Second review of "**Inter-comparison of snow depth retrievals over Arctic sea ice from**
radar data acquired by Operation IceBridge " by Kwok et al

I appreciate the efforts of the author in responding to my earlier comments. I feel some of the replies are unsatisfactory and thus request some more clarification and also clear additions to the manuscript following these responses, before I believe this to be ready for publication.

My new reviewer comments are in red, in response to the response document copied below. I only included comments that required a second response by me.

Main comments:
1. I think you need to discuss more the basin-scale differences between the products. I also think this should be moved further up the paper, as it is what motivates the whole exercise in my opinion, especially as you don't go into huge detail regarding the algorithm differences and echogram interface detection. It is also what most of the readers will be interested in seeing.
For example, I think the huge differences in the means across the products for the three ice classes should be its own figure near the start (although I also have issues with the use of age classes, see later main comment, so think this should be by region instead). There are differences of ~100% between some products in the earlier OIB years, while some products show strong regional trends and others don't. It's pretty crazy to me that this is not discussed more as a motivating factor to look more closely at the algorithms, and I feel these differences have been somewhat hidden later in the paper. The more detailed in-situ comparisons could then follow on to help understand why this is.
We appreciate the broader geophysical perspective although this was not the path taken by the inter-comparison project. Initially, the project was more interested in the quality of different retrieval approaches when assessed with *in situ* snow depth because that allowed for more quantitative evaluations of the procedures at the highest spatial resolution available (i.e., small scale variability). The merits of the basin-scale comparisons were recognized only after the results of the spatial and inter-annual differences were produced. We noted in the text that the robustness and adaptation of the retrieval procedures to changes in radar data quality over the IceBridge Mission are important considerations, in addition to the footprint-scale comparisons, in producing a long-term record. We have added to the text to provide a better description of the evolution of the project (i.e., from the small scale to the large, rather than the motivation suggested above) but we prefer to preserve the order of the discussion in the manuscript.

Can you indicate how you have done this please? I.e. what you added and where.

2. Any comments on how 'tuned' these data have been, especially to the other snow depth data included in this paper? I believe I'm right in thinking the different groups have all had access to these in-situ data and ERA-I snow depth fields (especially as the lead author has previously produced the ERA-I derived snow depth maps used in the inter-comparison) so is that one reason why some fits are better than others? I understand that

tuning happens and is often needed, but I think we need to understand this more to really understand if the differences are due to the choice of algorithm or other factors. Also, I think comments should be made if the individual algorithms were also compared against any other in-situ datasets in their respective papers and how good those fits were. The fact all authors are involved should make this easier.
The snow depth retrievals were contributed by different algorithm-developers. Thus, there was no control on the amount of 'tuning'. It was entirely up to the developers of the snow depth data sets. The level of maturity of the algorithms is different and depends on the amount of resources available to the developers. The aim of the work was not to the select the best algorithm, but rather to provide results that would serve to inform the development of the next-generation retrieval algorithm.

OK, but as all the developers are part of this paper, it should be possible to provide some factual statements regarding this and to include a discussion in the manuscript regarding the issue. This is a crucial point in terms of reproducibility and understanding how/why better correlations with in-situ/reanalysis data were found, which I feel is still inadequately treated.

3. As all the algorithm developers were part of the paper, I'm surprised a bigger comment was not made of what actually will happen next. Will one/multiple algorithms be scrapped or combined? Are there pros/cons of certain algorithms that will be adopted/used by the Operation IceBridge sea ice group? You do state in the paper that: "The aim of this paper is to examine these algorithms and to use the assessment results to inform the development of the next generation algorithm", but the path forward is unclear to me and I really hope we don't continue with multiple algorithms floating around that different groups/papers use for different reasons.
The next step is to develop an improved algorithm, for producing an OIB product, by integrating the experience gained from this work.

It is still not clear to me that this paper has made a significant step towards this other than highlighting the (albeit important) differences between the current algorithms, but it doesn't seem like a more concrete statement will be forthcoming.

5. Why were only these specific field campaigns chosen?
These were the only field campaigns over fast ice, where we did not have to deal with spatial registration issues related to sea ice motion.

Can you include this comment in the manuscript?

6. Why ERA-Interim for derived snowfall?
We considered MERRA2 as well but the snowfall from MERRA2 is known be biased (higher by ~30-40%) compared to climatology and ERA-Interim.

OK but there are other reanalyses available that might not have such a bias. Obviously some also don't provide snowfall which may be an issue? At least a comment on this would be useful (I don't expect you to add in any analysis on this at this stage).

7. Why were the Wavelet retrievals not available? The Newman et al., (2014) paper shows that data were produced in 2012..? This seems odd.
We used only those data sets that were available and provided by the algorithm developers at the time of this inter-comparison project. In the case of the Wavelet retrievals, the algorithm developers provided only retrievals from the flight over the Eureka field campaign.

This seems very unusual, although I don't expect you to change this at this stage.

Figure 9 - Confused by the numbers in Figure 9a. There is a lot of information being crammed in and I struggled to understand what it all means.
- *Why is this saturated at 15 cm?* 15 cm was selected because the threshold of detectability of the a-s interface is ~10 cm.

So this should be 10 cm then.

Multiple figures - The Jet color scale introduces false boundaries, isn't good for people with colorblindness, and should thus not be used in my opinion! Very bad for comparing geospatial data by eye.
It is somewhat difficult to control the quality of the figures in the pdf files generated by the publisher for review purposes. We have enlarged the tracks in Figure 9 so that they are easier to see. The quality in the final publication should be higher.

This is nothing to do with how the pdf is generated but the use of a bad color scale that makes it harder to interpret the figure data values, no matter how it's generated into the pdf. See e.g. here https://www.climate-lab-book.ac.uk/2014/end-of-the-rainbow/ for a discussion.