Reviewer comments from Report #2

The OIB ice thickness data set is derived from the ATM data (snow+ice surface elevations from laser altimetry). So, ice thickness is derived indirectly from sea ice + snow surface elevations, sea surface height estimation, and snow depth derived from the snow radar. And the conversion from snow freeboard to ice thickness multiplies the uncertainties in freeboard retrievals. In contrast, Air EM directly measures sea ice + snow thickness. Therefore, I doubt that OIB measurements are generally more reliable than Air EM or BGEP draft. Of course, also Air EM comes with uncertainties, and so do BGEP and OIB. And it could be that there is a problem with the AIR EM data here rather than with the OIB, but we don't know. I think we have to live with the fact that there are differences in the validation data sets due to the different retrieval methods, impact of snow, different surface types, surface roughness, footprints, etc.

So, what I am criticizing is that in the paper (and the mentioned line in the response letter), it comes across that OIB is "more reliable" because it fits better to the CryoSat-2 observations (and the model). But I don't think that this is a valid conclusion. Another CryoSat-2 product might fit better to the Air EM and BGEP data. So, I suggest reconsidering these statements (see above). I don't think it will change the main conclusions of the paper.

Author response:

Many thanks to the reviewer for their additional comments, and their useful explanation of the derivation of the independent validation datasets used in this paper. Reference to the “reliability” of the Air-EM data has been removed from the sections of the paper highlighted by the reviewer (and the abstract).

It should still be noted and discussed in the paper that the model and the CryoSat-2 observations fit better to the OIB data (and the BGEP data) than they do to the Air-EM data. Since there was unfortunately only one year’s worth of Air-EM observations that overlapped with the time period covered in this study, this yielded only 45 matchups with the model for Air-EM, compared to 547 for OIB. Therefore, it seems possible that, even considering the Air-EM dataset to be extremely reliable, there is likely to be an element of sampling uncertainty affecting the results.

Therefore, we agree with the reviewer’s criticism that the results do not indicate that the Air-EM observations are unreliable, and have taken this out, but have left in the discussion of sampling uncertainty potentially being the reason for the poorer results compared to OIB.

Changes made:

Lines 16-18:

“This may be evidence of uncertainty in the Air-EM validation observations, sampling error, noise in the SIT analysis, or uncertainties in the modelled snow depth or the assimilated SIT observations.”

Changed to:

“This may be evidence of sampling uncertainty in the matchups with the Air-EM validation dataset, owing to the limited number of observations available over the time period of interest. This may also

be evidence of noise in the SIT analysis, uncertainties in the modelled snow depth, in the assimilated SIT observations or in the data used for validation."

Line 466-470:

"However, the model standard deviation and correlation coefficient are poorer on assimilation of these data. It should be noted that there are many more OIB observations over both the Canadian Arctic and Beaufort Sea regions (Fig. 12), and these agree much better with the model output and the CryoSat-2 observations than do the Air-EM observations. This potentially indicates an uncertainty in the quality of the Air-EM observations. Uncertainty in the snow depth will also be contributing to this issue, although the difference between the CryoSat-2 and Air-EM observations is greater than the snow depth itself."

Changed to:

"However, the model standard deviation and correlation coefficient are poorer on assimilation of these data. It should be noted that there are many more OIB observations over both the Canadian Arctic and Beaufort Sea regions (compare Figs. 12 and 13; 547 matchups for OIB versus 45 for Air-EM), and these agree much better with the model output and the CryoSat-2 observations than do the Air-EM observations. This potentially indicates a sampling uncertainty in the Air-EM matchups. Uncertainty in the modelled snow depth will also be contributing to this issue, although the difference between the CryoSat-2 and Air-EM observations is greater than the snow depth itself."

Line 522-524:

"The BGEP data for March-April 2015-2017 (Fig. 15(a)) compares better with the CryoSat-2 observations than does the Air-EM data in the Beaufort Sea for April 2015 (Fig. 13(a)). This again suggests that the Air-EM data may be unreliable, as discussed in Sect. 5.2."

These lines have been removed.

Line 543-548:

"Validation against springtime airborne electromagnetic induction (Air-EM) combined SIT and snow depth observations (Haas et al., 2009) yields poorer results than for the OIB and BGEP datasets, despite covering similar locations. This may be evidence of uncertainty in the Air-EM observations, or sampling error owing to the limited number of matchups available from this dataset. It may also be a result of noise in the SIT analysis, uncertainty in the modelled snow depth, or uncertainty in the assimilated observations."

Changed to:

"Validation against springtime airborne electromagnetic induction (Air-EM) combined SIT and snow depth observations (Haas et al., 2009) yields poorer results than for the OIB and BGEP datasets, despite covering similar locations. This may be evidence of sampling uncertainty in the Air-EM matchups, owing to the more limited number of observations available from this dataset that cover the time period of interest. It may also be a result of noise in the SIT analysis, uncertainty in the modelled snow depth, in the assimilated observations, or in those used for validation."