedo” after “BATS and CLASS”

Line 340: replace “but it still has” with “but did not improve”

Line 340: replace “combinational” with “combined”

Line 341: “fully-implicit” should be “the fully-implicit”

Line 344: replace “worth to note” with “worth noting”

Line 364: I don’t think I expect much of a different between Jordan and BATS precipitation, but rather differences to be most pronounced when T thresh is set to 0C (Letcher et al. 2022 explores the 2.5 vs. 0C difference in this region).

Lines 367 and 370, please replace the word “Noah-MP” with “BATS” since Noah-MP has different options for precipitation phase partitioning, so the 2.5C threshold is specific to the BATS choice, not the Noah-MP model.