Revisions of: Assessing the impact of meteorological forcing and its uncertainty on snow modeling and reanalysis
By: Sun and Margulis

Benoit Montpetit

1 General Comments

This paper aims at analyzing the impact of forcing data variability on ensemble SWE estimates. Here, the objective is clearly not to demonstrate which forcing dataset is best overall, but rather show that by using multiple products, the multi-forcing ensemble show better performances than single-forcing ensembles separately. I really appreciated reading this manuscript, since this question of "which forcing dataset to use" is still up for debate in the cryospheric community and clearly, there is no "one-size fits all" answer yet. This paper is a step in the right direction to potentially standardize forcing datasets used for large domain studies.

The manuscript is well written and structured. The analysis structure with the four different experiments also helps readers to understand the impacts of the different forcing datasets and the impact of assimilating observations on the different forcings. Even though there are different Bayesian frameworks and downscalling methods that could perform better, I feel like this is not the objective of this study and the selection of the methodology is well described here, and well documented in previous publications, showing that it is applicable to this context.

That said, I do have some questions with regards to experiment 1a, where I feel like some information in missing to properly characterize where the variability of the different forcings and their uncertainties come from. This will be detailed in the next section.

Even though I call the latter "major comment", I still recommend the publication of this manuscript after what I consider minor revisions.

2 Major Comments

With the three watersheds included in this study and the differences in spatial resolution of the original input datasets, I feel like there should be a bit more information included, which could benefit and strengthen the discussion in section 3.1.

The first information to consider is land cover type for the different watersheds. Not being extremely familiar with all watersheds, it would be useful to know the different and dominant land cover types for each. Knowing that there are uncertainties linked to land cover type, this could help indicate why one forcing datasets might perform better for one watershed and not the other.

In combination with land cover, knowing the initial altitude to which the surface properties are given for the different datasets could help identify sources for the SD/SWE variability. Again, there are uncertainties in modelled SWE associated to its altitude.

Both information also provide discussion points with regards to rescaling the dataset to 150 m. I feel like this could explain some of the outputs we see in annual snowfall of Figure 2.

As mentioned by the authors, the objective of this addition is not to identify which forcing dataset is best, but in a context of analyzing various domains, this could indicate the importance of which datasets to use in a multi-forcing experiment.

3 Minor Comments

Figure 1: Would be interesting to add the land cover map, which is used in resampling (Figure 3).

Eqns. 1-3: Personally, I prefer the more compact ways of writing these equations. They could be shorten into one equation:

$$W_i = \frac{\prod_{j \neq i} \sigma_j^2}{\sum_i \prod_{j \neq i} \sigma_j^2} \tag{1}$$

Eqns. 4-6: Same as the previous equations, it could be summarized by:

$$N_i = W_i \times N \tag{2}$$

where $N = \sum_i N_i$

L. 321: Please clarify what is meant by "SWE variations". I assume it is the difference with ASO SWE observations.

Figure 4: it would be interesting to add error bars on the ASO measurements to show the spatial variability of SD and SWE across the different study areas and also the spread in SWE due to the mean ensemble density estimation. You show which forcing data works better on average, but the error bars could indicate whether the different forcing datasets provide estimates within the range of measured properties.

Table 3: Probably a precision error but weights do not add up to 1 for Merced.

Figure 7: I understand the idea behind keeping only the best and worst cases in this figure but I think it would still be relevant to include all three forcing cases. I would also keep the same order of presentation in the figure as previous figures to make it easier to compare. I also like to keep the y-axis range the same when comparing plots horizontally. It helps to identify which site has larger errors more easily.