The Cryosphere Discuss.,
https://doi.org/10.5194/tc-2017-131-RC1, 2017

The Cryosphere
Discussions
Open Access

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

Consistent biases in Antarctic sea ice concentration simulated by climate models

submitted to The Cryosphere by Lettie A. Roach, Samuel M. Dean and James A. Renwick

review by F. Massonnet

This study examines how Antarctic sea ice concentration (SIC) is simulated by modern climate models. The assessment goes beyond traditional metrics of model performance (total area/extent) by looking at how SIC regimes (loose/normal/compact ice) are captured. Multiple observational references are used to account for observational uncertainty in the assessment. The authors find that CMIP5 models overestimate the

fraction of loose ice all year long, while they underestimate the fraction of compact ice in summer. They hypothesize that these systematic biases might be due to the treatment of lateral melting. Running an ocean-sea ice simulation only, they investigate the sensitivity of SIC regimes to the assumed, constant floe size in a state-of-the-art sea ice model.

This work is much welcome and very timely as final developments for CMIP6 are taking place. The simulation of Antarctic sea ice by coupled climate models is arguably poorer than that of Arctic sea ice, hence it is very useful to trace back the possible source of model error. It is always instructive to exhibit systematic biases in large ensembles like CMIP5, because they are indicative of serious concerns in the physics of these models.

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. In fact, several CMIP5 models do not have such a lateral melting term in their formulation. 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. 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. 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).

I also have several methodological concerns.

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 loose 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). Information regarding observational uncertainty is lost. 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.

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 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.

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. 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.

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 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. 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.

Other issues and comments.

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.

p. 2, l. 8: evaluates –> evaluate.

p. 2, l. 28: ... Phase 5 (CMIP5; Taylor et al., 2012).

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.

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?

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 describing differences in algorithms between those products.

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.

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?

p. 3., l. 25: Why is the bin (0-10%] not in the list?

p. 4., l. 20: "A disadvantage of the IIAE is that it does not take into account the observational range, using 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.

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?

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 displaythe 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 Fig. 2 contain time-variability internal variabiltiy, 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.

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.

p. 7, l. 10. Notz (2014) "uses" or "used" but not "use".

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.

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 models simulate Antarctic ice drift could be helpful in that respect.

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.

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.

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.

Interactive comment on The Cryosphere Discuss., https://doi.org/10.5194/tc-2017-131, 2017.