Author Response to Reviewer 1

Note: Reviewer comments are reproduced below in blue text, with the author responses interspersed and in black text. Proposed revision actions are in *black italic text.*

Ritchie et al. evaluate a large suite of snow water equivalent datasets in the United States by comparing the performance of these data products, aggregated at the hydrofabric scale, to lidar-based SWE products from Airborne Snow Observatories. In my opinion, this is a valuable contribution to the community, as SWE is a highly valuable quantity and many products are used for different purposes, often without a rigorous consideration of how product choice may impact outputs.

In general, I think this is an important contribution to the field. However, I have major comments related to how this comparison is framed. In general, I think the authors should frame this as a direct comparison (in known basins) of many products to ASO, but without calling ASO the ground truth as it is not a direct observation. Further, it is somewhat unclear if the goal of the paper is to produce a highly quantified intercomparison of SWE products or to discuss how to make a protocol for such an intercomparison. In my opinion, the manuscript's strengths are in the former, and should focus on this, with notes on the intercomparison protocol appropriate in the discussion. I recommend that the authors consider the protocol content as a separate manuscript.

Response: We thank the reviewer for their consideration of this paper. The dual purpose of the paper is presented in the abstract and again in the major text (lines 93-101). In the abstract, we note that we conduct the intercomparison and use it to explore/demonstrate the potential of a shared community evaluation protocol. Of two objectives, the latter discussion (protocol topic) is the central motivating goal of the research and the paper, and would not be accomplished without the intercomparison (for demonstration purposes). There are multiple prior papers inter-comparing small assortments of SWE products, and while those have value, we wish to spur discussion over the potential value in establishing a broader community protocol that can be used by multiple studies, and the associated challenges. We expect that the example protocol discussed in this paper is likely not that which may be ultimately developed by the community (e.g., in the Western US), but we hope the paper spurs an advance in this direction. We therefore respectfully decline to separate these concepts and break the work into two papers. We believe the combination is compelling; in fact, since the preprint appeared, we have been approached by 3 other experimental SWE product groups that are interested in including their datasets in this comparison – which is consistent with the aims of the paper.

- The ASO SWE products are not an observation and are not universally available, nor is their actual accuracy well known. This is noted by the authors late in the work (L464) but does not seem to be integrated into the fundamental argument of the paper. I agree that the accuracy is likely quite high! However, it is not a direct observation and the framing of the comparison should be clear on this.

Response: The paper is careful not to introduce ASO SWE as an observation and has tried to characterize it as an ‘observational estimate’ or an ‘estimate’ in the early text. For instance, in the introduction (line 57), as ASO is introduced, we state that “It’s important to recognize that ASO SWE is not a direct observation (due to its combination of lidar-based depth and model density)”. In lines 144-149, we briefly discuss the uncertainties in the estimate, and refer to it as ‘quasi-observational’, and return to this caveat toward the end of the paper. ***We will provide a clearer description of ASO’s limitations at the earliest possible point in the paper so that it is not missed by the readers in the framing of the study. We will also drop the term ‘quasi-operational’.***

- The direct comparison to CAMELS is somewhat ill-founded in my opinion. This is not necessarily a flaw in the study, just a reframing. CAMELS is a dataset whose novelty was standardizing catchment attributes for a set of gaged US basins. This study is a SWE values intercomparison. I see the role of these two types of outputs as very distinct and find the comparison confusing. If the authors would like to use this comparison, I would want to see a better explanation of why this is an appropriate reference or structure for the current work. Again, I don’t think it is necessary for this paper to be like CAMELS in order to be a contribution to the field.

Response: Having witnessed the galvanizing effect of the CAMELS dataset in enabling the community to establish (and then extend) a common basins for intercomparing widely different methods of streamflow simulation – albeit organically, without it initially being called out as a ‘common protocol’ (until the Newman et al. benchmarking paper) – we feel it is a very useful reference point in framing this effort (and as co-author Wood also participated in the initial paper developing the CAMELS basin collection, this work reflects similar testbed-adjacent motivations. While there was novelty in the attribute aspect of CAMELS (added after the basin collection was first created), the current value of CAMELS is tied as much to the standard set of basins/locations as it is to the attributes (and often other sources of attributes are used). In this work, we also have hydrofabric catchment attributes, though those are not highlighted. The paper explains over lines 80-100 why CAMELS provides relevant framing. ***We will revisit this section to identify ways to improve this explanation, including clarification of any conceptual differences. In response to the 2nd reviewer, we will also greatly shorten this reference to CAMELS, and include other examples of standardized protocols in the introduction, such as from SnowPEX, to provide context.***

- All SWE products here are aggregated to smaller tiles within each basin (hydrfabric) before comparison. This accomplishes a spatial congruence between multiple sources, but also minimizes the value of higher-resolution products by removing small-scale spatial features. This allows products to be right for the wrong reasons.

Response: The standardization to a common spatial scale does cause a loss of information for the finer scale products. We note in the paper that comparison of the products on a scale that is finer than some of them places the coarser products at a disadvantage, but this can be expanded. ***We will provide additional discussion of the tradeoffs for the different products from adopting a single spatial scale for comparison.***

- The results section contains some discussion content throughout, for example in the end of section 4.3

  Response: Good point. ***We will review the results section for discussion content and move such elements that are not included mainly for clarification to the discussion section.***

- Increase metrics reported. For example, section 4.3 discusses comparisons of specific metrics (MAE, etc) but does not name any of them or reference a table that contains these metrics

  Response: ***We will extract metrics that are largely included in the figures into a more easily comparable/reviewable form of tabular summaries.***

- The authors should discuss the difference between aggregating up (e.g., ASO -> hydrofabric) and down (e.g., a coarse product with pixel size > hydrofabric size) when it comes to evaluating accuracy.

  Response: This point connects with the earlier comment about finer scale products. ***We will provide additional discussion of the tradeoffs for the different products from adopting a single spatial scale for comparison.***

Line edits:

Response: Thanks for these corrections (Line Edits and Figures). ***Unless noted below with a response, we will adopt all the suggested edits.***

L57: “it’s” to “it is”

L55: it is currently called Airborne Snow Observatories (slightly different name)

L58-59: phrasing in parenthetical is clunky, but noting that swe is a model product is important

L60: a main limitation of ASO is temporal sparseness (as noted) as well as coverage – only covers select US basins currently

L60: “quasi-observational” is a confusing category in my opinion. The SD products are directly observed, the SWE products are better introduced in your longer explanation in the previous phrases

L98: it is not appropriate to call the ASO SWE product an observation because it is not! There are significant modeling efforts behind the ASO SWE product, many of which are not public.

Further, ASO SWE are validated/calibrated by comparison to ground-based SWE observations, so they likely contain bias by prioritizing matching specific locations.'

Response: As noted above, the paper does explain that ASO is not an observation, and we will improve the early caveating of it as such. The Introduction describes "ASO SWE observational estimates (referred to as 'observations' [f]or brevity)". ***We can use the term 'observational estimate' throughout if required by the editor, though we do feel that if the concept that ASO is NOT an observation is explained sufficiently, and the use of the term 'observation' in the ASO context, then the readability of the paper can be improved by using the shorter 'observation' label.*** We agree that 'quasi-observational' is also not ideal.

L98: "for" typo

L249: missing subsubsection title

Section 2.4: these are standard metrics (R2 etc) and the formulas do not need to be reproduced. If they are reproduced, please define all variables.

Section 5.2: This is interesting, but discussion and comparison of protocol choices could be another paper and in my opinion does not align with the main point of this work.

Response: As discussed in an earlier response, the discussion of the value of a community protocol around areal SWE products, and the demonstration how this might be treated (including its challenges), building off the dataset intercomparison at the start of the paper, is the motivation and central goal of the paper. The intercomparison of datasets was conducted in support of this objective, although we recognize that it has a scientific value in and of itself. Other dataset intercomparison papers exist and we wrote this paper to tackle the larger question of how to create a structured/standardized focal point for the community so that it does not continue to rely solely on standalone dataset intercomparison papers that are published from time to time, and are difficult to synthesize. The paper will now also discuss other relevant protocol related efforts to provide more context for this objective.

Figures:

General: please add letter labels to subplots

Fig 1: it is unnecessary to have separate panels with and without the hydrofabric overlay (panels a and b)

Response: We feel that the sequence is an acceptable way to illustrate the transition from fine to coarse scale, and its current form does not detract from the communication of the process. In response to the second reviewer, however, we will be adding an additional part to the plot to illustrate the difference in relative scales for the hydrofabric target from a finer versus a coarser source gridded dataset – which should be useful in supporting the discussion of tradeoffs from this remapping (that is suggested earlier by this reviewer).

Fig 2 and Fig 6: are these showing essentially the same thing, except separated by state in Fig 6? If so, this is repetitive. If the point is to compare the states, consider another visualization that directly compares them, or use a numeric output.

Response: This is correct – Figure 6 gives a state-level breakdown of the Figure 2 data, to illustrate that such differences can exist. ***We will consider other ways to display the contrast of Figure 6 and whether the messaging of the Figure can be improved with such a revision.***