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

Recommendation Major Revisions:

The authors have put forth an interesting study that aims to quantify the impact of assimilating airborne gamma SWE measurements into the Noah-MP land surface model. This study focuses on forested regions in the Northeast United States, a region that is underrepresented in snow-hydrology research. Furthermore, the choice to assimilate airborne SWE measured from gamma radiation into Noah-MP and quantify the impacts is somewhat novel. The use the LIS, airborne gamma measurements and UA datasets is appropriate, and the paper is generally well structured. Finally, this topic of study is interesting and the results are of clear value to both the research community and regional stakeholders. Accordingly, this study fits well within the journal scope and is worthy of publication. However, I see two major deficiencies with in the work that need to be addressed prior to publication.

First, I do not see how any reasonable person would be able to replicate this work based on the information provided in the methods section and the data availability statement. I think this needs to be addressed, especially considering that many journals are moving towards an increased emphasis on replicability and open data practices. In my view, significant effort is required to modify the method section, and more data should be made available (if not large model output datasets, at least model config files and namelists) before this study could be considered open and transparent.

The second major deficiency is that the open-loop (OL) Noah-MP simulation performs so poorly that I'm left wondering if the model was configured or forced properly. This feeling is exacerbated by the fact that the authors provide very little information regarding the model configuration. Further, the authors make several vague and dismissive statements pinning the poor performance of the model on "model physics" with no supporting evidence. Considering that there are now dozens of studies that show that Noah-MP has rather good performance with respect to snow (including over the Northeast), the exceedingly poor model performance in this study is an outlier, and should be addressed. I suspect that the poor model performance is related to the model configuration chosen for this specific study rather than issues intrinsic to the model. So, prior to publication, the authors should revisit their baseline OL model configuration and track down some of the causes for the poor model performance. My suspicion is that in doing so, they will be able to attain much better accuracy within the OL simulation, and accordingly, better and more robust results regarding the impact of the SWE DA. Once these two issues are addressed, I think the study will easily meet the criteria for publication and make a wonderful additional to the snow-hydrology literature.

Major Specific Comments:

Line 175 – 190: This section could be improved with additional information regarding the Noah-MP configuration. There are several specific questions that I have regarding the model set up that may go towards explaining some of the results in the results section, particularly with regards to the melt season:

1) Was the TCF described in section 3.4 used to inform the Noah-MP vegetation fraction? And if not, was the Noah-MP vegetation fraction used in the model compared to the TCF for consistency?
2) Was forcing data downscaled to the terrain at all?
3) Similarly, was Noah-MP run on a grid covering a region? Or ONLY for "grid-cells" that would have assimilated data, i.e., along the flight-path + gridcells located within the localization radius?
4) What were the physics parameterizations? It's mentioned later on in the paper that the Jordan phase partitioning is used, but equally as relevant would be the snow albedo option (BATS vs. CLASS), the melt-factor used in determining the pixel snow-cover-fraction, the radiation_scheme option (specifically whether or not Fveg is used) and the temp_time_scheme option. (Cuntz et al, 2016 and You et al. 2022 illustrate the impact of some of these options on snow).
5) What went into the ensemble members? Presumably, this was an ensemble with perturbed initial / forcing conditions? Or was it a physics ensemble?

Towards this end, the data-availability statement of "To replicate the land surface model simulation and data assimilation, users can freely access …" is **entirely insufficient** for replicating the model simulation. At a *bare minimum*, any/all lis.config files used to run the simulations should be included somewhere as supplementary material. Further, was this in off-the-shelf version of LIS? Or one that was specifically modified to assimilate airborne gamma snow observations?

Lines 227 – 229: "*Figure 2 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)"*

The statement appears generally true looking at the time-series data, but that raises more pressing questions about the OL config. To me it looks like the OL model config has serious issues that cancel each other out in the bulk sense and leads a reasonable overall bias, were an assimilation of SWE can correct one of those issues, leading to worse results. Regardless, the fact that the OL simulation often shows a peak SWE over-estimate greater than 100mm (50-80% of observed) *and still* melts out 2-4 weeks before the observed melt-out date is a red flag. In particular, the time-series for WY 1989 and WY 1997 are concerning. I *strongly recommend* that the authors a) contextualize the poor performance of the OL simulation with other recent studies evaluating Noah-MP snow in this region (e.g., Letcher et al. 2022 or Sthapit et al. 2022), and/or b) attempt to track down and correct the source of the poor performance. My experience is that the Noah-MP performance is strongly tied to the quality of the forcing data, so I recommend at least doing some analysis comparing the MERRA forcing to in-situ data in the region. Really, any efforts to try and understand why the OL model is performing so poorly would improve the manuscript quality. My feeling is that, if Noah-MP is configured well and driven with good forcing data, the OL performance would be almost certainly be much better than this paper suggests. If the OL simulation is configured for good performance, the impact

of the DA will be more accurate and robust, so it is well worth the effort to try and improve the baseline simulation.

Line 243 – 248: This whole section needs additional analysis. The conclusions presented by the authors seem to make no effort in gathering evidence to support them. Rather, speculative comments like "probably due to limited model physics" or "may be attributed to model structure physics" are used to explain why the DA didn't help much during the spring. I don't think these types of statements really belong in a results section, especially when there are ample analyses that could have been done to support or refute them. Considering that there are several studies showing fairly decent Noah-MP performance, I don't think the poor performance found here is attributable to intrinsic issues with the model architecture.

"However, the assimilation of the airborne gamma SWE measurements was not able to improve the snow ablation timing probably due to temporally sparse gamma data as well as limited model physics"

I disagree with this statement; I and I see very little evidence in the paper supporting it. For instance, WY2015 seems to have a gamma-observation during the ablation season, and this does not improve the simulation at all.

General comment: Have the authors considered other metrics for quantifying snow improvement such as peak SWE amount and timing? Or melt out date? RMSE can be a tricky metric to quantify snow, since snow has a lower bound of zero (i.e., a lot of data points will simply be comparing zeros to each-other). See Rhodes et al. (2018) and Trujillo and Molotch (2014).

References:

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S. and Thober, S., 2016. The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model. Journal of Geophysical Research: Atmospheres, 121(18), pp.10-676.

Rhoades, A.M., Jones, A.D. and Ullrich, P.A., 2018. Assessing mountains as natural reservoirs with a multimetric framework. Earth's Future, 6(9), pp.1221-1241.

Sthapit, E., Lakhankar, T., Hughes, M., Khanbilvardi, R., Cifelli, R., Mahoney, K., Currier, W.R., Viterbo, F. and Rafieeinasab, A., 2022. Evaluation of Snow and Streamflows Using Noah-MP and WRF-Hydro Models in Aroostook River Basin, Maine. Water, 14(14), p.2145.

Letcher, T.W., Minder, J.R. and Naple, P., 2022. Understanding and improving snow processes in Noah-MP over the Northeast United States via the New York State Mesonet.

Trujillo, E. and Molotch, N.P., 2014. Snowpack regimes of the western United States. Water Resources Research, 50(7), pp.5611-5623.

You, Y., Huang, C., Yang, Z., Zhang, Y., Bai, Y. and Gu, J., 2020. Assessing Noah-MP parameterization sensitivity and uncertainty interval across snow climates. Journal of Geophysical Research: Atmospheres, 125(4), p.e2019JD030417.

Minor comments:

There are numerous instances of RMSD and a couple of instances of RMSE throughout the paper, so my assumption is that RMSD is preferred. I'm assuming that RMSD is Root Mean Square Difference? Either one is fine, but please be consistent, and consider spelling out RMSD in the paper when it's first used.

Line 206: Was the model snow LWC content updated proportionally? Or was any SWE added/subtracted through assimilation considered all snow? E.g., if the snowpack was 5% liquid, was SWE added -> LWC= LWC + 0.05*NEW and FROZ = FROZ + 0.95*NEW?

Line 225 -226: Did the authors mean: "closer to" instead of "closed to?" Since the authors include the median slope of the linear regression for the OL, would it make sense to also include it from the DA run?

Lines 226-227:
Was Figure 2 only for grid-cells that received DA updates? OR ALL gridcells (see previous comment on model config)?

Lines 226-227: I recommend rewording: "The lower
bias of the SWE estimates from the OL as compared to the DA in Figure 2 …" to be more specific: "The absolute SWE bias was higher in the DA simulation compared to the OL simulation (Figure 2b). However, this was a consequence of the fact …."

Line 237: Please replace the word "enhanced" with "improved"

Line 240: Recommend deleting "As shown in the figure"

Section 5.2 throughout:
I recommend replacing references to the "lower" and "higher" groups with "low" and "high" since this is how the groups were introduced.

Line 252: What is "differences in the 1:1 slope?" is it differences in the linear-regression slope between the OL and DA simulations? Or is it differences between the linear-regression slope AND the 1-to-1 line for each simulation? Please clarify.

Figure 4: I suggest increasing the font size on the legend to be more readable.

Line 253 – 255: Consider rewording: “In the figure, two groups of each land surface characteristics were determined by dividing the gamma flight lines into two (i.e.,low and high) groups of equal numbers of the flight lines.”

As:

“In this analysis, the land-characteristics sampled by the gamma flights were divided equally into two groups (i.e., low and high) such that the land characteristic value separating the two groups allowed for equal numbers of samples in each group.”

Line 260: what does “updated” mean here? Improved? Changed? Please clarify.

Line 264: “added value of the gamma SWE data on the model SWE estimates via assimilation” This is really awkward wording, please rephrase.

Figure 6: I’m a little confused how the mean slope for the 4km localized OL simulation is near 1-to-1 as compared to 1.8/1.9 shown in figure 3. I’m also confused as to why there is an increase in the error metrics in the OL simulation as the localization distance is increased? Does that imply that the model is better over the flight-lines, even without any DA? Am I just interpreting this analysis wrong? Is it just that the 4km localization provides fewer grid points, and therefore fewer opportunities for the model to accumulate large errors? Also, if the OL slope for 4km is close to one, how does that correspond to such high (50-60mm) bias? I don’t think the y-intercept would explain that.

My understanding is that the model is running Noah-MP along some flight-path, and some number of adjacent gridcells corresponding to various different radii of influence. Shouldn’t the OL metrics be independent of how many gridcells are in the simulation? Either way, the methods here are pretty vague, I recommend more clearly explaining the experiment to eliminate confusion.

Line 280: “closed” should be “close”

Line 307-308: “The overestimated SWE during the snow accumulation period was likely attributed to a precipitation phase partitioning method employed in Noah-MP.”

This could very well be true, but this is something that could be easily tested by running a couple of simulations with different temperature thresholding (e.g., PCP_PARTITION_OPTION = 3, instead of 1 in the namelist). That way the authors wouldn’t have to speculate here.

Lines 322 – 324: “Although the snowpack in Noah-MP can have up to three snow layers, it may not be enough to accurately reproduce the energy budget within the snowpack in the study area.”

Do you have a citation to support this statement? While there is probably at least some truth to it, I'm not sure that I agree 100%. In terms of simulating bulk SWE, I don't often see significant improvement with the addition of *more* snow-layers, just so long as the model has multiple discrete snow layers in it.

Line 344 – 345: "However, as proven in our findings, effective uses of the gamma SWE (e.g. localization function) will maximize the utility of the gamma SWE into the DA framework."

I would scale back the certainty in this statement: e.g., replace "proven in" with "supported by" and "will maximize" with "can enhance."

Lines 347- 348: "promising alternative"; Alternative to what?

Lines 356 – 358: "The added value of the gamma data on the model SWE estimates was greater for the relatively lower VCF range. For areas with higher topographic heterogeneity, the gamma-based DA SWE was still effective in reducing the errors."

This is very awkwardly worded, the first sentence is in reference to the vegetation fraction, and the second was for the topography heterogeneity. Specifically, the use of the phrase "still effective" is confusing since there is no "base-state" effectiveness of the DA for terrain heterogeneity described here in the conclusions. Please reword.

Lines 361 – 362: "uncertainties in the Noah-MP physics (i.e. precipitation partitioning and simplified snow layers)."

I really do not think that this statement belongs in the conclusions, since it's just speculation on the part of the authors.

Data Availability:
Definitely a lot to be desired here, by only pointing readers to publicly available datasets, and choosing not to include model configuration data, analysis code, or processed (example) data, there is simply no way that a reasonable person would be able to repeat the experiment and analysis described in the paper. At a bare-minimum, the model configuration files should be included, as should any code that processes the data prior to being fed into the model.

Author Contributions:
I'll leave this to the editors, but I'm not sure "provided the funding" is a good reason to include for co-authorship. While in this instance it doesn't matter since SVK and CMV provided other input that would merit co-authorship, explicitly listing "provided funding" seems like it could be problematic.