Final Author Comments

„Sea-ice extent provides a limited metric of model performance“

by D. Notz

I am very grateful to both referees for their very insightful, constructive, extensive and helpful comments. I am convinced that the present paper will gain substantially in both impact and clarity by taking these comments into account for a revised version.

Below, I have detailed how I will respond to the individual comments made by the two referees.

Anonymous Referee #1

The author is on the trail of finding arguments to assess models using the integrated ice area as well as integrated ice extent. There is a valid point in here that invites open discussion. The paper is too wordy and could make use of equations to define uncertainty terms for brevity and clarity. The main result, which is that use of both extent and area to constrain models (Fig 7) is underplayed whilst other issues relating to passive microwave retrieval inter-comparisons is going over old ground. Unfortunately the paper also includes distractions, such as fig 5, which lead nowhere. Even fig 3 and 4, although having the potential to tell a story, are rather pointless within the paper objectives – leading to just one (rather ambiguous) conclusion. This topic needs to be pursued separately within a modelling context. Finally, there is poor treatment of concepts for uncertainty both in models and observations. This is little discussion in comparison with algorithm or model intercomparison studies e.g. Comiso, Stroeve etc.

These are very helpful comments, which in particular show that the scope of the paper in its current form is broader than possibly implied by both its title and within the abstract. In a revised version, I will hence restructure the paper to make it both more focused and less repetitive compared to existing work. In doing so, I will in particular aim at better taking existing peer-reviewed publications on passive microwave retrieval inter-comparisons into account that make some of the points raised in this paper. My general impression is, however, that within the sea-ice modelling community still rather little awareness exists regarding biases of individual satellite algorithms. This in part reflects the, to my knowledge, still rather small number of papers that specifically address this point within the context of model evaluation. I do not agree with the reviewer that fig 3 and 4 are rather pointless within the context of this paper and will make their relevance to the paper's objective more explicit in a revised version.

Specific points.
Introduction : This contains messy concepts to the purpose of the paper. The issue surely is the transfer function between passive microwave (PM) observations and model. The model ice area is explicitly known (albeit with internal variability), but the observations are not – as it is just an empirical fit. Surely the issue is not to pick a particular algorithm which may be differently biased at different stages of the annual cycle, but to understand the observational error (including regridding, landmask etc). Because we are not in a position to forward model the ice area to PM space we must do the inverse. One does not use PM just to calculate the summer or winter extent

but to evaluate the phase and amplitude of the seasonal cycle (within observational error). Since over ice PM principally detects water, the only reason that Bootstrap may depict a more dense ice pack than NASA-Team is if it is less sensitive to water – it cannot escape detecting meltponds and cannot discriminate them from leads. The ASI algorithm uses the high frequency channels which are by their nature less sensitive to water. In spatial plots of PM discontinuous ice concentrations, which do not reflect real concentration variations, sometimes occur when the Bootstrap algorithm switches between polarization and frequency schemes. This is of course invisible when integrated. The discussion here on the difference between integrated extent and area is valid. However, integrated area does not assist in understanding the processes of albedo, heat and turbulent fluxes etc. For such studies one would do spatial plots of ice concentration or perhaps compare with observations of integrated albedo..

This comment again relates in part to the question of the scope of this paper, which I will specifically address within a revised version. The suggested way of dealing with the underlying source of the uncertainty seems very helpful and I'll restructure the introduction along the lines outlined by the reviewer.

3100:22-24. This statement is unclear as you do go on a lot about comparison between Bootstrap and Team throughout section 3. I suggest removing all such references in section 3 and literally just refer to Bootstrap.

Many existing studies use NASA Team as the sole satellite product for model evaluation. I hence find it useful to show which impact this choice has within this context. Hence, I will instead try to illustrate this point in a more direct way in section 3.

3103:29. No model yet has a prognostic floe-size distribution so it is the characteristic floe size that is prescribed. The total lateral melt may depend on the modelled extent of the MIZ, and vertical distribution of solar heat absorption by the ocean.

I agree that this sentence is distracting and will remove it in a revised version

3104:1-15. Not only do HadGEM2 and CCSM4 have multi-category ice but so do MIROC5 (but not MIROC4h) and NorESM1-M.

This is not a sensible method for discrimination between the models. The ice is just as compact if there is lots of 0% ice as well as 100% ice, so simply using a threshold on the 100% is overly simplistic. A better threshold would be on the gradient between the 90-100% and 80-90% bins. If you say that NASA-Team is wrong then this may provide a limit on the gradient threshold. However, a clear discrimination between models with multi-cat ice would still not be possible as MIROC4h does not have it.

Summer winds are rather light in the Arctic so this is unlikely to have a major effect on the ice distribution. More likely it is the initial ice thickness, which determines the ice strength, that has the major effect. Ice rheology may also have an effect as that determines the model spatial distribution of the ice in your rather limited polar zone.

Ironically to your argument here, many of the loose sea ice cover models are those which show best

agreement with the observed sea ice decline. This implies that they are loose because of some feedback characteristic.

I will look further into underlying processes that might help explain the differences in simulations. Given the reviewer's comment, I might have to conclude that even speculating on the underlying cause of this different behaviour is beyond the scope of this paper, in which case figure 5 would be removed and this discussion would be shortened substantially.

3105:1-15. This is an example of an emergent constraint from climate models and is a useful outcome of this study worth emphasising in the introduction and conclusions. (see Bracegirdle, Thomas J., David B. Stephenson, 2013: On the Robustness of Emergent Constraints Used in Multimodel Climate Change Projections of Arctic Warming.
J. Climate, 26, 669–678. doi: http://dx.doi.org/10.1175/JCLI-D-12-00537.1). Plotting in this fashion is far less confusing than the obtuse figures 5 and 6.

I was not aware of this reference, thanks for pointing this out. Regarding figure 5, this will either be removed or plotted along the lines outlined by reviewer 2 in a revised version. I honestly don't see why figure 6 is characterised as obtuse.

3105:4-6. The number of points inside the 'observed' error box appear to be larger for the 'diffuse' models than the 'concentrated' models!

This is true, but the main point here is that some diffuse models fall outside the error margin for area that still are within the error margin for extent. I will re-write this to make this clearer.

3106:7-26. Thin ice behaves differently than thick ice in the extent-vs-concentration regime. In particular, thin ice is likely in free-drift whereas thick ice still has an internal stress component.

I will include this comment as a possible explanation for the observed behaviour.

3109:1-2. The integrative means are still good for estimating the overall forcing of the ice to generate a seasonal cycle of the right amplitude and phase.. No modelling institute depends on these alone to access their model performance and spatial characteristics are consequently important. Here, however, you are looking for a means to rank model ice performances. Others have done this and sensibly resorted to thickness pattern as a better assessment. If you must stick with passive microwave, integrated extent or area, then temporal variability (eg. related to NAO) could be assessed against 'observations'.

A number of recent studies that evaluated CMIP5 model simulations focused almost exclusively on sea-ice extent for defining the quality of a model. While individual models in focused studies are indeed often evaluated based on a number of parameters, this is less so the case for model intercomparison. I will look into studies that have used thickness patterns for assessing CMIP5 model quality and will mention those as a more objective way to evaluate model performance.

3109:10-28. This section is poorly expressed. You are talking about initial condition ensembles and the mean value and the standard error on a 27 year mean. This mean includes a trend and hence is not strictly speaking isolating the internal variability as some models may have a large trend and

others not. Rather than quote upper and lower bounds for a single. There have been many studies on multidecadal oscillations in sea ice to quote (e.g. J J Day et al 2012 Environ. Res. Lett. 7 034011)

I agree that the wording here is confusing. I intended to use the term „internal variability“ to explain the range of trends that can be caused by a specific external driver. I will rephrase this section appropriately.

3109:17 "..mean September area. . ."

Will be changed.

3109:21-22. What do you mean "estimate of the truth"? Do you mean "observational uncertainty"? You really only have a ranger here rather than an 'uncertainty'. Do not understand why you are using a range in a model ensemble to justify this uncertainty. Observational uncertainty comes from the observations not the models. What does it matter that one ensemble member is close to the Bootstrap September extent?

I will try to make the wording of this section more clear. I am here referring to the fact that internal variability implies that an individual model simulation can be rather different to the observed trend without this implying anything on the quality of the underlying model.

3109:22. The so called "117 CMIP5 simulations" are not independent as many of these are ensemble members. In any case you need to mention the model uncertainty first as there is no point in this statement until that is done

I will address this in the rephrasing of this paragraph.

3109:25-27. No definition what 'close to Bootstrap' means quantitatively. Instead use 'a member which lies within the observational uncertainty range'. Does your quoted range refer to one specific ensemble or is the model uncertainty range across the subset of model ensembles, one member of which lies within the observed uncertainty range?

Again, this will be formulated more clearly in a revised version.

3110:1-2. This could be misinterpreted. What you mean is that all ensemble members from the CMIP5 archive, not the individual model means. Considering all ensemble members as a group biases any interpretation towards the models with large ensembles.

I will more specifically refer to ensemble members.

3110:4. Check your figures – just 6 lines earlier you have given a figure of 6.9 million km2 for Bootstrap extent.

Thanks for pointing this out.

3110:6-19. The discussion section will need to discuss why your analysis on trends is different from that of Stroeve et al (2012).

This is mostly because we consider different periods. I will clarify this point.

3110:20-23. It is not the case that more models lie outside the 'acceptable' range in trends for area than for extent. It seems to me that the take-home message from this section is that the trends are the same in area and extent, as each has its own internal consistency.

I was referring to the mean, not to the trend within this paragraph. Given that the preceding paragraph only deals with trends, I will be more specific on this point in a revised version.

3111:1-11. I assume that in this paper you have been using monthly mean observed ice concentration products to infer ice extent. This will then be consistent with the calculation from the models. However, if one were to use an ice extent product derived from daily from ice concentration then a comparison with the model would be in error as you describe.

Yes, everything here is monthly mean. I will clarify this fact.

3112:11-12. Since you have no verification of these retrievals from other than passive microwave 'observations' this is almost certainly an underestimate of the observational as consequently there will be seasonal systematic biases in the retrievals (eg. water cloud in summer, different snow cover characteristics in winter – possibly wet snow with arctic cycle seasons in spring, thin ice during freeze-up).

This is a valid point, and I will add this discussion to the revised version.

3114:10-13. This is not proven in the paper, and is essentially idle speculation. To demonstrate this would require access to diagnostics not available through PCMDI. Apart form my previously expressed objection to your definition of 'compact', if you wish to include this then I suggest it is rephrased as 'It is speculated that the difference in model summer ice distribution is associated with the partitioning of heat between lateral and vertical melting.' However, since you have demonstrated that model internal variability dominates in the error budget, it becomes increasingly unlikely that this suggested interpretation is valid.

As outlined above, I will reassess the underlying cause for the different model behaviour. Since models with multiple ensembles show consistent behaviour regarding their compactness, I don't think that internal variability plays an important role for this behaviour.

3114:20:23. Since models do not simulate trends which are outside the uncertainty band, this point is meaningless. Indeed, this is said in point 6.

This statement is true and independent of uncertainty bands. It relates only to the difference between trends in area and in extent, which is important since the observed trend (without any uncertainty band) is often taken as the main metric to evaluate model performance.

3115:8-10. This is weighted towards models with large ensembles. It would be more accurate to specify how many models have no ensemble member within the bounds.

This is a good suggestion that I'll take up for a revised version of this paper.

3115:13-14. Specify that this refers to passive microwave satellite retrievals. Also note that the observational error does not include systematic biases associate with PM observations.

Thanks for this helpful suggestion.

3115:19 Correction; 'area' rather than 'are'

OK.

Figure 4: Clarity: overlap of x-axis numbering

Will be improved.

Anonymous Referee #2

In "Sea-ice extent provides a limited metric of model performance" Dirk Notz provides a long overdue, comprehensive discussion on how different a metric sea ice area and extent are. He nicely presents the issues arising from focusing on either one in determining model skill. In the introduction he points out that sea ice area is the physically more meaningful measure; sea ice extent is, however, preferred by the remote sensing community since the estimation of the sea ice edge is associated with smaller errors than that of the overall ice concentration (or total area). The paper is well structured and written. The author provides an easy to understand introduction to the two different quantities sea ice area and extent, studies uncertainties associated with different remote sensing data sets, and provides ample proof of how misleading skill assessment can be if only ice extent is used. Although the paper basically provides no new scientific insight the discussed topic is very important. Model evaluation is essential and sea ice coverage has been in the focus due to the availability of extensive remote sensing coverage. However, as this study nicely demonstrates, sea ice extent is not the right metric for model evaluation although it is the preferred metric of the remote sensing community and has become publicly known as a measure of the annual sea ice minimum. It will be very helpful to have a decent study like this one as a reference.

I recommend the discussion paper for publication in The Cryosphere. Nevertheless, I have a few minor comments the author may want to consider before final publication.

Thank you very much for this positive, motivating assessment.

Detailed comments:
p.3097 l. 28 ". . . shift in the location or its spatial distribution of the sea-ice cover . . ."

Will be changed.

p.3099 l.19f ". . . time series of monthly mean sea-ice extent . . . from their monthly mean sea-ice concentration . . ."

Will be changed.

p.3100 l.6f please mention here that the time series based on the ASI algorithm cover a shorter time period; specifically state the covered periods.

Will be changed.

p.3102 l.19 I wonder if there is a more objective measure for "compact" and "loose" ice covers based on the red line in Fig. 7a, for example using a standard deviation or squared error to measure an acceptable deviation from this line.

I will look into this, though any criterion differentiating between the two will be subjective. Hence, I might at the end of the day stick to the current one because of its simplicity.

p.3103 l.4ff With respect to an earlier statement in the paper that part of the "open water" seen in

the NASA Team ice concentrations is actually meltwater-covered ice, I wonder about the value of the differentiated statement made here. I think it needs to be stated even more clearly, that lower summer ice concentrations in "loose" models vs. "compact" models really mean more open water whereas the difference between Bootstrap and NASA Team ice concentrations may mostly refer to wet ice/snow and water on ice but not open water—physically this has very different effect.

This is a valid point that I will take up in a revised version of this manuscript.

p.3107 l.16 I think that September 2007 is not a good example for testing the effect of grid resolution. If I am not mistaken grid resolution should matter most for a generally loose ice cover. However, summer 2007 featured a record low ice concentration because the atmospheric circulation compacted the otherwise loose ice pack to have a record low extent. Therefore, I assume that extent and area were not that much different in September 2007. I think a summer with a great difference in area and extent would suit the test better.

That is again a valid point and I will look into the impact of changing this analysis to a year with a more spread out, loose ice cover.

p.3107 l.24 what does "small" mean here, please give a number

Will be added in a revised version.

p.3109 l.10ff I am confused by the definition of "internal variability" and the yellow shading applied to Fig. 6. The text says that internal variability is measured in terms of the spread of ensemble members per model. However, the yellow shading seems to be constant across models, i.e. indicates the spread considering all 117 simulations at once. Both definitions could work as a measure of internal variability, just be clear.

The latter is what I did. This will be clarified.

p.3110 l.11 This sentence could be interpreted as a strong argument against anthropogenic (or CO2) forced climate change since the spread in trends is huge and the observed trends is within the range of modelled internal variability as are even positive trends shown by some simulations. This would not be the case if internal variability would be assessed for each model individually. Then some models would fail to reproduce the observed trend within their range of internal variability, which would be much more a statement about model quality, which the author is aiming for, I guess.

As outlined in my response to reviewer 1, this section was not formulated very clearly. I wanted to imply that internal variability would allow for a range of different trends given the current trend in the external forcing.

p.3110 l.21 replace "much" by "many"

OK.

p.3112 l.9 "(green curve in Fig. 13e)"

OK.

p.3111 l.24/27 (anomalies/offset), p.3112 l.21ff (absolute bias) In general I agree with the statements made in these lines. However, we should take into account that a fully coupled model system has many degrees of freedom and take into consideration that different models are designed for different purposes. I expect a good coupled model to correctly (say within the observed variability) simulate first the mean state of total ice area (!) and volume and second the observed variability (range of anomalies though not necessarily in the observed sequence) and third the observed trend. In this sense the absolute bias is a very hard to meet criterion that unnecessarily may make many models look to perform badly. If I want to study regional changes in particular with a focus on the observational period I would rather use a forced ice-ocean model, which I indeed expect to meet the absolute bias criterion. This said, I wouldn't say that the absolute bias is the most meaningful measure, it really depends on the purpose; it's probably the most precise measure of model performance.

This is a helpful discussion that I will add to this section

Figure 1 nice example! However, I would rather have a "top view" perspective than a "side view", i.e. ice floes should be fully contained in blue ocean boxes covering the indicated area; all boxes should be blue independent of having ice in them or not (currently left box in panel (a) and left and right boxes in panel (c) are white) to be related to open water.

This is a good suggestion for making this figure clearer, thanks.

Figure 2 legend: name satellites/sensors which are used for Bootstrap and NASA Team

OK.

Figures 2, 6, 7, 10, 13 use same colour key for satellite data in all four figures, i.e. red=Bootstrap, green=NASA Team, blue=ASI SSMI, and black=ASI AMSR

OK.

Figures 3 and 4 Would be great to note the ice-covered (>15%) area fraction in each panel (print a number in each panel). This would provide a quick access to differences in ice extent.

I had thought about this, but then found it difficult to make this compatible with an equal spacing up to 100 %. Will see what I can do about this.

Figure 5 This Figure is awfully dense and very difficult to read. Considering how it is used on p.3103 I suggest showing 4 scatter plots instead: concentration change vs. volume loss, thickness change vs. volume loss, concentration change vs. September concentration, and thickness change vs. September concentration with a single dot for each simulation. The "loose" models should then accumulate in a different part of each plot than the compact models.

This is a helpful suggestion. Given the discussion of my reasoning around this topic provided by reviewer 1, there's a chance that this figure will be removed from a revised version of this paper.

Figure 6 I suggest to group models by "compact" and "loose" ice conditions and then sort them

alphabetically within each group. As they are listed now I have a hard time to get a quick impression on how the one or the other group performs. Moreover, internal variability should be assessed per model not across models (see comment above).

I will re-group the models according to this suggestion. Regarding internal variability, many models have too few (i.e.: 1) ensemble members to get any estimate of internal variability. Hence, I prefer to remain very specific that this is just a very rough estimate of internal variability across all models.

Figure 7 I suggest to add a 1:1 line (slope 1) to both panels to get a quick sense of the difference between area and extent and for better orientation. And I recommend to use vertical and horizontal error bars centered on the red dot (or a gray shaded cross as in Figure 10) to indicate a reasonable deviation from the observation instead of the gray block on the right. Also, please mention in the caption that blue "loose ice" dots are only shown in panel (a).

I believe that a line with 1:1 slope can be misleading, since the absolute values of area and extent are different. I therefor prefer the current lines that connect all dots with the same percentage deviation. Horizontal and vertical error bars make it more difficult to assess which points lie within a certain error margin for, say, area but not for extent. I will, however, check if such plot nevertheless increases clarity. Panel (b) shows conditions in March, when all model simulations have compact ice. Hence, all simulations that are blue in panel (a) become red in panel (b). I will make this point more explicit in the caption.

Figure 8 I don't think this figure is necessary. I think it is obvious from Fig. 7a that dots (models) along the green line with excessive extent have a small bias in area and that dots at smaller than observed ice extents will have even greater area bias. This is very straight forward. In case the author should remove this figure, text between p.3105 l.16 and p.3106 l.6 should be shortened considerably; in case the author wants to keep this figure, I recommend to add a statement in the figure or its caption that the distance between the blue and read lines of Model 1 and Model 2 are the same (at least they should be).

I will critically assess whether this figure improves clarity in that it answers the question as to why the dots in fig 7a have the different biases that they have. If I should keep this figure, I will add the clarification that the reviewer suggests.

Figure 9 print names of "compact" models in red as in Figure 6

OK.

Figure 11 what is the threshold for "ice-covered" here, 15%? Please state this in the caption

I will clarify this in the caption.

Figure 12 typo: panel (c) is labeled (b)

Thanks for spotting this, will be corrected.