Reviewer 2

This manuscript performs an assimilation experiment over the European Alps to investigate several advances to the modeling and assimilation scheme. Specifically, they consider the assimilation of a snow depth machine learning product based on Sentinel-1 observations and the development of a dynamic observation error for their Ensemble Kalman Filter setup. I find this study to be very compelling with interesting results, and I only have a few minor comments.

We thank this reviewer for their thoughtful and helpful comments and we believe that our revised manuscript is much improved as a result. Our responses below, are in red, with manuscript text in dark red and *additions to the manuscript in italics.*

1) Are the snow depth ML estimates available everywhere and thus assimilated everywhere or are there flags to exclude assimilation in higher uncertainty areas where SAR does not perform well, like dense forests?

The final sentence in **Section 2.3 Data assimilation approach and experiments** explains the flags used to exclude assimilation under certain conditions. We will modify this sentence slightly so that these conditions are more obvious:

"We assimilated SD_{ML} estimates weekly each year from September 1 through March 31, *excluding assimilation further into the ablation period when wet snow complicates the S1 signal. Due to limitations of S1 in forested terrain*, and the unsuitability of the ML SD retrieval over glaciated terrain, we do not assimilate over forested areas or glaciers. Following De Lannoy et al. [2024], we also do not assimilate when the soil or vegetation temperature is above 5 °C."

2) How representative are the in situ observations of the surrounding ∼1km to match the modeled grid?

This is a good question. To address this, we will add the following section to the Discussion:

4.2 Limitations of site evaluation representativeness

Previous studies have shown that mountain snow is highly variable, and point-scale measurements don't necessarily well-represent the surrounding area, even at spatial scales as fine as 10 m [López-Moreno et al., 2011, Fassnacht et al., 2018]. Meromy et al. [2013] found that approximately half of the SNOTEL sites they analyzed where representative of the surrounding 1 km area, defining "representative" as snow station biases within 10% of the surrounding mean observed depth. More recently, Herbert et al. [2024] reported that roughly one-third of 476 paired lidar–station data observations were representative at the 1 km scale, with representativeness defined as in-situ measured snow within ±10 cm of the lidar-mean snow depth at that scale. However, they also showed little change between the 500 m and 1 km scales, with 35% of stations considered as representative at 500 m. Generally, in-situ snow stations

exhibit a high bias as these sites are often located in flat terrain that preferentially accumulates snow [Grünewald and Lehning, 2011].

In this study, we use in-situ snow depth and manual SWE measurements as the best-available reference in the European Alps that cover a range of terrain conditions and spans many years. Unlike in the western United States, where high-resolution spatial snow depth products from the Airborne Snow Observatory and NASA SnowEx missions are available, such products are extremely limited in the European Alps. As such, it is not feasible to assess the representativeness of all 588 snow depth measurement sites and 8211 manual SWE measurements at the 1 km scale, and these point-scale measurements provide the best available Alps-wide, multi-year data available for evaluation. Nevertheless, by leveraging a large network of sites that span a range of elevations and terrain types, we can mitigate some of the inherent limitations of using point measurements for evaluating the 1 km gridded product.

3) Are there any spatial datasets, like lidar, for evaluating modeled snow depth over a broader area?

Unfortunately, spatial datasets such as lidar are very limited over the European Alps. A number of airborne photogrammetry products exist for the Dischma Valley in Switzerland, but comparable, freely available datasets are non-existent elsewhere in the Alps. Moreover, most of these available spatial snow depth datasets (with the exception of two products analyzed in Dunmire et al. [2024]) were used in training the SD_{ML} product and therefore cannot be considered as independent evaluation sources. Given the extremely limited spatial and temporal coverage of these spatial datasets, we instead focus our evaluation on in-situ snow depth stations and manual SWE measurements. These datasets, collected across multiple nations, span a range of elevations and terrain conditions, and provide continuous multi-year records, offering a more robust basis for assessing the impact of the DA.

4) In Figure 2b, the OL performs better than either DA scenario compared to the in situ, though $\mathrm{DA_{var}}$ performs better than $\mathrm{DA_{const}}$. Is this similar behavior in other locations where the initial OL estimate is already close to the in situ truth?

Yes, inevitably there are locations where the OL model run is closer to reality than the observations and the assimilation leads to a deterioration. Figure 3a demonstrates how frequently this phenomenon occurs. We can see here that DA leads to a deterioration in MAE at 16% of Alps-wide measurement sites, with only $\sim$1% experiencing a deterioration > 125 mm. We will include this information in the text of the manuscript as follows:

"A measurement site with assimilated snow depths substantially greater than 1 m is demonstrated in Figure 2b. In this case, the observation uncertainty is smaller for $\mathrm{DA_{const}}$ than for $\mathrm{DA_{var}}$, resulting in stronger posterior state adjustments in $\mathrm{DA_{const}}$. *At this measurement site we see that the OL experiment is closer to the in-situ snow depth than the assimilated observations, leading to a deterioration in model performance when the DA is applied. This phenomenon occurs at 16% of all measurement site (Fig. 3a), with only 1% experiencing a deterioration in SD MAE greater than 125 mm.*"

5) How does the DA perform when snow needs to be added to the system? Figure 2a shows the DA reducing the modeled snow down to match the observations. Though in later results, it appears that DA mostly reduces snow in the model.

Here the DA does mostly reduce snow because the OL has a positive snow depth bias due to an overestimation of precipitation at lower elevations in ERA5. However, we can see in Figure 9 that both DA_{const} and DA_{var} generally add snow above approximately 2250 m. We will add the following paragraph in the Discussion section to discuss DA performance when snow needs to be added to the system:

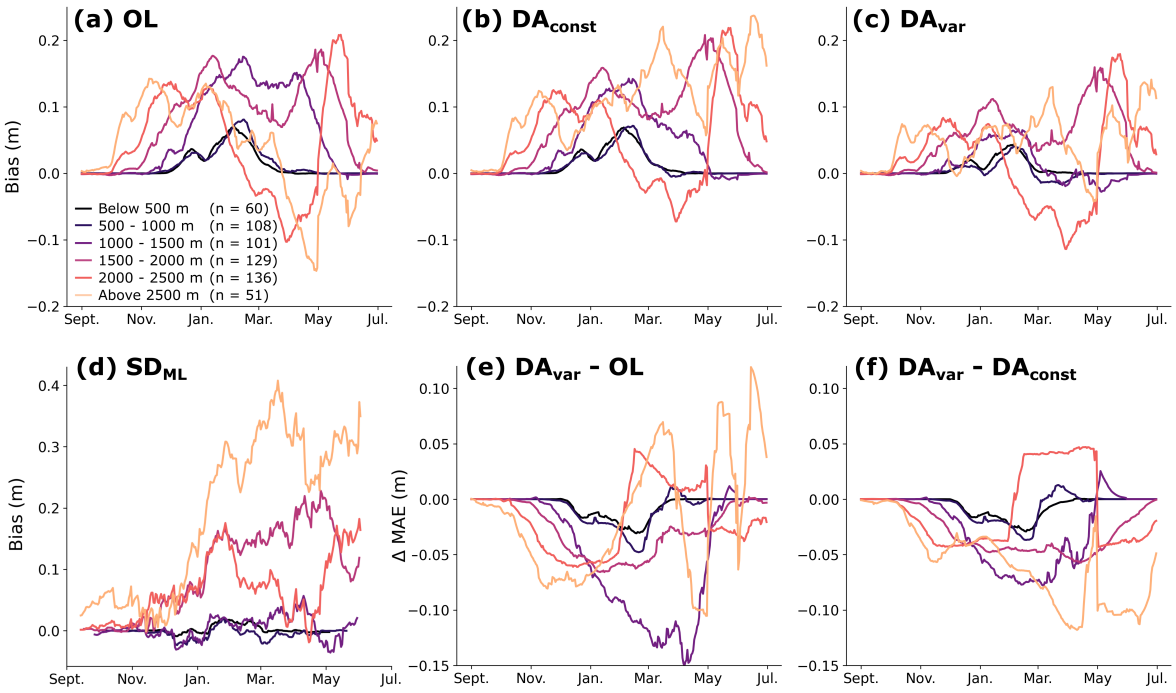
"Here, we highlight the implications of accounting for dynamic estimates of observation uncertainty and demonstrate that this system results in a more realistic modeled snow state. *Implementing the dynamic observation error generally improves performance in both places the DA adds and removes snow. In the OL experiment, snow depth has a positive bias at low elevations and a negative bias at high elevations (Fig. 4a). The* DA_{const} *experiment applies a static observation error that is relatively too large for shallow assimilated snow depths (e.g. Fig. 2a), limiting snow removal at lower elevations and leading to a still large positive bias at these locations. At higher elevations (above ∼1500 m), the assimilated observations exhibit a strong positive bias (Fig. 4d). The relatively small static observation error for deeper assimilated snow depths (e.g. Fig. 2b) leads to too much added snow in some cases, particularly above 2500 m (Fig. 4b). In contrast, in* DA_{var}*, less snow is added at high elevations (Fig. 9), resulting in improvements where snow needs to be added as well (Fig. 4f).*"

We will also add a new panel to Figure 4 (see Review Figure 1), showing the bias of the assimilated observations for the various elevation bands.

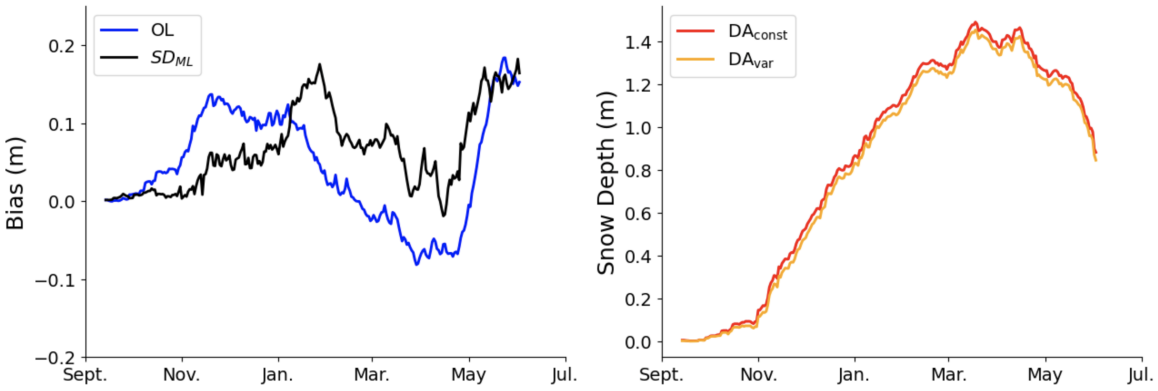
6) Lines 226-230: You mentioned that some of the increases in MAE at higher elevations could be due to limitations in the model during the melt season or biases from the assimilation. In your results, do you find that DA tends to add to remove snow at the higher elevations? How does the ML snow depth product perform at higher elevations where I assume the snow is deeper.

We can see from Figure 9 that both DA_{const} and DA_{var} both add snow above approximately 2250 m. To address this comment, we will add a subpanel, depicting the bias of the SD_{ML} assimilated observations for different elevation bands, to Figure 4 (See Review Figure 1). This panel indicates that the assimilated observations exhibit minimal bias below 1500 m and a positive bias above 1500 m. We will modify Lines 226-230 as follows: "However, an increase in MAE at high elevations during the melt season (March onwards) suggests a tendency for the DA experiments to retain snow for too long, which could be due to limitations in the modeled melt processes or biases introduced by the assimilated observations at higher elevations *(e.g., Figure 4d).*"

7) Figure 4e: Do you have an idea why DA_{var} degrades performance around peak accumulation for the 2000-2500 m elevation band? Aside from that line, it appears like DA_{var} improves upon DA_{const} for each elevation band throughout the winter



Review Figure 1: (Revised Figure 4 in manuscript) Seasonal evolution of bias and mean absolute error (MAE) stratified by elevation. Panels (a)-(d) show the seasonal snow depth bias for the (a) OL, (b) DA_{const}, and (c) DA_{var} experiments, and for *(d) the assimilated observations (SD_{ML}).* Bias is computed relative to in-situ snow depth measurements and is grouped by elevation bands (indicated by different colors). Panels (e)–(f) show the change in MAE between the OL and DA_{var} experiments (e) and between the DA_{const} and DA_{var} experiments (f). Negative values in (e)–(f) indicate improved performance (decreased MAE). Statistics are computed for each day, averaged over the entire 8-year period (2015–2023). A 14-day smoothing is applied to each timeseries and the number of in-situ measurement sites within each elevation band is provided in the legend.



Review Figure 2: Seasonal evolution of snow and bias for sites in the 2000-2500 m elevation band. (a) Seasonal snow depth bias for the OL (blue) and the assimilated observations (black). Bias is computed relative to in-situ snow depth measurements from sites within this elevation band. (b) Mean snow depth at these same in-situ measurement sites from DA_{const} (red) and DA_{var} (orange).

Yes, Review Figure 2 shows side-by-side comparison timeseries of the bias for the OL and the assimilated observations for sites between 2000 and 2500 m elevation. We see that the OL has a positive snow depth bias until approximately February and then a negative snow depth bias until May, while the assimilated observations have a positive bias throughout the entire snow season. However, in the early season (before January), the assimilated observations have a lower magnitude bias than the OL experiment. As discussed throughout the manuscript, DA_{var} leads to more effective snow reductions due to its lower assumed observational uncertainty for shallow assimilated snow depths. As such, in February, DA_{var} has a lower average snow depth than DA_{const} for sites between 2000 and 2500 m (Review Figure 2). This indicates that a lack of early season corrections in DA_{const} can propagate to better simulated late-season snow, although this impact is limited to specific to sites where the snow depth bias is not consistent throughout the snow season.

We will discuss this phenomenon in the following text in the Discussion section:

"However, we see that the DA_{var} *experiment performs worse than* DA_{const} *between February and May within the 2000-2500 m elevation band (Fig. 4f). In this range, the OL experiment has a positive snow depth bias until approximately February, followed by a negative snow depth bias until May (Fig. 4a).* DA_{var} *more effectively reduces this early season positive bias, resulting in lower mean snow depths later in the season, and poorer performance during the period when the OL is negatively biased. This suggests that a lack of early-season corrections in* DA_{const} *can, in some cases, propagate to more accurate late-season snow depths, although this effect is likely limited to locations where the snow depth is not consistently positively or negatively biased throughout the season."*

8) Figure 5: It appears like the modeled SWE almost reaches an asymptote around 600-700mm of SWE. Any ideas why the model is underestimating at the deepest SWE values?

This apparent asymptote is likely primarily due to limits with both the precipitation forcing and the observations. Indeed, Raleigh et al. [2015] demonstrate that forcing bias is the dominating uncertainty source in snow modeling. The atmospheric forcing used in this study comes from the relatively low resolution ERA5 atmospheric reanalysis product (31 km horizontal resolution). As such, the precipitation forcing for the land surface model has a coarse horizontal resolution and is unable to resolve orographic precipitation, resulting in an underestimation of precipitation at high elevations and a corresponding underestimation at high SWE. From Figure 2b in Dunmire et al. [2024] we see a similar asymptotic behavior of SD_{ML} for recorded snow depths above ∼3.5 m, likely attributed to these deep snow measurements being underrepresented in the ML training. To discuss this in the manuscript we will add the following paragraph in the Discussion section

"In the OL, we see an overestimation of SWE at measurement sites with low recorded SWE, and an underestimation of SWE at measurement sites with high recorded SWE (Fig. 5c). Previous work has demonstrated that forcing bias is the dominant source of uncertainty in snow modeling [Raleigh et al., 2015]. Here, we use ERA5 atmospheric forcing, which has a relatively coarse spatial resolution (31 km). While we apply a standard lapse-rate correction to downscale the near-surface air temperature forcing, precipitation is not downscaled, and therefore is unable to resolve orographic precipitation, resulting in relatively low precipitation and SWE spatial variability, and an underestimation of high SWE values. Furthermore, the SD_{ML} product has also been demonstrated to underestimate deep snow, likely due to these measurements being underrepresented in the ML training [Dunmire et al., 2024]. As such, the assimilation of this product is unable to fully correct the negative SWE bias for measured SWE > ∼800 mm, as can be seen in Figure 5d/e."

9) Figure 7: The difference plot for DA_{var} is much jumpier than that from DA_{const}. Can you explain that? Does DA_{var} have more ephemeral snow that comes and goes throughout the winter and spring? I assume much of that is from low elevation snow?

Yes, in DA_{var} snow at low elevation and early in the snow season (October, November) is often lower than in DA_{const} and therefore melts away more quickly, resulting in the jumpier nature of the DA_{var} difference timeseries. We will add the following sentence to the manuscript to describe this phenomenon: *"The relative difference in snow-covered area between* DA_{var} *and the OL fluctuates more than for* DA_{const} *(Fig. 7a), primarily due to the shallower early-season and low-elevation snowpacks in* DA_{var} *which melt out more quickly."*

10) Lines 264-268: Do you have any time series with snow cover comparisons to IMS or Copernicus?

It would be difficult to compare snow cover time series with the Copernicus product. The Copernicus Fractional snow cover product is not gap filled, meaning that the data is not spatiotemporally continuous. The model's strong overestimation of snow cover (described in Section 3.3, and in L360-371) compared to the IMS product is temporally consistent and therefore a timeseries comparing snow cover from our experiments with the IMS product would not provide substantial new information beyond that which is already presented in Figure 8, Supplemental Figure S4 and the text.

11) Figure 8: Does it make more sense to have the y axes on these plots be the percentage of sites within each elevation band so it is easier to compare against 25%, 50%, etc of sites?

Thanks for the suggestion. We will change the y-axis to *'Cumulative fraction of sites with snow disappearance'* instead of 'Cumulative sites with snow disappearance'.

12) Lines 307-320: I think this would be better in the methods. It feels odd to introduce a new dataset in the discussion section.

Agreed, we will introduce the DA_{S1} dataset in a new subsection of the methodology:

2.4.4 Comparison to SD_{S1} DA

To compare with previous work that assimilates snow depth retrievals from the S1 change detection algorithm (SD_{S1}; Lievens et al. [2022]), we compared output from our two DA experiments with DA output from De Lannoy et al. [2024] (experiment DA_{S1}). This DA_{S1} experiment utilized the same DA setup as in DA_{const}, with a static observation uncertainty ($\sigma_{obs} = 0.3$ m), but assimilates SD_{S1} retrievals instead of SD_{ML}. Here, we utilized 4548 manual SWE measurements collected within the Po River basin (the study domain of De Lannoy et al. [2024]) to compare SWE MAE between the DA_{const}, DA_{var}, and DA_{S1} experiments.

References

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.

S. R. Fassnacht, K. S. J. Brown, E. J. Blumberg, J. I. López Moreno, T. P. Covino, M. Kappas, Y. Huang, V. Leone, and A. H. Kashipazha. Distribution of snow depth variability. *Frontiers of Earth Science*, 12(4):683–692, 12 2018. ISSN 2095-0195. doi: 10.1007/s11707-018-0714-z.

Thomas Grünewald and Michael Lehning. Altitudinal dependency of snow amounts in two small alpine catchments: can catchment-wide snow amounts be estimated via single snow or precipitation stations? *Annals of Glaciology*, 52(58):153–158, 9 2011. ISSN 0260-3055. doi: 10.3189/172756411797252248.

Jordan N. Herbert, Mark S. Raleigh, and Eric E. Small. Reanalyzing the spatial represen-

tativeness of snow depth at automated monitoring stations using airborne lidar data. *The Cryosphere*, 18(8):3495–3512, 8 2024. ISSN 1994-0424. doi: 10.5194/tc-18-3495-2024.

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.

J. I. López-Moreno, S. R. Fassnacht, S. Beguería, and J. B. P. Latron. Variability of snow depth at the plot scale: implications for mean depth estimation and sampling strategies. *The Cryosphere*, 5(3):617–629, 2011. ISSN 1994-0424. doi: 10.5194/tc-5-617-2011.

Leah Meromy, Noah P. Molotch, Timothy E. Link, Steven R. Fassnacht, and Robert Rice. Subgrid variability of snow water equivalent at operational snow stations in the western USA. *Hydrological Processes*, 27(17):2383–2400, 8 2013. ISSN 0885-6087. doi: 10.1002/hyp.9355.

M. S. Raleigh, J. D. Lundquist, and M. P. Clark. Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework. *Hydrology and Earth System Sciences*, 19(7):3153–3179, 7 2015. ISSN 1607-7938. doi: 10.5194/hess-19-3153-2015.