Response to Referee Dr Massonnet

Format key:

Normal text is considered general discussion

Italic text quotes the Referee

Bold text indicates additions to the manuscript

~~Bold text~~ indicates deletions from the manuscript

We wish to thank Referee Dr Massonnet for his thorough and constructive review of our submission. Dr Massonnet clearly spent a lot of time and effort meticulously reviewing our submission which certainly benefits the manuscript, and this is very much appreciated.

Dr Massonnet's review is in two sections.

Firstly he has three *"main comments"* about the merits of the manuscript. This is followed by three drawbacks to the manuscript followed by three suggestions about how we may wish to address these three drawbacks which we found to be very helpful.

Secondly Dr Massonnet has *"other comments"*; these are more detailed line by line suggestions of amendments, additions and re-phrasing of various sentences throughout the text and to the figures. Again we found this very informative and appreciate the time taken to create these improvements.

This Response will address both these main and other comments in this order. The "*main comments*" because of their nature will consist of a discussion about the manuscript's drawbacks and our attempts to rectify or justify these points as appropriate. With regards to the "*other comments*" a confirmation that the amendment has been performed will be provided for the majority of cases.

Changes to the manuscript (here in bold) will also be visible in the tracked changes .pdf mark up.

Main comments

1) One of the drawbacks of using SIT instead of SIC is that SIT is much less constrained by observations. In fact, there are no long-term and spatially homogenous observations of SIT. The authors work around this by using PIOMAS. PIOMAS is the best we have for this type of study, but we shouldn't forget that PIOMAS is primarily a model output where some observations (no SIT observations) are assimilated following a very simple scheme (nudging). The paper by Lindsay et al. (2014, doi: 10.1175/JCLID-13-00014.1) and/or Zygmuntowska et al. (2014, doi:10.5194/tc-8-705-2014) could be cited in addition to the others in the manuscript to reflect how uncertain PIOMAS is with respect to observational products.

Suggestion

1) To ensure a balanced and more objective introduction to PIOMAS in section 2.1, consider citing the two papers listed above and briefly discuss how current estimates of SIT, including PIOMAS, are uncertain. Everyone knows that PIOMAS is the best we have, but no one should forget that it is not free of errors. I stress that PIOMAS is first and foremost a model output!

We have edited section 2.1 to introduce PIOMAS more critically:

> **As a reanalysis PIOMAS is constrained by the quality of the assimilated observations, Lindsay et al. (2014) forces PIOMAS with four different atmospheric reanalysis products producing differing results. Schweiger et al. (2011) found biases in PIOMAS of 0.26 m in autumn and 0.1 m in spring when compared with ICESat (Zwally et al., 2002) although the spring bias is within the range of uncertainties found by Zygmuntowska et al. (2014). Larger differences are found in areas of thickest ice north of Greenland and the Canadian Archipelago with ICESat retrievals around 0.7 m larger than PIOMAS. However in this region PIOMAS agrees better with in situ data (Schweiger et al., 2011). Zygmuntowska et al. (2014) suggests that this discrepancy is due to the choice of sea ice density in ICESat, and they support this explanation by finding lower discrepancies between PIOMAS and CryoSat-2 (Laxon et al., 2013) which utilises an alternative sea ice density value.**

2) I am more doubtful about the physical validity of the recalibration. When recalibrating for the mean and for the variance (but not the trend in SIT), the evolution of SIT might be physically incompatible with the mean state over the calibration and future periods. In other words, the recalibration would be physically robust if the trends in SIT wouldn't depend on the mean state, but just on the external forcing. There is evidence from the observational record that the September sea ice extent (SIE) is following a quadratic rather than a linear evolution. There is also evidence from CMIP5 models (Fig. 4 of Massonnet et al., 2012, cited in the manuscript) that SIE trends are nonlinearly related to the mean SIE. I don't know whether this is the case with SIT, too. If so, the rate of SIT loss might be biased after recalibration and this could affect the conclusions.

Suggestion

2) The second point is touched in the conclusion (p. 3838, ll 13-17), but it'd be good to know how the trend in SIT relates to the mean SIT in different grid points of CMIP5 models. If there is no dependence (constant trend), then a simple recalibration of the trend would be enough - although large uncertainties exist. If the link is nonlinear, then even recalibration of the trend over the historical period wouldn't be sufficient. I'm not asking to change the recalibration method, but simply to investigate how valid the additional recalibration of trends would be for projections.

> While we agree with Dr Massonnet's concerns and indeed point this out ourselves in section 3, we feel that much of this is outside the scope of this manuscript and the MAVRIC method. We do not wish to apply a trend correction for various reasons: primarily it is not clear that trends calculated from PIOMAS would be a robust estimate of the forced trend. We agree that the work suggested here would be interesting and likely be significant and need to be taken into account IF a trend correction had been applied; we feel that as we do not attempt to perform a trend correction exploring this aspect falls outside the scope of this manuscript. It may even warrant a separate study

akin to Blanchard-Wrigglesworth and Bitz (2014) with regards to the mean state dependence of variability.

3) The link "lower spread in projections –> more confidence in these projections" is not as straightforward as the authors suggest. It is undeniable that the spread in projections shrinks after the bias-correction method is applied (Fig. 9 of the manuscript). As a matter of fact, models that are forced to look alike in the present will also look alike in the future. The question is whether this recalibration method does not itself introduce systematic biases in the updated projections. This would be the case if PIOMAS was overly thick/thin in some regions (point 1) above) or if the response of SIT would be mean-state dependent in CMIP5 models (point 2) above). In other words, it is "easy" to narrow uncertainties in projections by recalibration, selection or many other methods; but it should be kept in mind that another source of uncertainty (related to the recalibration/selection method itself) is introduced but does not appear on the final plots.

Suggestion

3) For the last point, I have a suggestion. The authors did train their recalibration method by splitting the PIOMAS period in two parts; while the results are satisfactory, the problem is that the training and testing periods are very short and close to each other. My suggestion is the following: apply the MAVRIC correction on 5 GCMs by taking as reference one of the member of the 6th one (i.e., replace PIOMAS by one member of one GCM). This "sister" experiment could allow to verify that the 5 GCMs are properly constrained to track the evolution of SIT of the 6th one, and in particular the dates of sea ice disappearance. I know that this requires some (technical) work, but I think that a positive result would strengthen the validity of this method a lot!

Dr Massonnet here agrees with Prof. Flato's opinion that our assertion that reduced spread intrinsically leads to increase confidence is too enthusiastic. We will add **potential** or an equivalent to "reducing uncertainty" and "increased confidence" in the manuscript to highlight our slight hesitation with such claims.

An additional source of uncertainty that we neglect here is the PIOMAS calibration uncertainty emerging from the choice of reanalysis and model tuning. This could be assessed by sampling the different versions of the PIOMAS reanalysis described in Lindsay et al. (2014). They find the different versions are broadly similar and can be accounted for by appropriate tuning of the ice model component. This uncertainty in PIOMAS itself will introduce systematic biases to the MAVRIC projections. This bias is not a flaw in MAVRIC however but a limitation intrinsic to the observational dataset one is correcting to.

In the following sections, we equate reducing model spread with reduced uncertainty. While this is true in the sense that some of the outlier simulations of SIT are now more similar to the multi-model mean, this doesn't necessarily equate to reduction in uncertainty. For example the initial selection of GCMs may not have been representative, or all of the GCMs from CMIP5 may have some inherent systematic biases, reducing the spread of which wouldn't help sample future observations.

Dr Massonnet also points out a limitation in the MAVRIC validation method we use in Sect 4 and Fig 4. We also state our reservations about the temporal length of both the calibration and validation period.

An additional limitation to this method is that the calibration and validation periods are very close to each other.

We did at first consider using "*sister experiment*". Although this would provide a rigorous test of the MAVRIC method, we deemed that in practice it is unnecessary for the reasons given below.

We also have other reservations about the necessity for a fully-fledged 'sister experiment'. As Dr Massonnet points out, the test of whether the method adequately constrains the other five GCMs will be that they all reach the ice-free date at similar times. Even if we conducted this experiment on our data this would not be seen. This is because the ice-free date is primarily dependant on each GCM's own ice loss trend. The MAVRIC method intentionally does not correct this trend and so would 'fail' this test.

As a compromise however we feel that we execute a comparable experiment using the MAVRIC model dataset itself. This is because all the models have effectively gone through a 'sister' type experiment as they are all constrained to the same 'sister', i.e. PIOMAS.

Appendix C Additional MAVRIC performance analysis

To highlight whether the estimated uncertainties are reliable, we examine the errors in the projections when considering one member as 'truth'. As all ensemble members are constrained by PIOMAS one individual ensemble member out of sample should fall with in the distribution of the remaining ensemble members. This principle should hold true for all ensemble members out of sample in turn.

The root mean square error (RMSE) is calculated using the following formula:

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\mathbf{18}}\sum_{n=1}^{18}(E_n - \overline{E_{15}})^2}$$

where E_n is the ensemble member between 1 to 18, $\overline{E_{15}}$ is the mean of the 15 ensemble members from the models of which E_n is not a member.

Figure C1 shows the advantage of the MAVRIC method in this out of sample RMSE test. A decreasing RMSE means that the models are initially biased though are converging to a common value (as we expect in this case as the models trend towards being ice-free). An increasing RMSE means that the models are diverging as they have different ice loss trends.

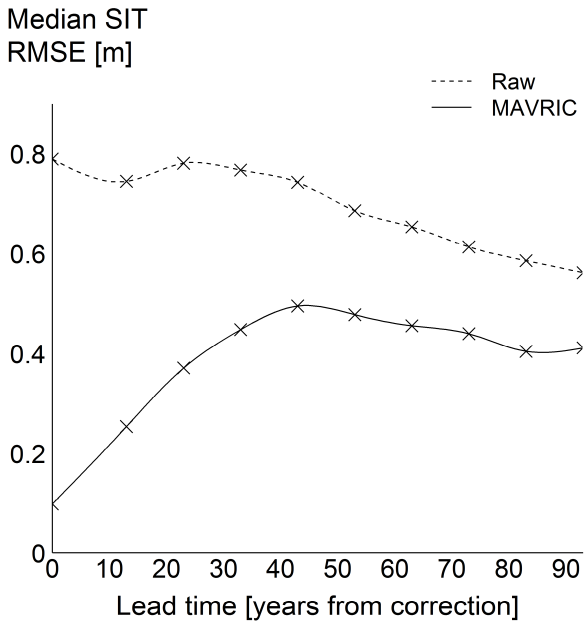

Figure C1. Multi-model ensemble out of sample September median SIT RMSE]

The MAVRIC ensemble trained on every individual ensemble member within MAVRIC results in a RMSE of 0.1 m initially and up to a maximum RMSE of 0.5 m. The fact that the Raw RMSE decreases (as opposed to increases) highlights that the models have biases. The 0.1 m in the MAVRIC RMSE indicates that initially the MAVRIC ensemble members differ only in internal variability. The RMSE then grows due to differing ice loss trends which is expected as no attempt to correct the trends in this study.

To find the dispersion of the MAVRIC multi-model ensemble we repeat this style of experiment with the standard error (SE) metric, using the following formula:

$$SE = \frac{E_n - \overline{E_{15}}}{\sigma_{15}}$$

where E_n is the ensemble member between 1 to 18, $\overline{E_{15}}$ is the mean of the 15 ensemble members from the models of which E_n is not a member. σ_{15} is the standard deviation of the 15 ensemble members of which E_n is not a member. This is repeated for all 18 ensemble members giving 18 SEs of how different each ensemble member is to the rest of the multi-model ensemble set. The SD across these 18 SEs is the dispersion of the multi-model ensemble. A perfectly dispersed ensemble set will have a dispersion of one. Numbers less than one mean the ensemble set is under-dispersed and hence predictions/projections from that set will be under-confident as the SD is too large. Values greater than one indicate that the system is over-dispersive and hence over-confident.

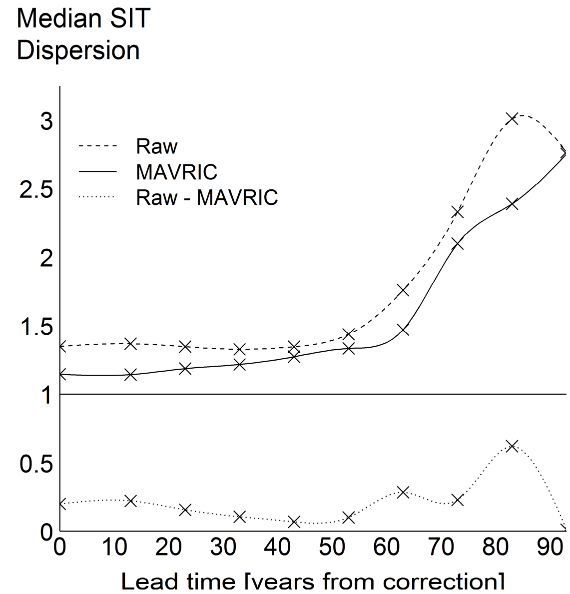

Figure C2. Multi-model ensemble out of sample September median SIT dispersion

The results of the dispersion calculation are shown in Fig. C2. The MAVRIC ensemble is approximately 15 % - 30 % over-dispersed for lead times of up to 60 years. This means that the ensemble is slightly over-confident and thus has slightly too little overall variance. The rapid increase in dispersion from 60 years is solely due to the CSIRO GCM, specifically it's comparatively slow ice loss trend. This was tested by repeating the dispersion experiment omitting CSIRO (not shown). At this lead time many models are starting to be ice-free in September while CSIRO retains ice. It is to the merit of MAVRIC that it is less over-dispersed than the Raw output, hence more reliance can be placed on MAVRIC than the Raw output as it's ensemble distribution is more representative.

Other comments

Listed as Page Number / line

3822/5 Drop "spatial and temporal": biases is enough.

~~Spatial and temporal~~

3822/12 Replace "sea ice internal variability" by "climate internal variability on SIT uncertainty"

Replaced ~~sea ice internal variability~~ with **climate internal variability on SIT**. The word SIT is dropped as already mentioned in this sentence, uncertainty is omitted as the implications are more general than this

3823/1 Replace "SIT" by "SIT evolution"

Evolution

3823/15 "[In the case of SIT], Model bias makes a contribution to model uncertainty". Even for well-behaved (statistically speaking) variables like SST, model bias contributes to uncertainty: working with anomalies does not guarantee that other quantities such as sea water density, or air-sea fluxes, will be consistent after the bias has been removed. I would drop this last sentence.

~~Since absolute values are used, model bias makes a contribution to model uncertainty.~~

3823/19 "BC has not previously been applied to projections". See the papers of Boé et al. (2009, doi: 10.1038/NGEO467), Wang and Overland (2009, doi:10.1029/2009GL037820; 2012, doi:10.1029/2012GL052868), Zhang (2010, doi:10.1111/j.1600-0870.2010.00441.x), Mahlsteing and Knutti (2012, doi:10.1029/2011JD016709). The present manuscript is novel in that it recalibrates SIT, and does it locally.

SIT ~~sea ice~~; this manuscript is novel in that it recalibrates SIT, and does it locally. References useful and are cited as appropriate.

3824/10 As I wrote above, PIOMAS is a model based estimate of SIT constrained by some observations. Consider changing "observationally based" by "model based".

~~Observationally~~ reanalysis

3824/19 Same as previous comment.

"For an observationally based estimate of SIT, we use the PIOMAS reanalysis. PIOMAS is a coupled ice-ocean model that is forced with the National Centers for Environmental Prediction (NCEP) atmospheric reanalysis, and assimilates satellite observed sea ice concentration (Lindsay and Zhang, 2006) and sea surface temperature (Schweiger et al., 2011)."

This section is clear that PIOMAS is reanalysis and a model that assimilates observations so we feel the langue is not misleading here.

3825/1 Sea ice thickness has two usual definitions: sea ice volume divided by sea ice area ("in-situ thickness") or sea ice volume divided by grid cell area ("mean thickness"). In CMIP5 models, mean thickness is reported. Did the authors check that PIOMAS also reports mean thickness and not in-situ thickness? This is to ensure consistence when recalibrating CMIP5 models.

Zhang and Rothrock (2003) quote "mean thickness", volume used is per area the same as CMIP5, hence the correction is constant.

3825/17 Delete sentence "The thickest ice is located north...". This is more descriptive than informative.

~~**The thickest ice is located north...**~~

3826/1 The criteria chosen to screen the full CMIP5 ensemble are rather subjective ("have a reasonable spatial resolution", "comprise at least one ocean channel in the Canadian archipelago"). Is there a particular reason why these criteria were applied? Other criteria based for instance on sea ice extent would directly eliminate the CSIRO model. Did the authors also apply the MAVRIC method on rejected CMIP5 models? There is no fundamental reason why models without a channel in the Canadian Archipelago would give worse bias corrected SIT in the central Arctic, for instance. How are the results sensitive to the initial choice of CMIP5 models?

We are not solely interested in model performance versus observations. For example the fact the CSIRO performs 'poorly' for some metrics is beneficial to the manuscript as it provides a rigorous test of the MAVRIC method. The MAVRIC method is only trained on the models listed in Table 1. The criteria is rather subjective, the stipulations where made for the benefit of a paper currently in preparation that assesses the future of Arctic transit shipping, as such reasonable resolution and an at least partially resolved north west passage are seen as a necessity.

have a reasonable spatial resolution, and ~~**at least one ocean channel through the**~~ **a somewhat resolved** Canadian archipelago**. A consistent spatial distribution of land is needed for realistic and spatially complete multi-model means.***3826/5 I suspect that CMIP5 models were interpolated onto a common grid to make the grid-point recalibration feasible. The authors should indicate which reference grid was used (PIOMAS's? A regular 1x1?).*

Information added to Appendix A Supplementary MAVRIC methodology details:

For model biases to be calculated a common grid needed to be used, hence all MAVRIC calculations took place on the CMIP5 models native grid. This means that PIOMAS was converted to the CMIP5 model grid for each GCM's bias calculations. This choice was made as it only involves interpolating one of the two fields each time and generally it is PIOMAS that has the higher resolution.

3826/15 Change "observed" to "PIOMAS"

PIOMAS ~~**observed**~~

3826/16 Change "there is only one realization of the past" by "PIOMAS only yields one realization". In fact, PIOMAS was run with many atmospheric forcings (see Lindsay et al. (2014, doi: 10.1175/JCLI-D-13-00014.1)) but only makes one publicly available. Applying MAVRIC with other versions of PIOMAS wouldn't sample uncertainty related to internal variability, but at least to the atmospheric forcing used to generate PIOMAS.

PIOMAS only yields one ~~**there is only one**~~

(see Lindsay et al. (2014) for discussion of PIOMAS forced with alternative atmospheric forcings).

3826/17 I'm a bit confused here, because I think two ideas have to be expressed separately. First idea: the calibration period is short, hence internal variability pollutes the recalibration method. Second idea: even if the recalibration was done on a very long period, it is not sure that the future evolution of SIT would be correct because of the possible dependence of SIT on the mean state.

It is Dr Massonnet's first idea here that we discuss. The second idea we completely agree with in principle although we never claim any recalibration is "correct" only we argue in this manuscript that the SIT distribution and variance are more like PIOMAS. There is a complex mean state dependence in the models, to adequately rectify this would require active bias correction to many variables as the GCM is run, something far beyond the purpose of a simple post-processing bias correction technique like MAVRIC.

3826/24 Change "observations" by "PIOMAS"

In this case we are talking about bias correction methods in general in which case it is not appropriate to quote a specific data set.

3826/27 I don't understand the following sentence, explaining why trends are not corrected: "Our reasoning is to keep this as prescribed by the different models because the response of the SIT to future warming is unknown and GCMs are designed to give an estimate of this". Do the authors mean that it is useless to correct the trends over the PIOMAS period because the trends might anyway be different in future periods? If so, please rephrase.

Our reasoning is to keep this as prescribed by the different ~~**models**~~ **GCMs** because the response of the SIT to future warming is unknown and **likely non-linear and** GCMs are designed to give an estimate of this.

We are also cautious of over fitting. If we correct the mean, variance and the trend the resulting product will likely be woefully under-dispersed. Out of the mean, variance and trend we feel that given the nature of the data we can improve the mean and variance in GCMs but the trend is far more uncertain thus we leave this to the individual GCMs to resolve.

We are cautious of over fitting; applying a trend correction would potentially result in an over-confident projection.

3827/7 The toy model uses an AR1 process with declining linear trend. How was this choice made? What are parameters of the AR1 model? Did the authors check the auto-correlation properties of CMIP5 SIT evolution to design this toy model? When SIT approaches zero, negative values are reset to zero? All this information would be welcome to be able to reproduce the results.

The purpose of the toy model was to test different bias correction methods in a simplified time series so the effects of the different methods can be clearly seen. An AR1 model struck a good balance between being realistic enough that the system retains some memory (versus random numbers) or a more complex model where some of the differences between the methods may be harder to distinguish from a complex timeseries. To pick the parameters of the AR1 model timeseries auto-correlations where indeed consulted so that the toy model we used had similar properties. The AR1 parameter is 0.3, the standard deviation is model dependant and varies between 0.3 to 0.9. Negative values are reset to zero.

"produced using a first order auto-regressive (**with an AR(1) parameter of 0.3 chosen to be broadly representative of CMIP5 SIT auto-correlation)** model imposed on a declining linear trend **with negative numbers reset to zero**,"

3827/22 Replace "mean" by "time-mean"

Time-mean

3828/12 Sections 3.1-3.3, illustrating the limitations of simple recalibration methods, could cite the paper of Blanchard-Wrigglesworth and Bitz (2014, cited elsewhere in the manuscript) where the mean-variance relationship of SIT is clearly illustrated.

We choose not to mention aspects that have not yet been introduced. In sections 3.1-3.3 a variance correction has yet to be introduced. It would more appropriately appear in section 3.4 however we feel that it is more appropriate to the discussion section of the manuscript where the mean-variance relationship is discussed and Blanchard-Wrigglesworth and Bitz (2014) is there cited for that purpose.

3829/13 Add "thickness" between "sea ice" and "variance"

~~Sea ice~~ SIT

3829/15 The authors should define "ice-free" at this point of the manuscript. This concept is defined elsewhere in the manuscript, but it'd be good to have it where it is first introduced.

Ice-free is now defined at first occurrence in Section 3 in line with an earlier suggestion.

3830/12 CSIRO also has too much ice areal coverage, this could be added here.

The ice in CSIRO generally has too much ~~ice~~ **areal coverage** and too little variability

3832/10 How did the authors find that the shift towards earlier ice-free dates is attributed to the change in the variance rather than the mean? Is it a speculative statement or were tests done with and without mean or variance correction in MAVRIC?

Fig. 5c shows that the means between the Raw and MAVRIC time series are very similar (6% different) whilst the change in SD is far larger (176%) therefore it is clearly the variance term in MAVRIC that accounts for the 15-46 year difference in projected ice free date.

3832/13 I wouldn't use the term "projections" over the historical period, rather "simulations"

~~Projections~~ simulations

3835/23 What is the asterisk in SIV? I couldn't find where this points to.*

Edited for clarity, the * is also explained in the last line of the Fig. 10 caption

"Figure 10 shows the raw and corrected CMIP5 subset SIV* projections until 2100 using the 18 multi-model ensemble members in each scenario as before. ~~**The SIV**~~**(*** calculated here does not consider SIC as it is not bias corrected**)**."

3835/25 The assumption of 100% SIC in September is questionable. Have the authors looked at SIC in CMIP5 models in September for future periods? It is likely that models simulate values much lower than that. Did the authors try other baseline values for SIC? That is, can the sentence "this assumption should only have a relatively small effect" be supported by objective arguments?

The 100% SIC was used for consistency. As per Dr Massonnet's suggestion, future SIC has been analysed. We use take the mean (of the non-zero gird cells) September SIC in CCSM4 RCP8.5 and find a typical SIC of approximately 50% for 2006-2100. We then recalculate SIV* using 50% instead of 100%. We also reduce our ice-free threshold to 1×10^3 km^3 as opposed to 2×10^3 km^3 and this is has the benefit of now being directly comparable with the often used ice-free threshold for SIC of 1×10^6 km^2. This also means the timings remain the same. Fig 10. has been updated to reflect these changes.

~~**Instead, 100 % SIC is assumed throughout.**~~ **To find a representative SIC for the SIV* calculation we use the September SIC in CCSM4 RCP8.5 and find a mean (non-zero) SIC of approximately 50% for 2006-2100.** ~~**It is worth noting that SIV is heavily influenced by the thicker ice to the north of the Canadian archipelago where the true SIC is near 100 %, so this assumption should only have a relatively small effect.**~~

3836/9 Magnitude is always positive. Delete "absolute", unless you want to oppose it to relative magnitude.

~~**Absolute**~~ **relative**

3841/11 Did the authors check the residuals ($T^2 = M^2 + I^2 + S^2$) to quantitatively verify that the independence between the three sources of uncertainty can safely be assumed?

"We note that the variances calculated above do not always sum exactly in this way due to small interaction terms **(approximately 10%)** which we ignore."

However the figures are scaled to 100% so the relative magnitudes remain representative

Fig. 1 The colorbars (and colorbars of all subsequent figures) have a bin that goes below zero. This is a bit disturbing, as we know that sea ice thickness is always positive. Following the colorbar conventions, dark blue areas must be ice-free (SIT=0) grid cells, but then white areas *must be grid cells with SIT>=2.25 m, following the same convention. Another person could interpret white areas as 2m<SIT<2.25m, though. There might be confusion.*

We do not regard the colorbars as confusing. With regards to Fig 1, the dark blue area represents areas of SIT = 0 m hence water. The next darkest blue bin then represents SIT greater than 0 m to 0.25 m, the next bin is then greater than 0.25 m to 0.5 m. This follows to the white bin which is regions for SIT greater than 2 m.

Fig. 2 Same comment as for Fig. 1

Same logic above applies here and throughout.

Fig. 3 Please add units of SIT along the y-label.

SIT **[m]**

Fig. 4 Same as for Fig. 3

SIT **[m]**

Fig. 5 Same as for Fig. 3

SIT **[m]**

Fig. 6 Same as for Fig. 1. Also, adding the PIOMAS SIT fields would be insightful to report the improvements.

The PIOMAS fields are in Fig 1 and hence not duplicated here.

PIOMAS SIT fields shown in Fig 1.

Fig. 7 Same as for Fig. 1. Also, a map with differences (corrected minus raw) would be very helpful to interpret the benefits of the bias-correction method. In the current version of the figure, it is really difficult to see where the corrections occurred. A blue-red set of maps with positive-negative changes in SIT could be added as a third row.

Agreed. A row of MAVRIC – Raw is added and adds a lot of information. A green to purple colour bar is used to avoid confusion of a blue-red: less to more versus cold to hot contradiction.

“Figure 7. September **multi-model ensemble mean (three members from each model)** mean SIT from the CMIP5 subset, using the raw data (top row) and after MAVRIC (~~**bottom**~~ **middle** row). ~~**The multi-model ensemble mean (three members from each model) is shown.**~~ **The bottom row shows (MAVRIC – Raw) and hence green areas are where MAVRIC has reduced SIT and purple areas are where MAVRIC has increased SIT.”**

Fig. 8 Same as for Fig. 1. Also, make clear that you define the "sources of SIT uncertainty" as the standard deviation of the detrended SIT.

Figure 8. September 2015-2024 sources of SIT uncertainty from the CMIP5 subset **(SD of the detrended SIT)**.

Fig. 9 I would change "Uncertainty" by "Variance" in panel (a), because "uncertainty" has been used interchangeably with "standard deviation" in the rest of the text. Alternatively, you can choose to show the standard deviation but then loose additiveness.

~~Uncertainty~~ Variance

We again thank Referee Dr Massonnet for his thorough and constructive review of our submission.

Kind Regards,

N. Melia, K. Haines and E. Hawkins

References

Blanchard-Wrigglesworth, E. and Bitz, C. M.: Characteristics of Arctic Sea-Ice Thickness Variability in GCMs, J. Clim., 27, 8244-8258, doi: 10.1175/Jcli-D-14-00345.1, 2014.

Zhang, J. and Rothrock, D.: Modeling global sea ice with a thickness and enthalpy distribution model in generalized curvilinear coordinates, Mon. Weather Rev., 131, 845-861, 2003.