Response to Referee Prof. Flato

Format key:

Normal text is considered general discussion

Italic text quotes the Referee

Bold text indicates additions to the manuscript

We wish to thank Referee Prof. Flato for his quick response and constructive review of our submission.

Prof. Flato suggests two separate point of consideration:

*"I have a slight quibble with the use of the word 'uncertainty' in this paper, and in particular the extent to which reducing 'spread' is equivalent to reducing 'uncertainty'. This may be largely a semantic issue, but it is not obvious to me that reducing spread *necessarily* reduces uncertainty (in the sense of the confidence one has, or should have, in a prediction or projection). Spread is of course directly related to uncertainty, and the partitioning suggested by Hawkins and Sutton yields considerable insight into the sources of uncertainty and how these change over time. But I think one has to be a bit careful in equating reduced spread with reduced uncertainty (and by extension, enhanced confidence) as is done here. One can readily construct schemes that reduce spread (e.g. discarding all models but one), but don't really reduce uncertainty. Perhaps a few sentences on this topic could be added?"*

Firstly, our use of the word uncertainty in this context is perhaps a little enthusiastic as we do indeed equate a reduction in model 'spread' with a reduction in 'uncertainty'. Prof. Flato states a simple example where this would not be a valid statement which we consider a helpful point from which to adjust and clarify our terminology. But also note the additional comments in our reply to Dr Massonnet about testing our uncertainty estimates which, within the limitation of the models examined, appear largely reliable.

We will add **potential** to "reducing uncertainty" and "increased confidence" to highlight our slight hesitation with such claims. We will clarify our use of the word 'uncertainty' with the following sentences located in Sect. 4.4 of the manuscript:

An additional source of uncertainty that we neglect here is the PIOMAS calibration uncertainty emerging from the choice of atmospheric reanalysis and ice model tuning. This could be assessed by sampling the different versions of the PIOMAS reanalysis described in Lindsay et al. (2014).

In the following sections, we equate reducing model spread with reduced uncertainty. While 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.

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.

..

Secondly, Prof. Flato rightly points out two points of confusion in Fig. 3:

"I did note two things related to Figure 3 however: – the caption states that 'ice-free' is defined as the "first occurrence ... below 0.15m", but the legend gives a range of years of 'ice-free year'. I didn't understand this.

– the legend indicates no change in 'ice-free' year for the high-mean (blue) example when the multiplicative correction is applied (compare Fig 3a and 3c) even though the curve is obviously shifted downward. I suspect a typo in the legend. The same applies to Fig 3b and 3d where again the ice-free year for the blue curve is unchanged."

The reason for the confusion in the first point is primarily due to inadequate explanation of what the dates below 'ice-free' represent in this figure. This is rectified by adding the following sentence to the caption:

"Ice-free" is here defined as the first occurrence of an ensemble member below 0.15 m. **Shown is the "ice-free" ensemble range, i.e. the year of the first ensemble member to be "ice-free" to the last ensemble member to be "ice-free".**

Secondly Prof. Flato also notes that the "ice-free" statistics are identical on comparing Fig 3a with 3c, and Fig 3b with 3d. This is in fact not a typo and a true representation of our "ice-free" criterion, this is partly coincidence and partly due to how the four correction methods shown manipulate the time series. While Prof. Flato rightly points out that the curves have obviously shifted, the "ice-free" date remains the same. This is shown by examining when the thin coloured lines cross 0.15 m. This is an important point that Prof. Flato observes, the following paragraph is added to Sect. 3.4 to highlight this behaviour:

Comparing the ensemble range in projected ice-free date between the correction methods it is apparent that although the shapes of time-series have qualitatively changed this does not always result in a different range in projected ice-free date. For example on comparing the high mean – high variance GCM (blue) between (a) to (c) and (b) to (d); this is partly coincidence and partly due to how the four correction methods shown manipulate the time series. The MAVRIC method (e) results in a unique set of ice-free dates. This is an important attribute that the MAVRIC method displays, as the ice-free date is of vital importance to stakeholders in the Arctic and more basic methods of bias correction fail to appropriately impact on this parameter.

We again thank Referee Prof. Flato for his quick response and constructive review of our submission.

Kind Regards,

N. Melia, K. Haines, and E. Hawkins

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:

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

where $\boldsymbol{E_n}$ is the ensemble member between 1 to 18, $\overline{\boldsymbol{E_{15}}}$ is the mean of the 15 ensemble members from the models of which $\boldsymbol{E_n}$ 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.

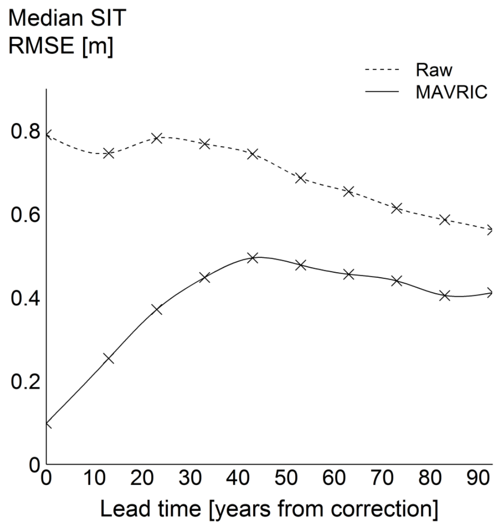


Figure **C1**. Multi-model ensemble out of sample September median SIT RMSE]

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:

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

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

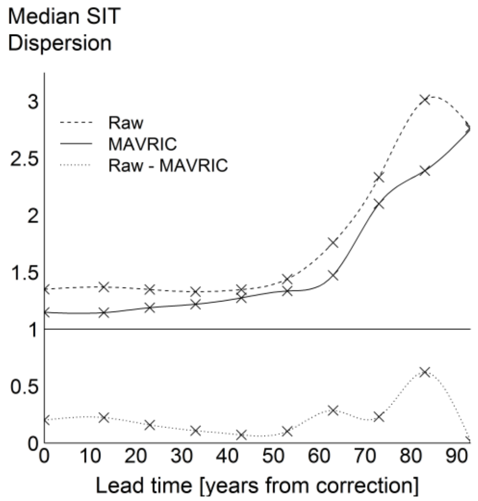


Figure C2. Multi-model ensemble out of sample September median SIT dispersion

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.

Improved Arctic sea ice thickness projections using bias corrected CMIP5 simulations

N. Melia[1], K. Haines[2] and E. Hawkins[3]

[1] {Department of Meteorology, University of Reading, Reading, United Kingdom}

[2] {National Centre for Earth Observation, Department of Meteorology, University of Reading, Reading, United Kingdom}

[3] {NCAS-Climate, Department of Meteorology, University of Reading, Reading, United Kingdom}

Correspondence to: N. Melia (n.melia@pgr.reading.ac.uk)

Abstract

Projections of Arctic sea ice thickness (SIT) have the potential to inform stakeholders about accessibility to the region, but are currently rather uncertain. The latest suite of CMIP5 Global Climate Models (GCMs) produce a wide range of simulated SIT in the historical period (1979 – 2014) and exhibit various ~~spatial and temporal~~ biases when compared with the Pan-Arctic Ice Ocean Modelling and Assimilation System (PIOMAS) sea ice reanalysis. We present a new method to constrain such GCM simulations of SIT to narrow projection uncertainty via a statistical bias correction technique. The bias correction successfully constrains the spatial SIT distribution and temporal variability in the CMIP5 projections whilst retaining the climatic fluctuations from individual ensemble members. The bias correction acts to reduce the uncertainty in projections of SIT and reveals the significant contributions of climate internal variability ~~sea ice internal variability~~ in the first half of the century and of scenario uncertainty from mid-century onwards. The projected date of ice-free conditions in the Arctic under the RCP8.5 high emission scenario occurs in the 2050s, which is a decade earlier than without the bias correction, with potentially significant implications for stakeholders in the Arctic such as the shipping industry. The bias correction methodology developed could be similarly applied to other variables to narrow uncertainty in climate projections more generally.

1 Introduction

Global Climate Models (GCMs) are the primary tool for making climate predictions on seasonal to decadal time scales, and climate projections over the next century (Flato et al., 2013). In a warming climate, changes to sea ice thickness (SIT) are expected to lead to significant implications for polar regions and beyond. A reduction in SIT will likely open up the Arctic Ocean to economic diversification including new marine shipping routes (Smith and Stephenson, 2013) and extraction of natural resources, as well as changes to the Arctic ecosystem and potential links to mid-latitude weather (Francis and Vavrus, 2012). Many of these economic opportunities may rely on SIT evolution, but current projections have considerable uncertainty. SIT is also much more informative than sea ice concentration (SIC), especially in the central Arctic, where future thinning can occur without major changes in the local SIC.

The GCMs from the Coupled Model Intercomparison Project, phase 5 (CMIP5) (Taylor et al., 2012) exhibit a large range in sea ice volume (SIV), spatial SIT distribution, and temporal SIT variability under present day forcing conditions (e.g. Blanchard-Wrigglesworth and Bitz (2014)). For September sea ice extent, Swart et al. (2015) showed the uncertainty in CMIP5 projections over the next few decades is dominated by these differences between models, termed model uncertainty by Hawkins and Sutton (2009, 2011). Uncertainty in climate projections arises from three distinct sources: (1) model uncertainty, (2) internal variability, and (3) scenario uncertainty, as discussed by Hawkins and Sutton (2009, 2011) for temperature and precipitation respectively. In contrast to projections of temperature where the anomalies are often used, the absolute value of SIT is important – for example, ships have critical SIT thresholds above which their use is not possible (Stephenson et al., 2013). ~~Since absolute values are used, model bias makes a contribution to model uncertainty.~~

Bias correction (BC) of GCM simulations has the potential to reduce the model uncertainty and hence increase confidence in near term climate projections. The importance of BC in impact based climate change studies was described in a special report of the IPCC (Seneviratne et al., 2012), but BC has not previously been applied to projections of SIT~~sea ice~~; this manuscript is novel in that it recalibrates SIT, and does it locally. There are many different types of proposed BC techniques, (e.g. Boe et al. (2009); Christensen et al. (2008); Ho et al. (2011); Mahlstein and Knutti (2012); Vrac and Friederichs (2014); Watanabe et al. (2012), and references therein), which have mainly been applied to temperature and precipitation. However, these existing methods need refining for sea ice as SIT is a particularly challenging variable. This is due to its positive semi-definite nature, and the spatial and temporal occurrence of zeros, in observations and projections of SIT.

This study addresses the development of a new BC technique that constrains both the mean and variance of SIT in GCMs to an estimate of the observed statistics. It is important to correct the mean as this corrects the spatial SIT distribution. Variability in SIT also has a significant impact on the range of regional ice-free dates, something of great interest to stakeholders, and the CMIP5 GCMs exhibit a wide range in their SIT variability. The study also uses multiple ensemble members from the same model when performing the BC, something that is often not utilised in other studies. This is important as it enables an assessment of the role of internal variability in future projections to be made. The techniques described in this paper are not limited to SIT, and would work for many climate variables. The exact implementation used in this study should also be calibrated to the user's needs based on factors such as the length of reliable observations and number of ensemble members.

In this paper we use the Pan-Arctic Ice Ocean Modelling and Assimilation System (PIOMAS) (Zhang and Rothrock, 2003) as a~~n~~ ~~observationally~~ reanalysis based estimate of recent SIT, along with climate projections from a subset of six GCMs from the CMIP5 archive (Sect. 2). We first test the performance of increasingly complex BC approaches in a toy model environment (Sect. 3) and then apply our favoured method to the subset of CMIP5 GCMs in Sect. 4. We test the BC method by splitting the historical PIOMAS data, and then explore how the uncertainty in SIT projections is reduced using these techniques (Sect. 4) and summarise and discuss the results in Sect. 5.

2 Climate simulations and observations

2.1 PIOMAS

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). It does not however assimilate sea ice thickness (SIT), although this has been attempted using the NASA Operation IceBridge and SIZONet campaigns of 2012 (Lindsay et al., 2012).

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.

We choose PIOMAS to represent observations of SIT as satellite observations are limited in their spatial and temporal range. For example, data from ICESat are only available between October and March 2003 – 2008 (Kwok et al., 2009). More recently Cryosat-2 ~~(Laxon et al., 2013)~~ has started producing real-time SIT datasets but only for the non-summer months (Tilling et al., 2015). This is also not ideal as it is the summer months when the ice is thinnest that are most relevant for potential economic activity. The spatial consistency, temporal length and completeness of the data are important considerations when computing climatological means and variances as the longest time series possible is needed to validate the statistics. It is for this reason primarily that PIOMAS has been chosen to represent observations in this study. Several studies (e.g. Laxon et al. (2013), Schweiger et al. (2011), Lindsay and Zhang (2006), and Stroeve et al. (2014)) have compared PIOMAS to satellite and in situ observations and models and find it a suitable estimate of observed SIT. PIOMAS is also deemed realistic enough to initialise numerical models for seasonal forecasts e.g., the Sea Ice Outlook (Blanchard-Wrigglesworth and Bitz, 2014) where the accuracy of the initial conditions is vital.

Comment [NM1]: Reference added

Figure 1 shows the mean September SIT and temporal standard deviation (SD) after linear detrending for PIOMAS over the satellite era (1979 – 2014). ~~The thickest ice is located north of the Canadian archipelago and Greenland.~~ In the heart of the Canadian archipelago, ice thickness is up to 1.5 m, in the central Arctic it is about two meters, and it is between zero and one meter along the north Russian coast. The SIT is most variable around the edge of the ice pack and especially near land. An effective BC should ensure that the simulations replicate these patterns of mean SIT and SD over this recent period.

2.2 Global climate models

This paper utilises a subset of six GCMs from CMIP5. Since a large part of this work assesses SIT variability, it is necessary for each GCM to have multiple ensemble simulations in the historical period and for each of the representative concentration pathways (RCPs) 2.6, 4.5 and 8.5 for future scenarios (Van Vuuren et al., 2011). In addition, the GCM mean spring thickness must fall within the 10th and 90th percentile of PIOMAS (Stroeve et al., 2014), 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. The six GCMs that comprise this CMIP5 subset are listed in Table 1.

For the CMIP5 subset the historical simulations are used for the period 1979 – 2005. In most of the analysis for the period post-2005 the RCP8.5 scenario is used, which ramps up the amount of greenhouse gases to have a cumulative effect of increasing the direct radiative forcing by 8.5 Wm^{-2} (approximately 1370 ppm CO_2 equivalent) by 2100 (Van Vuuren et al., 2011). The impact of other scenarios is assessed later in the analysis. Figure 2 shows the 1979 – 2014 ensemble-mean September SIT for the CMIP5 subset, highlighting the considerable differences between the model simulations, and indicating that model bias is likely to be the dominant uncertainty in near-term projections.

The aim of the SIT BC outlined in this paper is to correct the mean and variance in the CMIP5 subset shown in Fig. 2 to the ~~observed~~ PIOMAS statistics. Although this should improve short-term predictions, a caveat to this approach is that PIOMAS only yields one ~~there is only one~~ realisation of the past (see Lindsay et al. (2014) for discussion of PIOMAS forced with alternative atmospheric forcings)~~,~~. ~~and w~~We have to assume that the relatively short period over which we have observations (36 years) captures a representative sample of the behaviour we expect from the climate system. In the short term, this is probably a reasonable assumption, as the GCMs will not have evolved far from their corrected state of the recent past; this assumption is explored further in Sect. 4.

3 Bias correction methodology

Bias correction methods effectively aim to reduce model uncertainty by constraining GCMs to observations. There are two components to model uncertainty: the overall mean difference (or bias), and differences in the amplitude of response to specified forcings. We have deliberately chosen not to try and correct the simulated ice loss trend to that which is currently observed. 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. It is also doubtful how well the current trend can be determined from 36 years of data given the high noise to signal ratio for trends, especially on grid point scales. It is also unclear how much of the recent ice loss seen in the observations can be attributed to changes in external forcing as opposed to internal variability (e.g. Day et al. (2012); Kay et al. (2011); Swart et al. (2015)). We are cautious of over fitting; applying a trend correction would potentially result in an over-confident projection.

To test the performance of different possible BC methods a 'toy model' was used as proxy ensemble timeseries (representing SIT at a single grid point for the same month each year for the period 1979 – 2100). The timeseries are shown in Fig. 3a for a high mean - high variance model (blue) and a low mean - low variance model (red), where the black line shows the "truth" observations with one realisation over the historical period only. The time series were all produced using a first order auto-regressive (with an AR(1) parameter of 0.3 chosen to be representative of CMIP5 SIT auto-correlation) model imposed on a declining linear trend with negative numbers reset to zero.~~,~~ ~~with~~Each model has five separate model ensemble members (thin coloured lines) and the thick lines representing the ensemble means. The statistics in all the legends are calculated over the observation window (1979 – 2014). 'Ice-free' in Fig. 3 is here defined as the first occurrence of an ensemble member below 0.15 m. Shown is the ice-free ensemble range, i.e. the year of the first ensemble member to be ice-free to the last ensemble member to be ice-free. A successful BC method should transform the individual ensemble members (thin red and blue lines) to match the mean and variance of the observations (black line), producing matched statistics. We test various approaches for such a bias correction. The mathematical notation for the following equations is in Table 2.

3.1 Additive correction

A basic additive correction, which has previously been used for temperature projections, is shown in Fig. 3b. This approach simply corrects the time-mean by subtracting the difference between the historical model ensemble-mean time-mean, $\langle\overline{M_h}\rangle$, and observation time mean, $\overline{O_h}$, from each of the model ensemble members, M.

$$\text{Additive corrected thickness} = M - (\langle\overline{M_h}\rangle - \overline{O_h}) \quad (1)$$

However, as the low ice model is adjusted up by the addition of a constant, it equilibrates at a positive value in the future rather than zero. Likewise the high ice model equilibrates at negative values. Neither of these properties are sensible.

This study makes use of multiple ensemble members from the same model, raising the question of how to treat ensemble member statistics when calculating a particular GCM's bias. For calculating the mean SIT, each GCM's ensemble mean is used because it is the GCM's mean bias that we wish to correct. This is important because a particular ensemble member's deviation from the ensemble mean is retained; it allows an individual ensemble member's time mean to be different to the observations over the historical period, but not the ensemble mean. The treatment of ensemble members for the SD calculation is described in section 3.4.

3.2 Multiplicative correction

If a multiplicative correction is used (Fig. 3c), where the ratio of the observed time mean and model ensemble-mean time-mean, $\overline{O_h}\,/\langle\overline{M_h}\rangle$, is multiplied as a factor to the model ensemble members, M, then the corrected thickness is:

$$\text{Multiplicative corrected thickness} = M\frac{\overline{O_h}}{\langle\overline{M_h}\rangle} \quad (2)$$

Multiplicative methods effectively preserve the future zero ice year, which is potentially an important value for a wide range of stakeholders. However, when applied as above this approach has the undesired effect of distorting the variances by the same factor as the mean correction, as visible in Fig. 3c.

3.3 Mean multiplicative correction

To avoid altering the variances, the mean multiplicative correction can be introduced (Fig. 3d), where the multiplicative mean correction, $\overline{O_h}\,/\langle\overline{M_h}\rangle$, is applied only to the 11-year-centred running-mean ensemble-mean, $\langle\widetilde{M}\rangle$. This corrects the model mean evolution without corrupting the sub-decadal variance as $\langle\widetilde{M}\rangle$ is smoothed. The model anomalies for each ensemble member, $M-\langle\widetilde{M}\rangle$, are then added back to the corrected mean evolution:

$$\text{Mean multiplicative corrected thickness} = \left(M-\langle\widetilde{M}\rangle\right)+\langle\widetilde{M}\rangle\frac{\overline{O_h}}{\langle\overline{M_h}\rangle} \tag{3}$$

This works to correct the mean SIT and does not suffer from any peculiarities of the previous two methods. The model variance now remains unchanged but the approach opens up the possibility of correcting the variance towards that observed in the historical period. Note that by using the ensemble mean, $\langle\overline{M_h}\rangle$, for all these corrections we ensure that each ensemble member is corrected in the same way, thus preserving certain ensemble properties into the future.

3.4 Mean and variance correction

The GCMs from CMIP5 show a large range in ~~sea ice~~SIT variance, and the magnitude of these variations is a significant factor determining when regions of the Arctic may first become accessible (when one ensemble member may first become ~~ice free~~ice-free). Therefore a variance correction is incorporated into Eq. (3) by taking the ratio of the temporal standard deviation of the detrended observations, $\sigma_{\widehat{O_h}}$, to the square root of the ensemble mean of the variance of the detrended model ensembles, $\langle\sigma_{\widehat{M_h}}\rangle$ (detrended mean ensemble SD), over the historical period. The detrending in the models is calculated using each model's ensemble mean linear trend. This has some similarities to the approach of Ho et al. (2011) in application to temperature projections for Europe. Also see Appendix A for some further discussion of the choices made.

To incorporate the variance correction, the mean multiplicative correction (Eq. (3)) is first de-trended, the variance correction applied, and the trend re-applied. This creates the Mean And VaRIance Correction (MAVRIC), shown in Eq. (4):

$$\text{MAVRIC} = \left(M-\langle\widetilde{M}\rangle\right)\frac{\sigma_{\widehat{O_h}}}{\langle\sigma_{\widehat{M_h}}\rangle}+\langle\widetilde{M}\rangle\frac{\overline{O_h}}{\langle\overline{M_h}\rangle} \tag{4}$$

Fig. 3e shows the MAVRIC does a near perfect job of correcting both the mean and variance to the observed statistics while still retaining the individual ensemble members' own climate fluctuations, but fractionally scaled by the variance ratio.

Comparing the ensemble range in projected ice-free date between the correction methods it is apparent that although the shapes of time-series have qualitatively changed this does not always result in a different range in projected ice-free date. For example on comparing the high mean – high variance GCM (blue) between (a) to (c) and (b) to (d); this is partly coincidence and partly due to how the four correction methods shown manipulate the time series. The MAVRIC method (e) results in a unique set of ice-free dates. This is an important attribute that the MAVRIC method displays, as the ice-free date is of vital importance to stakeholders in the Arctic and more basic methods of bias correction fail to appropriately impact on this parameter.

4 Bias corrected sea ice thickness projections

Figure 3e illustrates that the MAVRIC successfully corrects the mean and variance in a toy model environment. Before proceeding to investigate the impact of the MAVRIC on SIT projections it is prudent to test whether the MAVRIC can improve GCM performance by validating with real observations. We use CSIRO-Mk3.6.0 (CSIRO) as the GCM to test. The ice in CSIRO generally has too much ~~ice~~areal coverage and too little variability and is a CMIP5 outlier model with regards to SIT (Stroeve et al., 2014). However, CSIRO benefits from having 10 ensemble members, increasing the robustness of the statistics. For these two reasons, it is considered a thorough test of the MAVRIC's performance within a real GCM.

The test uses a data denial method where we train the MAVRIC on a subset of PIOMAS observations, 1979 – 1999, termed the calibration window. From this we examine how the MAVRIC predicts the observations for 2000 – 2014, termed the validation window. A limitation with this method is the length of observations: the period over which the MAVRIC calibration takes place must be long enough to capture a robust measure of the observed statistics. The validation period must also be long enough to be able to draw robust conclusions. It is not clear whether either the 21 year calibration or the 15 year validation windows are long enough for robust method calibration and results verification, but we are limited by the data available. An additional limitation to this method is that the calibration and validation periods are very close to each other.

Figure 4 shows the performance of the MAVRIC at three grid points for September. The raw CSIRO ensembles (grey) are bias corrected via the MAVRIC using the PIOMAS observations (black) over the calibration window, producing the MAVRIC corrected ensembles (green) for the validation window. If the MAVRIC can produce plausible predictions, the characteristics of PIOMAS should be indistinguishable from individual corrected ensemble members in the validation window. It is clear from the validation beanplots (right), that the distribution from the corrected ensembles resembles PIOMAS much more closely than the raw distribution, e.g. non-zero probability of zero ice. We do not expect the distribution from PIOMAS to match the corrected distribution perfectly as PIOMAS only has one realisation (15 data points) while CSIRO has 10 realisations. We can tentatively accept that this test demonstrates the validity of the MAVRIC approach.

In the following sections the MAVRIC is applied to the CMIP5 subset of six GCMs used in this study (Table 1). PIOMAS estimates of Arctic SIT are available from 1979 – 2014. This 36 year window is the period over which statistics are calculated in the observations, and in the CMIP5 subset (using historical runs for 1979 – 2005 and RCP8.5 for 2006 – 2014). Each model, month, and grid point has its own specific correction which is applied to all years (1979 – 2100). However, separate ensemble members from the same GCM are treated with the same correction, as we wish to correct the model bias and retain the ensemble spread. Results are shown for September, initially only for CSIRO and later for all six models combined to form the ‘CMIP5 subset’ used for this study.

4.1 Temporal perspective example

Figure 5 shows the impact of the MAVRIC in September in CSIRO at the same three grid points as Fig. 4 but for the entire calibration window (1979 – 2014). The East Siberian Sea in CSIRO has about double the SIT and half the SD of PIOMAS (Fig. 5a). The correction therefore reduces the mean SIT whilst increasing the variance. This brings forward the range of first year ice-free conditions (the first occurrence in each ensemble member of a SIT below 0.15 m) from after 2100 to 1981 – 2032. Ice age (and hence strength) correlates well to ice thickness (Maslanik et al., 2007), and values below 0.15 m correspond to young and grey ice categories, and operations in this ice regime require no specific ice strengthening of vessels (Transport Canada, 1998). Similarly in the Beaufort Sea (Fig. 5b) the SD needs to be almost tripled, and the correction results in the first ice-free year coming over 100 years earlier. In the Fram Strait (Fig. 5c) CSIRO and PIOMAS have similar SIT requiring only a small mean adjustment, however CSIRO requires a big increase in variance. The MAVRIC moves the first possible ice-free date about 30 years earlier and increases the ensemble uncertainty range from 32 to 63 years. It is worth noting that the dominant cause of this shift to earlier ~~ice free~~ice-free date at this location is due to the variance correction term in the MAVRIC rather than the mean correction term. This highlights the importance of correcting the variance in addition to the mean. Figure 5 demonstrates that the MAVRIC can lead to ~~projections~~ simulations that look significantly more like reality in the historical period and have an impact on regional ice-free projections.

4.2 Historical spatial perspective

In addition to examining the MAVRIC in a temporal sense, it is important to evaluate the results spatially to see where the MAVRIC is having the most effect and if it works at all locations. Figures 2 and 6 show that the mean September SIT distribution is very different in HadGEM2-ES and CSIRO. After the MAVRIC has been applied, the mean SIT fields are almost identical for the historical period (Fig. 6). It is important to note there are still differences when considering individual years and ensemble members i.e. the year-to-year variability and ensemble spread is preserved (although adjusted by the MAVRIC).

Figure 6 also shows the SD before and after the MAVRIC. The SD shown is the detrended mean ensemble SD as before. CSIRO has too low variability in the majority of locations although correctly places the maximum SD near the edges of the ice pack similarly to PIOMAS. HadGEM2-ES exhibits about the same magnitude of variability as the observations but the variability is too high in the centre of the ice pack and too low at the edges. After the correction the SD fields in both GCMs now look more similar to each other with the highest variability located at the edge of the ice pack and at coastal locations. They are now also both similar to the estimate from PIOMAS (Fig. 1).

4.3 CMIP5 subset multi-model sea ice thickness projections

The bias corrected SIT from each GCM can be brought together to form the multi-model mean CMIP5 subset, computed using three ensemble members (the maximum available across all models) from each of the six GCMs for the historical and future decadal periods (Fig. 7). It is remarkable how the raw multi-model mean product for the historical period is not too different from PIOMAS in Fig 1, showing that the location and magnitude of model biases cancel out to a considerable degree, at least with this subset of models. Given this result it is not so surprising that the raw and corrected fields are fairly similar for the future projections also.

Nevertheless, even in this multi-model multi-ensemble framework the MAVRIC is still making some discernible differences. These differences are most apparent in the Canadian archipelago and the Russian Arctic seas, where the correction leads to a reduction in SIT of approximately 1 m in both regions. Both the raw and bias corrected fields predict a SIT loss of about 0.25 m per decade.

The fact that the MAVRIC is still making a significant difference on the regional scale is critical, e.g. for ship route availability. Currently studies that assess the future opening of Arctic shipping routes, which critically depend on the absolute value of SIT, do not yet account for such factors and will need to be reassessed.

4.4 Sources of uncertainty in projections of sea ice thickness

The uncertainty in climate projections can be partitioned into three distinct sources: (1) model uncertainty: for the same radiative forcing different models simulate different mean distributions and temporal changes. (2) Internal variability: the natural fluctuations of the climate present with or without any anthropogenic induced changes to radiative forcing. (3) Scenario uncertainty: uncertainty in future radiative forcing resulting from unknown future emissions. Hawkins and Sutton (2009, 2011) assessed these sources of uncertainty in global and regional temperature and precipitation projections, and here we quantify the sources of uncertainty in SIT, utilising the CMIP5 subset multi-model ensemble. Crucially we use the absolute values of SIT rather than considering anomalies as is often done for other climate variables. The methodology for partitioning these sources of uncertainty is detailed in Appendix B. An additional source of uncertainty that we neglect here is the PIOMAS calibration uncertainty emerging from the choice of atmospheric 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 bias 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 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.

The MAVRIC method outlined in this study acts to eliminate the model bias (and hence potentially reduce the uncertainty) in the MAVRIC calibration period (1979 – 2014). After this period the model uncertainty grows due to the GCM's differing responses to changes in external forcing. The sources of uncertainty for SIT for the decade 2015 – 2024, immediately following the MAVRIC calibration period, are shown in Fig. 8. The total uncertainty in the corrected CMIP5 subset is strikingly lower than in the raw CMIP5 subset. Closer analysis reveals that this is due to the substantial reduction in model uncertainty owing to the MAVRIC. The other sources of uncertainty do not change as much.

The temporal evolution of these sources of uncertainty is shown in Fig. 9a by taking the median variance from each of the panels in Fig. 8 for this and other periods. There are three competing factors for how the uncertainty will change with time. First, the SIT is decreasing, and this will reduce the uncertainty as the range of values of which the SIT can occupy shrinks. Second, the separate GCM's simulated SIT responses due to external forcing will differ from each other, causing GCMs to drift apart over time. Thirdly, sea ice at the grid point scale becomes more mobile and vulnerable to external factors as it thins. This will increase variability, initially at least (Sou and Flato, 2009). All of these factors are involved in the evolution of the uncertainties.

The raw CMIP5 subset exhibits a decrease in total uncertainty with time (dashed black in Fig. 9a). This is primarily due to the reduction in model uncertainty (dashed blue), likely because the mean SIT is reducing. The corrected total uncertainty is lower than the raw uncertainty until at least the end of the century. This means that the MAVRIC can reduce uncertainty and increase confidence in climate projections of SIT throughout this period. The corrected model uncertainty increases for the first three decades, as the models start from a similar state and subsequently diverge because of differing responses to the changes in external forcing. Later the corrected model uncertainty reduces as the mean SIT decreases towards zero.

The total uncertainty is the sum of model uncertainty, internal variability, and scenario uncertainty (see Appendix B for more details). The other panels in Fig. 9 illustrate the relative importance of these sources of uncertainty in terms of the percentage total variance explained, for the raw data, and after the MAVRIC.

Fig. 9b illustrates that in the raw projections, model uncertainty remains the dominant (> 50 %) source of uncertainty until at least 2100, whereas it only becomes dominant for a few decades mid-century after the MAVRIC (Fig. 9c). The absolute magnitude of internal variability, and its contribution to the total uncertainty, decreases with time because SIT also decreases with time. In the corrected projections, the internal variability is the major contributor to the total uncertainty for the first 25 years, compared to a maximum contribution of only 26 % in the raw projections. This highlights the importance of correcting the variance to realistic magnitudes and also the key role of natural variations in predicting the near future evolution of sea ice. The scenario uncertainty accounts for less than 10 % of the total uncertainty for the first 50+ years. Additional analysis metrics on the improvement the MAVRIC method affords can be found in Appendix C

4.5 Reducing~~ed~~ uncertainty in timing of ice-free conditions

By reducing the model uncertainty~~, confidence in SIT projections is improved as~~ the range of possible outcomes has been reduced, this potentially leads to greater confidence in SIT projections. 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 ~~sea ice concentration (~~SIC~~)~~ as it is not bias corrected). ~~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 (of the non-zero grid cells) 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.~~

The thick coloured lines are the multi-model scenario mean and the coloured regions represent the 16 – 84 percentiles (equivalent to 1σ around the mean of a Gaussian distribution) of the ensemble members. To account for the large range in SIT at any particular time in the CMIP5 subset, we use a method similar to that of Massonnet et al. (2012) to calculate first ice-free conditions. We postulate that SIV for ice-free conditions is

~~2~~1 $\times 10^3$ km^3, which is in agreement with previous studies calculating first ice-free dates (e.g. Massonnet et al. (2012) and Overland and Wang (2013)), and is equivalent to ~~two~~ one meter thick ice for an ice extent of 10^6 km^2.

The MAVRIC reduces the total SIV, but the ~~absolute~~ relative magnitude of this reduction decreases as SIV declines. The 16 – 84 % range has also been vastly reduced, particularly for the near future. For example, in 2025 the MAVRIC has reduced the 16 – 84 % range from ~~12~~6 $\times 10^3$ km^3 to 2.5~~5~~ $\times 10^3$ km^3. It is this reduction in the plausible range of SIV that leads to potential increased confidence in projections of SIT and SIV. To assess when the Arctic will first display ice-free conditions, we focus on RCP8.5, the most realistic scenario from the last 10 years (Fuss et al., 2014). The cumulative number of ensemble members having satisfied the ice-free criterion as a function of time is shown in Fig. 10c. If uncertainty in this parameter has reduced, this will be shown by the gradient of the line increasing after MAVRIC, and this is clearly seen. Figure 10d further illustrates the uncertainty reduction with boxplots, where the line represents the median (9th) ensemble member to go ice-free. This occurs in 2052 with the MAVRIC, nine years earlier than before. The box represents 16 – 84 % of the ensemble members, this range has been reduced by about 20 years; dates after 2085 can now be eliminated.

Corrected results from the other emission scenarios show similar features but with later ice-free dates, as expected for lower emissions, and some ensemble members fail to go ice-free by 2100. For RCP4.5 the MAVRIC makes a profound difference with the median ice-free date occurring 35 years earlier in 2060. For RCP2.6 there is uncertainty reduction mid-century but the CMIP5 subset before and after the MAVRIC are in good agreement by the end of the century, with projected ice-free dates around 2090.

5 Summary and discussion

5.1 Summary

This study has developed a bias correction methodology for simulations of sea ice thickness (SIT). By constraining CMIP5 simulations with the PIOMAS reanalysis we have demonstrated that:

- GCMs simulate a wide range of SIT in the historical period and exhibit various spatial and temporal biases when compared with the PIOMAS reanalysis. This model uncertainty is the dominant source of uncertainty in CMIP5 future climate projections of SIT.
- The Mean And VaRIance Correction (MAVRIC) technique outlined in this paper significantly reduces the total uncertainty in future projections of SIT out to 2100 by reducing model uncertainty. Correcting both mean and variance of models is found to be critical for improving the robustness of the projections.
- The MAVRIC results in internal variability being the dominant source of uncertainty until 2022, and model uncertainty is dominant thereafter. From mid-century onwards, scenario uncertainty becomes increasingly important and as influential as model uncertainty by 2100.
- The MAVRIC results in projected September ice-free conditions in the Arctic under RCP8.5 occurring up to 10 years earlier (2050s) than without the correction, and with a much narrower uncertainty range, e.g. excluding post 2085 dates.

5.2 Discussion

Without the MAVRIC, the true magnitude of the internal variability and scenario uncertainty in projections of SIT is concealed by the dominant model uncertainty. This demonstrates that time invested in running many ensemble members to sample internal variability in SIT may be more beneficial than running many future emission scenarios for near term projections. These findings implicate that there is room for improvement in GCMs at least for 50 year projections where the scenario differences are negligible. However, for projections at the end of the century, the scenarios become more important.

The MAVRIC bias correction technique developed in this study results in a significant improvement in model simulations of SIT with respect to observations. In future projections, the MAVRIC results in a substantial reduction in uncertainty of SIT, potentially leading to increased confidence in climate projections. As absolute values of SIT are utilised, this reduction in uncertainty potentially has important implications for stakeholder sectors operating in Arctic waters such as shipping. The application of the bias correction results in a 60% reduction in the likely range (16 – 84 percentiles) of sea ice volume in September 2025.

There are a number of caveats to these findings. No attempt is made to constrain the trend in the GCMs. This would be difficult because of the short time scale over which observations are available, raising serious questions about the robustness of calculated historical trends. However future studies could consider this further and assess the feasibility of a trend correction to GCMs. In addition, it is important to recognise that PIOMAS, used here as observations, will also have errors. It would be possible to reduce the multiplicative weightings in Eq. (4) to reflect some uncertainty in the historical data constraint. Other temporally and spatially complete sea ice reanalyses could also be used in future to address this issue.

The simulations tend to show an increase in variance as the sea ice thins, before subsequently declining as the thickness approaches zero (Goosse et al., 2009). Blanchard-Wrigglesworth and Bitz (2014) assessed the relationship of this mean state dependant variance in 19 GCMs, including five of the six used in this study, in addition to PIOMAS. They find a relationship between mean thickness variability and mean thickness in models, i.e. models with thicker SIT depict more variable SIT. In the 19 GCMs assessed, PIOMAS sits on the trend line for the correlation between mean thickness variability and mean thickness. However, in the developed MAVRIC, the change in variance is decoupled from the applied change to the mean state. This aspect could be further developed, but only by making additional assumptions about future changes in SIT variability.

Studies should make use of the~~is~~ MAVRIC in assessing the impact on potential stakeholders sensitive to SIT and a paper utilising the MAVRIC to investigate the opening of the Arctic sea routes is in preparation. We also ~~intend to~~ make the bias corrected SIT fields freely available online for further investigations. DOI: xxxx http://

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. The BC shown in Eq. (4) contains two terms for the representation of the variance in both observations $\sigma_{\widehat{O_h}}$ and models $\langle\sigma_{\widehat{M_h}}\rangle$. Over the 36 year period of observations the magnitude of the ice loss trend ~~is~~ can be significant. To accurately calculate variances this externally forced trend should first be removed to leave the variance due to internal variability. Here a choice needs to be made about how best to remove the externally forced trend. For the PIOMAS observations we choose to linearly detrend the monthly data. A smoothed detrending was considered, however this might remove longer time scale variability which is undesirable. Using similar reasoning it is possible that the linear detrending is removing some variability on the multi-decadal timescale. This is assumed to be significantly less than variability on smaller timescales, and much of the trend is attributed to be externally forced over the 36 years, hence should not be included as internal variability. The performance of a smoothed detrend was tested in a theoretical framework and resulted in a 10 % loss of accuracy in the standard deviation correction due to describing variance as trend.

The calculation of variance in the models is more complicated due to the fact that there is more than one realisation. It is obvious that the required variance should be calculated from the individual ensemble members rather than the ensemble mean. The variance should be calculated in each ensemble member and then the mean taken. There is another choice to make, i.e. whether each ensemble member should be detrended with its own trend, or should the ensemble mean trend be used? We propose that the ensemble mean trend should be used as this is the models response to the changes in forcings. The model detrended ensemble mean standard deviation, $\langle\sigma_{\widehat{M_h}}\rangle$, was calculated by calculating the detrended ensemble variances, then taking the square root of their mean.

The running mean for the future model correction term $\langle\widetilde{M}\rangle$ is calculated over an 11 year period of the ensemble mean, this window hence starts at 1975 for the historical calculations. The chosen period must be long enough to adequately smooth the time series, whilst still being able to capture variations in the sea ice decline trend. This was also tested and found to outperform a 21 year period.

Appendix B Partitioning sources of uncertainty

The sources of uncertainty in Sect. 4.4, Figs. 8 and 9 are calculated for each decadal period (2005 – 2014, 2015 – 2024, etc.) separately as follows. Three ensemble members from each of the six GCMs are utilised for three different emission scenarios (RCP2.6, 4.5, and 8.5). This results in each decade having 6(GCMs) × 3(ensemble members) × 3(scenarios) × 10(years) = 540(fields).

- The total uncertainty is the variance calculated across all 540 fields.
- The internal variability is calculated similarly to the total variability except instead of the absolute values the anomalies from the models' decadal-mean ensemble-mean for each scenario are used.
- To calculate the model uncertainty, each of the six models' decadal-mean ensemble-mean is calculated, resulting in six fields. The variance is then calculated across these six fields, and repeated for all three scenarios separately (to eliminate differential model dependent responses to the different emission scenarios). The model uncertainty is the square root of the mean of these three fields.
- The scenario uncertainty is calculated in a similar way. For each model, each of the three scenarios decadal-mean ensemble-means are calculated resulting in three (scenario-dependant) decadal-mean ensemble-means for each of the six models. The variance is then calculated through these three scenario mean fields for each of the six models, resulting in six fields of the variance in each model. The square root of the mean of the six models scenario uncertainty is the scenario uncertainty.

To create Fig. 8b and c it is assumed that the total variance (total uncertainty, T^2) is the sum of the variance due to model uncertainty (M^2), internal variability (I^2), and scenario uncertainty (S^2), formally:

$$T^2 = M^2 + I^2 + S^2 \tag{B1}$$

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.

Appendix C Additional MAVRIC performance analysis

Formatted: Heading 5,Appendix

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 Eq. (C1):

Formatted Table

$$RMSE = \sqrt{\frac{1}{18}\sum_{n=1}^{18}(E_n - \overline{E_{15}})^2} \quad \text{(C1)}$$

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 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.

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 Eq (C2):

$$SE = \frac{E_n - \overline{E_{15}}}{\sigma_{15}} \quad \text{(C1)}$$

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.

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.

Author contribution

N. M., K. H., and E. H. designed the methodology and experiments.

N.M. developed the code, and performed the experiments.

N. M., K. H., and E. H. wrote the manuscript.

Acknowledgements

We thank Dr Steffen Tietsche for the conversion of the PIOMAS data, Prof. Daniel Feltham and Prof. Ellie Highwood for comments on a pre-submission draft. We thank Referees Prof. Gregory Flato and Dr Francois Massonnet for their quick responses and thorough and constructive reviews. We thank Dr Robert Darby and the University of Reading Research Data Archive for facilitating the hosting of the MAVRIC data set. All statistical analyses and figures were accomplished using the R language and environment for statistical computing and graphics. For more information, see http://www.r-project.org/.

N. M. and E. H. are funded by the APPOSITE project (grant NE/I029447/1), funded by the UK Natural Environment Research Council as part of the Arctic Research Programme. E. H. is also funded by NERC Fellowship. K. H. is partly funded by the National Centre for Earth Observation NCEO.

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.

Boe, J., Hall, A., and Qu, X.: September sea-ice cover in the Arctic Ocean projected to vanish by 2100, Nat. Geosci, 2, 341-343, doi: 10.1038/ngeo467, 2009.

Christensen, J. H., Boberg, F., Christensen, O. B., and Lucas-Picher, P.: On the need for bias correction of regional climate change projections of temperature and precipitation, Geophys. Res. Lett., 35, L20709, doi: 10.1029/2008gl035694, 2008.

Day, J. J., Hargreaves, J. C., Annan, J. D., and Abe-Ouchi, A.: Sources of multi-decadal variability in Arctic sea ice extent, Environ. Res. Lett., 7, 034011, doi: 10.1088/1748-9326/7/3/034011, 2012.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 741–866, 2013.

Francis, J. A. and Vavrus, S. J.: Evidence linking Arctic amplification to extreme weather in mid-latitudes, Geophys. Res. Lett., 39, L06801, doi: 10.1029/2012gl051000, 2012.

Fuss, S., Canadell, J. G., Peters, G. P., Tavoni, M., Andrew, R. M., Ciais, P., Jackson, R. B., Jones, C. D., Kraxner, F., Nakicenovic, N., Le Quere, C., Raupach, M. R., Sharifi, A., Smith, P., and Yamagata, Y.: Betting on negative emissions, Nat Clim Change, 4, 850-853, doi: 10.1038/nclimate2392, 2014.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., and Vertenstein, M.: The community climate system model version 4, J. Clim., 24, 4973-4991, 2011.

Goosse, H., Arzel, O., Bitz, C. M., de Montety, A., and Vancoppenolle, M.: Increased variability of the Arctic summer ice extent in a warmer climate, Geophys. Res. Lett., 36, L23702, doi: 10.1029/2009gl040546, 2009.

Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, Bull. Am. Meteorol. Soc., 90, 1095-1107, doi: 10.1175/2009bams2607.1, 2009.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in projections of regional precipitation change, Clim. Dyn., 37, 407-418, doi: 10.1007/s00382-010-0810-6, 2011.

Ho, C. K., Stephenson, D. B., Collins, M., Ferro, C. A. T., and Brown, S. J.: Calibration Strategies: A Source of Additional Uncertainty in Climate Change Projections, Bull. Am. Meteorol. Soc., 93, 21-26, doi: 10.1175/2011bams3110.1, 2011.

Jungclaus, J., Keenlyside, N., Botzet, M., Haak, H., Luo, J.-J., Latif, M., Marotzke, J., Mikolajewicz, U., and Roeckner, E.: Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM, J. Clim., 19, 3952-3972, 2006.

Kay, J. E., Holland, M. M., and Jahn, A.: Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world, Geophys. Res. Lett., 38, L15708, doi: 10.1029/2011gl048008, 2011.

Kwok, R., Cunningham, G. F., Wensnahan, M., Rigor, I., Zwally, H. J., and Yi, D.: Thinning and volume loss of the Arctic Ocean sea ice cover: 2003-2008, J. Geophys. Res. Oceans, 114, C07005, doi: 10.1029/2009jc005312, 2009.

Laxon, S. W., Giles, K. A., Ridout, A. L., Wingham, D. J., Willatt, R., Cullen, R., Kwok, R., Schweiger, A., Zhang, J., Haas, C., Hendricks, S., Krishfield, R., Kurtz, N., Farrell, S., and Davidson, M.: CryoSat-2 estimates of Arctic sea ice thickness and volume, Geophys. Res. Lett., 40, 732-737, doi: 10.1002/Grl.50193, 2013.

Lindsay, R., W., Haas, C., Hendricks, S., Hunkeler, P., Kurtz, N., Paden, J., Panzer, B., Sonntag, J., Yungel, J., and Zhang, J.: Seasonal forecasts of Arctic sea ice initialized with observations of ice thickness, Geophys. Res. Lett., 39, L21502, doi: 10.1029/2012gl053576, 2012.

Lindsay, R., Wensnahan, M., Schweiger, A., and Zhang, J.: Evaluation of Seven Different Atmospheric Reanalysis Products in the Arctic*, J. Clim., 27, 2588-2606, doi: 10.1175/jcli-d-13-00014.1, 2014.

Lindsay, R. W. and Zhang, J.: Assimilation of Ice Concentration in an Ice–Ocean Model, J. Atmos. Oceanic Technol., 23, L21502, doi: 10.1029/2012gl053576, 2006.

Mahlstein, I. and Knutti, R.: September Arctic sea ice predicted to disappear near 2°C global warming above present, J. Geophys. Res. Atmos., 117, D06104, doi: 10.1029/2011jd016709, 2012.

Maslanik, J. A., Fowler, C., Stroeve, J., Drobot, S., Zwally, J., Yi, D., and Emery, W.: A younger, thinner Arctic ice cover: increased potential for rapid, extensive sea-ice loss, Geophys. Res. Lett., 34, L24501, doi: 10.1029/2007gl032043, 2007.

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C. M., Philippon-Berthier, G., Holland, M. M., and Barriat, P. Y.: Constraining projections of summer Arctic sea ice, The Cryosphere, 6, 1383-1394, doi: 10.5194/tc-6-1383-2012, 2012.

Meehl, G. A., Washington, W. M., Arblaster, J. M., Hu, A., Teng, H., Kay, J. E., Gettelman, A., Lawrence, D. M., Sanderson, B. M., and Strand, W. G.: Climate change projections in CESM1 (CAM5) compared to CCSM4, J. Clim., 26, 6287-6308, 2013.

Overland, J. E. and Wang, M.: When will the summer Arctic be nearly sea ice free?, Geophys. Res. Lett., 40, 2097-2101, doi: 10.1002/grl.50316, 2013.

Rotstayn, L., Jeffrey, S., Collier, M., Dravitzki, S., Hirst, A., Syktus, J., and Wong, K.: Aerosol-and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: a study using single-forcing climate simulations, Atmos. Chem. Phys, 12, 6377-6404, doi: 10.5194/acp-12-6377-2012, 2012.

Schweiger, A., Lindsay, R., Zhang, J., Steele, M., Stern, H., and Kwok, R.: Uncertainty in modeled Arctic sea ice volume, J. Geophys. Res. Oceans, 116, doi: 10.1029/2011jc007084, 2011.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., and Rahimi, M.: Changes in climate extremes and their impacts on the natural physical environment. In: Managing the risks of extreme events and disasters to advance climate change adaptation, edited by: Field, C. B., Barros, V., Stocker, T. F., Qin, D., Dokken, D. J., Ebi, K. L., Mastrandrea, M. D., Mach, K. J., Plattner, G. K., K., A. S., Tignor, M., and M., M. P., A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), Cambridge University Press, Cambridge, UK, and New York, NY, USA, 109-230, 2012.

Smith, L. C. and Stephenson, S. R.: New Trans-Arctic shipping routes navigable by midcentury, Proc. Natl. Acad. Sci. U.S.A., 110, E1191–E1195, doi: 10.1073/pnas.1214212110, 2013.

Sou, T. and Flato, G.: Sea Ice in the Canadian Arctic Archipelago: Modeling the Past (1950–2004) and the Future (2041–60), J. Clim., 22, 2181-2198, doi: 10.1175/2008jcli2335.1, 2009.

Stephenson, S., Smith, L., Brigham, L., and Agnew, J.: Projected 21st-century changes to Arctic marine access, Clim. Change, 118, 885-899, doi: 10.1007/s10584-012-0685-0, 2013.

Stroeve, J., Barrett, A., Serreze, M., and Schweiger, A.: Using records from submarine, aircraft and satellite to evaluate climate model simulations of Arctic sea ice thickness, The Cryosphere, 8, 1839-1845, doi: 10.5194/tc-8-1839-2014, 2014.

Swart, N. C., Fyfe, J. C., Hawkins, E., Kay, J. E., and Jahn, A.: Influence of internal variability on Arctic sea-ice trends, Nat Clim Change, 5, 86-89, doi: 10.1038/nclimate2483

2015.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bull. Am. Meteorol. Soc., 93, 485-498, doi: 10.1175/Bams-D-11-00094.1, 2012.

The HadGEM2 Development Team, Martin, G. M., Bellouin, N., Collins, W. J., Culverwell, I. D., Halloran, P. R., Hardiman, S. C., Hinton, T. J., Jones, C. D., McDonald, R. E., McLaren, A. J., O'Connor, F. M., Roberts, M. J., Rodriquez, J. M., Woodward, S., Best, M. J., Books, M. E., Brown, A. R., Butchart, N., Dearden, C., Derbyshire, S. H., Dharssi, I., Doutriaux-Boucher, M., Edwards, J. M., Falloon, P. D., Gedney, N., Grey, L. J., Hewitt, H. T., Hobson, M., Huddleston, M. R., Huges, J., Ineson, S., Ingram, W. J., James, P. M., Johns, T. C., Johnson, C. E., Jones, A., Jones, C. P., Joshi, M. M., Keen, A. B., Liddicoat, S., Lock, A. P., Maidens, A. V., Manners, J. C., Milton, S. F., Rae, J. G. L., Ridley, J. K., Sellar, A., Senior, C. A., Totterdell, I. J., Verhoef, A., Vidale, P. L., and Wiltshire, A.: The HadGEM2 family of Met Office Unified Model climate configurations, Geosci. Model Dev., 4, 723-757, doi: 10.5194/gmd-4-723-2011, 2011.

Tilling, R. L., Ridout, A., Shepherd, A., and Wingham, D. J.: Increased Arctic sea ice volume after anomalously low melting in 2013, Nat. Geosci, 8, 643-646, doi: 10.1038/ngeo2489, 2015.

Transport Canada: Arctic Ice Regime Shipping System (AIRSS). Transport Canada (Ed.), Ottawa, 1998.

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., and Lamarque, J.-F.: The representative concentration pathways: an overview, Clim. Change, 109, 5-31, doi: 10.1007/s10584-011-0148-z, 2011.

Vrac, M. and Friederichs, P.: Multivariate—Intervariable, Spatial, and Temporal—Bias Correction, J. Clim., 28, 218-237, doi: 10.1175/jcli-d-14-00059.1, 2014.

Watanabe, M., Suzuki, T., O'ishi, R., Komuro, Y., Watanabe, S., Emori, S., Takemura, T., Chikira, M., Ogura, T., and Sekiguchi, M.: Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity, J. Clim., 23, 6312-6335, doi: 10.1175/2010jcli3679.1, 2010.

Watanabe, S., Kanae, S., Seto, S., Yeh, P. J. F., Hirabayashi, Y., and Oki, T.: Intercomparison of bias-correction methods for monthly temperature and precipitation simulated by multiple climate models, J. Geophys. Res. Atmos., 117, doi: 10.1029/2012jd018192, 2012.

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.

Zwally, H. J., Schutz, B., Abdalati, W., Abshire, J., Bentley, C., Brenner, A., Bufton, J., Dezio, J., Hancock, D., Harding, D., Herring, T., Minster, B., Quinn, K., Palm, S., Spinhirne, J., and Thomas, R.: ICESat's laser measurements of polar ice, atmosphere, ocean, and land, Journal of Geodynamics, 34, 405-445, doi: 10.1016/S0264-3707(02)00042-X, 2002.

Zygmuntowska, M., Rampal, P., Ivanova, N., and Smedsrud, L. H.: Uncertainties in Arctic sea ice thickness and volume: new estimates and implications for trends, The Cryosphere, 8, 705-720, doi: 10.5194/tc-8-705-2014, 2014.

Table 1. List of models used: the CMIP5 subset and observations.

Institution	Model name	Ensemble members*
Commonwealth Scientific and Industrial Research Organisation (CSIRO)	CSIRO Mark version 3.6.0: ***CSIRO-Mk3.6.0*** (Rotstayn et al., 2012)	10
Met Office Hadley Centre	Hadley Centre Global Environment Model version 2-Earth System: ***HadGEM2-ES*** (The HadGEM2 Development Team et al., 2011)	4
National Center for Atmospheric Research	Community Climate System Model, version 4: ***CCSM4*** (Gent et al., 2011)	6
National Center for Atmospheric Research	Community Earth System Model, Community Atmosphere Model, version 5: ***CESM1-CAM5*** (Meehl et al., 2013)	3
Model for Interdisciplinary Research on Climate (MIROC)	MIROC version 5: ***MIROC5*** (Watanabe et al., 2010)	3
Max Plank Institute for Meteorology (MPI)	MPI Earth System Model, low resolution: ***MPI-ESM-LR*** (Jungclaus et al., 2006)	3
Applied Physics Laboratory (University of Washington)	Pan-Arctic Ice Ocean Modelling and Assimilation System: ***PIOMAS***** (Zhang and Rothrock, 2003)	1

*multi-model statistics are calculated (Sect. 4.3 onwards) using the first 3 ensemble members.

**used as observations.

Table 2. Notation key

Notation	Description
M	Model
O_h	Observations
x_h	x over the historical period (1979 – 2014)
$\bar{x}$	Time mean of x over historical period
$\langle x \rangle$	Ensemble mean of x
$\tilde{x}$	Running time mean (11 years) of x
$\hat{x}$	Temporally detrended x over the historical period
σ	Standard deviation

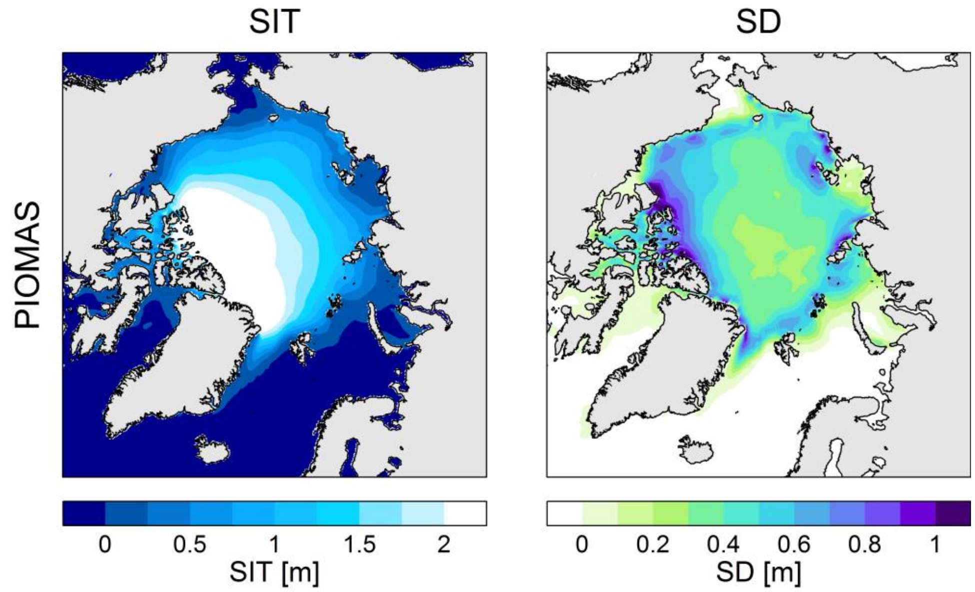


Figure 1. September 1979 – 2014 mean SIT and standard deviation (SD) from the PIOMAS reanalysis. SD is calculated after removing the linear trend.

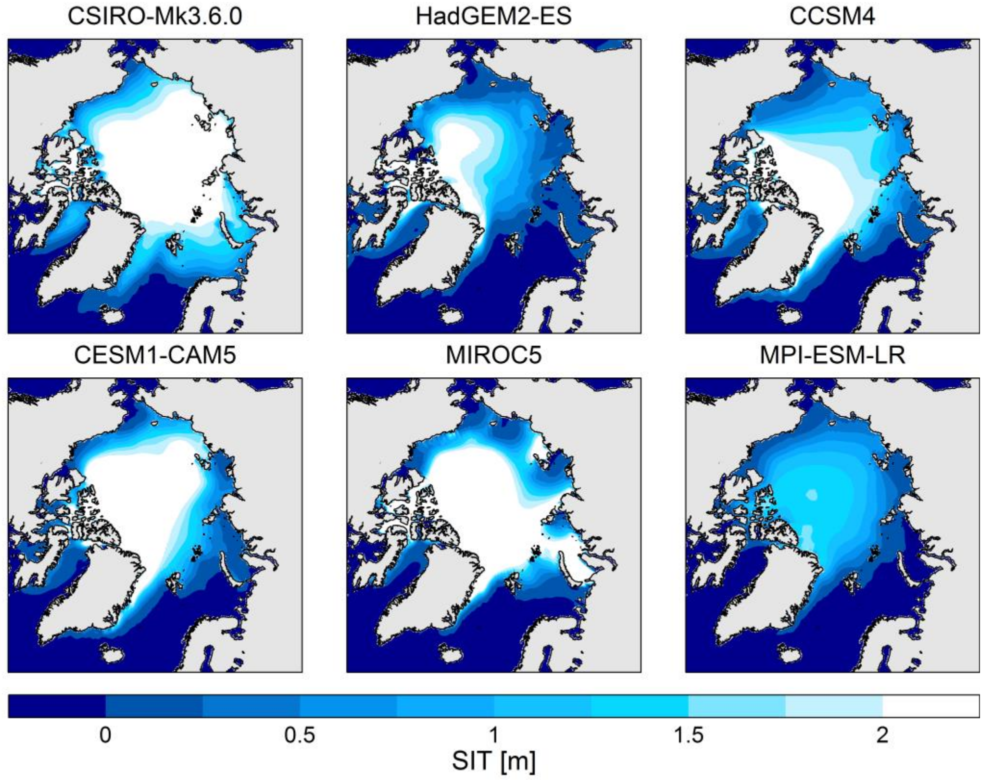
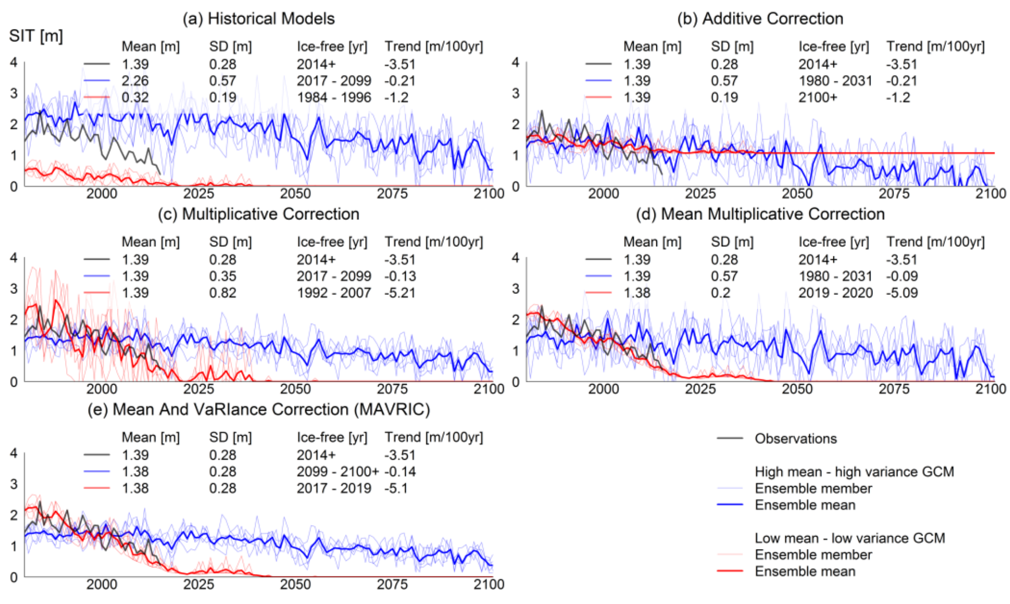


Figure 2. Mean September SIT for each of the six GCMs considered, averaged over the period 1979 – 2014.



Comment [NM2]: Added units to y-axis

Figure 3. Performance of different SIT BCs for one particular month at a hypothetical grid point in a toy model. Mean, SD (detrended) and trend legend statistics are calculated over the observation period (1979 - 2014). 'Ice-free' is defined as the first occurrence of any ensemble member below 0.15 m. Shown is the ice-free ensemble range, i.e. the year of the first ensemble member to be ice-free to the last ensemble member to be ice-free. The black line represents 'observations', the blue and red lines represent high and low ice models respectively. The thin coloured lines represent ensemble members, and the thick lines are the ensemble mean.

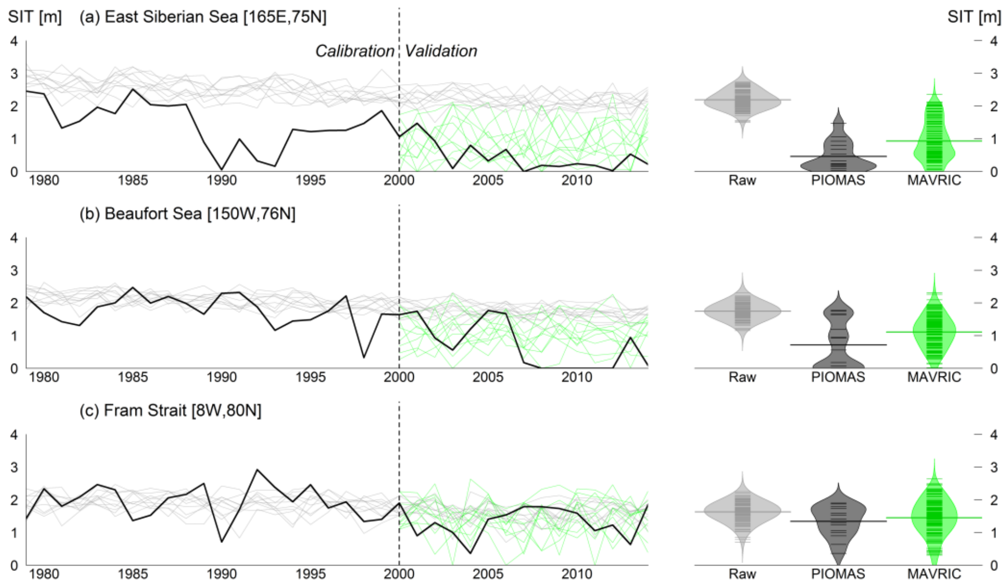


Figure 4. September SIT at three grid point locations in the Arctic, from PIOMAS (black) and CSIRO-Mk3.6.0 historical (1979 – 2005) and RCP8.5 (2006 – 2014) raw output (grey) and post MAVRIC (green). The raw CSIRO ensembles (grey) are bias corrected via the MAVRIC using the PIOMAS observations (black) over the calibration window, producing the MAVRIC ensembles (green) for the validation window. Beanplots (right) show the distribution of the SIT for the validation period. Small horizontal lines show every SIT value, the frequency of which is illustrated by the width of the shaded region. Thick horizontal line is the mean.

Comment [NM3]: Added [m] to y-axis “Corrected” changed to “MAVRIC”

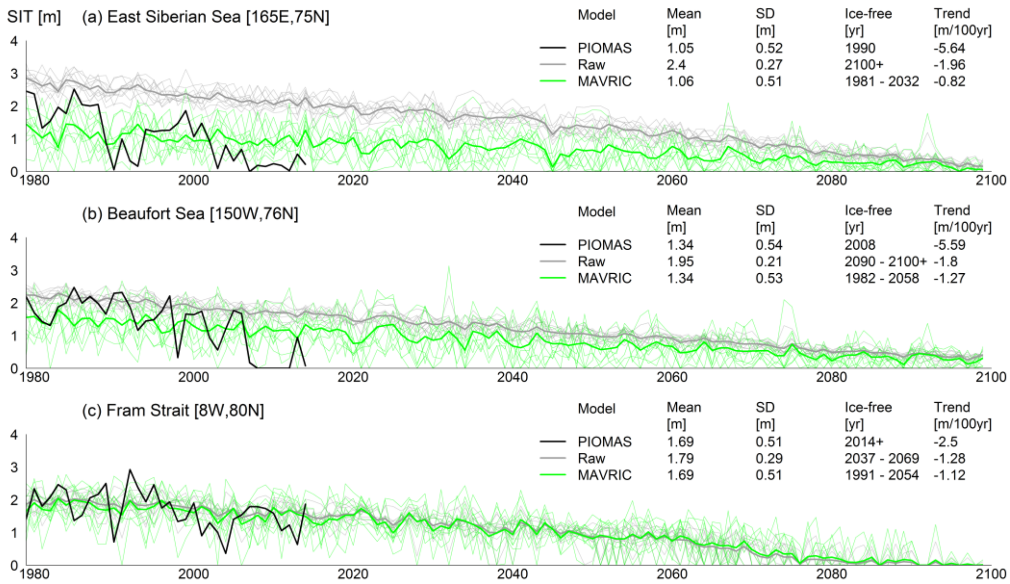


Figure 5. September SIT at three grid point locations in the Arctic, from PIOMAS (black) and CSIRO-Mk3.6.0 historical (1979 – 2005) and RCP8.5 (2006 – 2100) raw output (grey) and post MAVRIC (green). Thin lines are individual ensemble members, thick lines are the ensemble means. Mean, SD and trend legend statistics calculated over the period of observations (1979 – 2014). The SD is the detrended mean ensemble SD. Ice-free is the range of the first occurrence of the first and last ensemble member below 0.15 m.

Comment [NM4]: "Corrected" changed to "MAVRIC"

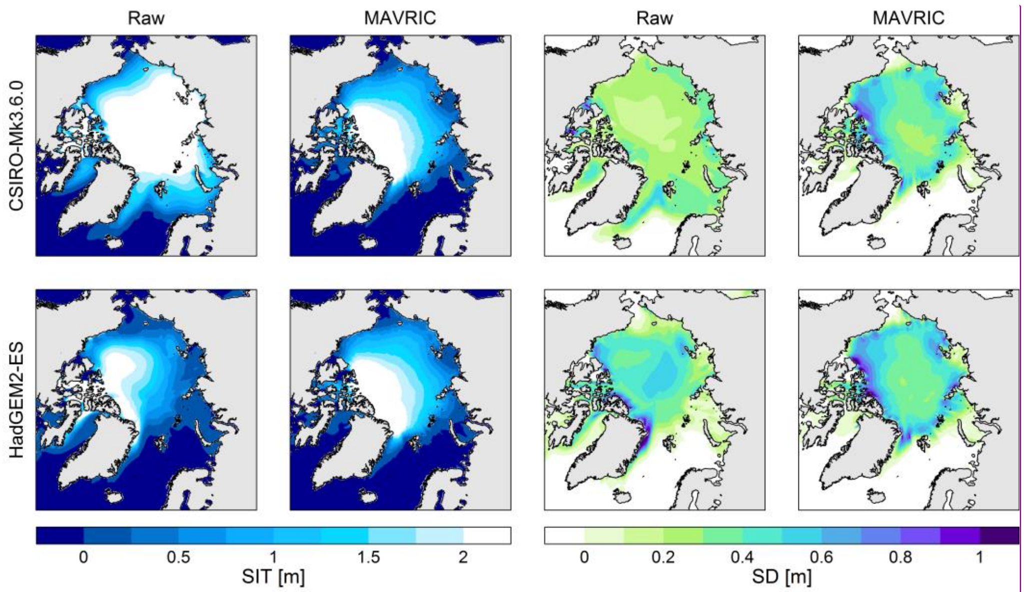


Figure 6. CSIRO-Mk3.6.0 and HadGEM2-ES, September 1979 – 2014 ensemble mean SIT and SD (detrended). The raw columns are the model solutions as found in the CMIP5 archive. The corrected columns show the distribution after the MAVRIC has been applied. PIOMAS SIT fields shown in Fig 1.

Comment [NM5]: "Corrected" changed to MAVRIC

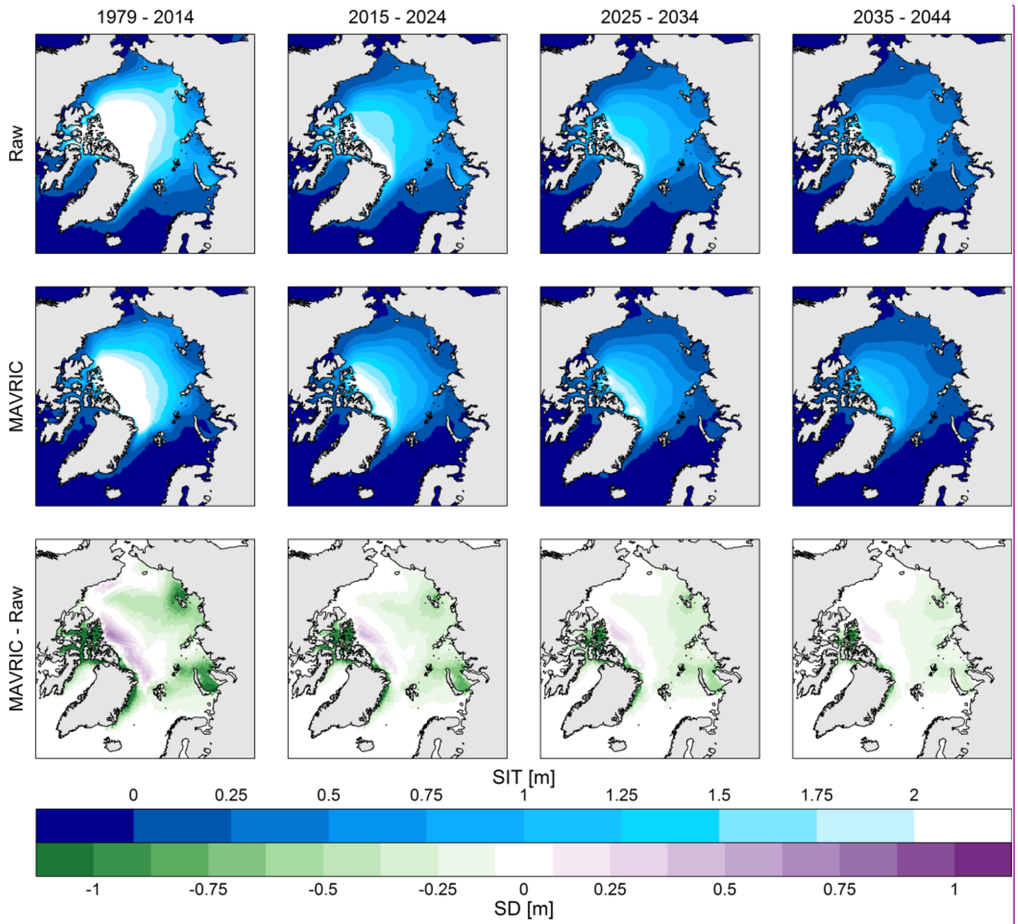


Figure 7. September multi-model ensemble mean (three members from each model) mean SIT from the CMIP5 subset, using the raw data (top row) and after MAVRIC (~~bottom~~ middle row). ~~The multi-model ensemble mean (three members from each model) is shown.~~ The bottom row shows (MAVRIC – Raw) and hence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

Comment [NM6]: Added bottom row

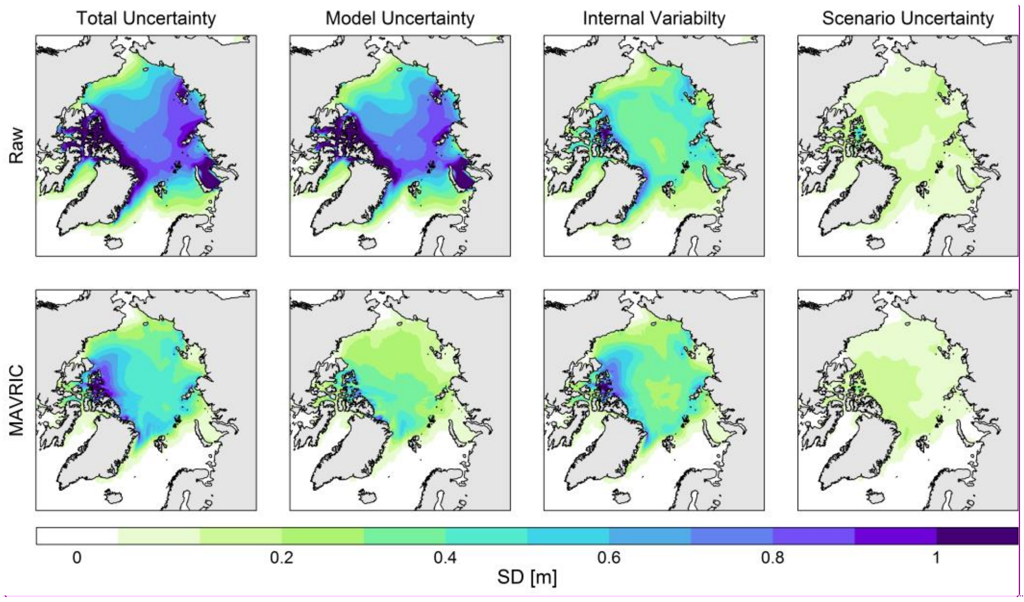


Figure 8. September 2015-2024 sources of SIT uncertainty from the CMIP5 subset (SD of the detrended SIT). The multi-model ensemble mean (three members from each) is shown when comparing raw (top row) and after MAVRIC (bottom row).

Comment [NM7]: Replaced "Corrected" with "MAVRIC"

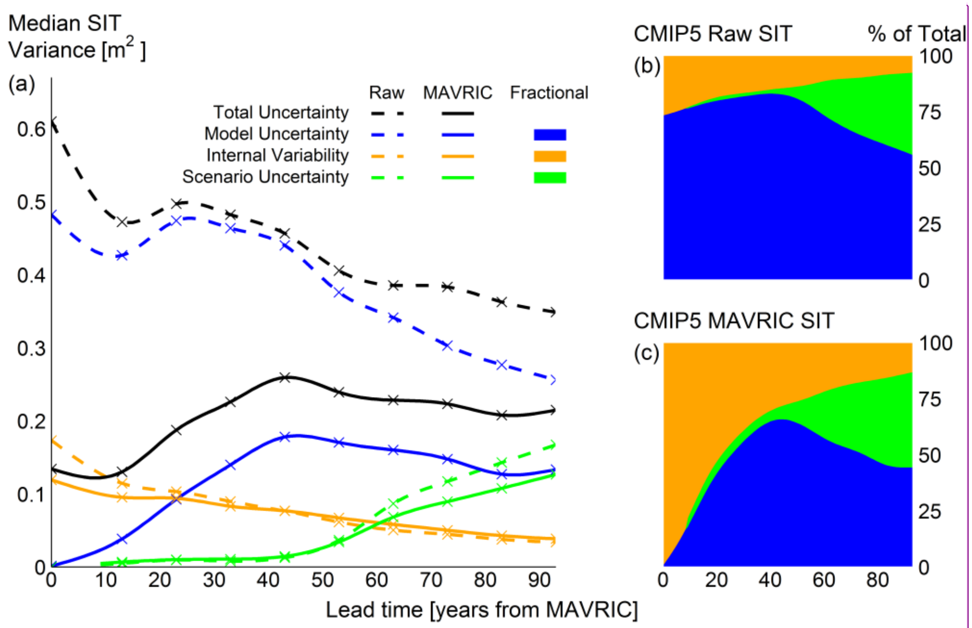


Figure 9. The evolution of the sources of September SIT uncertainty in the CMIP5 sub-set with lead time. Year zero is the MAVRIC window mid-point (1997) and the emission scenarios (RCPs) start in 2006. Panel a shows the change in magnitude of the different sources of uncertainty. The uncertainty shown is the median SIT variance and hence the lines scale additively. The dashed lines are for the raw model output and solid lines are for post MAVRIC. Contributions of model uncertainty, internal variability and scenario uncertainty as a fraction of total uncertainty are shown for the raw output (b) and post MAVRIC (c).

Comment [NM8]: Replaced "Uncertainty" with "Variance" and "Corrected" with "MAVRIC"

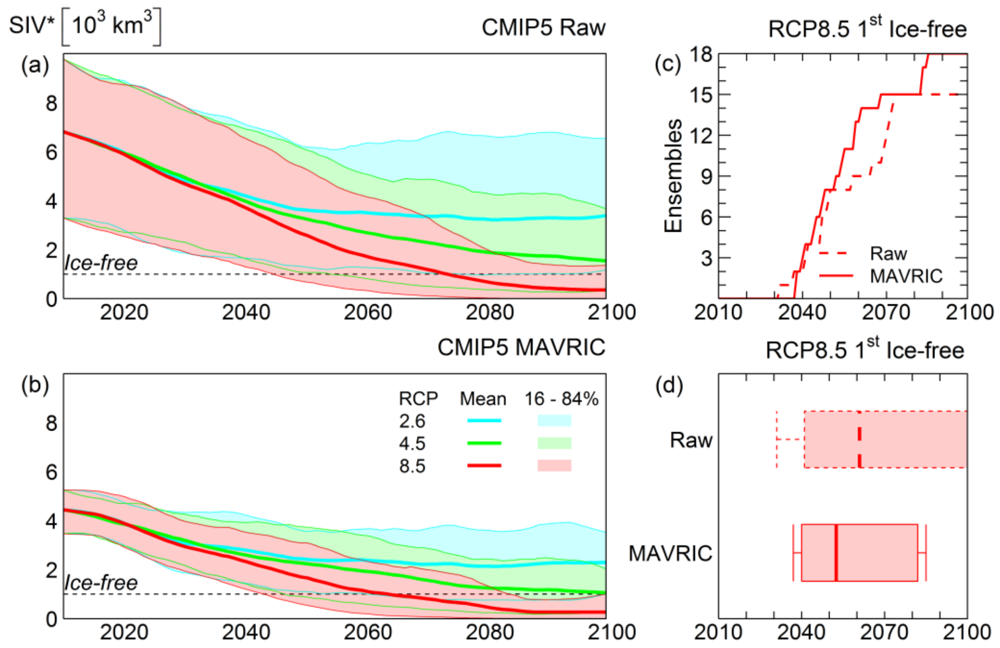


Figure 10. CMIP5 subset sea ice volume (SIV*) projections and first ice-free conditions. Panels a and b show the projected SIV* from all six models (18 ensemble members total) in both the raw and corrected GCMs (11 year running mean), and shaded regions are the 16th – 84th percentiles. Panel c shows the number of ensemble members having passed the ice-free threshold. Panel d shows the statistics of c, with the whiskers representing the range (1st and 18th ensemble member ice-free), the box capturing the 16th – 84th percentiles, and the bold line showing the median (9th ensemble member). Ice-free is defined as the first year the pan-Arctic SIV* dips below 1~~2~~ $\times 10^3$ km^3 for a particular ensemble member. *Volume (SIV*) is calculated using a constant~~100~~ 50 % ~~sea ice concentration (~~SIC~~)~~ throughout.

Comment [NM9]: Updated for 50% SIC

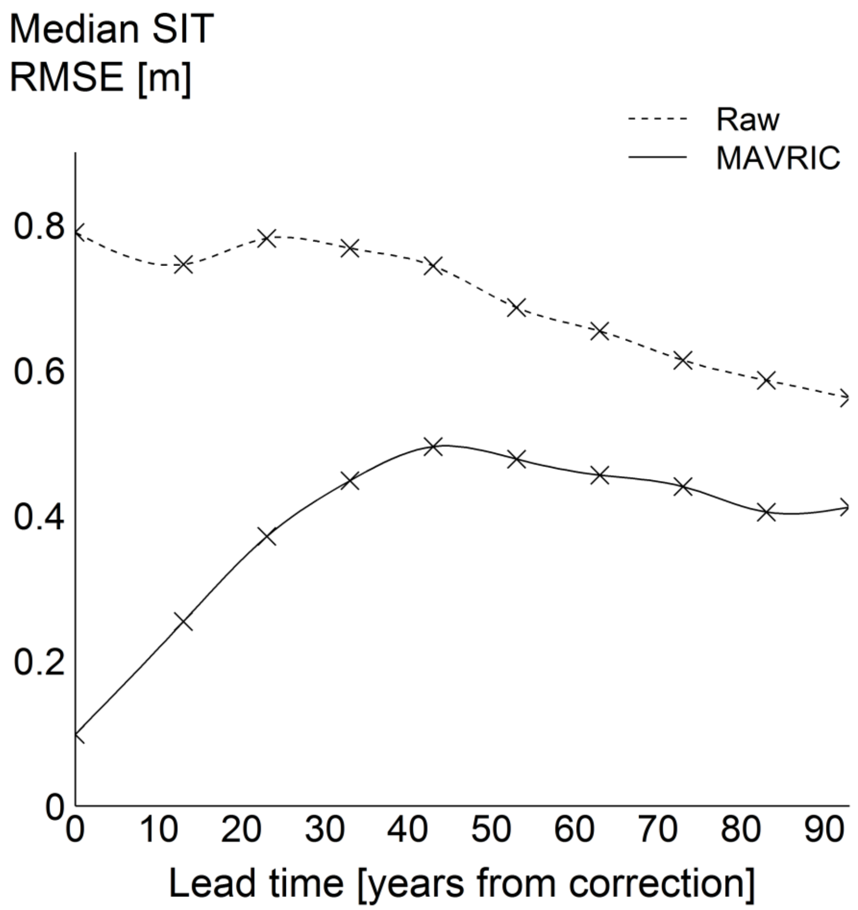


Figure C1. Multi-model ensemble out of sample September median SIT RMSE

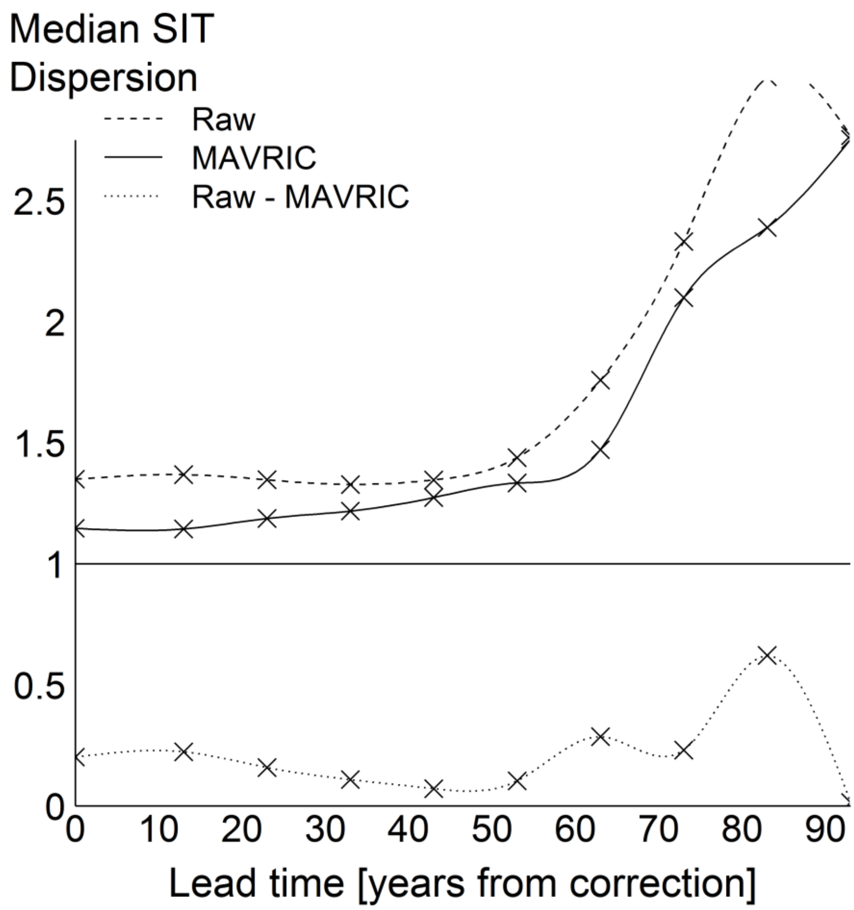


Figure C2. Multi-model ensemble out of sample September median SIT dispersion