New perspectives on the skill of modelled sea ice trends in light of recent Antarctic sea ice loss

Holmes *et al.*

Second response to reviewers, 22nd August 2024

We thank both reviewers for their insightful comments regarding our interpretation of the results, and we thank the editor for the opportunity to further revise the paper. We respond fully to the reviewer's comments below.

Our view on the main message of the paper has been shaped by the reviewers' input over the previous two rounds. The key conclusion we now emphasise in the paper is now that, if we consider linear trends in models and observations, then we find that the level of agreement varies over time (which includes of course the agreement for the most recent period). This is the core scientific result of the paper (and is now the opening line of our conclusions), whose robustness is not contentious, and we believe this is an important result in itself, worthy of publication. We are not comfortable with leaving the literature in its current state, with no recognition of this result. The agreement over the most recent period is then a key aspect of this result.

We fully agree with the reviewers that there are several areas in the interpretation of this result that are quite philosophical and highly nuanced, including the applicability of linear trends as a metric, and how we should interpret this changing assessment of model skill. We have incorporated a discussion of these elements, and we believe the paper is now much more nuanced and improved as a result.

As detailed below, we have thoroughly revised the paper to reflect this viewpoint. In particular, we have suggested re-naming of the paper to 'New perspectives on the skill of modelled sea ice trends in light of recent Antarctic sea ice loss'. We recognise that the original title implies an unqualified approval of model performance, and so we think the revised title is a more appropriate summary.

From the editor

Please focus on the following concerns:

Reviewer 1 requests that you **reconsider the implications of your results in the context of model performance across time periods.**

We thank reviewer 1 for highlighting this point. We agree that the deficiency in the models remains based on the mismatch in ~2010-2018, and that this has implications for our confidence in the models. We recognise we have not been clear enough on this point, and in particular how we interpret it, in our discussion. We have responded in detail to the reviewer below. We have edited the paper such that the title and abstract are softer in their conclusions, and made edits to the discussion section such that statements on confidence are carefully qualified and the timescale question is carefully considered.

Reviewer 2 raises **important concerns regarding the use of linear trends as an appropriate metric for assessing Antarctic sea ice loss given the step change behaviour of the observed time series.**

We thank reviewer 2 for their careful reasoning on the question of linear trends. We agree with many of the points the reviewer makes regarding interpretation of linear trends in this case, but ultimately

we do believe there is a justification for assessing the models' linear trends against observations. We provide a fuller argument below as to why we believe linear trends are appropriate, but we also consider their limitations in much more detail, specific to this case. We have edited the manuscript substantially to clarify our motivation in using linear trends and the question we are trying to answer; the strengths and weaknesses of this evaluation; and their interpretation.

Reviewer 1 (Will Hobbs)

The authors have largely addressed my technical comments about the CMIP6 data. I think they slightly misinterpreted my comment about the spread of SIA variability (e.g. Roach et al 20202 Fig 2) across the different models as a comment on each models' ensemble size, but I don't think this affects the analysis in any way.

I do however disagree with the authors Conclusion, based on their own results. They are correct that the assessment of "low confidence" in the CMIP models (e.g. IPCC SROCC) was based on trend analysis to 2005, as well as on mean state biases. However in that period the observed total SIA trend was not statistically different from the models after accounting for uncertainty/internal variability. These results show that this was not the case for trends ending in ~2010-2018. That surely implies a deficiency in the models (since the model ensemble should capture the obs at any time), and that deficiency is still there, even if recent extreme events have wiped out the observed trend. As disappointing as it is, in fact these results are stronger evidence of the "low confidence" assessment than previous studies.

In summary I think the analysis itself is robust and appropriate, but I think the authors need to reconsider the implications of their results

We thank the reviewer for confirming that the analysis is in order. We agree that the deficiency in the models remains based on the mismatch in ~2010-2018, and that this has implications for our confidence in the models. While this was mentioned in the abstract and results (lines 146-148, 164-168), we recognise we have not been clear enough on this point, and in particular how we interpret it, in our discussion.

Regarding the suggestion that these results are stronger evidence of a 'low confidence' assessment than previous studies, we think our findings suggest that the model—observation comparison is subject to different issues - and therefore reveals different levels of confidence - on different timescales. Our results do indeed add to the evidence for issues representing decadal to multidecadal timescales, and thus do not lend us any more confidence in model processes on those timescales. That is an important result of our analysis, and we now highlight that further.

However, we also suggest that the longest timescales may be most relevant to anthropogenic trends on centennial timescales and therefore, our results may constitute evidence for model projections of centennial decline being more realistic than previously thought. Naturally, this is a tentative conclusion and greater confidence in this latter interpretation would require more years of data or different analysis.

We have

- Changed the title and the abstract to constitute a discussion of the implications of our results, rather than an unqualified statement of higher confidence.

- Emphasised the key new result that the assessment of model skill varies with the timescale considered e.g. lines 150-151, 158.

- Changed the final line of the Introduction (L57-59) to reflect the new framing of the paper, in particular the changing assessment of skill and the implications.

- Restructured and rewritten the discussion section to enable us to cover the points made by both reviewers.

    - No comment is made on ‘confidence’ until the fourth paragraph (L178-192)

    - We carefully discuss the implications of our results on different timescales (L178-192). We highlight the remaining low confidence at shorter timescales (L182-183)

    - We have removed the statement (formerly in the final paragraph) about confidence in attribution

    - The statement “we should now have some level of greater confidence” is amended to “**one interpretation may be that we can** have some level of greater confidence in projections of substantial **centennial** decline” at line 184-185, thus is more specific to long timescales and is more cautious.

    - We comment on the need for new data / analysis to gain confidence in the correct interpretation of our results (line 189-190)

[A very minor pedantic point. On line 114, strictly speaking this is "a failure to reject the null hypothesis" not an acceptance of the null hypothesis]

- We thank the reviewer for spotting this; it is corrected in the revised version.

Reviewer 2

I appreciate the authors reading and responding to my previous round of comments. Their responses fully satisfy some of my previous comments, including their addition of a new Fig B1 with time series plots and a more recent reference regarding the "two-timescale" response to stratospheric ozone forcing. However, they have not satisfied my main concern in any meaningful way.

Quoting my previous review: "However, I do not find the conclusions to be compelling. The issue is that although we often use linear trends to compare models with observations, the observed Antarctic sea ice area evolution looks strikingly unlike a linear trend plus any straightforward type of noise."

As the first panel in the new Fig B1 indicates, the observed Antarctic sea ice area evolution is characterized by a long and gradual expansion followed by abrupt contractions during the last several years. This does not resemble a linear trend plus any straightforward type of noise.

The authors are correct to note that although the ensemble mean CMIP6 result indicates steady decline, many of the individual models show more complicated or noisy behavior. But I do not think that this justifies comparing the linear trends between models and the observations. The authors defend their choice to use linear trends as follows: "This wide range of behaviours [in the individual CMIP6 model time series results] further justifies the linear trend metric; without a hypothesis of what shape we expect sea ice evolution to take, the linear metric is the best ‘first test’ of the data."

I respectfully disagree. For example, we do not normally fit the observed ENSO nino3.4 time series to a linear trend, nor would we use this to assess how well a GCM captures the observed ENSO variability. The reason we do not do this is that nino3.4 has variability that looks nothing like a linear trend in time.

On the other hand, under global warming a lot of climate indices do approximately resemble linear trends in time, and hence we match the observed trends with the modeled trends.

The observed Antarctic sea ice area roughly resembled a linear trend in time until about 2015, and many studies focused on its linear trend. However, it strikingly does not resemble a linear trend when the more recent years are included (see first panel of the new Fig B1).

Quoting my previous review again: "In my opinion, it is misleading to use the similarity between these linear trends to conclude that 'we should now have some level of greater confidence' in the model simulations and that therefore 'projections of substantial future Antarctic sea ice loss may be more reliable than previously thought' (quotes from the manuscript). I am not sure what meaningful information can be gleaned from noting this similarity in linear trends between the observations and GCMs during 1979-2023."

I see nothing in the reviewer responses to compel a change in this view. And it noteworthy that in the revised manuscript, the text that is quoted in my previous comment (above) remains unchanged.

We thank the reviewer for these insightful comments. We think there are two aspects to this point: whether the linear trend assessment is worthwhile, and how we interpret the results of this assessment. While we respectfully argue that we think the linear trend assessment does have value (see below), we agree that caution should be applied when interpreting the results, and that we did not address this topic sufficiently in previous versions of the manuscript. We thank the reviewer for raising a lot of points that have clarified our thinking on the issue and enabled us to improve the paper.

Our rationale is based in the underlying motivation for this study. In response to the changing Antarctic sea ice, one question we hear asked by both scientists and laypeople is ‘does this mean models aren’t so bad then?’ and this is the question we wish to approach: ‘Has the low sea ice area in recent years changed how we should judge models’? This is why the existing linear evaluation is our starting point. We are specifically interested in what the new years of observation tell us and how things have changed compared to previous influential studies.

It is, of course, possible that what the new observations tell us is that a linear model looks increasingly invalid for observations; this then is part of the answer to our question. This conceptual point was missing from our previous drafts; our answer should be not only about how the statistical results of the evaluation have changed, but also how appropriate that assessment is in light of the apparent step change of ice loss in the observational record. As long as this point is made clear, we believe that the linear trend evaluation is an important piece of evidence that needs to be reflected in the published literature.

Our detailed arguments are as follows:

We agree that linear trends are a crude metric in general (as stated in our first revision) and appear particularly so when taking into account the apparent step-change behaviour seen in observations (Purich and Doddridge 2023, Hobbs et al, 2024). Therefore, the recent abrupt change does make the application of a linear trend to the observations appear less valid. We are not convinced that linear trends are a sufficient metric of model skill and we do not recommend over-interpreting them without caution. However, there are several reasons why an updated view of linear trends is an important addition to the literature:

- Many studies have used linear trends in the past for Antarctic model evaluation. While we recognise that new data makes this approach seem less valid, these previous studies have made claims which have been impactful (e.g. informing IPCC statements). Therefore it is critical to re-examine those results, and therefore linear trends, in the light of recent changes,

but with appropriate discussion of the limitations that are now increasingly apparent. These papers include: Hobbs et al 2015 (fingerprinting technique using linear trends as the metric of change), Polvani and Smith 2013, Zunz et al 2013 (evaluate whether linear trends are consistent between models and observations in the light of modelled variability), Roach et al 2020, Shu et al 2020 (comparing modelled and observed linear trends in CMIP6); Rosenblum and Eisenman 2017 (comparing linear trends in both sea ice and in global mean temperature). Other papers (Schneider and Deser 2018, Zhang et al 2019) have used linear trends as their framing for examining the mismatch between observations and models.

- Many of the simulations (although not all, as noted in our original response) could be argued to appear consistent with a linear trend superimposed with noise. Therefore, it is a valid exercise to compare these modelled linear trends to observations since the predominance of these linear trends suggests that linear evolution is a reasonable hypothesis and moreover, it is the models we seek to test.

- We argue that an anthropogenically forced linear trend superimposed with a strong multidecadal variability is one plausible explanation for the observed behaviour, and thus that a regime shift is not the only interpretation. This interpretation- a trend superimposed with strong variability- is consistent with the findings of Zhang et al (2019) about the relevance of the phase of multidecadal variability for the trend over the satellite era. Moreover, the 1905-2020 sea ice reconstruction by Fogt *et al.* (2023) implies steady sea ice decline throughout the 20th century before the satellite era, which the 2016/17 sea ice area value was consistent with (Fogt *et al.* did not cover 2023). This is also supported by early satellite data, which suggest that the ice area was more variable in the 1960s (Gallaher et al, 2014; Meier et al, 2013) and dropped rapidly immediately before the onset of continuous coverage in 1979 (Zwally et al, 1983; Cavalieri et al, 2003). These results are consistent with the interpretation that the slight increase during 1979-2015 may have been an unusual phase of variability superimposed on a longer-term underlying forced trend of sea ice decline. In this case, evaluating linear trends on increasingly long timescales would capture more of that underlying forced trend. References:

    - Cavalieri, D. J., C. L. Parkinson, and K. Y. Vinnikov. 2003. 30-year satellite record reveals contrasting Arctic and Antarctic decadal sea ice variability. Geophysical Research Letters 30(18), doi: 10.1029/2003GL018031.
    - Gallaher, D. W., G. G. Campbell, and W. N. Meier. 2014. Anomalous variability in Antarctic sea ice extents during the 1960s with the use of Nimbus data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(3):881-887, doi: 10.1109/JSTARS.2013.2264391.
    - Meier, W. N., D. Gallaher, and G. G. Campbell. 2013. New estimates of Arctic and Antarctic sea ice extent during September 1964 from recovered Nimbus I satellite imagery. The Cryosphere 7(2):699-705, doi: 10.5194/tc-7-699-2013.
    - Zwally, H. J., J. C. Comiso, C. L. Parkinson, W. J. Campbell, F. D. Carsey, and P. Gloersen. 1983. Antarctic sea ice, 1973-1976: Satellite passive-microwave observations. Report No. NASA-SP-459. Washington, DC: National Aeronautics and Space Administration.
    - Fogt, R.L., Sleinkofer, A.M., Raphael, M.N. et al. A regime shift in seasonal total Antarctic sea ice extent in the twentieth century. Nat. Clim. Chang. 12, 54–62 (2022). https://doi.org/10.1038/s41558-021-01254-9

- This interpretation- that periods of slow change and periods of rapid change arise from the superposition of variability and forcing- is common in the literature for other time series, even including global mean surface temperature, where there is a well-known ‘hiatus-and-surge’ behaviour. While this hiatus and surge behaviour of global mean temperature is not nearly as dramatic as the sea ice evolution we are discussing here, we believe it is a tenable view that the same conceptual interpretation is plausible.

We therefore believe that we should examine what the linear trend assessment alone tells us (that there is apparently increasing agreement between the models and the observations) and then we and the community may decide how to interpret this conclusion. Emerging evidence for a regime shift in other studies may imply that the recent rapid sea ice loss means that a linear paradigm is not the best one. However, it is rather early to conclude this, and future studies establishing the nature and drivers of the recent change, and the role of multidecadal internal variability, should investigate this. For example, work is already underway elsewhere showing that models only very rarely capture sea ice changes as large as observed and generally only when they have variability biases (Diamond et al 2024).

On the reviewer's final point of confidence in projections, we believe it is a tenable view that the recent sea ice loss could be the first evidence that the ice decline predicted by models is starting to manifest and that the agreement of linear trends on the longest timescales we are able to assess suggests that the forced *centennial* trends in the models *may* be slightly more reliable than previously thought. While worth raising, however, we do think this statement requires extreme caution. First, it is one possible interpretation, and future analysis will be required to assess its veracity. Second, even if the forced centennial trend in the models were accurate, biases in modelled variability could still result in future sea ice evolution that looked quite unlike anything found in any CMIP6 models.

Edits to Paper

We have now edited the paper substantially to cover the above points. We have [line numbers refer to the edited version]

- Changed the title as noted above to 'New perspectives on the skill of modelled sea ice trends in light of recent Antarctic sea ice loss'
- Referred to the discussion about the implication of the linear trend in the abstract, such that this issue is foremost in the reader's mind.
- Removed the quote '…projections of substantial future Antarctic sea ice loss may be more reliable than previously thought' from the abstract
- L44-45: introduced our motivation for the paper as the question whether the new observational data should change our judgement of model skill.
- L45-46: highlighted, and given the motivation for, our use of linear trends
- L58-59: Added a sentence to indicate that there are limitations of linear trends
- L105-106 (Methodology): Commented on possible problem with linear trend assessment
- Extended the Discussion section significantly, and restructured it significantly from lines 178 onwards. The discussion from this point is now reframed in light of the question 'does the recent decline change our assessment of skill and confidence in the models?' and is structured as, first, a discussion of how we interpret the linear trend results at face value (incorporating response to comments by Reviewer 1); and second, a discussion of the justification of the linear trend metric and what the new observational data could imply about the interpretation of the linear trend metric evaluation.
- Added a new paragraph (L194-209) about linear trends, covering the following points
    - There is an apparent nonlinearity of the observed time series, and recent papers present evidence that this represents a regime shift (Purich and Doddridge, 2023; Hobbs et al, 2024).
    - The use of linear trends in previous assessments does require us to assess the changes that have occurred using the same metric, to see if the conclusions of those studies still hold.

    - (L199-202) Comparing models and observations is important for interpreting model output, and many simulations do look like a linear trend with noise, which justifies a comparison.
    - (L202-209) A regime shift is not the only interpretation, and multidecadal variability superimposed on a forced linear trend is an alternative one. The arguments of Zhang et al (2019) and also evidence for pre-satellite-era sea ice variability (Fogt et al, 2023; Gallaher et al, 2014; Meier et al, 2013; Cavalieri et al, 2003) are consistent with this interpretation.
- The statement “we should now have some level of greater confidence…” is now at line 184-185 and reads “one interpretation may be that we can have some level of greater confidence in projections of substantial centennial decline”. This is also in the first of our rewritten paragraphs, which we explicitly state is taking linear trends at face value, and is followed by caveats to this interpretation.
- Finally, the final paragraph expands on the interpretation of the linear trend assessment.
    - We state that the linear model looks less valid for observations in light of the recent rapid decline (L212) and that this means the linear trend evaluation results should not necessarily imply more confidence in models (L212-213)
    - We state that the rapid decline highlights new areas in which models and observations disagree and which should be the focus of future studies, and highlight the need for more work probing the nature of the recent change (L213-220)
    - We no longer conclude with a statement about increased confidence.