General Comment:

The work has improved from the last draft; however, several conclusions in this paper overstate the magnitude of improvement, terminology drifts into subjective language, and a few structural/citation issues need attention.

We thank the reviewer for their further comments and commitment to improving this manuscript. Our responses below, are in red, with manuscript text in dark red and *additions to the manuscript in italics.*

Major Comments:

I don't agree with the overall conclusion of the paper that DAvar provides substantial improvement over DAconst. The claimed performance gains of DAvar over DAconst are statistically significant but small in absolute terms (<10 cm and, in places, ∼2 mm), and not uniform across sites. Please change the framing in the Result/Discussion/Conclusions to emphasize the limited magnitude and spatial inconsistency and to discuss whether the added complexity of DAvar is justified by these gains.

We will modify the results section to be as quantitative and objective as possible, removing language such as "substantial", and adding quantification where necessary. We will clarify that improvements are "*small, but significant*" in places where we test statistical significance. We will further add a few sentences on spatial and temporal inconsistencies in the first paragraph of the discussion:

"The snow depth estimated from this ML model has been shown to possess superior accuracy compared to prior S1 snow depth retrieval work by Lievens et al. [2022] (SD_{S1}) [Dunmire et al., 2024], which has previously been assimilated into the Noah-MP land surface model using an Ensemble Kalman Filter [De Lannoy et al., 2024, Brangers et al., 2024]. *Recent work by Mirza et al. [2025] has questioned the utility of assimilating S1 snow depth retrievals, highlighting inconsistencies in temporal and spatial errors of the SD_{S1} in the Western United States, where less regular S1 data are available. Despite advancements made by SD_{ML}, the quality of the ML-based observations assimilated in this study also varies across space and time, which can lead to localized degradations in DA performance (e.g., Fig. 3). Although improving mountain snow depth estimation is an active area of research, progress is limited by the current suite of satellite sensors, which are not specifically designed for snow depth or SWE retrieval. Future DA efforts that incorporate more reliable snow depth or SWE products should reduce these spatial and temporal inconsistencies, improving overall DA performance.*"

Furthermore, in the previous iteration of review, we added the following text to discuss spatial inconsistencies. We are grateful for the reviewer's previous comments and believe this discussion has improved the manuscript. However, we believe that the current level of discussion sufficiently addresses spatial inconsistencies.

"While $\mathrm{DA_{var}}$ improves performance at most snow depth and SWE measurement sites, some locations see little benefit, or even a deterioration in performance (approximately 12% of snow depth sites and 20% of SWE sites). These degradations are more likely to occur where the SD_{ML} product is less accurate than the OL experiment, *and the DA_{var} experiment more strongly corrects to these inaccurate observations.* To account for known limitations of SAR-based snow depth retrievals, we did not assimilate the SD_{ML} product over dense forests or glaciers, and after March 31. Nevertheless, SD_{ML} remains inaccurate in some places, leading to localized deterioration when these observations are assimilated. Locations with minimal differences between $\mathrm{DA_{const}}$ and $\mathrm{DA_{var}}$ typically occur where the observations already agree well with the OL, or where $\sigma_{obs} >> \sigma_f$, thus the DA increments are small, and the model receives limited benefit from the observational information."

We will also add the following text to the conclusions to acknowledge the small and spatially inconsistent improvements in $\mathrm{DA_{var}}$:

"*However, given limitations of the assimilated satellite-based snow depth product, improvements from the DA, or from the specific implementation of a dynamic observation error in* $\mathrm{DA_{var}}$*, are limited in magnitude and not spatially consistent.* As most snow DA work and operational snow DA systems assume that the observational uncertainty is constant in space and time, this work highlights the impact of these assumptions, and the importance of observation uncertainty considerations when designing a DA system. *Future studies should put effort into the consideration of observation uncertainties and the parameterization of observation uncertainty should depend on study goals, the DA system used, and specific characteristics of the assimilated observations.*"

Finally, regarding whether the added complexity of $\mathrm{DA_{var}}$ is justified by these gains: very recent work by Gichamo et al. [2025] demonstrates improvements of similar, and even slightly smaller magnitude (SD MAE reduction of 12 mm) when switching from an Optimal Interpolation to a EnKF for snow data assimilation (DA) in NOAA's NWP system, the Global Forecast System (GFS). The authors conclude that "the results are encouraging and motivate implementing ensemble methods for the snow data assimilation in NWP systems, in place of the current OI-based systems". This implementation of ensemble systems would be a much larger step in increased complexity than going from $\mathrm{DA_{var}}$ to $\mathrm{DA_{const}}$ in our work (which resulted in a SD MAE reduction of 22 mm). In light of recent work highlighting the value of improvements of this magnitude in snow DA systems, we do not believe that achieving comparable, in fact slightly larger, gains should be downplayed here, especially given the minimal added complexity of $\mathrm{DA_{var}}$.

Some results are currently summarized with site averages that can be skewed by a few poor sites; compare medians for DAconst vs. DAvar and, if feasible, repeat significance testing on medians.

It is true that the distribution of SD and SWE MAE is right-skewed. As such, we have used a Mann-Whitney U test (an alternative to the two-sample independent t-test when the data is not normally distributed) to test for significantly different distributions in the evaluation metrics. We will specify our use of the Mann-Whitney U test in the results section. Where relevant, we will also report median MAE values and include significance tests on the medians. Below are the instance where these significance tests will be included:

- For snow depth: "While improvement in MAE from the OL experiment is not significant for $\mathrm{DA}_{\mathrm{const}}$ (*Mann-Whitney U test p-value = 0.59, median-test p-value = 0.68*), the MAE improvement is *small, but* significant for $\mathrm{DA}_{\mathrm{var}}$ (*Mann-Whitney U test p-value = 0.001, median-test p-value = 0.03*)".
- For temporal ACC: "The improvement in temporal ACC for $\mathrm{DA}_{\mathrm{var}}$ from both the OL and $\mathrm{DA}_{\mathrm{const}}$ is statistically significant (*$p < 0.01$ for both a Mann-Whitney U test and median-test*, Fig. 3d)."
- For SWE: "Compared with 8,211 manual SWE measurements from 231 different measurement sites across the Alps, the $\mathrm{DA}_{\mathrm{var}}$ experiment also *offers small, but significant* improvements for SWE MAE compared to both the OL and $\mathrm{DA}_{\mathrm{const}}$ experiments (*$p << 0.001$ for both a Mann-Whitney U test and median-test*)."

Avoid subjective terms such as "substantial" where differences are on the order of millimeters; replace with exact values.

We will remove subjective language such as "substantial" and quantify where necessary.

The statement in discussion line 435 that modest snow/SWE improvements translate to streamflow gains is out of context here; withdraw or support it with streamflow evidence.

We will remove this sentence from the discussion.

Abstract can also be one or two lines with clear results. Right now, the abstract is vague without stating a clear outcome of the study.

We will add the following sentences to the abstract to summarize the results:

"*The* $\mathrm{DA}_{\mathrm{var}}$ *experiment offers small, but significant improvements to snow depth and snow water equivalent (SWE) mean absolute errors (MAE), and slightly reduces snow cover, thereby better matching satellite-based snow cover observations. Compared to an open loop (no DA) experiment (OL), and an experiment with an assumed static observation error* ($\mathrm{DA}_{\mathrm{const}}$)*,* $\mathrm{DA}_{\mathrm{var}}$ *reduces SWE MAE by 25% and 13%, respectively, compared with over 8000 manual SWE measurements.* This work demonstrates the benefits of machine learning based snow depth retrievals and the impact of incorporating dynamic observation errors in EnKF-based snow DA."

Specific clarifications.

Where a “15 mm improvement” is cited (Lines 301–306), specify the metric (likely bias) and state the exact value and sign.

We will specify that the 15 mm improvement refers to SWE MAE.

Replace “standard WY 2016/2017” with an unambiguous convention (e.g., “Water Year 2017”) to avoid seasonality confusion.

We will replace “2016/2017” with *“Water Year 2017”*. We have also replaced “2017/2018” (L299) with *“Water Year 2018”* and “2020/2021” (L302) with *“Water year 2021”*.

Across Lines 295–314, report concrete numbers rather than qualitative characterizations.

We will add a quantitative assessment to Lines 295-314. This section will be modified to the text below:

“Across all experiments, SWE typically peaks during the first week of March (March 1–7). Water Year 2017 recorded the lowest modeled SWE in our OL experiment, and correspondingly saw the largest SWE increases in $\mathrm{DA_{var}}$ prior to early March, particularly in the Central Alps and Austrian Alps (Fig. 6a). However, $\mathrm{DA_{var}}$ SWE improvements were mixed during this year. Of the 41 manual measurements taken between March 1 and March 7, 2017, only 24% demonstrated improved SWE MAE of more than 15 mm in $\mathrm{DA_{var}}$. While the DA led to more accurately estimated SWE at some sites (e.g., Supplemental Fig. S3b,d), it resulted in an overestimation of SWE at others (e.g., Supplemental Fig. S3c,e,f). *For example, three measurement sites in Italy (dark pink dots in Fig. 6a) experienced an average increase of 101 mm in added SWE in* $\mathrm{DA_{var}}$ *relative to the OL. The average SWE MAE at these sites increased by 134 mm in* $\mathrm{DA_{var}}$*, indicating that the assimilated* SD_{ML} *observations overestimate snow at these locations. The degradation is even larger in* $\mathrm{DA_{const}}$*, where the SWE MAE increases by 193 mm compared to the OL. This stronger deterioration arises from the lower assumed* σ_{obs} *in* $\mathrm{DA_{const}}$ *at these locations, which leads to stronger corrections toward the observations. A time series of modeled and observed SWE at one of these sites is shown in Supplemental Fig. S3e.*

The largest SWE reductions from the OL to the $\mathrm{DA_{var}}$ experiment occurred during Water Year 2018, particularly in the Bavarian Alps, Swiss Alps, and French Alps (Fig. 6b). In general, the reduced SWE in $\mathrm{DA_{var}}$ aligns more closely with in-situ observations (e.g., Supplemental Fig. S4). *The average SWE MAE for in-situ measurements taken between March 1-7, 2018 decreases from 164 mm in the OL, to 137 mm in* $\mathrm{DA_{const}}$ *and 116 mm in* $\mathrm{DA_{var}}$*. In* $\mathrm{DA_{var}}$*, SWE MAE is improved by more than 15 mm in 59% of the 68 manual measurements taken between March 1 and March 7, 2018.*

Water Year 2021 also experienced a large SWE reduction between the OL and $\mathrm{DA_{var}}$ experiments, especially in the Swiss Alps and Eastern Dolomites. In the Dolomites region, *where SWE reductions are often greater than 100 mm*, a lack of in-situ observations makes it difficult to assess whether these reductions are realistic. However, limited measurement sites along the Italy-Austria border suggest that the SWE reductions may be too strong (e.g., Supplemental Fig. S5d). *For instance, two in-situ measurements sites along the Italy-Austria border (indicated with yellow circles in Supplementary Fig. S5a) have an average SWE decrease of 142 mm in* $\mathrm{DA}_{\mathrm{var}}$*, and a corresponding degradation in SWE MAE of +113 mm. Meanwhile, southwest of these locations, 8 measurement sites in Italy (black box in Supplementary Figure S5a) demonstrate contrasting improvements in* $\mathrm{DA}_{\mathrm{var}}$ *SWE MAE. At these eight sites, SWE decreases by an average of 100 mm in* $\mathrm{DA}_{\mathrm{var}}$*, with a corresponding 74 mm reduction in SWE MAE. This result highlights some of the spatial inconsistencies of the DA improvements, which are likely due to spatial and temporal variation in the quality of the assimilated observations.*

In the concluding sections (Lines 385–391), explicitly acknowledges that DAvar's advantage over DAconst is slight and not pervasive, and suggests that method choice should depend on study goals and acceptable complexity.

In the conclusion, we will add the following text to acknowledge the limited improvement and spatial inconsistencies. We will also suggest that the design of future studies should depend on study goals, the DA system used, and characteristics of the assimilated observations.

"*'However, given limitations of the assimilated satellite-based snow depth product, improvements from the DA, or from the implementation of a dynamic observation error in* $\mathrm{DA}_{\mathrm{var}}$*, are limited in magnitude and not spatially consistent.* As most snow DA work and operational snow DA systems assume that the observational uncertainty is constant in space and time, this work highlights the impact of a these assumptions, and the importance of observation uncertainty considerations when designing a DA system. *Future studies should put effort into the consideration of observation uncertainties and the parameterization of observation uncertainty should depend on study goals, the DA system used, and specific characteristics of the assimilated observations.*"

Line-by-line:

Line 73: add PBS downscaling citations (Bachand 2025; https://doi.org/10.1175/JHM-D-24-0131.1).

We will modify the sentence in L73 and add the suggested citation: "In particular, particle batch smoothers have been commonly applied to create snow reconstructions [Margulis et al., 2015, Baldo and Margulis, 2018] *or to downscale model variables such as precipitation [Girotto et al., 2024, Bachand et al., 2025]*."

Lines 75, 100, 114: insert https://doi.org/10.5194/egusphere-2025-978 wherever relevant to substantiate recent usage and performance.

We will add the above citation to the following sentences:

- "Recent studies have used both particle batch smoothers and the EnKF to assimilate

SAR-based snow depth retrievals from Sentinel-1 (S1), thereby improving modeled snow depth, SWE and streamflow compared to in-situ measurements [De Lannoy et al., 2024, Brangers et al., 2024, Girotto et al., 2024, Mirza et al., 2025]."

- "Recent work by Mirza et al. [2025] has questioned the utility of assimilating S1 snow depth retrievals, highlighting inconsistencies in temporal and spatial errors of the SD_{S1} in the Western United States."
- "ERA5 has previously been used as atmospheric forcing in other snow DA studies [Pflug et al., 2024, De Lannoy et al., 2024, Mirza et al., 2025]..."

Section 2.1: Section 2.1 name is misleading: the title ("Noah-MP land surface") suggests a model description, but the text mixes model and forcing details. Rename to "Model setup and data" or similar, and keep model vs. forcing clearly separated

As suggested, we will rename Section 2.1 *"Model setup and data"*. To clearly separate model and forcing we will further add two subsections: *"2.1.1 Noah-MP land surface model"* and *"2.1.2 Atmospheric forcing for Noah-MP"*.

Line 260: compare medians (and re-test significance on medians if earlier tests used means).

We will report the median SD MAE values and test significance on these medians. As described above, we also use a Mann-Whitney U test to test for significantly different distributions of non-normal data.

"Both the $\mathrm{DA_{const}}$ and $\mathrm{DA_{var}}$ experiments improve these metrics, with site-average MAE values of 0.237 m and 0.215 m (*median values of 0.207 m and 0.185 m*), RMSE values of 0.292 m and 0.268 m, and biases of 0.106 m and 0.055 m, respectively."

"While improvement in MAE from the OL experiment is not significant for $\mathrm{DA_{const}}$ (*Mann-Whitney U test p-value = 0.59, median-test p-value = 0.68*), the MAE improvement is slight, but significant for $\mathrm{DA_{var}}$ (*Mann-Whitney U test p-value = 0.001, median-test p-value = 0.03*)."

Lines 295–300: replace "substantial" with exact mm values and note that benefits must be weighed against DAvar complexity.

We will modify this text with: "In the OL experiment, we observe a positive bias for low observed SWE and a negative bias for high observed SWE (Fig. **??**c), similar to the bias patterns seen for snow depth. The $\mathrm{DA_{var}}$ experiment reduces both biases, with the largest improvements occurring for low observed SWE values. *For instance, for in-situ SWE below 200 mm, the bias is reduced by 52% in* $\mathrm{DA_{var}}$ *compared to the OL (OL bias = +166 mm,* $\mathrm{DA_{var}}$ *bias = +80 mm), meanwhile the bias in-situ SWE measurements above 600 mm is reduced by 7% in* $\mathrm{DA_{var}}$ *(OL bias = -362 mm,* $\mathrm{DA_{var}}$ *bias = -335 mm).*"

Lines 301–314: report actual numbers; avoid "substantial."

As discussed above, we will modify to text in these lines to include a more quantitative analysis, avoiding subjective language.

References

Claire L. Bachand, Lauren C. Andrews, Tasnuva Rouf, and Manuela Girotto. The Utility of Satellite Snow Depth Observations for Downscaling Hydrologic Variables over the Indus Basin Mountain Ranges. *Journal of Hydrometeorology*, 26(5):555–575, 5 2025. ISSN 1525-755X. doi: 10.1175/JHM-D-24-0131.1.

Elisabeth Baldo and Steven A. Margulis. Assessment of a multiresolution snow reanalysis framework: a multidecadal reanalysis case over the upper Yampa River basin, Colorado. *Hydrology and Earth System Sciences*, 22(7):3575–3587, 7 2018. ISSN 1607-7938. doi: 10.5194/hess-22-3575-2018.

I. Brangers, H. Lievens, A. Getirana, and G. J. M. De Lannoy. Sentinel-1 Snow Depth Assimilation to Improve River Discharge Estimates in the Western European Alps. *Water Resources Research*, 60(11), 11 2024. ISSN 0043-1397. doi: 10.1029/2023WR035019.

Gabriëlle J. M. De Lannoy, Michel Bechtold, Louise Busschaert, Zdenko Heyvaert, Sara Modanesi, Devon Dunmire, Hans Lievens, Augusto Getirana, and Christian Massari. Contributions of Irrigation Modeling, Soil Moisture and Snow Data Assimilation to High-Resolution Water Budget Estimates Over the Po Basin: Progress Towards Digital Replicas. *Journal of Advances in Modeling Earth Systems*, 16(10), 10 2024. ISSN 1942-2466. doi: 10.1029/2024MS004433.

Devon Dunmire, Hans Lievens, Lucas Boeykens, and Gabriëlle J.M. De Lannoy. A machine learning approach for estimating snow depth across the European Alps from Sentinel-1 imagery. *Remote Sensing of Environment*, 314:114369, 12 2024. ISSN 00344257. doi: 10.1016/j.rse.2024.114369.

Tseganeh Z. Gichamo, Clara S. Draper, and Michael Barlage. Improving NOAA's global NWP snow data assimilation by updating to an Ensemble Kalman Filter. *Journal of Hydrology*, 660:133301, 10 2025. ISSN 00221694. doi: 10.1016/j.jhydrol.2025.133301.

Manuela Girotto, Giuseppe Formetta, Shima Azimi, Claire Bachand, Marianne Cowherd, Gabrielle De Lannoy, Hans Lievens, Sara Modanesi, Mark S. Raleigh, Riccardo Rigon, and Christian Massari. Identifying snowfall elevation patterns by assimilating satellite-based snow depth retrievals. *Science of The Total Environment*, 906:167312, 1 2024. ISSN 00489697. doi: 10.1016/j.scitotenv.2023.167312.

Hans Lievens, Isis Brangers, Hans-Peter Marshall, Tobias Jonas, Marc Olefs, and Gabriëlle De Lannoy. Sentinel-1 snow depth retrieval at sub-kilometer resolution over the European Alps. *The Cryosphere*, 16(1):159–177, 1 2022. ISSN 1994-0424. doi: 10.5194/tc-16-159-2022.

Steven A. Margulis, Manuela Girotto, Gonzalo Cortés, and Michael Durand. A Particle Batch Smoother Approach to Snow Water Equivalent Estimation. *Journal of Hydrometeorology*, 16(4):1752–1772, 8 2015. ISSN 1525-755X. doi: 10.1175/JHM-D-14-0177.1.

Bareera N. Mirza, Eric E. Small, and Mark S. Raleigh. Evaluating the Utility of Sentinel-1 in a Data Assimilation System for Estimating Snow Depth in a Mountainous Basin. *Cryosphere Discussions*, 3 2025. doi: 10.5194/egusphere-2025-978.

Justin M. Pflug, Melissa L. Wrzesien, Sujay V. Kumar, Eunsang Cho, Kristi R. Arsenault, Paul R. Houser, and Carrie M. Vuyovich. Extending the utility of space-borne snow water equivalent observations over vegetated areas with data assimilation. *Hydrology and Earth System Sciences*, 28(3):631–648, 2 2024. ISSN 1607-7938. doi: 10.5194/hess-28-631-2024.