Interactive comment on "Inter-comparison of snow depth retrievals over Arctic sea ice from radar data acquired by Operation IceBridge" *by* Ron Kwok et al.

Anonymous Referee #2 (Referee comments are in italics)

Received and published: 27 July 2017

The paper provides an inter-comparison of five different NASA OIB snow depth products that have been developed in recent years. The OIB products are compared against in-situ data from two field campaigns and snow depth estimates derived from the ERA-Interim reanalysis and a modified snow climatology.

This work addresses an important issue for the Arctic sea ice community what is the snow depth over Arctic sea ice and which (if any!) OIB derived snow depth product should we be using? The effort in bringing together these various snow groups and datasets is laudable. I believe the study should be published we need to see these differences and have a baseline for snow inter-comparison discussions - but I believe the paper has some shortcomings that need to be addressed first.

We thank the reviewer for the detailed reading of the manuscript.

We would like to note that there is only one standard product (archived at NSIDC) provided by the OIB project, and other snow depth estimates used here (Wavelet, SRLD, JPL) were contributed by scientists interested in the retrieval process. We now include a table (Table 2) that shows the availability of retrievals for each algorithm at the time of the inter-comparisons were carried out.

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.

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.

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.

4. I'm not a huge fan of using the ice type mask to delineate the results. The comment on Page 14: " MYI that advected into this region, which was used in the construction of the modW99 fields but is absent in the ERAI-sf fields (because the MYI is not used in the estimates). " implies that the presence of MYI doesn't mean much for snow depth. I believe other cited studies (from co-authors) have come to similar conclusions (e.g. recent King and Webster papers). The modified Warren climatology seems just plain wrong in my opinion so I would be tempted to drop that entirely unless you want to make the point that some groups are using this now and we need to explore its potential biases.

There is merit in using multiyear ice (i.e., ice that survived the summer in this case) as a gross indicator of the chronological age of the ice measured from the 'beginning' of the growth season; one expects more snow to accumulate on older ice (on average) and hence a thicker snow cover on this ice type. We have clarified this in the text as this usage is somewhat different than the way we typically think about ice age (i.e., in terms of years rather than from the beginning of the season).

Regarding modified Warren climatology, the modification adapts to the thinner snow cover over seasonal ice. Some form of modified climatology is used by most of the sea ice thickness algorithm at different institutions.

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.

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.

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.

Specific Comments:

P2, L14 - maybe 'needs to be inferred by other methods' instead of left to be measured or modeled

We prefer the way it is currently phased because it suggests the two alternatives for obtaining snow depth.

P2, L14 - I think (if I've interpreted this right) that you should say why snow density matters before saying we need routine measurements of it.

Added a note about snow density.

P2, L16 - you say hence, but then start by discussing forecasting, which seems odd.

Re-ordered.

P2, L20 - I think more should be made of the fact that people still use this climatology, despite it being many decades old!

Added: "…The *W99* climatology is still widely used in ice thickness retrievals."

P2, L23-25 - 'of about several centimeters' and 'broadly consistent' seems pretty loose. Drop or further clarify.

The discussion is a broad summary of the results from the list of papers provided at the end of this sentence.

P2, L26 - mention that this is predominantly the western Arctic, (except for 2017 which isn't included in this study).

Revised to read: "…repeat surveys of the early spring snow and ice conditions in different parts of the western Arctic.

P2, L29 - add something like 'from the OIB snow radar..'

Revised to read: "…These snow depth datasets from the OIB snow radar…"

P5, L30 - this sentence is poorly worded.

Reworded.

P6, L4 - why and how did it vary with ice topography? Just because it was older do we think this is an exhaustive list of retrieval algorithms?

Modified to read: "Transect variations in density were conservative, with a mean of 306 kg m^{-3} and standard deviation of 50 kg m^{-3}, comparable to the assumed climatological mean of ~320 kg m^{-3} near the end of the winter."

P8, L12 - unsure of the comment " The initial application to existing OIB snow radar data from various campaigns (2009 - 2012) and the need for it to be applicable to future campaigns, required a process that would adapt to the data and not be dependent on fixed thresholds in the radar return signal ". Why can't you apply thresholds and update these each year when you process the data?

The comment emphasizes the need for a procedure that is independent of fixed thresholds is desirable (and arguably necessary) for several reasons. For the SRLD algorithm, there is no need to change threshold values depending on which data set is being analyzed, whether it is for the latest campaign, or the multiple data sets within an archive of previous campaigns (e.g. 2009-2015). More significantly, there is no need to determine those threshold values, which would need to be done empirically by analyzing each data set beforehand. Additionally, there is no need to determine when to determine new threshold values. That is, a change in calibration could conceivably occur within a campaign for example.

P9, L6 - how is it robust? It seems we are testing that in this paper, no? What do you mean by this?P9 - Does the removal of deformed ice from the Wavelet algorithm introduce a bias compared to other algorithms?

For a more detailed description of the Wavelet algorithm, the reviewer is referred to Newman et al. (2014).

P9, L20 - is this the only thing that has been removed from the algorithm?

Yes.

P10, L11 - this should be Figure 3.

Corrected.

P10, L15 - I'm a bit confused by this. the resolution of each radar footprint is around 5-10 m, right? So how does a 20 m radius circle correspond to 9 radar spots again?

There are approximately nine radar footprints (sampled at 5-meter intervals) in a 40-m along track segment. The aim is to reduce the geophysical variability as well as the sensitivity to accommodate for uncertainties in the spatial overlap between the snow-radar footprint and the point samples from the field measurements. This is discussed in the text.

P10, L18 - you mean the mean AND standard deviation, right? In which case I don't get how using an averaging window changes the mean value. An average of an average should produce the same average..? It should obviously change the shape of the distribution though (reducing the tail).

It is the mean standard deviation (i.e., the mean of the distribution of standard deviations), not the mean AND standard deviation.

Added to clarify:"… (i.e., mean of the distribution of σ_f).."

P11, L25 - can you list them here? It looks like the JPL and Wavelet correlations decrease. Any comment on how this compares with the htopo parameter? i.e. do you expect surface roughness to co-vary with the snow depth variability?

1. The changes in the correlation values are now listed in the text.

2. These comparisons do not make use of specific metrics provided by each algorithm.

P12, L14 why only four of five algorithms? Pretty interested to see the NSIDC differences, especially as this is probably the most commonly used..?

The NSIDC products are averaged spatially and do not provide snow depth estimates for each echogram, which was needed for the plots shown in Figure 3b.

P12, L27 - this seems the fundamental tenet of the whole paper, no?!

Yes!

P14, L20 onwards - this should go at the start of the section in my mind as it's a pretty key point.

We prefer order of the current list as most of points are important things we have learned in this process.

Interpreting Table 2 and 3 was pretty painstaking at first. Can you make it more obvious that the diagonal elements are taken from Table 2 and maybe draw a box around these?

Yes – we have clarified this in the caption. Also, quantities from Table 2 are now italicized.

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.

- *Why are the NSIDC panels missing the repeat tracks and distributions?*

  The repeat tracks are not processed by the NSIDC algorithm.

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.

Figure 10 - this seems pretty pointless so I would be inclined to drop it. Figure 12 and 13 should be split up and made more readable.

This figure shows the differences between ERAI-sf and modW99. We believe that it is a useful illustration of the spatial differences of the two fields. Figures 12 and 13 are now in four separate pages.