**Arctic sea ice thickness loss determined using subsurface, aircraft, and satellite observations**

Response to reviews of version 1.

**Referee #1**

General Comments:

This is an interesting paper. The authors attempt to combine disparate data sets to tease out the spatial and temporal trends in Arctic sea ice thickness. The primary assumption in their regression is that there is one number that describes the bias of a given data set relative to a selected reference data set. And, once these biases are identified then all the estimates from different instruments could be combined into a 'rationalized' data set for understanding spatial and temporal trends.

I think, this assumption is only valid if there were no space- and time-varying biases in the reference data set. That is, the reference data set has to be internally consistent: no relative biases in time or space within the data set itself. Take ICESat:G for example, if the thickness of  multiyear ice were systematically under-estimated compared to the thickness of seasonal ice and in addition their relative biases changes in time (t) and space (X) due to snow loading, then a single parameter will not adequately describe these biases, since $b_j$ has to be represented as $b_j$(X, t) = const + $f_j$(X, t). And, this is probably the case for ICESat:G. In these cases, the time- and space-varying biases are subsumed by other terms in the regression, which could cause biases in spatial and temporal trends, and in the magnitude of the relative biases. In fact, the authors alluded to this in the discussion of systematic biases in Section 4.1 that the intercomparisons in one region may not be applicable elsewhere.

In the end, I am not entirely convinced that $b_j$ is adequate for describing the biases of each data set relative to the reference data set. And, the data set may not be rich enough to assume that $f_j$(X, t) somehow averages to zero. In the conclusion section, the authors cautioned that the bias term should be interpreted only in a relative sense, I would add that that there are hidden (unmodeled) biases in a regression analysis of this sort although the magnitude is unknown.

In summary, this represents an interesting attempt at using all available data set, the results seem reasonable. I recommend publication after revision. I should add that I had fun reading this paper.

R. Kwok

Thank you for the careful review of our paper.  It is true that the differences between data sets undoubtedly have a complex spatial and temporal structure and that by finding just one number to characterize this difference is a simplification. However, recognizing this limitation, the one-number bias does provide information that can be useful in assessing observational or model differences and invites further investigation about the causes of this bias. It therefore is just a beginning and meant to prompt further research.   For example, when there is good overlap between data sets, such as with the two ICESat products, it is possible to tease out more of the structure of the differences.  When the overlap is lacking or sporadic it is much more difficult to do so and we hope to motivate the research community to come up with solutions.  We have added more comments about this issue and have moved the regional fits section, which address the issue directly, to the Error Assessments section, where it fits better.

Detailed Comments:

Page:Line number

4547: 23 – Kwok et al., 2008 offer a preliminary examination of four campaign and does not cover 2003-2008. Please use Kwok et al., 2009.

done

4549: 13 – re:Submarines. For the years after 2000, substantial portions of the submarine cruises are not under multiyear ice. Should/would the biases due to beam widths and depths be different (time and space-varying)? Perhaps a short discussion is warranted because the impact of these biases could be significant when compared with other data sets. Some of these issues are discussed in Section 5.

We added a short comment here about the submarine bias issue.

4550:3 re: Air-EM. The uncertainty of the ice thickness from Air-EM is not clearly stated as uncertainty in snow depth from PIOMAS has not been assessed. I suppose the impact of errors in snow depth should be relatively small for the purposes here. In any case, this should be addressed the NCEP-NCAR precipitation could be biased. (why not use Warren?)

A comment about the PIOMAS snow depth uncertainty has been added. We have used the PIOMAS snow depth instead of Warren to obtain better spatial and interannual variability. It is biased slightly lower than the Warren climatology (it averages between 1 cm greater in May and 7 cm less in October). We estimate the uncertainty in the PIOMAS snow depth to be on the order of 0.10 m. These comments are added to the text and a short discussion of the contribution of snow depth errors is added to the errors section.

4551:1 re: BGEP data: there is no comment on data quality?

Krishfield *et al.,* (2014) say the uncertainty is less than 10 cm. This now noted.

4551: IceBridge data – the error is dependent on the snow depth as well and could be a significant source of error.

We added a short comment about the uncertainty in the OIB snow depth. Wish we could say more.

4551: IOS-CHK, IOS-EBS: should comment on the data quality. Melling claims 0.1 cm.

The draft uncertainty for each was added.

4552: ICESat-J. A clarification:

The results in Kwok and Cunningham, 2008 were based on a preliminary analysis and do not include the improvements in the estimates as discussed in Kwok et al., 2009.

The primary reference is changed to Kwok 2009.

Within the 25-km segments, open water samples are included in the creation of the gridded field. However, in there is no accounting for the overall ice concentration within a grid cell after data accumulation and interpolation.

This clarification was added.

Question: would it be useful to tabulate the expected error (well, our best understanding) of each data source even though they are not used in the regression.

Since the errors of the average measurements are so difficult to determine from the reported point errors due to unknown error correlations, it seems to us that such a table would be based on guesses.

4553:10 Comments:

f(m) is probably lower over first-year ice. (reduced accumulation)

This is now noted.

The lower density of older ice (could be down to 880 kg m-3) would introduce variability in your thickness estimates.

This is also now noted.

So, what is the overall uncertainty in the submarine-derived ice thickness estimates after all this and including the uncertainty in ice draft? adding to the uncertainty of ~0.25 m in ice draft?

The point is that we don't really know the uncertainties. Figuring out what this is has a long and not entirely convincing history. Here we take a different approach and intercompare all measurements to arrive at relative corrections. We hope that our results prompt further investigations on more detailed characterization of errors for each measurement source.

4554:20 In the regression, the uncertainties of the measurement source are not considered?

No, the uncertainties are not included, largely because they are poorly characterized. This is now noted.

4553:23 I think Kwok et al. (2009, not 2008) pointed out this was unrealistic. The following is probably better phrased: "…. this assumption is unrealistic (Kwok et al., 2009) since the sea ice….'

Thank you. This now included.

4554:25 It would be useful to show all the terms T(x,y,t) used in the regression.

All of the terms for the full basin fit are shown in Table 2. Showing all of the terms for the other regional fits seems a bit tedious for the reader.

Table 2. The indicator coefficients are NOT ordered by the magnitude of the coefficients.

Sorry. Fixed.

Figure 3. It would be interesting to show the RMS error of the fits here as well.

The RMS error can be guessed from the residuals and the total error in the fit for each sub-region is shown in Table 3, however we agree that showing the RMS error for each source would indicate which systems have a larger error in the fit, so we added the RMS error for all points from each system to Table 4.

4557:18-27 The submarine ice thickness used here is different from submarine-derived ice thickness used in Kwok et al. (2009), in which the snow was not accounted for. With the snow loading in your sub data, it would make sense that the resulting comparison would be different (signs). There is no inconsistency in the two comparisons.

Right. The sentence about the signs is removed.

So, what do you make of the difference between the comparison between the ITRP ice thickness and ICESat:J vs between ICESat:J and ICESat:G?

The difference between the ICESat-G and ICESat-J values may be related to the different techniques of determining the sea level in order to obtain the freeboard or to the different treatments for snow. (this was added)

It seems that if the indicators were correct that I should expect the difference between ICESat:J (b = 0.42 m) and submarine (b = ~ -0.1 m) to be more like 0.5 m, but the difference is only ~0.1-0.2 m if directly comparisons were made. I am somewhat puzzled. So, there must be spatial gradients or some other differences between the two data sets that caused the differences?

Good point. We had noticed this but had not pursued it as we should have. The larger difference between the submarines and the ICESat-J estimates for the entire basin, 0.47 m, stems from the inclusion of the 2000 submarine cruise when there is no overlap with the ICESat data. If the analysis period is chosen as 2001-2013, with all sources included, the ICESat-J product is just 0.05 m +/- 0.09 m thicker than the submarine-based estimates. The ICESat-J coefficient changes very little, so this exercise shows that the 2000 sub cruise may have sampled much thinner ice than indicated by the full fit. This was added to the discussion of the basin-wide fits.

4557:18-27 Aren't these comparisons (2000-2013) dominated by the large populations of ICESat:J and ICESat:G (only 2005-2008) estimates and thus just a measure of the difference between the two?

The ICESat data sets where subsampled to avoid this. The two ICESat data sets contribute half of the points, 1500 out of 3000. However, it is true that much of the spatial pattern in the computed biases is established by these measurements. Their mean difference is accounted for by the indicator coefficients while the spatial pattern of the fit is approximately the average between the two.

And if the trend in 2000-2013 was not exactly linear, the ITRP would pick a trend that fits through ICESat:G? Looking at Figure 1, between 2000 and 2013, the population is dominated by the satellite retrievals. There is submarine data only in 2005 and 2013. If this were correct then it would explain my remarks in my previous point.

The trend is established by all of the data. In an experiment where all of the ICESat data are excluded, the trend remains except in the Beaufort Sea. The most recent sub data are in 2000 and 2005.

4558:20 …. as well as the size of the sample population?

Correct. This is added.

4558:21 My remarks above for the other regions as well: Beaufort, North Pole, Lincoln Sea.

Figure 4. It was indicated that observations were adjusted for the bias of each. This was not discussed, but perhaps should be mentioned, in Section 4.2.

We have now removed all of the data points.

4563:15 Withholding data is a good idea although it does not change the fact that the sample populations are biased to a given sensor.

Correct. Withholding one type of data is an attempt to show what the effect of sampling bias of each sensor type is on the overall result.

**Referee #2**

The authors present a study into systematic differences between Arctic sea ice thickness datasets obtained from various observation systems. Using a least-squares multiple regression model, they find close agreement between five of the systems while others give significantly thinner or thicker ice. Combining all observations, the authors derive substantial negative trends for annual mean ice thickness. The study is an interesting and timely attempt at solving the issue of comparing and combining sea ice thickness measurements made at different locations and different times. The manuscript is well written but could be improved especially in the error assessment. I recommend moderate revisions.

Thank you for your careful review of the manuscript.

General comments: - I miss a comment/discussion on the use of a mean value for ice thickness. Given the usual non-symmetric shape of an ice thickness distribution, the mean can be misleading and the mode of the distribution is more reliable. How does this affect the biases in the different datasets and the systematic differences? –

While the mode may be more reliably measured than the mean by some systems the mode is not directly related to the total ice volume and is somewhat dependent on the bin size of the thickness distribution and may have additional problems for some distributions. The mode reflects the thermodynamic growth or melt of the dominant thickness class, but does not represent the ridging processes. While the ITRP could be applied to the mode, the greater interest is in characterizing the mean ice thickness. Also ICESat-J does not provide a distribution from which to determine the mode.

All datasets are heavily affected by how snow is treated during the processing or in the conversion of draft, freeboard, or total snow&ice thickness to ice thickness. In the descriptions of the datasets, the authors mention the different snow thickness estimates that have been used for the different datasets, but do not include a discussion on the implications of using the different snow data for the systematic differences (e.g. using the Warren climatology which is likely not representative for the Arctic sea ice regime in the 2000s). While I agree that a full discussion is beyond the scope of the study, it is an important point that deserves at least a qualitative, short discussion in the error assessment. It might be worth including the Webster et al., JGR 2014 paper. –

A short discussion of the errors associated with snow has been added to the section on errors.

On a similar note, I think the authors underestimate the effect of open water in the ice thickness estimates of some of the datasets, and the footprint issue, especially when creating and comparing 50km averages from measurements with very different footprint size and shape. While a full statistical analysis of both is needed which is obviously beyond this study, a proper acknowledgement of these sources of errors should be included in the discussion.

All of the data sets, with the exception of ICESat-J, include open water in the mean thickness estimates as far as we know. Many hundreds of samples are included in the 50-km or one-month aggregate means, so footprint size and shape are not directly related to the aggregate uncertainty, unless of course the footprint of the point estimates contributes in some way to the bias.

Specific comments:

The title does not really reflect the main component of the paper, ie. the assessment of systematic difference between the datasets –

We prefer to keep the emphasis in the title on the geophysical results, rather than on the methodology. Though the paper naturally emphasizes the methodology, in light of prior results using similar data which suggest a reduction of the ice thickness decline towards the end of the period, we believe this is justified.

p. 4546, l. 8: make clear that you mean the sources used in this study – there are lots and lots of on-ice measurements (e.g. Renner et al., 2014) - Introduction: It should at least be mentioned and acknowledged that thickness has been measured for a long time and in many regions using drilling and on-ice methods (e.g. ground-based electromagnetics, buoys) –

Good point. We added a comment about point measurements to the introduction.

p. 4547, l. 24-26: There is a CryoSat-2 ice thickness product available from AWI at http://www.meereisportal.de/de/datenportal.html ? -

Yes, we have seen it. As far as we know it is a preliminary version and nothing has been published about it yet, so we did not think it would be fair to CryoSat to include it at this time. This is not the same data set as Laxon et al (2014) wrote about. His data are unavailable.

p. 4549, l. 3-5: The averaging is unclear: I guess the mooring have been averaged only in time? Did e.g. IceBridge or airborne EM flights get averaged over a month too? Also, how do you deal with oversampling by overlapping footprints in the averaging? –

Correct, mooring data were averaged for one month and not averaged spatially. The airborne data were averaged both spatially and temporally, in that nearby flights from less than a month apart are averaged together. This was added to the section 2 introduction.

p. 4549 ff: I assume the submarine and the mooring data do not include open water in their thickness estimates? The Air-EM data have an open water bias when open water is in the measurement footprint. What about the other measurements? There's some inconsistency in the dataset descriptions; would be good to have the same information (footprint size, known biases like the underestimation of ridges, distance between measurement points etc) for all datasets. –

Open water is included in all of the mean thickness estimates. This is added to the section 2 introduction. We are not aware of the AIR-EM open water bias. Is there a reference?

p. 4550, l. 28: What are the characteristics of this profiler?

The Ice Profiler is a 420 kHz ULS instrument with a 1.8° beam width, a precision of 0.05 m, and a sample rate of 2 seconds. This has been added.

p. 4551, l. 15 (and other occasions): What do you mean by "clustered"? –

"Clustered" means point data within a given spatial and temporal range are aggregated to form an average. We changed the term to "aggregated", perhaps that will be clearer.

p. 4552, l. 7-11: I'm not sure I understand your argument here. The numbers of observations are still highly variable between the different datasets, simply because of the varying spatial and temporal resolution and coverage. Why do you need to subsample here and not in other cases? –

The satellite data would overwhelm the other data sets if they were all used. If they were without errors one would of course want to use all of them. However, our goal was to tease out biases, therefore we needed weight them approximately the same as the others.

p. 4556, l. 4: There seems to be some spatial structure in the residuals with high values close to coast lines. Any thoughts why that is? Issues with the reference datasets due to the proximity to the coast? –

The reference data set has no more impact on the shape of the function or on the residuals than the other sources, it just establishes a reference for the indicator coefficients. The high and low values near the coast may be due to the more variable ice pack there as well as to the inadequacy of a simple polynomial to capture all of the spatial variability.

p. 4557, l. 10: Here and throughout the paper: Given the uncertainties in the datasets which often are around 10 cm or more, does it really make sense to include the second decimal in the analyses of the differences? –

While the uncertainty in the point measurements are on the order of 10 cm or more, the error does not remain that high when many points are averaged. Indeed the sigma values for some of the indicator coefficients are just 0.02 or 0.03 m. This not always the case, but two decimal places seem appropriate.

p. 4559, l. 2: How do the biases change regionally?

The point of looking at the different regions was to attempt to evaluate how the biases change and therefore provide an assessment of the sensitivity of our procedure to the spatial sampling by different observation systems. New Figure 7 tries to show this graphically. This section has been moved to the error assessments section.

p. 4559, l. 6: The ICESat data around the North Pole are not really observations, are they? It seems strange to me to include them in this part when the thickness estimates most of the area of the North Pole are based on interpolations, not actual measurements. –

Yes, it is a bit strange. But since the publically available JPL data includes this interpolation and this is how other users (e.g. modelers making comparison) will use them, we think it is best to retain them.

p. 4564, l. 13-17: Do the relative magnitudes also change a lot when other datasets are used as reference (IceBridge, Air-EM)? Otherwise this makes me wonder how reliable the spatial distribution in the submarine data is… –

The relative magnitudes of the indicator coefficients remain unchanged if the reference data set is changed. This is mentioned in the methodology section.

p. 4564, l. 27: replace ”does” with ”do” –

done.

p. 4565, l. 2: This would be one of the reasons to use the mode of the thickness distribution. –

The mode also has problems of interpretation, for example when there is more than one prominent peak in the thickness distribution or if the peak is very broad. There are no physical laws for changes in the mode analogous to the thickness distribution equation that we are aware of.

Sometimes thermodynamic growth is shown by changes in the mode, but not always.

p. 4564, l. 19 – p. 4565, l. 5: At least note that there is in situ data? –

This has now been noted.

p. 4565, l. 18: the largest negative value – relative to ICESat-G –

This is also now noted.

p. 4566, l. 13: delete ”from”

done.

Comments to the figures: -

Figure 1: The maps are too small to recognize anything. I realize that that is partly due to the formatting of TCD, but even if blowing up the figure on the computer screen it is difficult to spot details. Also, which ice thicknesses are plotted? All observations or averages? It looks like all observations and I wonder if it really makes any sense to plot measurements from different seasons on the same map. This creates patterns that are due to the seasonal cycle and not the geographical distribution. To me, the maps do not add any information that is not covered by the graphs in the right column and Fig. 2 a.

We have tried to improve the presentation by dividing the one figure into two separate figures, one for the maps and one for the times. The maps no longer show ice thickness, as the times of the measurements makes it confusing, as you say. In old Figure 2a it is very difficult to make out where the moorings are, so we prefer the separate presentation of the source locations.

Figure 2: Continuing on from the previous comment, Fig 2 a gives valuable information about the geographical distribution of the different datasets, however, it is impossible to spot the moorings. In panels a, b, d, and f, it is very difficult to distinguish the colours (almost impossible for colour blind readers). The two greens for the IceSAT datasets are too similar, and same for IOS-CHK and IceBridge. I suggest pulling out Fig 2 a into a separate, larger figure, and use different symbols and colours for different datasets. Regarding Fig. 2 b a similar question as for the maps in Fig 1 applies: are these average thicknesses?

We have dropped the top two panels. Their information is redundant with the new Figures 1 and 2. The colors for the observations systems in the time series have been reduced to just four: satellite, airborne, moorings, and submarines. We choose retain the thickness plot (old Fig 2c) but acknowledge the confusion of time and space in the two plots.

Figure 3: Again, colours are difficult to distinguish on the map; black and dark blue look almost the same when only a thin line. Typo in the legend (SCICEX Box)

Good point. We changed the dark blue to a red and fixed the typo.

Figure 4: The dots in Fig 4 b are so tiny, it is difficult to see them at all. Instead of a cloud of dots, it would be more interesting to see distributions, e.g. annual. That would also give more information how ”representative” the mean (vs the mode) is.

We removed all the dots…they are confusing at best because their spatial locations are not shown. We added the seasonal minimum and maximum ice thickness lines.

Same comment to Figure 5.

**We made the dots a little bigger, but we retained them so that the temporal distribution of the available data is easily seen.**