Overall Comment:

This manuscript explores the assimilation of a machine learning-derived Sentinel-1 snow depth product into a NOAH land surface model using a dynamically varying observation error. While the approach is innovative and addresses an important limitation in current snow data assimilation systems, the demonstrated improvements over the static error method are relatively modest/marginal and not consistent across space or time. Given the added complexity of implementing a dynamic error model, I do not fully agree with the authors' conclusion that this approach provides a clear performance advantage. The manuscript has potential, particularly if it reframes the findings to emphasize that the benefits of dynamic error observation methods are highly dependent on the pattern and variability of observation errors. One has minimal leverage on the other, and a more thorough error characterization is essential for selecting an appropriate DA strategy for the problem at hand.

Minor Comments:

Comment 1- Introduction:

Since this paper focuses on evaluating different data assimilation (DA) approaches, it would strengthen the manuscript to include a more thorough overview of existing DA algorithm literature, for context and completeness. I recommend adding this near Lines 44–51, where the background on DA methods is introduced.

Comment 2 – Methods/Results:

Justify the use of 0.05 to 1.05 bounds and increasing error with time. It would be helpful to include spatial and temporal figures (e.g., from representative sites) comparing the input observations against independent validation data. This could illustrate the spatial and temporal variability of observation errors and help justify the bounds selected for dynamic error modeling (0.05 to 1.05). For instance, do errors vary systematically with snow depth, such as being lower in shallow snow and higher in deeper snow? As noted by Alonso-González et al. (2024), no data assimilation algorithm is universally superior; performance depends on the data and task at hand. Thus, this figure can further support the choice of the Ensemble Kalman Filter (EnKF) and observation error technique.

Major Comments:

Comment 1 - Result:

While the authors claim that the Dvar experiment outperforms DAconst in terms of snow depth and SWE estimation, the practical improvement is marginal at best. For snow depth, the spatial ACC increases only slightly—from 0.72 (DAconst) to 0.73 (DAvar)—a mere 0.01 gain (Figure 3c), which is unlikely to represent a meaningful enhancement in most applications. The temporal ACC comparison (Figure 3d) shows that just ~52% of sites improve with DAvar while nearly half do not, and 11% of sites degrade by more than 0.02. Similarly, the MAE reduction from DAconst to DAvar occurs at only ~51% of sites, with 12% worsening. These statistics reveal that the advantage of DAvar is not robust or generalizable. The pattern is echoed in the SWE evaluation, where only 56% of sites see improvement in MAE under DAvar compared to DAconst. In other words, nearly half the sites experience no benefit or deterioration, which raises questions about the reliability of the variable uncertainty approach. This is consistent with SDD and SCF evaluation (Figures 7 and 8). It is also surprising that RMSE was not reported, as it is a standard metric in snow modeling and data assimilation evaluations, and helps better in error magnitude in snow depth and SWE.

Despite the narrative of statistical significance (e.g., $p < 0.001$), these results suggest that the magnitude of improvement is small, the spatial consistency is weak, and the operational gain may not justify the added complexity. The authors should contextualize these findings more carefully, perhaps by comparing the computational cost or exploring why improvements are minimal across the full domain.

Comment 2 - Discussion:

The discussion attempts to cover many important aspects of the study. However, it overemphasizes statistical significance while underplaying the marginal and spatially inconsistent nature of the improvements. Standard metrics like RMSE are missing to give a clear picture. Key ideas like bias treatment and dynamic observation error comparison with SDs1 are introduced without prior mention or clear connection to the results or existing literature.

Line to Line Comments:

Lines 17–29 would be more effective as a single cohesive paragraph. The current break into two paragraphs disrupts the logical flow and makes the message harder to follow.

Lines 30-43 discuss various SWE estimation methods, but some other approaches such as spaceborne laser altimetry (e.g., ICESat/ICESat-2), Sentinel-2, and MODIS, are missing and should be acknowledged for completeness. Additionally, the paragraph uses SWE and snow depth somewhat interchangeably (line 42). It would strengthen the clarity to note explicitly that snow density is required to convert snow depth into SWE.

Line 58-59: particle batch filters and smoothers.. can be more computationally expensive given the large number of particles required." While this is generally true for particle filters, it is not necessarily the case for particle batch smoothers (PBS). For example, Alonso-González et al. (2024) showed that PBS was less computationally expensive than EnKF and other particle-based methods across multiple particle counts (100, 200, 300). The authors should either revise the statement to reflect this variability or cite relevant studies to justify the statement.

Line 67-86: Report the accuracy metrics from Lievens et al. compared to those from Dunmire et al. (2024). Additionally, please clarify that higher accuracy was achieved when evaluated over the European Alps only, as this distinction is crucial for understanding the geographic limitations of these methods.

Line 62-73: Clarify that the primary goal of this work is to assess the utility of incorporating dynamic observation errors versus static ones, because there are studies that have compared the performance of different algorithms already.

Line 101-102: specify the reported accuracy metrics (same as 67 and 86). What does better mean?

Line 173-184: The evaluation section introduces Snow Disappearance Date (SDD) and Snow Cover Fraction (SCF) without prior mention or justification in the earlier sections of the manuscript. To improve clarity and coherence, it would be helpful to introduce these variables earlier in the manuscript, explain their relevance to the study objectives.

Figure 2:

- Why were these two sites chosen? State reason (representativeness, topography, etc) either in the figure description or the result section.
- Axis can be shared for better readability (general comment for all figures).

Line 240-243: State the bias-corrected numbers for the constant as well and compare them to the variable. “Both DAconst and DAvar also substantially improve the Pearson correlation coefficient…” report numbers.

Line 325: The claim that dynamic observation error estimation “had not yet been explored” overlooks prior work that has implemented such approaches (e.g., Alonso-González et al., 2022). While these studies may not compare dynamic and constant errors directly, they do demonstrate prior use of dynamic error treatment in snow science. I suggest rephrasing to acknowledge existing efforts and clarify this study’s specific novelty.