The Cryosphere Discuss., 8, 2179-2212, 2014
www.the-cryosphere-discuss.net/8/2179/2014/
doi:10.5194/tcd-8-2179-2014

Julienne Stroeve 8/5/14 12:06 PM
Formatted: Bottom: 1.03"

Review Status
This discussion paper is under review for the journal The Cryosphere (TC).

Using records from submarine, aircraft and satellite to evaluate climate model simulations of Arctic sea ice thickness

J. Stroeve, A. Barrett, M. Serreze, and A. Schweiger

Dear reviewer, please see our reply to your helpful comments below. Our responses are given in blue italics.

Best regards,
Julienne Stroeve

General comments

The paper proposes a detailed evaluation of CMIP5 modeled Arctic sea ice thickness over the past four decades, using a hierarchy of observational sea ice thickness data. To my knowledge, no earlier study has ever engaged in such a comprehensive evaluation of models using sea ice thickness data, and this paper is in this respect very welcome. The authors find that CMIP5 models simulate the average sea ice thickness reasonably well, but that only few models simulate spatial patterns of Arctic sea ice. Finally, the authors discuss trends in modeled sea ice volume as compared to the PIOMAS sea ice reanalysis. The multi-model mean trend is found to be underestimating the PIOMAS trend, but to lie within the range of uncertainty of the PIOMAS statistic.

The paper has a several positive points. First, it is novel in the use of so many observational thickness data. Second, it proposes to use sea ice thickness to evaluate coupled models, which I agree with the authors is a more physical metric than sea ice extent,and certainly important for constraining projections. Third, The authors do not hesitate to test some of their hypotheses extensively, as e.g. the use of multiple atmospheric reanalyses data set to examine the skill of sea level pressure in CMIP5. Finally, the paper is well structured and has a clear scope.

We appreciate the positive comments on the value of the study.

Using sea ice thickness to evaluate models has certainly a better physical basis than evaluations based on sea ice extent alone; however, the price to pay is that thickness products are subject to larger uncertainties: sampling in time and space is not uniform, instrumental and methodological errors are large (Zygmuntowska et al., 2014, doi:10.5194/tc-8-705-2014; the authors should cite this very informative study); PIOMAS is certainly useful but is a model, for which long-term trends can be sensitive to the atmospheric forcing used. On top of that, natural variability is pronounced and makes the evaluation a delicate task, especially for short periods of time. The authors are aware of these individual sources of uncertainty as discussed nicely in the text. However, understanding the interplay between all these sources of uncertainty, and how large is the resulting total uncertainty, is key to making a clean model evaluation. In the diagnostics, the uncertainties in observational data are probably underestimated because not treated as a whole: for instance, in Fig. 3, the authors co-locate the model and observations thickness in space (very good choice) but do not co-locate model and observations in time (models statistics span 1981-2010, one of the products spans 2011-2013, the other 2004-2005, ...). In addition, it is not clear if instrumental uncertainty and methodological uncertainties (related e.g. to assumptions on snow load, snow and ice densities when ice thickness is retrieved) are taken into account. This introduces additional uncertainty in the comparison, which is not displayed in the error bars. If the authors think it is not the case, they should then argue why. I have listed below (in the Specific Comments) several places where I think uncertainties could be larger than displayed. Thus, in my opinion, statements such as "The climate models as whole also tend to underestimate the rate of ice volume loss from 1979 to 2013" (Abstract) must be tempered by the recognition that uncertainties are much larger than for the less-physical, but more reliable ice extent metric. I agree 100% with the authors that model evaluation based on sea ice thickness and its distribution in space and time has a clear physical meaning and would be a good choice to constrain projections. Yet, the conclusion that the CMIP5 models have low ability to replicate sea ice thickness is, to me, too strong given the large cumulative uncertainties in observational data or reanalyses of sea ice thickness and volume.

As the reviewer notes, using ice thickness to evaluate model performance is likely a better metric than sea ice extent used in many previous papers, including some of our own. However a long-term observationally-based thickness data record is not available like we have for sea ice concentration/extent. We acknowledge in our paper that there are large uncertainties, both from methodology and instrumental approach to spatial and temporal sampling that make it difficult if not impossible to produce a consistent long-term record of sea ice thickness. It is outside the scope of this study to produce such a data record or to assess the cumulative uncertainties resulting from choice of snow/ice density, snow depth, freeboard uncertainties, etc. Instead we use published sea ice thickness data sets to compare to the CMIP5 data. We agree that it is difficult to compare mean ice thickness between CMIP5 and the limited observations because (1) we cannot expect the models to be in phase with the observations in terms of natural climate variability and (2) the biases in each individual sea ice thickness data set are not well-quantified. The main usefulness is (1) provide an evaluation of the spatial pattern of the thickness distribution and (2) if PIOMAS does provide reasonable ice thickness estimates, this could be used to evaluate CMIP5 model performance in terms of total ice volume changes as it spans several decades. While the individual observational data records are not produced consistently, the spatial pattern of ice thickness is a long-term climate feature of the Arctic, and one would expect (hope) that the CMIP5 models can also reproduce this basic spatial pattern. This unfortunately is not accomplished by the majority of the models and is an important result. We have tried to clarify this in the revised text (see specific examples below).

Figure 3 is a useful comparison of the distributions of mean thickness fields in each of the models. We could show the models in a single panel and not include observations if the editor feels that would be better.

I list below several comments related to my main point. I also list several other points that deserve more detailed information in the text (I pointed several inconsistencies that need to be looked at in more detail). In particular, I would not be able to replicate several figures myself just based on the information given in the text, so that some clarifications are needed. I hope that my review of this paper will help the authors. If my comments/questions are addressed and my remarks taken into account, I strongly recommend the paper for publication.

Specific comments

1.p. 2180, line 19: Please cite the source for the trends reported, and include uncertainties.

There is no peer-reviewed reference for this statement, this is based on NSIDC data. Instead we rewrote the sentence to read: "Using data from the NSIDC Sea Ice Index (Fetterer et al., 2002) the linear trend for September, as calculated over the 1979 through 2013 period, stands at −14.0% dec^{-1}, or −895 300 km^{2} dec^{-1}."

Fetterer, F., K. Knowles, W. Meier, and M. Savoie. 2002, updated daily. Sea Ice Index. Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.

2.p. 2181, line 1 : The September 2013 sea ice extent anomaly is thought to be "partly a result of anomaly cool summer conditions". Are there studies that have been investigating the causes for this unusually high minimum compared to the trend line? If so, could you refer to those studies?

This is based on the NSIDC statements as part of the Sea Ice News and Analysis, and evaluation of the summer 2013 melt season performed by the first author. We recently published a similar statement in GRL related to the ARCUS Sea Ice Outlook. Thus, we add that reference here (Stroeve et al., 2014)

3.p. 2181, line 15: In order to stick to the CMIP3 assessment made in line 13 (67% of the models...), I would not use "most" here, but rather a quantitative estimate as well.

Done, the text has been amended to read: "However, historical trends from 85% of the model ensemble members examined remain smaller than observed, and the spread in simulated extent between different models remains large."

4.p. 2183, line 20: The CMIP5 database is complete since more than one year now; why are only 27 climate models analyzed (out of 39 available)? Did the authors apply a first filtering on the models before the analysis was conducted? Could the conclusions be sensitive to the inclusion of the models not taken into account?

We started with 27 climate models as those were the ones available when we started this analysis. No filtering of models was performed.

We now have 33 models in our data base. While we already have a good sample and the conclusions are not sensitive to increasing the number of models, we went ahead and updated the analysis to include 33 models. Note that some of these models are similar versions of models we already had access to.

5.p. 2184, lines 21-26 and Fig. 1: The results are extremely interesting, and probably worth investigating (perhaps not in this paper!). It appears from first-order inspection of Fig. 1 that the three models with the most intrinsic variability in sea ice thickness comprise an ice-thickness distribution (ITD) framework, and the three others don't. That is, it looks like models that resolve the statistical sub-grid scale distribution of sea ice thickness (EC-Earth, CCSM4, HadCM3) produce grid-cell thicknesses that are more likely to be influenced by natural variability than models without ITD. Could there be a physical reason for that? Anticipating that most models of the next generation will include sea ice models with an ITD, the evaluation of mean thickness will perhaps be even more difficult in CMIP6 than it is today with CMIP5.

We looked at generating a table showing sea ice components and physics in the models (see below). However, it is not complete because some of this information is not available in publications or websites. Thus, we leave it to the editor to decide if we should include this table, albeit incomplete, in the manuscript. We do mention the reviewer's observation in the revised manuscript that models with ITD may be more sensitive to natural variability than models without. The reviewer seems to suggest that HadCM3 also uses a ITD framework, although we have been unable to verify that.

Model	**Sea Ice Model**	**Physics**
ACCESS1-0	**CICE v4**	**Energy conserving thermo, EVP, ITD**
ACCESS1-3	**CICE v4**	**Energy conserving thermo, EVP, ITD**
BCC-CSM1-1	**SIS**	**Semter 3-layer, EVP Rheology, ITD**
CanCM4		
CanESM2	**CanSIM1**	**Cavitating fluid**
CCSM4	**CICE v4**	**Energy conserving thermo, EVP, ITD**
CESM1-CAM5	**CICE v4**	**Energy conserving thermo, EVP, ITD**
CESM1-WACCM	**CICE v4**	**Energy conserving thermo, EVP, ITD**
CNRM-CM5	**GELATO v5**	**EVP, ITD**
CSIRO-Mk3-6-0		**3-layer, Cavitating fluid**
EC-EARTH	**LIM2**	**Semter 3-layer + brine pockets, VP, virtual ITD**
FGOALS-g2	**CICE v4**	**Energy conserving thermo, EVP,**

		ITD
FIO-ESM	**CICE v4**	**Energy conserving thermo, EVP, ITD**
GFDL-CM3	**SISp2**	**Modified Semter 3-layer, EVP, ITD**
GFDL-ESM2G	**SISp2**	**Modified Semter 3-layer, EVP, ITD**
GFDL-ESM2M	**SISp2**	**Modified Semter 3-layer, EVP, ITD**
GISS-E2-H		**4-layer, VP**
GISS-E2-R	**Russel Sea Ice**	**4-layer, VP,**
HadCM3		**Semter 0-layer, Free-drift**
HadGEM2-AO	**Sea ice component of HADGOM2**	**Semter 0-layer, EVP, ITD**
HadGEM2-CC	**Based on CICE**	**Semter 0-layer, EVP, ITD**
HadGEM2-ES		**Semter 0-layer, EVP, ITD**
Inmcm4		
IPSL-CM5A-LR	**LIM2**	**Semter 3-layer + brine pockets, VP, virtual ITD**
IPSL-CM5A-MR	**LIM2**	**Semter 3-layer + brine pockets, VP, virtual ITD**
MIROC-ESM	**Sea ice component of COCO3.4**	**Semter 0-layer, EVP, 2 ice categories**
MIROC-ESM-CHEM	**Sea ice component of COCO3.4**	**Semter 0-layer, EVP, 2 ice categories**
MIROC4h		**Semter 0-layer, EVP, 2 ice categories**
MIROC5	**Sea ice component of COCO3.4**	**Semter 0-layer, EVP, 2 ice categories**
MPI-ESM-LR	**Component of MPI-OM**	**Semter 0-layer, VP rheology, ITD**
MPI-ESM-MR	**Component of MPI-OM**	**Semter 0-layer, VP rheology, ITD**
MRI-CGCM3	**MRI.COM3**	**2-layer, EVP, ITD**
NorESM1-M	**CICE v4**	**Energy conserving thermo, EVP, ITD**

6.p. 2184, line 27-29: The spatial correlations of thickness between individual ensembles are found to be very high (>0.9). The authors infer that evaluation based on thickness patterns is not too much affected by natural variability. This statement relies on the hypothesis that the models simulate the correct natural variability; was this hypothesis tested, and how? In line with my previous comment, models comprising more realistic sea ice physics simulate more spatial variability. Does that mean that the other models may underestimate the natural variability in sea ice thickness? Given the short period of time of the ICESat campaigns (a few years) that are used for the evaluation of spatial patterns (Fig. 5), is this evaluation really robust and free of impacts from natural variability?

These figures show that the spatial climatology is robust and that the variability in thickness from year to year does not impact this. We do not agree that the correlation rests on the models getting natural variability correct. Natural variability in the models (from the different ensemble members) does not appear to impact the spatial pattern of the *thickness distributions. In the correlations discussed, models with ITD do not necessarily have lower correlations between ensemble members. Note that the correlations in Figure 1 are from 1981 to 2010 (this is now stated in the text).*

Finally, the point that the spatial pattern from ICESat may not be used to evaluate the spatial pattern in the models is well taken. Six years of data is too short, especially as the CMIP5 model time is not commensurate with actual time. However, in general the spatial pattern is persistent between all the observations. ICESat does show thinner ice out towards the Chukchi and East Siberian seas which is not as well defined for example in the CryoSat or ERS-1 data. The spatial pattern correlation during the ICESat time-period remains similar as during the longer time-period. We performed an additional comparison between the models and the PIOMAS thickness fields and found the correlations are generally higher between CMIP5 and PIOMAS than for the ICESat data. This is not surprising because PIOMAS tends to show thicker ice towards the Chukchi and East Siberian seas, which is also a feature seen in the climate models (see Table below). However, this is not a comparison between models and observations but rather a comparison between models. Simply put, PIOMAS does not necessarily reflect the true observed thickness distributions. Compared to observations, biases in PIOMAS are similar to those for the models in the CMIP5 archive.

	IceSat	PIOMAS
ACCESS1-0	-0.05	0.64
ACCESS1-3	-0.09	0.63
bcc-csm1-1	0.18	0.44
CCSM4	0.65	0.62
CESM1-CAM5	0.27	0.79
CESM1-WACCM	0.34	0.67
CNRM-CM5	0.35	0.59
CSIRO-Mk3-6-0	-0.14	-0.09
CanCM4	-0.30	0.09
CanESM2	-0.21	0.26
EC-EARTH	0.15	0.74
FGOALS-g2	0.15	-0.01
FIO-ESM	0.06	0.37
GFDL-CM3	0.22	0.83
GFDL-ESM2G	0.34	0.74
GFDL-ESM2M	0.34	0.68
GISS-E2-H	-0.07	0.49
GISS-E2-R	-0.07	0.33
HadCM3	-0.07	0.57
HadGEM2-AO	0.27	0.82
HadGEM2-CC	0.33	0.87
HadGEM2-ES	0.25	0.83
inmcm4	-0.17	0.35
IPSL-CM5A-LR	-0.14	0.49

IPSL-CM5A-MR	0.04	0.66
MIROC-ESM-CHEM	-0.01	0.46
MIROC-ESM	0.08	0.55
MIROC4h	0.35	0.67
MIROC5	0.65	0.80
MPI-ESM-LR	0.28	0.74
MPI-ESM-MR	0.25	0.73
MRI-CGCM3	0.58	0.31
NorESM1-M	0.30	0.60

7.p. 2188, lines 7-9: The satellite thickness fields were regridded using a drop-in-the-bucket approach. Please specify how you treated instrumental/methodological uncertainties (related, e.g., to assumptions on snow and ice densities when thickness is retrieved), how you propagated uncertainties from the 25km level to the 100 km during this interpolation, and whether you accounted for these uncertainties in the evaluation. These uncertainties are maybe much lower than the interannual variability, in which case they can be ignored as a first approximation, but then please show that this is the case.

We did not propagate the uncertainties in the observational data sets during regridding. We used published data sets of ice thickness and while these are not necessarily consistent with each other, at the moment this is the best we have. At present, no consistently processed data set (i.e. consistent approach for ice/snow density and snow depth and fully characterized freeboard uncertainty) is available that spans all observational data sets. We also stress that no previous comparisons between models and observations (i.e. work by Kwok with CMIP3) have dealt with observational uncertainty of ice thickness.

8.p. 2190, lines 9-11. I would temper this statement. I can accept that PIOMAS estimates for the mean sea ice thickness compare well with observational estimates (as seen in Fig. 2, and discussed in Laxon et al., 2013 or Schweiger et al., 2011). That the trends in PIOMAS volume may be used with confidence to evaluate CMIP5 trends should be tempered by the recognition (i) that the PIOMAS trends are sensitive to the atmospheric forcing used (Lindsay et al., doi:10.1175/JCLI-D-13-00014.1), but also that the evaluation is strongly impacted by natural variability. If these two sources of uncertainty are independent, the error bars displayed in Fig. 8 are probably larger than depicted.

The "error bars" are the 2-sigma standard error of the model trends. We think that what the reviewer is getting at is that the uncertainty in PIOMAS could be larger than we show. In which case, more models would have trends included in the "acceptable" category. In addition natural variability might not be completely included in the reanalysis forcing because the sea-ice model could introduce uncertainty – e.g. internal variability. We now include a statement to that effect. Note that the PIOMAS integrations are sensitive to the forcing fields but the $1x10^3$ km^3 uncertainty estimate is consistent with those.

9.p. 2190, line 19: "uncertainty of decadal PIOMAS trends of1103km3": the units are confusing for characterizing trends. Write "uncertainty in PIOMAS trends of 1103km3/dec"?

This should have been km^3 dec^{-1}.

10.p. 2190, lines 19-21: "Given the large observed volume trend ..., PIOMAS is a suitable tool for assessing long-term trends in CMIP5 models". I don't understand the logical articulation of this sentence. The suitability of a reanalysis to assess models is not related to the magnitude of the trend, rather to the confidence we have in this trend.

The reviewer does have a point that if we want to validate a particular trend then the magnitude has little to do with it. Conversely, the ratio of the magnitude of the trend over the uncertainty is a measure of the signal to noise ratio and if we had a trend that is much smaller than the uncertainty then we wouldn't want to use it for validation of models or other purposes.

The Reviewer also points out that the PIOMAS integrations are sensitive to the forcing fields, they are, but the 1x10^3 km^3 uncertainty estimate is still consistent with those.

p. 2190, line 20: Remove "observed". PIOMAS is a model.

Done

p. 2191, lines 16-17: What is meant by "spread"? The 10-90% interval, the range, ...?

The spread is the 10-90% interval.

p. 2192, line 17: "PIOMAS facilitates more robust comparisons". Again, I would temper this sentence (see my comment [p.2190, lines 9-11]): using PIOMAS brings the advantage of long and homogeneous records, at the expense of using a model instead of observations.

We changed the sentence to read: "On the other hand, the fairly long PIOMAS record (30 years) brings the advantage of a long and homogeneous data record to compare with the model data".

p. 2193, lines 6-13: This diagnostic is extremely interesting. If I follow the authors and inspect Fig. 5, models resemble more each other than they resemble observations. Is this an indication that models share the same biases (rheology, thermodynamics, winds)?

We think that it is an interesting observation but this is outside the scope of the present work.

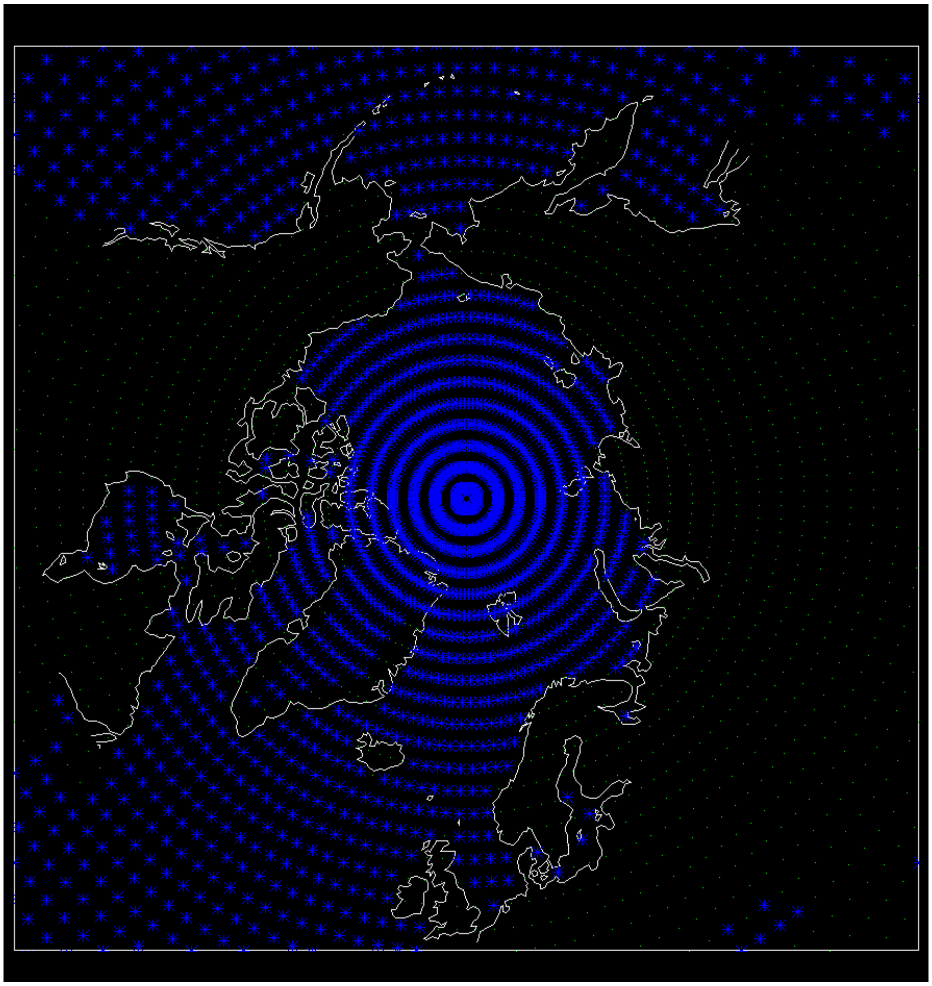
p. 2194, line 24: The authors mention the range of 14470 km3 to 87000 km3 for simulated ice volume in March and refer to Fig. 7 - dashed lines. The dashed lines in Fig. 7 are at the

19000 km3 and 43000 km3 levels and are supposed to represent the minimum and maximum volumes in the model ensemble. Did I miss something?

This was a mistake on our part. The range is 18,000 km^3 (CanESM2) to 48,000 km^3 (CESM1-WACCM)

p. 2194, line 24: The value of 87000 km3 for GISS-E2-R is clearly unrealistic.
It turns out that the GISS-E2-R model output has sea ice thickness of 1 m and sometimes more over a large fraction of Northern Hemisphere continents. Did the authors correctly mask the continents when calculating sea ice volume? What is the impact on the multi-model mean volume/trends?

It turns out that the GISS data does not designate the interior of the Greenland, Ellesmere Island or Svalbard as land. Moreover, much of N. Eurasia and N. America has ice in the sea ice fields. We ended up generating our own landmask for the GISS data.



p. 2195, line 23-25: "The majority of ensemble member trends ... can therefore be considered compatible with PIOMAS". If the null hypothesis is "H0: CMIP5 trends are consistent with PIOMAS" (as stated p. 2195, line 12), then the fact that the majority of CMIP5 2 sigma ranges overlap with the PIOMAS does not allow to reject H0. Thus, I would turn the sentence in "The majority of trends cannot be considered incompatible with PIOMAS".

The sentence has been changed as suggested.

p. 2196, line 1: The individual ensemble members are averaged together to produce the multi-model ensemble mean trend in March ice volume. If I understand well, more weight is thus given to model with more ensembles. Is there a particular reason for that? Why was the evaluation of mean thickness carried out by giving equal weight to each model by pre-averaging members (p. 2184, line 4)?

That was a typo on our part. The individual ensemble members are averaged together to first produce an ensemble mean for each model and then those are used to generate the multi-model ensemble mean trend in ice volume. Thus, each model has equal weight. We clarified this by rewriting the sentence to read: "Averaging together the individual ensemble means from each model yields a multi-model ensemble mean trend..."

p. 2196, lines 14-15: Units are 103km3/dec, not 103km3.

Corrected.

p. 2198, lines 1-2. I cannot follow the sequence of arguments here. It is said that only two models have the correct spatial thickness patterns but have very different trends in sea ice volume, so that constraining models based on sea ice thickness patterns is not promising. I think it is, as the distribution of ice thickness has been shown to be a source of spread in projections (Holland et al., 2010, doi:10.1007/s00382-008-0493-4). It is, probably, not sufficient to filter projections based on thickness patterns only. Is that what the authors meant?

We agree that filtering based on thickness patterns is insufficient. We are currently working on another paper addressing whether it is valid to constrain models at all. We rewrote the sentence for clarity.

p. 2202, Table 2: In the caption: "Mar">"March".

This seems to have been in an old version of the Table, this is corrected in the revised manuscript.

p. 2202, Table 2: In the caption: "Trends are listed as km3" should be replaced by"Trends are listed as 103km3/decade" or "Trends are listed as 102km3per year"(according to the table header).

Done.

23.
p. 2202, Table 2: "NorEMS1-M">"NorESM1-M"

Done.

p. 2202, Table 2: I suggest to include a brief description of the sea ice model used in each CMIP5 model. Since the paper evaluates sea ice thickness, it seems important to me to specify what thermodynamic scheme is used, whether the model includes the sub-grid scale ice thickness distribution or not, and the type of rheology that is used. To increase the impact of this paper and help subsequent groups identifying how biases in sea ice thickness relate to the physical sea ice model used, this step seems instructive to me.

See the response and table to major comment 5. The table is not complete so if the editor feels it is still useful to include it we can do that, noting that we do not have the information for every model.

Comments on the figures

1.Fig. 1: Over which time period are the "stddev" and "average" statistics computed? For what month are the diagnostics shown (March, September, annual average)? For a given model and given grid cell, how is computed "stddev": by first averaging thickness in time for each member, then taking the standard deviation over members, or by first taking the standard deviation of thickness over members for each year and then averaging over years? The order has an importance.

Statistics were calculated for the period 1981 to 2010. Time mean fields of thickness were calculated first for each model ensemble member. Coefficients of variation were calculated using the model ensemble means and standard deviations. Results for March are shown in Figure 1. The figure caption has been changed to reflect this information.

2.Fig. 3: In the "IceSat" panel (third from the top), at least 10% of the data was sampled in open-water since the 10% percentile line (green) is super- imposed on the zero-line. Returning to the paper of Kwok et al. (2009, oi:10.1029/2009JC005312). I can read that IceSAT samples with ice draft less han 10 cm are considered to be open water. Is that the explanation, or the>10% of data with ice thickness equal to 0 m are really open water? In the former case,did you also mask the model output below 10 cm to ensure consistency in the comparison?

Reading through Kwok et al. 2009, there is no mention of setting ICESat thicknesses less than 0.1 to open water. For most of the observations, there is no open water. However, ICESat and PIOMAS do have open water, even when we regrid to the 100 km resolution. Including open-water in the regridding is consistent with how the CMIP5 models report ice thickness: the mean thickness for a cell, including open water.

3.Fig. 5: It would be good, at least for the correlations, to specify which ones are significantly greater than 0. Given that a large number of grid points is used to compute the correlations (the grid resolution is 100 km by 100 km, the area covered is approximately 10x106km2 so I would expect about 1000 grid points) the correlations are probably significant even for low values. Providing the significance would also allow to point out which models have a totally unrealistic sea ice thickness.

The reviewer is correct that almost all of the correlations are significant, even at the 99% level. However, we have added filled and hollow circles to indicate correlations that are significant at the 99% and 95% level.

4.Fig. 5: There are only 25 models evaluated in this figure but in the text 27 models are presented. That is, correlations and RMSE scores are not shown for CanCM4 and GFDL-ESM2M in Fig. 5. Why leaving these models aside?

We have updated all the related figures with 33 models.

5. Fig. 7: The title ("March") is cropped.
This has now been fixed.

6.Fig. 8: In the legend, please change "Observed" by "Reanalyzed" or "PIOMAS". PIOMAS is a model.
*The legend does not include the word "observed", but rather states: **Figure 8.** March ice volume trends from 1979 to 2013 for all 92 individual CMIP5 model ensembles as well as the multi-model ensemble mean (shown in black) with confidence intervals (vertical lines). The 1σ and 2σ confidence intervals of PIOMAS trends are shown in dark gray shading (1σ) and light gray shading (2σ).*

7.Fig. 8: How is the confidence interval for the multi-model mean constructed? Is its width equal to the average width of all confidence intervals, or is its width calculated directly from the time series of multi-model mean sea ice volume? Referring to my comment [p.2196, line 1], is this confidence interval biased towards models with more members?

The bars on are 2 sigma confidence intervals. For the multi-model ensemble mean, the standard deviation is calculated as the standard deviation of the model ensemble mean trends following Santer (2008) equation 9.

4Technical corrections, wording, typos, etc.
1.p. 2181, line 5: I think "but" is not necessary

Done.

2.p. 2182, lines 25-26: What do you mean by "mean distribution of sea ice thickness"? As I understand from criterion (1), it is rather the "(statistical) distribution of mean thickness". In the abstract, the wording "mean thickness distribution" is used; is the meaning equivalent?

We changed the wording to "statistical distribution of mean ice thickness fields".

3.p. 2187, line 2: "similar same""similar", or "same"

Done.

4.p. 2191, line 17: "fall">"falls" ("the spread... falls")

Done.

5.p. 2191, line 18: "Fig. 2">"Fig. 3".

Done.

6.p. 2193, lines 9-10: "Fig. 5, top" and "Fig. 5, bottom" should be replaced by "Fig. 5, left" and Fig. 5, right", respectively.

Done.

7.p. 2193, lines 23: "annual mean annual">"annual mean"

Done.

8.p. 2194, line 1: "FGOALS">"FGOALS-g2"

Done.

9.p. 2194, line 13: "the decline" is not necessary in the sentence "sea ice volume is declining faster than the decline in ice extent"

Done.

10.p. 2197, line 16: "maybe become">"may become" ?

Done.

The Cryosphere Discuss., 8, 2179-2212, 2014
www.the-cryosphere-discuss.net/8/2179/2014/
doi:10.5194/tcd-8-2179-2014

Review Status
This discussion paper is under review for the journal The Cryosphere (TC).

Using records from submarine, aircraft and satellite to evaluate climate model simulations of Arctic sea ice thickness

J. Stroeve, A. Barrett, M. Serreze, and A. Schweiger

Dear reviewer, please see our reply to your helpful comments below. Our responses are given in blue italics.

Best regards,
Julienne Stroeve

Anonymous Referee #2
Received and published: 1 July 2014
This study concerns the efficiency of climate models to simulate Arctic sea ice thickness. Assessment of a suite of CMIP5 numerical models is conducted via comparison with ice thickness observations collected from a variety of platforms. The authors find that although the mean thickness and volume for the Arctic Ocean appear to be well represented by many of the CMIP5 models, the spatial patterns of sea ice thickness are poorly represented. Indeed the range of spatial patterns, presented in Fig 4, is a useful contribution that delineates the current limitations of global climate models in representing Arctic sea ice state. The authors find that the model deficiencies in predicting the prevailing atmospheric circulation over the Arctic can partly explain the errors in the prediction of the geographical location of the thickest and thinnest sea ice in the Arctic. The authors find that these deficiencies reduce confidence in the reliability of future projections based on the CMIP5 climate model suite.

A range of sea ice thickness observations spanning 1986-2013 are presented in this study and are used to assess the ability of a subset of CMIP5 models in reproducing both the mean sea ice thickness and the spatial gradient in ice thickness across the Arctic basin. There are a number of concerns with regard to the analysis of the model predictions using novel observational data. The treatment of the observational data sets used to validate the model simulations is unfortunately unsatisfactory since the uncertainties associated with each observational data set are neither discussed sufficiently (P2188 L27-L3) nor taken into account when comparing the observations to the model predictions.

We added a discussion of recent papers detailing such uncertainties.

For example, an assessment of the errors within and between the observational data sets has not been completed. The utility of the observational thickness datasets and associated data quality must be considered carefully, before these data are used to quantify the merits of other data or model predictions. The assessment is premature and impacted by biases across thickness estimates from different sensors. Many of the observational data utilized in this study have specific limitations that can result in ice thickness estimates with a range of uncertainties related to particular measurement conditions. Specific examples include uncertainty in the penetration of the snow pack by satellite radar altimeters and uncertainty in the snow depth and snow and ice densities used to convert altimeter measurements of sea ice freeboard to sea ice thickness. While these conditions and limitations are detailed in the publications (and/or documentation) associated with each dataset/technique, the authors of this manuscript make no reference to specific biases in the observational data or reference data filtering approaches that might be utilized to standardize the thickness estimates derived from the range of platforms used here. This is a serious limitation of this study that should either be addressed, or remedied by using e.g. one (or two?) specific datasets in the model assessment. The ULS data are likely the most convenient since these are likely to have the lowest associated uncertainty.

Further it is difficult to follow exactly how the observational data have been used in this study. First it appears that all observational data were treated equally (eg., P2182 L26 "aggregating all available data", P2183 L3 "combined thickness records", ...).

What we meant is that all available data for each observational data set are aggregated together to compare with the model data as an independent data set. We have clarified this by stating: "how well they replicate the observed mean sea ice thickness based on aggregating all available data across the Arctic for each observational data set;" for P2182 L26, and removed the word "combined"

Later it appears that some (or all?) datasets are compared to the models on an individual basis. Fig 3 compares each of the observational datasets to the CMIP5 model predictions. However the temporal sampling of the observational data is variable and does not match the time period of the CMIP5 model runs (1981-2010). The widespread decline in Arctic thickness and volume that occurred in the period between 2002 –2008, as documented by many authors (e.g., Giles et al. 2008; Kwok and Rothrock 2009; Laxon et al., 2013) will give rise to a thickness bias due to the difference in temporal sampling of the particular observational data set vs the time period of the models were run (fig 3.). This also raises the question of the usefulness of assessing a mean field spanning 1981-2010, considering the rapid decline in mean thickness during the last decade.

We agree with the reviewer. The reason for the longer time-period for the models vs the observations is because we cannot expect the models to be in phase with the observations and show the rapid decline in ice thickness in recent years. For example, year 1996 in the observations does not correspond to year 1996 in the CMIP5 models.

Figure 3 is a useful comparison of the distributions of mean thickness fields in each of the models. We could show the models in a single panel and not include observations if the editor feels that would be better.

The treatment of the observational data and the averaging approaches used should at least be clarified. The authors state on P2189 L10-11 that "the combined records show a decline through time in thickness" but this is not fully supported by the thickness maps in fig 2. For example the thin ice in the IceSat field is 0.5 – 1.0 m thinner than the sea ice in the same areas in the CryoSat field. The IceBridge data span a similar period to CryoSat but indicate the ice is 1-1.5 m thicker.

We agree that the statement is not fully supported. We address some of the concerns by rewriting the paragraph to address uncertainties in the observations and now state:

"Along with temporal sampling problems, the various thickness records have a range of biases due to differences in sensor types and retrieval approaches. Radar and laser technologies use different wavelengths and footprints, and different techniques have been used to estimate snow depth and snow/ice density, which in turn impacts ice thickness retrievals. This creates additional challenges in generating a consistent sea ice thickness time-series as differences in snow/ice density and snow depth values used can lead to large biases in ice thickness [e.g. Zygmuntowska et al., 2014]. For example, difference in ice density used by Kwok et al. [2009] and Laxon et al. [2013] range from 925 to 882 kg m^{-3}, respectively, for multiyear ice. According to Kurtz et al. [2014], this could lead to a thickness difference of 1.1m for a typical multiyear ice floe of 60 cm snow-ice freeboard with a 35 cm deep snow cover. Similarly, given an ICESat freeboard of 0.325 m with an estimated 0.25 m of snow (density 300 kg m^{-3}) atop the ice (density of 900 kg m^{-3}), we would compute a sea ice thickness of 1.5m. Yet if there had been only 0.15 m of snow, the ice would be 2.2 m thick, a change of 0.70 m or 46% of the original estimate.

*At present, there is no consistent long-term sea ice thickness data set that applies these parameters in a consistent manner regardless of which instrument is used. It is nevertheless encouraging that all of the records show similar spatial patterns of ice thickness [**Figure 2: left column**], which while lending confidence to the data, also demonstrates persistence of the general spatial pattern of Arctic sea ice thickness from 1979 to present. Mean thicknesses are greater along the northern coasts of the Canadian Arctic Archipelago and Greenland where there is an onshore component of ice motion resulting in strong ridging. Mean thicknesses are lower on the Eurasian side of the Arctic Ocean where there is a persistent offshore ice motion and ice divergence, leading to new ice growth in open water areas. When viewed as a whole for the Arctic, the combined records show a decline through time in ice thickness, though this must be tempered by differences in physical assumptions used to retrieve thickness [Zygmuntowska et al., 2014]."*

The section (P2186/2187, L22-L3) describing the technique used to derive sea ice thickness from altimeters is unclear. For example how is the sea surface height derived from gravity models? The authors mention that altimeters measure the height of snow and ice surfaces relative to the reference ellipsoid (L23) and then mention the heigh of these surfaces above the geoid (L25), but the geoid and reference ellipsoid are not the same. The authors could refer to the published literature on this technique in order to improve the description in this section of the paper.

We have edited the text as suggested and simplified the discussion, this section now reads as:

Unlike submarine sonar, satellite and aircraft radar and laser altimeters measure the height of bare-ice, snow-covered ice and snow surfaces above the ocean surface, depending on instrument characteristics and ice-surface conditions. By identifying leads between the ice floes, the freeboard (the height of the snow or ice surfaces above sea level) can be derived. Ice freeboard is converted to ice thickness using Archimedes principle in a similar way as the conversion of submarine ice draft to ice thickness and using estimates or assumptions of snow and ice density and snow depth.

Minor comments:
Grammatical error in title - satellite should be pluralized since ice thickness fields from a number of satellite missions are discussed.

Done.

P2180 L1: "sea ice thickness distributions from models" - do the authors refer to the distribution of ice thickness across a range of thickness classes within CMIP5 models, or do they in fact refer to spatial mean / gradient across the Arctic?

We refer to both the spatial patterns and mean thickness that correspond to the observational data sets.

P2181 L9: why only "in part"? If GCMs cannot reproduce realistic ice thicknesses then there can be little confidence in the projection of an ice free Arctic.

We removed "in part".

P2181 L27-29: statement should also address limitations in temporal coverage and latitudinal limit of IceSat's orbit.

Done.

P2182, L6; P2187 L6-9; P2188 L9: The ERS-1 mission ended on 10 March 2000 – can the authors provide further details and references for the mean field that spans to 2001?

This is based on Laxon's 2003 data, which provides an 8-yr period 1993–2001 mean thickness field up to 81.5° N from both the ERS-1 and ERS-2 satellites. We have now made this clear.

P2812 L8-10: CryoSat was launched in 2010.

Corrected.

P2182 L10: how do you define "sufficient coverage"?
Sentence was rewritten to state: "Together, these data provide a valuable source of information for the validation of spatial patterns of sea ice thickness."

P2182 L15: define meaning of "biases" in CMIP5 models. Potential for confusion with biases in the observations, which should be addressed.

It is unclear to us what the reviewer is getting at. We mean the bias in the thickness/volume in the models compared to the observations and we state that in the sentence.

P2182 L16-18: how is this evaluation accomplished?

This is detailed later in the methods section. Here we are only introducing the goal of the paper. As we state later, we evaluate mean ice thickness for each observational data set in comparison to that in CMIP5 for the same spatial coverage that each observational data set corresponds to. Ice volume is compared with that from PIOMAS. The spatial patterns in the observations are consistent among data sets, and variability between model ensembles in the spatial pattern is small, and therefore the limited ice thickness data (i.e. from ICESat) can be used to evaluate if the CMIP5 models accurately simulate the spatial pattern in ice thickness.

P2182 L18-19. How does snow melt influence the radar measurements? IceBridge data are currently only available in March/April.

The penetration depth of the radar signal depends on snow properties. Beaven et al. (1995) conclude from laboratory experiments, that a Ku-band radar signal at normal incidence reflects at the snow-ice interface if sea ice is covered by dry, cold snow (though this depends on the absence of internal ice layers and other snow properties).
The radar signal does not penetrate into the snow layer if it's wet, instead it reflects from the air-snow interface (Hallikainen 1992). Melt onset over the regions sampled does not start until May/June/July.

P2183, L20: What governed the choice of CMIP5 global model assessed? i.e. why was a subset of 27 models used in this study?

This was only because we originally only had available 27 models when we started this analysis. We have now updated using 33 models.

P2187 L8-10: What is the source of the ERS-1 single mean field ice thickness? Is it publicly available/reproducible?

It was provided by the late Seymour Laxon.

P2187 L21 Thickness retrievals are not detailed in Kurtz and Farrell (2011) since their study concerns airborne snow depth retrievals from IceBridge.

The reference was removed as suggested.

Table 1 how do the results presented for Icesat and CryoSat-2 compare to the results presented in Laxon et al 2013? The correlation between PIOMAS and CryoSat seems a lot lower in this study.
Laxon et al (2013) only present maps of ICESat, CryoSat-2 and PIOMAS, so we cannot compare our correlation with theirs. We also use CryoSat-2 thickness data from AWI. Laxon et al use thickness data processed in a lightly different way. However, comparing maps of thickness in our Figure 2 with maps of thickness in Figure 1 in Laxon et al suggests a good agreement between the two CryoSat-2 products, as well as between IceSat data.

Fig 2: can the authors comment on the large scatter (3-4 m) in ice thickness in the scatterplots for IceSat and CryoSat, if the data were averaged to 100 km grid cells?

We have discussed this issue at length when preparing the first version of the manuscript but cannot come up with a good explanation. One obvious possible explanation is biases in both PIOMAS and the satellite products combined with the greater spatial coverage provided by the satellites.