Interactive comment on "Consistent biases in Antarctic sea ice concentration simulated by climate models" by Lettie A. Roach et al.

F. Massonnet (Referee)
francois.massonnet@uclouvain.be

Received and published: 22 August 2017

	Reviewer Comment	Author Response	Author intended action
1	"The paper essentially consists in two parts, each conveying its own message: (1) CMIP5 models have consistent biases in the simulation of SIC regimes, (2) The prescribed ice floe diameter affects the simulated SIC regimes The hypothesis of the authors is that (1) is largely caused by (2), hence motivating the development of specific lateral melt processes. However, while I appreciate the efforts of the authors to work on parts (1) and (2) individually, my impression is that they are not able to close the loop and directly prove that the deficiency in simulated SIC regimes in CMIP5 is caused by lateral melt parameterizations."	Our hypothesis is that an under-estimation of lateral melt may be a contributor to the biases noted in the low-concentration regime, but cannot account for biases in the high concentration regime. We have undertaken modelling simulations that support this hypothesis. It was not our intention to suggest that the poor quantification of lateral melt in models is the only possible cause of the biases presented. It almost certainly isn't. We have perturbed that lateral melt in one model sea ice model to show that we can affect a change of the right sign, but have not explored other plausible causes. The language used in the manuscript may have not made this clear enough. We therefore intend to check all wording to clarify that we are suggesting a possible influence, rather than claiming a dominant cause.	Check all wording to clarify that we are suggesting a possible influence, rather than claiming a dominant cause eg. P1, L8 and P8, L13: replace 'partially determined' with 'partially influenced.' Include citations referencing other possible explanations (see author intended action #3). See author response #2.
2	"In fact, several CMIP5 models do not have such a lateral melting term in their formulation." "To test formally whether (2) explains (1), the authors should separate the CMIP5 subset into two groups: those who use a lateral melting term, and those who don't. If the two sub-groups simulate SIC regimes differently in a statistical sense, then this represents a major step forward to confirming the initial hypothesis of the authors."	We thank the reviewer for this suggestion and we will begin our analysis in the second part of the paper by including such a figure. We did not expect to see a very large difference between models which do not include lateral melt and models which do, given that no model includes a representation of floe diameter. All models must parametrize the impact of lateral melt in some way - otherwise there would be no change in sea ice concentration due to melt – but it may be included explicitly or implicitly. However, using Table 1 to divide the models into two groups (and using the additional models that we can incorporate by using monthly data), we find significant differences in sea ice concentration regimes between the two groups. There is a clear tendency to overestimate the fraction of low concentration ice in models without the explicit lateral melt term. The inclusion of this analysis, together with our modelling simulations which artificially enhance lateral melt, strengthens our hypothesis.	Include a figure showing the sea ice concentration distribution from models with and without an explicit lateral melt term and discuss the outcome.

3	"In addition, there may be many processes, including dynamical ones, that could explain (1). An important paper on the topic (Lecomte et al., 2016; https://doi.org/10.1016/j.ocemod.2016.08.001) should be cited."	We completely agree with this point and will cite the recommended paper. The recommended paper is consistent with our suggestion that model physics (including dynamical processes) are not currently suitable for the low-concentration regime.	P9, L12: Modify final two sentences to '*Note that we have tested only the impact of lateral melt on this bias. A number of other physical processes, including dynamical ones, may also contribute. Lecomte et al. (2016) find systematic wind-driven biases in sea ice drift speed and direction at the exterior of the Antarctic ice pack. Errors in surface winds could contribute to poor simulation of low-concentration sea ice. However, we find a very strong over-estimation in low-concentration sea ice in the NEMO-CICE model, which is forced by a reanalysis atmosphere and so should not have very unrealistic winds. The dynamical response of sea ice to winds at the edge of the ice may be poorly represented, as we would expect sea ice dynamics to be floe-size dependent. Alternative rheologies (such as a granular rheology (Feltham et al., 2005)) may be better suited to this domain. Concentrations could also be reduced by mechanical interactions between floes. However, we cannot test the impact of such floe-size dependent processes without access to sea ice models that include them*.' P11, L17. Insert *'Given the possible contribution of dynamic processes to model biases in the sea ice concentration distribution, a full exploration of sea ice dynamics for all CMIP5 models using the sea ice concentration budget decomposition of Uotila et al. (2014) would be welcome*.'
4	"In summary, I'm confident that (2) has a significant effect on the simulation of SIC regimes, but I'm challenging the idea that the CMIP5 ensemble is appropriate to test this hypothesis, and that (2) is the dominant cause for (1)."	See author response and intended actions #1 and #3. The CMIP5 ensemble is the only way we can test this hypothesis until models with floe size information are available. The results presented here motivates developing such models so that the hypothesis can be tested more robustly	See author intended actions #1 and #3.
5	First, I appreciate the use of multiple observational datasets in the analyses. This is becoming a standard in the community, and this study is fully in line with that approach. What I find worrying is that the three observations are averaged to make a "super observation". This is a problem when non-linear metrics such as the Integrated Ice Area Error (IIAE) are used, because we lose the sense of uncertainty. If the IIAE was conducted separately on each observational reference first and then reported three times, then the impact of observational uncertainty would be immediately visible. Here, instead, the observational mean is used as a reference giving only one IIAE (e.g., Fig. 3).	Although there is precedent for averaging observations and using their mean as a reference, as the true value is not known (eg. Ivanova et al., 2014), we agree that averaging three observations is unsatisfactory. This occurred in the first draft only in the calculation of the integrated ice area error (IIAE) and in Fig. 8 to show a concentration difference. P3, L16 was misleading in suggesting that we did this throughout the analysis. Calculating the IIAE for each observational data set is a good suggestion and we are happy to do this. Calculating three separate IIAEs means that we cannot show original Fig. 1 in order of increasing IIAE, as the ordering differs slightly according to the observational product used. We could instead show the original Fig. 1 plots ordered alphabetically and	Remove P3, L16. Remove IIAE from original Fig. 1 Add an additional figure showing the IIAE for each model relative to each observational data set Show ice errors relative to each observational product separately in original Fig. 3 Show the sea ice concentration difference relative to each observational product separately in original Fig. 8

	Information regarding observational uncertainty is lost.	add an additional figure which plots the IIAE for each model from each observational product. This may allow us to separate models into 'good', 'middle' and 'poor' categories, without an overall ordering. For the original Fig. 3, we propose showing the ice errors for models relative to each observational product separately. Preliminary analysis suggests that this makes very little difference to the conclusions drawn from this figure. For the original Fig. 8, we propose showing the SIC difference relative to each observational product. Again, this makes very little difference to the conclusions drawn from this figure. For the original Fig. 9, we propose showing the SIC difference relative to one of the observational products. Preliminary analysis suggests that this makes very little difference to the conclusions drawn from this figure - all three observational products show that very few models simulate high enough concentrations in the Weddell Sea.	Show the sea ice concentration difference relative to one observational product in original Fig. 9
6	For other figures (e.g., Fig. 2) it would be good to display all three observation products individually in order to gauge how much of the variability is due to the product, and how much is due to sub-seasonal and interannual fluctuations.	We show observational products individually in the original Fig. 4, to show variability due to the product separately from time variability for the normalized SIC distribution. These show that variability in observational products is more significant than interannual fluctuations in the high-concentration regime throughout the year, This a good suggestion for the original Fig. 2. We propose adding additional panels, with one panel showing boxplots for each model and observational product individually. Preliminary versions of these plots show that the lower to upper quartile ranges for each model (representing only inter-annual variability) are generally smaller than inter-model differences. They also show that the variability due to differences in observational product in sea ice area is more significant at the winter maximum than the summer minimum. The adjacent panel would show the population of all models and the population all observations, as in the original draft.	Add additional panels to Fig. 2 to show all observational products and models individually, as well as the synthesized populations, and discuss the results in the text
7	Second, I sometimes have the impression that the metrics used are overly complicated. The best example is in Fig. 5. There are so many levels of processing that it becomes difficult to understand	Since the Discussion paper was submitted, from personal communication with D. Notz we have found that our 'binned fractional sea ice extent' is calculated in exactly the same way as the 'normalized sea ice concentration distribution' in Notz (2015).	Replace 'fractional binned sea ice extent' with 'normalized sea ice concentration distribution' throughout text and figures.

	the meaning of the metrics shown on the figure (see also my comment on Fig. 5 below). For the "binned" diagrams (Figs 4-5-6), why not use only three classes: 0-15%, 15-90%, 90-100%? This would simplify the figures much without removing the useful information that deficiencies in the simulation of SIC happen at the edges of the distribution. In general, I have the impression that I could not replicate the figures if I had the original data.	When we read Notz (2015), we were unsure how he normalized the distribution, but it is normalized by grid cell area, as we have done. We therefore intend to adopt the name 'normalized sea ice concentration' from Notz (2015), instead of 'fractional binned sea ice extent.' We hope that is more intuitive for the reader than our original name for this metric. We would prefer to keep the bins at 10% spacing as this is consistent with Notz (2015) and gives more of a sense of the whole distribution. Also see response to comments #20 and #30 below.	Refer to Notz (2015) in the Metrics section and say that we use the same approach. See also author response #20 and #30 below.
8	Third, the authors introduce new metrics without real justification and make inconsistent choices. For example, they adapt the Goessling et al. "Integrated ice edge error" by moving from a "extent-like" definition to an "area-like" definition. This is initially a good idea, since area is a more physical measure than extent. However, the decomposition into an "absolute error" and a "misplacement error" is not as straightforward, as I'm showing now: consider a model A with uniform SIC of 80% in a number of grid cells, compared to an observational reference with 70% of ice at the same points. Consider also model B, which has half as much ice as A (40% in the grid cells). Even though ice has not been misplaced, the misplacement error term will increase when going from A to B! (While it won't if the Goessling definition was followed). One of the ideas of the Goessling's approach is precisely to use a threshold at 15% to be able to separate total and misplaced errors. The area-like version of that metric looses that property.	We agree that the misplacement area error does not have a physical meaning and are happy to remove it from the paper. This does not affect any of the conclusions of our work. We also agree that the integrated ice area error is a good physical measure, so will continue to use this. We believe that presenting the decomposition of an integrated ice extent error into misplacement extent error and absolute extent error using the Goessling et al. (2016) approach is useful here, and propose showing them in the updated version of Fig. 3.	Remove misplacement area error from paper. Adapt Fig. 3 to show (a) The integrated ice area error (IIAE), (b) the integrated ice extent error (IIEE), (c) the absolute extent error as a fraction of the IIEE, and (d) the misplacement extent error as a fraction of the IIEE. (These will be shown relative to each observational product separately, see author intended action #5).
9	I also don't understand why the authors compute per-bin sea ice extent, and not sea ice area in Figs. 4-5-6. Sticking to ice area would be a more natural choice given the adaptation made earlier of the IIAE.	See response to comment #7 above. We believe that consistency with Notz (2015) is preferable, as it allows comparison with his results for the Arctic.	
10	Finally, I don't fully understand why the authors used daily data in their analyses. This restricts the number of CMIP5 models available, and adds	We used daily data as some aspects of marginal ice zone behaviour, which is more variable than compact ice, may be visible in the daily data and not in the monthly data. Sea ice models do simulate	Conduct analysis using CMIP5 monthly output and monthly output from the Bootstrap and NASA Team observations (via the Climate Data Record archive). ASI-SSMI observations are only available as daily

	considerable variability to the metrics. By design, CMIP5 models are not supposed to capture synoptic variability in sea ice extent. Using monthly output would partially average out this variability, and would better allow to exhibit the significant biases of the CMIP5 models as the signal-to-noise ratio would increase.	synoptic scale variability in sea ice extent, but it is certainly not a focus of development within CMIP5 class models We agree with the reviewer's point that using monthly data rather than daily data has the advantage of allowing us to include more models in our inter-comparison. It also allows us to extend the time period of the analysis. We will switch the paper to using monthly mean data for the CMIP5 models. Initial analysis suggests that using monthly data and more models makes no substantial difference to the conclusions drawn from our model-observation comparisons. This strengthens the robustness of results.	output; we will average the sea ice concentration fields each month (using Climate Data Operators, `cdo monavg`). Update these methodological details in the text
11	As I'm writing below in a comment, error bars are so large that it is easy to play the devil's advocate and claim that observational references and CMIP5 models are, in the end, not so inconsistent.	We do not show error bars in this paper. We present the data as box plots, indicating the, median, the interquartile range (IQR) and whiskers indicating 1.5 times the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR. Box plots have the advantage of being non-parametric and can indicate the degree of dispersion and skewness in the data. We have currently suggested that where we see lower quartile – upper quartile ranges that do not overlap, the populations being compared are significantly different. There are a number of statistical tests that could be applied to examine the differences. We have applied t-tests, but this is a not a particularly useful test as it only considers whether two distributions have different means. Using the test of overlapping confidence intervals as the reviewer hints at would require the assumption of normality and only pass if 95% of the data were different in the two samples. For normalized SIC distributions, which are bounded on the interval [0,1], this is a very difficult test. After further thought on how to robustly assess inconsistency, we propose including the Kolmogorov–Smirnov test as a robust statistical test of whether the distributions are different. The K-S test is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples. Preliminary analysis shows that all model – observation comparisons fail the K-S test at the 95% confidence level. However, the p-value obtained from the K-S test, which represents the confidence that the two populations come from the same distribution, is a useful tool to quantify the degree of disagreement. For example, in the original Fig. 5 (a), the p-values from the K-S test	Include further discussion of statistical tests in the text. Include results from the Kolmogorov–Smirnov test for all two-population comparisons, eg. as annotated text on original Figures 2 and 5, and on the comparison between standard NEMO-CICE and NEMO-CICE with a reduced floe diameter on original Figures 6 and 7. Discuss these results in the text.

		comparing models to observations in each bin are lowest for the 90-100% and the 10-20% bins. This allows us to make the objective conclusion that model-observation difference is most significant for low-concentration and high-concentration sea ice. Finally, it's worth noting that our main aim, and what we find most interesting about the paper, is the identification of consistent behaviours amongst the CMIP5 models, ie. tendencies in one particular direction – rather than any claim of simple model-observation disagreement.	
12	p. 2, l. 3: The statement that CMIP5 has improved in the simulation of Arctic sea ice compared to CMIP3 is strong, and perhaps too strong. Rosenblum and Eisenman (2016, http://dx.doi.org/10.1175/JCLI-D-16-0391.s1), for example, suggest that this could be due to the omission of volcanic forcings in several CMIP3 models. Also, it is unclear if the CMIP5 to CMIP3 differences actually reflect improvements, or changes in tuning strategy (e.g. Notz, 2015, http://dx.doi.org/10.1098/rsta.2014.0164). Please nuance this statement.	We agree with this comment	Replace '*Advances in Earth system modelling have improved simulation of Arctic sea ice compared to the previous intercomparison project (CMIP3) (Stroeve et al., 2012)*' with '*Advances in Earth system modelling have somewhat improved simulation of Arctic sea ice compared to the previous intercomparison project (CMIP3) (Stroeve et al., 2012), although this may reflect changes in forcings (Rosenblum & Eisenman, 2016) or tuning strategy (Notz, 2015) rather than changes in model physics*.'
13	p. 2, l. 8: evaluates –> evaluate.	Agree	p. 2, l. 8: evaluates –> evaluate.
14	p. 2, l. 28: ... Phase 5 (CMIP5; Taylor et al., 2012).	Agree	p. 2, l. 28: ... Phase 5 (CMIP5; Taylor et al., 2012).
15	p. 2, l. 30: Years prior to 2000 are neglected in the model evaluation. Have the authors still conducted the evaluation on 1979-2000, for which the CMIP5 output and the observations are readily available? Do conclusions of the study hold? Please elaborate. It would be useful to present the results of such analyses in the supplementary material, to test the stability of the metric developed over time.	Our use of daily data meant that were limited (in memory) in our analysis. Use of monthly data means that we can consider a longer time series, and is indeed a more sensible solution. The ASI-SSMI observations begin in 1992, so we plan to do analysis over 1992-2014. Initial analysis suggests that including a longer timeframe of data does not alter the conclusions.	Conduct analysis using model output and observational data from 1992-2014. Swap the order of Subsections 2.1 and 2.2 in Methods. Add a sentence to stay that ASI-SSMI observations begin in 1992, so we conduct analysis over 1992-2014.
16	p. 2., l. 31: Could the authors conduct their diagnostics on two members of the same climate model, in order to gauge how much internal variability affects the metrics developed in this study?	It is correct that we have only used one ensemble member from each model at this point. We do not believe this is problematic for the results of the paper. But we agree that the reviewer has proposed an interesting question, and as part of revising the manuscript we will do an analysis of ensemble members from one model to see how much spread is attributable to internal variability. We could use the K-S test to quantify this difference.	Will undertake analysis and comment on, or include, result.
17	p. 3., l. 3: Three observational products were used in the study. Since part of the study describes the differences between sea ice concentration in those products, it would be useful to have a few lines	Agree	Replace P3, L1-14 with: *'Passive microwave radiometers deployed on satellites measure the brightness temperature of the Earth's surface, and can be used to calculate sea ice concentration. Various observational data sets apply different algorithms to convert passive-*

	describing differences in algorithms between those products.		*microwave signals into sea ice concentration, reflecting the uncertainty in satellite observations (Bunzel et al., 2016). As summarized by Ivanova et al. (2014), differences between algorithms are caused by 1. choice of radiometer channels; 2. tie-points, which are the brightness temperatures used to identify different surfaces; 3. sensitivities to changes in physical temperature of the surface; and 4. weather filters, which correct for atmospheric effects falsely indicating the presence of sea ice.* *To account for some of this product uncertainty, we use three observational data sets: the Bootstrap algorithm (Comiso, 1986), the NASA Team algorithm (Cavalieri et al., 1984) and the ASI algorithm (Kaleschke et al., 2001; Spreen et al., 2008). We do not consider datasets that merge different observation methodologies. Bootstrap uses cluster analysis of brightness temperatures from two channels (19 GHz and 37 GHz vertical polarization in the Antarctic), applies an ocean mask and is available from 1979 at a resolution of 25 km. NASA Team uses ratios of brightness temperatures (which tends to cancel out physical temperature effects) from three channels (19 GHz in the vertical and horizontal, 37 GHz in the vertical), removes weather contamination based on certain spectral gradient ratios and is available from 1979 at a resolution of 25 km. The ASI algorithm uses the difference in brightness temperatures between horizontal and vertical polarization at 85 GHz, uses lower frequency channels at lower resolution to filter atmospheric effects (which are more apparent at 85 GHz than lower frequencies), and is available from 1992 at a resolution of 12 km. We choose to conduct our analysis over 1992-2014.* *Differences between the three selected data sets are large: in the Antarctic, the NASA Team algorithm shows the marginal ice zone (defined as the extent of sea ice with concentration between 15 % and 80 %) to extend over 2 million km more than the Bootstrap algorithm (Stroeve et al., 2016). NASA Team is more sensitive to clouds and wind over open water than the Bootstrap mode (Anderson et al., 2006), while the high-frequency ASI algorithm is also sensitive to such atmospheric effects (Spreen et al., 2008). Bootstrap is more sensitive to physical temperature changes than NASA Team, and may underestimate concentrations at low temperatures, such as near the Antarctic coast (Comiso et al. 1997). For low concentrations, atmospheric effects, which generally lead to falsely increased sea ice, become increasingly important (Anderson et al., 2006). The weather filters/ocean masks used to correct these differ between the different algorithms.'*

18	p. 3., l. 5-7: To what season does the statement on marginal ice zone area difference apply? I think it's winter, please specify.	It's winter, September and October specifically	Insert 'in the winter months' in P3, L7
19	p. 3., l. 22: "Sea ice area is the sum of the area of all grid cells with more than 15 % sea ice concentration multiplied by the sea ice concentration in each grid cell". Following the conventional NSIDC definition I would have thought that area is just the product of ice concentration by grid cell area, summed over the domain (https://nsidc.org/cryosphere/seaice/data/terminology.html). Why only considering the grid cells with > 15% of ice?	Other locations on the NSIDC website show the definition we used in the original draft. For example see: https://nsidc.org/data/docs/noaa/g02135_seaice_index/#comp_area where it states, 'The monthly average sea ice area calculation is performed through simple pixel-by-pixel arithmetic of multiplying the daily concentration by the size of the grid cell[1], for all grid cells which satisfy the 15 percent threshold and then averaging them together for a month'. Also consider: http://nsidc.org/arcticseaicenews/faq/#area_extent which states "Area takes the percentages of sea ice within data cells and adds them up to report how much of the Arctic is covered by ice; area typically uses a threshold of 15%." We note that Notz (2015) and Turner et al. (2017) do not use a 15% cutoff, while Ivanova (2016) does use a 15% cutoff. We conclude that either definition is acceptable, as long as it is stated clearly. We also must be consistent - a 15% cutoff should be used for the IIAE if it used for the SIA. As we use the IIEE and associated MEE and AEE, which by the Goessling definition use a 15% cutoff, we think that the most consistent approach is to use the 15% cutoff.	
20	p. 3., l. 25: Why is the bin (0-10%] not in the list?	As shown in Ivanova et al. (2016) (Fig. 2d), the CMIP5 multi-model mean and the NASA Team observations have a high fraction of ice below 10% sea ice concentration in the summer. We find that the fraction of <10%-concentration ice varies in the models from 0.005 to 1.0 (when models are essentially ice-free) in the summer. It consists of up to around a third of the ice in other seasons for some models. Including these very low concentrations therefore heavily skews the normalized SIC distribution towards low concentrations. It obscures behaviour at higher concentrations. Our aim is to look for consistent model behaviour; with such variance between different models and between different observations at very low concentrations, it's difficult to conclude anything about model tendencies.	Update Fig. 1 to show 0.1-10% sea ice concentration. Explain that some models (refer to Fig.1 and Ivanova et al., 2016) have very large numbers of cells with very small concentrations (0.1-10%), noting that the fraction varies greatly between models, and substantially between the three observational products. Including these very low concentrations heavily skews the normalized distribution, so we exclude them from the SIC distributions.
21	p. 4., l. 20: "A disadvantage of the IIAE is that it does not take into account the observational range, using	See response to comment #5.	See response to comment #5.

	only the observational mean as the 'true' state". That's not really a disadvantage of the metric, but rather a methodological issue. The authors could repeat the IIAE taking successively the three observations as references. They would obtain three IIAE's, which would give a sense on the uncertainty associated to the products. Why didn't the authors go this way? See also my first comment on methodology.		
22	p. 6, l. 17. "[Ocean-sea ice] Model output is analysed on its native grid". Does that mean that the observations were then interpolated onto the NEMO-CICE grid? At page 9, line 18 and in Fig. 8 the NEMO-CICE model is evaluated using the IIAE metric, this means that at some point an interpolation must take place, correct? The model output and the observational reference need to be on the same grid for Eqs (2) and (3) to be evaluated. Why didn't the authors interpolate the NEMO-CICE output on the same target grid as all CMIP5 models, to ensure consistent analysis?	Following comments by other reviewers (anonymous review #2 and C Holmes), as well discussion with others in the community, we have thought more carefully about the interpolation. We believe it is preferable to avoid interpolation as much as possible. This is possible for sea ice area and the normalized sea ice concentration distributions, which we thus propose to calculate on original grids. Preliminary analysis suggests that model-observation differences in the normalized sea ice concentration distributions at low concentrations are slightly reduced when conducting the analysis on the native grids. We intend to state that the normalized sea ice concentration distributions show some sensitivity to grid interpolation in the Metrics subsection. Integrated ice errors and sea ice concentration differences between models and observations must be calculated on the same grid. In this case, we propose interpolating onto a regular 1 degree grid using Climate Data Operators bilinear interpolation function and state this in the manuscript. This may cause some smoothing of the ice edge. This will have a negligible effect on integrated ice errors and sea ice concentration differences at high concentrations (original Fig. 9). It may impact sea ice concentration differences at low concentrations (original Fig. 8). We shall investigate whether using bilinear or nearest-neighbour interpolation results in differences in the original Fig. 8.	Calculate sea ice area and normalized sea ice concentration distributions on original grids and state this in the manuscript. State in Subsec. 2.3 that there is some sensitivity to grid interpolation for the SIC distributions. Calculate integrated ice area/extent errors and sea ice concentration differences after interpolation onto a regular 1 degree grid using an appropriate Climate Data Operators (https://code.mpimet.mpg.de/projects/cdo/) interpolation function and state this in the manuscript.
23	p. 6, l. 20 and Fig. 2. "Sea ice area at the annual minimum is consistently biased low". Here I'm playing the devil's advocate. The blue boxes in Fig. 2 display the distribution of the three observational references, which are three times the same climate realization plus noise due to the retrieval algorithm. Hence these blue boxes embody time-variability and product variability. By contrast, the green boxes in	See response to comment #11. In summary, to more robustly assess inconsistency of distributions we will include the Kolmogorov–Smirnov test as a robust statistical test of whether the distributions are different. The K-S test is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.	See author intended action #11. See author intended action #40. See author intended action #6

	Fig. 2 contain time-variability internal variability, and model error. So, the whole question is whether these observational references are incompatible in a statistical sense with the models. Put differently, could the observations be the (N + 1)th CMIP model? Judging from Fig. 2b, the observations lie in the range [1.5 * IQ_75%, IQ_75%] and they could be one of the CMIP models. Or couldn't they? A more quantitative test would be welcome.	Also see response to comment #6. Separating boxplots would allow us to discuss the contributions of model/observational product variability and time variability	
24	p. 7, l. 5-8. The sentence "Such total errors..." is unclear. First, there are many more possible reasons than just ocean/atmosphere temperature biases to explain differences in total area (which is captured by the AAE). Second, it's not clear why looking at the bias per concentration bin would help isolate the role of the sea ice component.	The case we wanted to make is that sea ice extent will be different in a model with a normal ocean compared to one that is 1 degree warmer everywhere. However, both models could still stimulate an appropriate normalized sea ice concentration distribution. Therefore we would argue that a normalized sea ice concentration distribution depends less on overall ocean/atmosphere temperature biases than sea ice extent.	Replace P7, L5-8 with: '*We now consider sea ice concentration distributions from observations and models, which provide a more detailed assessment than hemisphere-integrated measures. A normalized sea ice concentration distribution may help isolate the role of the sea ice component, as models with a constant temperature bias in the atmosphere or ocean, resulting in a biased sea ice area or extent, may still simulate the relative fraction of different concentration regimes successfully.*'
25	p. 7, l. 10. Notz (2014) "uses" or "used" but not "use".	Agree	Correct to Notz (2014) "uses"
26	p. 7, l. 22-35. The two paragraphs deliver somewhat contradictory messages. The first one finishes by "that the sea ice components of CMIP5 models are somewhat successful at simulating the distribution of sea ice concentration" while the second paragraph says "large discrepancies between models and observations in the highest and especially the lowest concentration bins". This would need better rephrasing, saying for example in the first paragraph that the "big pictures" are consistent but that this is mostly thanks to cancellation of errors, as explained in the second paragraph.	We agree with the reviewer and will reword accordingly. The use of the K-S test to quantify the differences between the two population is useful here. As explained above, the p-value from the K-S test, which represents the confidence that the two populations come from the same distribution, is highest for the 90-100% and 10-20% bins in DJF.	Reword and include the outcome of the K-S test in the presentation of these results
27	p. 10, l. 4-6 and Fig. 9. The authors conclude that the systematic underestimation of highly concentrated ice in the Weddell Sea is related to melt or break-up processes. Why is the possibility of a systematic misrepresentation of dynamics ruled out? It could be that all models have deficiencies in capturing the Weddell gyre dynamics. It could be that models are neutral to divergent while observations are in convergent motion. I haven't tested this hypothesis myself, but I don't have enough information from the results of the paper to rule out properly this alternative hypothesis. An exploration of how the	We agree that dynamic processes are a possible cause of the underestimation of highly concentrated ice. The findings of Lecomte et al. (2016) are particularly relevant here. They suggest that models with high ice drift speeds in coastal areas simulate a faster sea ice retreat. These high drift speeds may be influenced by sea ice rheology as well as wind speeds.	Discuss the possible contribution of sea ice dynamics to this bias, with reference to Lecomte et al. (2016). Also see author intended action #3

	models simulate Antarctic ice drift could be helpful in that respect.		
28	p. 10, l. 29: 2. Three observational products are considered in this study. While I appreciate this effort, it looks sometimes like the authors assume that observational errors are random and that the mean of all three products is representative of the truth. It could be that the three observational products have a systematic bias with respect to the truth, which could explain model-obs mismatch on top of model error. The authors don't seem to explore this possibility in the assessment. For example it is known that most algorithms underestimate sea ice concentration as ice becomes very thin. Could this explain the model-obs differences, in particular differences in binned sea ice extent? It is also known that wet snow has a brightness temperature that makes sea ice concentration retrievals higher than they should be. Could this have an impact? More discussion on observational systematic errors would be welcome, in order to place the CMIP results in perspective.	As discussed above, we plan to do as suggested and avoid use of any observational mean in this study. We agree that including different observational products will give some estimation of error arising from differences in processing satellite data, but there is still the possibility of systematic errors common to all three observational products. In the Discussion, we did briefly discuss systematic errors in the observations: *'Accounting for the observational range, we find that models overestimate the extent of low-concentration sea ice throughout the year, while underestimating the extent of high-concentration sea ice in summer. This common behaviour across diverse models with varying physics is a result not previously highlighted and warrants further attention. We note that using the observational range as an uncertainty estimate neglects biases that are common to the three different satellite observations. As mentioned above, satellite observations of sea ice are most uncertain in summer. However, we see the bias in low concentration ice from CMIP5 models throughout the year, and observed summer high-concentration ice is unlikely to be affected by the melt processes that complicate satellite retrievals. The suggestion that the NASA Team algorithm overestimates low-concentration ice (Steffen & Schweigher, 1991) would further strengthen the contrast between models and observations in this regime.'* We agree that this is worthy of more comprehensive discussion within the paper and will expand on this discussion point.	Add the following paragraph to subsection 2.2: *Besides structural uncertainty in observational algorithms, systematic biases common to all three products are possible. Lack of validation data (Ivanova et al., 2014) mean it is difficult to quantify this, but accuracy is understood to be lower in the presence of melt ponds or other surface melt effects (Ivanova et al., 2014), which may act to lower retrieved concentrations; large fractions of thin ice (Cavalieri, 1995); and stormy conditions near low concentrations (Anderson et al., 2006). Transitions between ice type can cause differences in emissivity (Grenfell and Comiso, 1986), but because models do not simulate ice types such as grease ice, this issue should not impact model-observation comparisons.* Add to the discussion: *As mentioned above, sea ice concentrations are considered to be most uncertain during melt conditions, for large fractions of thin ice and at low concentrations during storms. In the context of the results from the model-observation comparison for normalized sea ice concentration distributions, we suggest that the impact of uncertainty of melt conditions is limited as the high bias in low-concentration ice from CMIP5 models is visible throughout the year. The low bias in high-concentration ice during the melt season would be strengthened if observations were underestimating ice concentrations in this season. Inclusion of both NASA Team and Bootstrap algorithms, with the former tending to cancel out physical temperature effects, will sample some of this uncertainty. The underestimation of sea ice concentrations in areas of thin ice (<35 cm) (Ivanova et al., 2015) may cause a bias at any concentration in the observed normalized sea ice concentration distribution from observations, with the possibility of a positive bias in the very lowest concentrations. Stormy conditions near the ice edge lead to false sea ice concentrations near the ice edge; weather filters may accurately remove these, leave them uncorrected (Anderson et al., 2006) or erroneously remove real sea ice. The latter may underestimate low concentrations (personal communication, S. Kern). Spreen et al. (2008) suggest the filter method used in ASI observations may result in a positive bias in the marginal ice zone, and Steffen & Schweiger (1991) found that the NASA Team algorithm overestimates low-concentration ice when compared to Landsat imagery. Considering all this evidence we suggest that the magnitude or sign of any systematic biases in satellite radiometer observations is unclear when comparing with climate models. This is particularly true for low concentrations. Here the*

			use of different approaches to weather filters within the different algorithms may assist in sampling observational uncertainty. Development of sea ice satellite emulators, which use climate model output to calculate brightness temperatures (eg. Tonboe et al., 2011), may help to reduce uncertainty when comparing models to observations in the future.
29	Fig. 1. Interestingly, it is possible that two CMIP5 models with similar IIAEs (e.g., MIROC5 (4.6 Mkm2) and FGOALS-g2 (4.85 Mkm2)) have drastically different sea ice concentration patterns (one with not enough ice and one with way too much ice; panels ac and ad). In the same vein, two models with similar patterns (e.g. CMCC-CMS and HadGEM2-CC) may have very different IIAEs. This is because of the definition of IIAE which penalizes over- and underestimation in the same way. The authors should comment on that aspect (which I see as a weakness of that metric). Although there is no definition of what a "good" metric is, we could expect that it satisfies properties of continuity in some sense: two models close to each other should have similar metric values.	Is there are reason to favour over-estimation or under-estimation? We don't see that one is better than the other, so we don't see the lack of distinction between the two in the IIAE to be an issue.	Explicitly state in the text that the IIAE does not favour over-estimation or under-estimation
30	Fig. 5. This is one example where I would have difficulties in reproducing the result. If I follow correctly, from Fig. 5 caption and from the text: (1) Grid cells are binned according to their concentration (2) The total sea ice extent is computed in the three observational references, in each CMIP5 model, for each day of each year. (3) The extent per bin is normalized by the total extent (for Fig. 5 panels a-d) (4) The normalized extent of each CMIP5 model is compared to the normalized extent of each observation (or to the mean of them?) to give a fractional deviation. I doubt that, out of 10 readers, more than one can replicate Fig. 5 exactly. The authors should detail their approach in a supplementary material, or simplify the metrics.	See response to comment #7 above. Our calculation of fractional binned sea ice extent/normalized sea ice concentration distribution is the same as Notz (2015). We propose explicitly staying the steps involved in the Methods section. We agree that Figs (e-h) are confusing and are happy to take a different approach. The aim of Figs(e-h) in the first draft was to show the biases independent of scale, but we are happy to simply remove (e-h) from the manuscript.	See author intended response #7 above. Add further detail in Methods to explicitly explain the steps used to calculate the normalized sea ice concentration distribution. '*The sea ice concentration distribution for each model or observational product is calculated by binning grid cells according to their concentration at a 10%-spacing. The distribution is then normalized by the area of grid cells*.' Remove Figs(e-h)

Interactive comment on "Consistent biases in Antarctic sea ice concentration simulated by climate models" by Lettie A. Roach et al.
Anonymous Referee #2

Received and published: 23 September 2017

	Reviewer Comment	Author Response	Author intended action
31	Page 2, Line 30: Why limit the analysis to 2000-2014, when longer observational and model timeseries are available? Longer time series would make the analysis more robust.	This comment and the comment below are connected – our use of daily data meant that we were limited (in memory) in our analysis. Use of monthly data means that we can consider a longer time series, and this is indeed a more sensible solution. The ASI-SSMI observations begin in 1992, so we plan to do analysis over 1992-2014. See author response to comment #15 above	See author intended action #15
32	Page 2, line 32: Why is daily sea ice concentration used here? This should be explained, as many more models provide monthly than daily output, and it looks like the authors proceed to average the daily output to seasonal averages.	See author response to comment #10 above	See author intended action #10 above
33	Page 3, Line 15-17: While I agree that one needs to consider the observational uncertainty, I am not convinced that averaging several products is the best way to do that. First of all, they could all have consistent biases, and hence their range still would not account for the observational uncertainty. Secondly, one of them might be a lot better than the others, and so the combined data might be further from the truth than the best one. So while I am not suggesting that the authors perform an evaluation of the three observations, which is best done by the creators of these data sets, I would encourage the authors to add a sentence or two here to highlight the potential shortcomings of this approach they are using.	We agree with your points. Please see author response to comment #5 above regarding averaging observational products. We do combine the three sets of observations in original Fig.s 5-7 for the concentration distributions. It is not clear to us from the literature that any of the three datasets is better than the others. Evaluation of the products is indeed beyond the scope of this manuscript. Further, Ivanova (2014) states that 'we cannot establish an absolute ranking of the performance of the algorithms because of the lack of good validation data,' and recommends constructing an ensemble of different observational products.	See author intended action #5 P3, L15: Replace the final paragraph in original subsection 2.2 with: '*In this study, for some of the analysis we consider the three observational data sets individually. In order to compare the sea ice concentration distribution from the set of models against observations, we create an ensemble of the ASI, Bootstrap and NASA Team observational products. Combining the observational products in this way does have limitations, as different algorithms are likely to perform better for certain sea ice conditions and seasons. However, it is not clear from the literature where exactly the strengths of the various algorithms lie, and evaluation of the different algorithms is beyond the scope of this manuscript. The difficulty in ranking various observational algorithms is noted by Ivanova et al. (2014), due to a lack of validation data. They recommend constructing an ensemble of different observational products*.'
34	Page 3, Line 19: Why was the sea ice output re-gridded, rather than analyzed on the model grids? This can introduce additional errors that have nothing to do with the physics of the model. So there needs to be a good reason to re-grid the model output, otherwise the analysis should be re-done on the original grids. And if the authors have a good reason to do the re-gridding, please include information on how exactly the regridding was done, so it can be replicated by others.	See author response to comment #22 above	See author intended action #22 above

35	Page 3, Line 27: Why are concentrations below 10% not included? Others included them, so please explain why you would not. For loose sea ice, wouldn't it be important to look at below 10%?	See author response to comment #20 above	See author intended action #20 above
36	Page 8, line 15, Table 1: Since the authors have the information on whether and how lateral melt is included in the CMIP5 models, do they find any difference between models that include it or not? That would provide an important argument for the hypothesis of the authors that the too loose sea ice concentration is a result of deficiencies in lateral melt.	See author response to comment #2 above	See author intended action #2 above
37	Page 10, line 24-28: Please remove this entire paragraph. It is pure speculation what modeling centers look at during model development, and this speculation does not add anything to the arguments or results presented in the paper.	Agreed. We are happy to remove this.	Remove lines 24-28
38	Page 10, Line 29: The observational range is not necessarily fully counted for, as discussed earlier. This should be reflected here.	See author response to comment #28 above	See author intended action #28 above Replace P10, L29 'Accounting for the observational range' with 'Accounting for the range in three observational products.'

Interactive comment on "Consistent biases in Antarctic sea ice concentration simulated by climate models" by Lettie A. Roach et al.

C. Holmes
calmes@bas.ac.uk

Received and published: 12 September 2017

	Reviewer Comment	Author Response	Author intended action
39	Figure 1 and throughout: I'm unsure of the relevance of DJF; although the traditional meteorological austral summer season, it's arguably not particularly relevant for sea ice, particularly since you do not link analyses to atmospheric variables. However I recognise it's not obvious what the best season would be. I'd suggest showing the minimum or maximum, or if to use DJF, please give some justification (in particular why summer not winter) and mention any sensitivities to season, if found.	Our interest in summer stems from Fig. 2 (SIA), where we looked at sea ice area from models versus observations and concluded that the minimum showed more disagreement with observations than the maximum. This is further supported by the analysis in Fig.3 (IIAE). We therefore chose to examine the months leading up to the summer minimum (DJF) in more detail. We chose to show DJF, MAM, JJA, SON in original Fig. 5 as we wanted to include data from all months. The normalized SIC distribution for DJF shows largest differences from observations at the high and low ends of the distribution. We propose looking at the low (10-20%) and high (90-100%)	Explore the seasonality of results (for original Fig. 5) or suitably justify our interest in a particular time period (for original Fig. 1 and results from the lateral melt experiments)

		concentrations bins throughout the year, as there are some sensitivities to season. In response to the other reviews, we are significantly updating the figures. Wherever relevant, we will explore the seasonality of processes or suitably justify our interest in a particular time period. For example, for the second part where we investigate the impact of lateral melt, we could calculate the month(s) where the impact of lateral melt is greatest, and show results for this time period	
40	Figure 2: This combines spread in information from different years and from different observational data sets. In particular the conclusion in the main text that there is 'no clear bias' at maximum is a little confusing as climatologies are not shown (I would think of 'biases' as referring to climatologies), and the discussion of this figure in the text is very brief; half a sentence or so. Also panel a) appears to be missing outliers? I suggest separating the panels into separate boxplots, particularly for observations, clarifying the multi-model vs multi-yr distinction (if possible), checking the figure caption, and expanding the discussion of this figure a little (it need not be much)	Fig. 2 shows that the interquartile range of the CMIP5 models overlaps that of the observations for the sea ice area maximum, but it does not for the sea ice area minimum. We conclude that there is a tendency for models to underestimate the sea ice area minimum, but there is not such a significant tendency at the sea ice area maximum. This conclusion can be quantified using the K-S test, as discussed above. We agree with the suggestion of separating out the boxplots, see author response #6. The data in Fig. 2a does not have outliers when the whiskers are set to 1.5 of the interquartile range.	P6, L20 Replace '*While sea ice area at the annual maximum has a large inter-model and inter-annual spread with no clear bias compared to observations, sea ice area at the annual minimum is consistently biased low.*' with discussion of the separated boxplots and use of the K-S test to quantify the degree of difference between models and observations. Replace 'clear bias' with 'significant tendency.' See author intended action #6
41	Methodological note: Please say how the regridding is performed (the method and the package used). I have certainly seen cases myself and at meetings (sorry I cannot bring a citation!) which suggest that it can affect results particularly since you are concerned with distributions rather than aggregate measures. Such methodological details are rarely stated in papers about CMIP5, but for reproducibility it they should be!	The comment on the impact of regridding is correct. See author response to comment #22 above	See author intended action #22