Response to reviewers:

We thank the reviewers again for their helpful comments. The response below contains all information from our online author comment, with additional clarification ***in bold italics*** *next to each 'intended adjustment' stating how exactly we have edited the manuscript. Line numbers refer to the tracked changes document.*

We thank the three reviewers for their helpful comments, and for highlighting several aspects where we needed to expand our analysis of the data to ensure the robustness of our results. We have conducted several further analyses – detailed below – and we feel that with this additional evidence we can satisfactorily address all the reviewer comments. We are happy to interact further with the reviewers on any of these points if needed.

Reviewer 1 Will Hobbs

The authors reassess the apparent disagreement between observed Antarctic sea ice trends and CMIP models, in light of recent extreme low sea ice events and the impact on the observed trend field. This is a valuable exercise that gives important context to the reliability (or otherwise!) of coupled models for the Antarctic climate system.

We thank the reviewer for their support and agree that this topic is important to our understanding of modelled Antarctic climate.

Unfortunately I think the execution needs to be improved before recommending publication:

1) the treatment and presentation of the literature on Antarctic SIA trends isn't really adequate, and this does have an impact on the interpretation of the results/discussion. It's not really clear from the opening paragraph whether the author's are claiming a discrepancy between modelled and obs trends to 2014, but the language implies that (e.g. "...consistent with observations."). That's not really true in the literature – at least not for total SIA, when the model internal variability is properly accounted for (zunz et al 2013, polvani and smith, 2013) – it's only when spatial trends were considered that the model trends were incompatible with obs (Hobbs et al 2015). I think the Intro needs a very clear statement from the authors about what they mean by agreement with obs (also in the Discussion section), and a bit more nuanced outline of the literature to-date.

We deliberately kept the introduction short, as there is a wealth of literature on this topic and for a Brief Communication there is a limitation on how much we can cover. Specifically on the topic of whether a discrepancy exists between modelled and observed historical trends, this is a nuanced topic and the conclusion drawn depends upon the philosophy of the analysis method used. Irrespective of the earlier studies, however, our methodology shows that for trends calculated up to end dates between 2011 and 2021, there is a discrepancy in overall SIA trends between models and observations (see below).

In retrospect we agree that a much clearer discussion of previous studies is warranted, and we will expand the introduction and discussion in the revised manuscript. While some studies have found modelled trends to be consistent with observations, those which conclude this most strongly tend to be those which had much less satellite data available. For example, the three studies cited above relied upon historical simulations in CMIP5 models and therefore used data to 2005. This increases the chances of model—observation agreement because the shorter period i) increases the spread of model trends (due to internal variability) and ii) did not capture the strongest observed positive trends. Figure R1 below shows that for trend end dates up to 2010 inclusive, we would accept the null hypothesis that the observed trend could be drawn from the model distribution of trends. Thus our results are consistent with the earlier studies. In the revised paper we will discuss the earlier studies in the context of the changing trend duration.

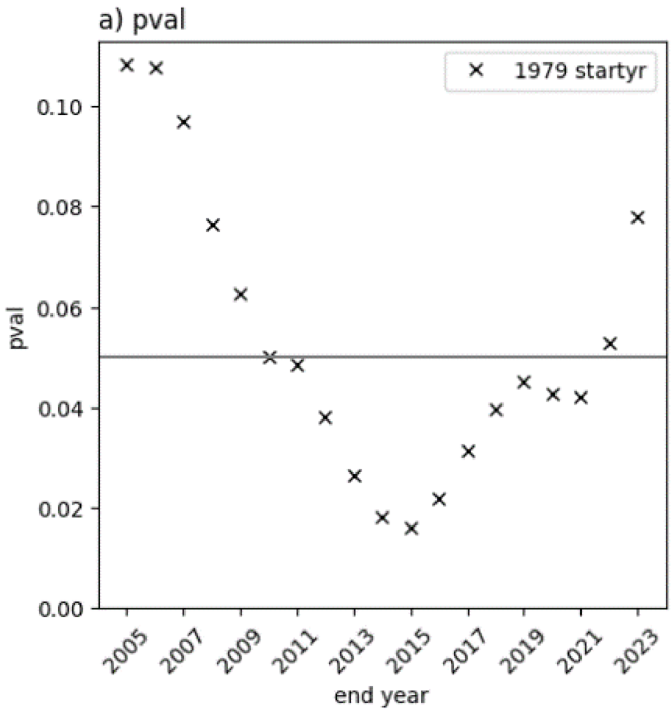


Fig R1: Analysis as in paper but with end dates as early as 2005 (for fixed start date trends only, since we cannot calculate a 35-year trend in observations ending before 2008).

INTENDED ADJUSTMENTS:

- Add to the introduction that 'Some studies found that the observed pan-Antarctic trends lay within the distribution of modelled trends (Polvani and Smith, 2013; Zunz et al,2013) and that only regional trends could be concluded to differ (Hobbs et al, 2015). However, these studies considered data to 2005 only and over this much shorter 27-year period the role of internal variability is larger than with later end dates. ' ***This text has been added at lines 23-26.***
- We will add values for 2005 through 2012 end dates (as in Fig R1) to manuscript Fig 1c, showing that our results show i) consistency between models and observations prior to 2011 (consistent with the papers referenced above) but ii) a clear model vs observation discrepancy for end dates between 2011 and 2021. ***We have actioned this point with an updated panel 1d) and an update to Appendix Figure C1 where the contribution of observed and modelled changes is examined, and with a new histogram showing 2005 end date in Figure 1a).*** We think this additional analysis substantially improves the paper, as it adds further nuance to the temporally evolving conclusion that model trends disagree with observation, and also clarifies the relationship between the present paper and previous published results. We will state in the final paragraph of the introduction that we consider end dates from 2005. ***We have added a new sentence to the end of the Introduction (L51-54) which states this point and also clarifies our goals.***
- Rephrase the final paragraph of the introduction to explicitly clarify what we mean by 'consistent with observations': "we consider whether the distribution of trends simulated by the

CMIP6 models allows for a trend of the observed magnitude, and thus whether observed trends are consistent with the multi-model ensemble.” ***This sentence has been added.***

- Reflect the above three points in the revised Discussion, emphasising clearly that the comparison of observed trends to the model ensemble of trends varies with time (as internal variability reduces due to increasing trend length, and the evolution of the real Antarctic sea ice changes becomes apparent) and placing this clearly in the context of previously published results. ***We have added mention of early end dates to the first sentence of the introduction and extended Results and Discussion sections to include all three periods.***

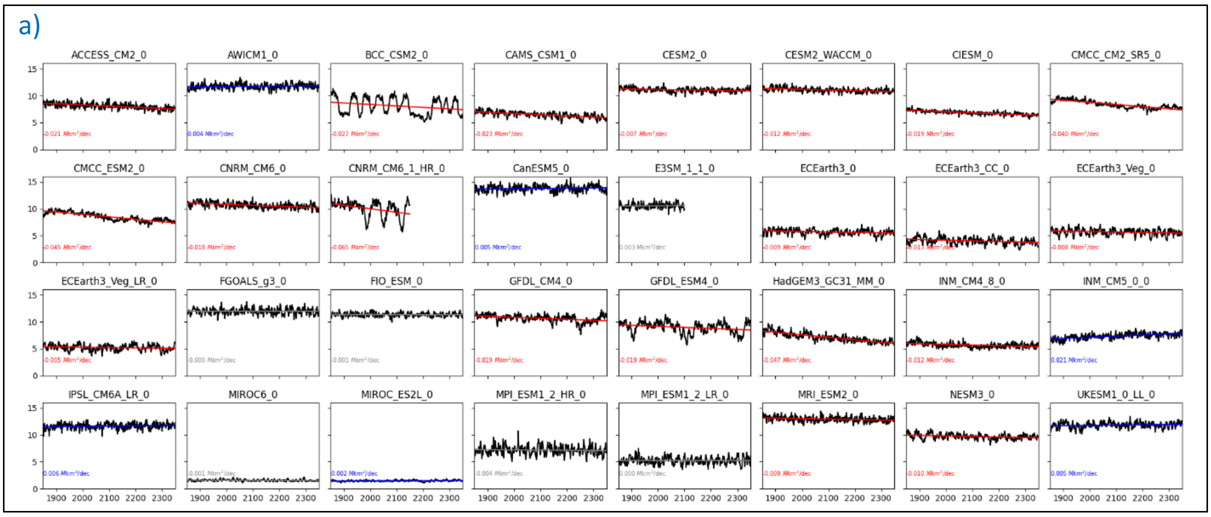
My biggest concern is that I don’t think the handling of the CMIP6 data is adequate.

2a) Firstly there is no drift correction which is essential for dealing the historical simulation trends. Most models will have little drift in SIA but some on the list (e.g. MIROC) are known to have quite large drifts. I think at the very least proof from the piControl experiments that spurious trends in SIA are small is required.

We agree it is worth examining pre-industrial drift. We have calculated the linear trend over the full pre-industrial period available (for the 32 models with data available), henceforward referred to as ‘drift’. In the UHH dataset considered these periods range from 150 to 500 years in length. Figure R2a displays time series of pre-industrial SIA with linear trends and their statistical significance indicated.

In all cases, drifts are an order of magnitude smaller than trends for years 1979-2023 (Fig R2b, note the different scales on the axes with the one-to-one line shown for clarity) so we conclude that drifts are negligible for our study. This can also be seen in that the corrected satellite era trends (Fig R2b, pink crosses) are only slightly modified. That drifts are negligible in this context is consistent with findings for CMIP5 (Gupta et al., 2013). We also note there is no significant inter-model relationship between the drift in a model’s pre-industrial simulation and the ensemble mean of linear trends in that model (p=0.48).

INTENDED ADJUSTMENTS: We intend to include the pre-industrial trend values in the data table in the paper, and state in the paper that there is no evidence that drift is impacting the historical trends, referencing these values. ***We have added the pre-industrial linear trends to the table (Table B1), and stated our conclusion (that drifts are not affecting our results) at the end of section ‘2.2 Model Data’.***



b)

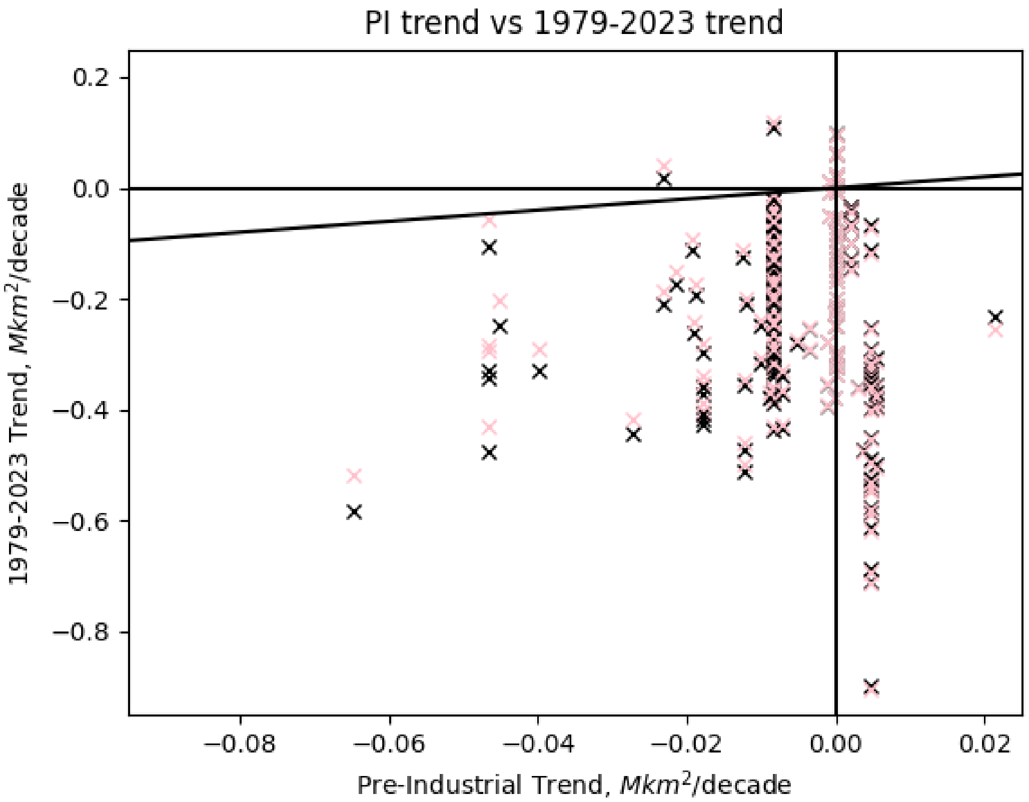


Fig R2: Pre-industrial trends and their relationship to historical era trends. a): Pre-industrial time series for each model analysed, with trend indicated in grey (not statistically significant), red (statistically significant reduction), or blue (statistically significant increase). Models are sorted alphabetically. b) Scatter plot of the pre-industrial trend for each model (x-axis) with the 1979-2023 trend for all ensemble members for the same model (y-axis, scale is ten times than on x-axis), with one-to-one line in black. Pink crosses indicate the 1979-2023 trend with the pre-industrial trend removed.

2b) internal variability in the models isn't properly dealt with. The method used implicitly assumes that the models all have similar internal variabilities but this assumption isn't stated and isn't really valid – some models (e.g. GFDL) have some pretty large multidecadal internal variability (Zhang et al 2018) that differs greatly from other models. Even by truncating the max contribution of each model to 6 ensemble members, there's still a weighting towards those models with more members.

We agree that the use of different numbers of ensemble members assumes the models have similar internal variability. However, when only the first ensemble member of each model is used, the conclusions are unchanged (Compare Figure R3a with R3b). We also tested the robustness of this result by randomly resampling the single ensemble member from each model to be used (Figure R3c) and the mean p-value for a 2023 end date is very similar to our original result. The result is marginally altered for earlier end dates, but this is because we have substantially reduced the size of the ensemble; the ensembles with one member per model have N=39 while the full ensemble (up to 6 members per model) has N=98. We choose to use the larger ensemble (up to 6 members per model) because in this way we have many more samples of internal variability.

INTENDED ADJUSTMENTS:

- We will modify the paper to state that the use of up to 6 ensemble members was aimed to maximise sampling of internal variability. ***We have added a sentence to lines 75-76 in the section 'Model Data' that makes this point clearer.*** However, this does lead to uneven sampling of models, which have different internal variability; we have therefore checked that if we randomly resample one ensemble member per model, results remain on average the same for 2023 end dates. ***This discussion has been added to the Appendix under 'Sensitivity to Treatment of CMIP6 models', L257-260.*** We will add Fig R3c to the existing appendix 'Sensitivity Tests'. ***This has been done and is panel d) "pval resampled first member" of new figure Figure C2.***
- At the point in the discussion where we state that a discrepancy in trends could be due to different forced trends or different variability, we will explicitly state that "Indeed modelled variability has been shown to exceed observed variability (Zunz et al 2013, Roach et al 2020), although these results were before recent increases in the observed variability (Hobbs et al, 2024). In addition, modelled variability varies strongly between models (e.g. Roach et al, 2020), with some models containing large centennial variability (Zhang et al, 2018)"
    - Hobbs, W., and Coauthors, 2024: Observational Evidence for a Regime Shift in Summer Antarctic Sea Ice. J. Climate, 37, 2263–2275, https://doi.org/10.1175/JCLI-D-23-0479.1.
    - ***A more succinct version of this text about modelled and observed variability has been added to the discussion (L177-179)***

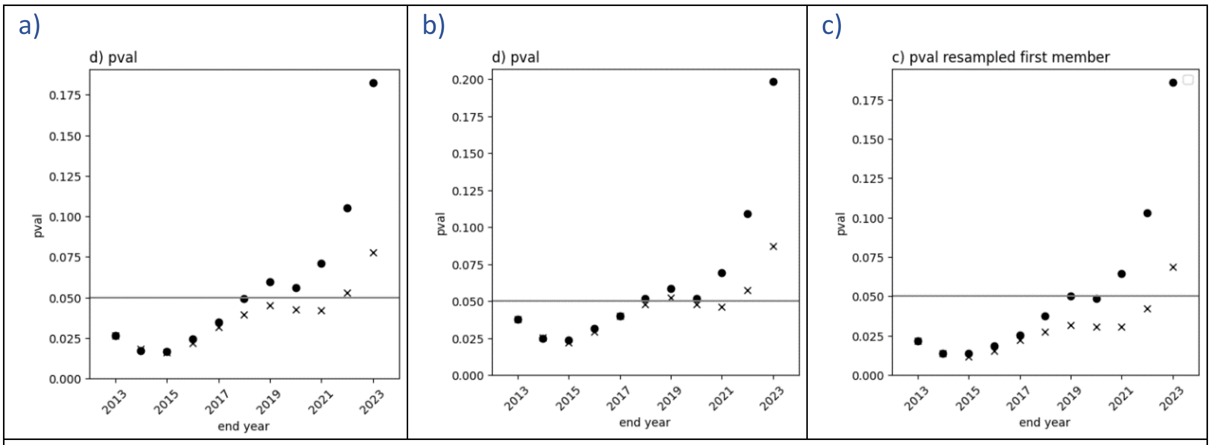


Figure R3: Sensitivity to truncating contribution of each model. LH panel: First 6 ensemble members from each model (version in original submission). Middle panel: First ensemble member from each model. Right panel: 1 random ensemble member from each model, resampled 10000 times; mean of pvals.

As for the model drift correction, interrogating the piControl experiments is the correct way to represent modelled internal variability.

The piControl timeseries (Fig R2a) demonstrate that multidecadal to centennial variability in some models is large, which will impact historical trends (as discussed by Zhang et al, 2018, for example, for the CMIP5 GFDL model). However, this is already accounted for in our interpretation and conclusions. The internal variability is present within the historical simulations analysed, and our ensemble of simulations considered is sufficiently large (N=98) that the range of internal variability is well-sampled across the ensemble. Our aim in this paper is simply to analyse whether or not the modelled and observed time series, as quantified by a linear trend, are consistent. We compare the observed trends to the range of modelled trends, which includes the role of internal variability and forced responses and associated model structural uncertainty. Furthermore, with this methodology we are considering the internal variability that occurs under the influence of present-day anthropogenic forcing, which would not be the case if we were studying piControl runs.

INTENDED ADJUSTMENTS:

- State explicitly in methods that we are using the 98 members of the historical multi-model ensemble to sample internal variability under the influence of anthropogenic forcing, consistent with previous studies. ***A comment to this effect has been added to methods, lines 75-76.***

The existing discussion section explicitly states that within our methodology, a mismatch in observed and modelled trends can be due to a difference in forced trend or a difference in internal variability. However, the results section does contain two explicit references to modelled forced anthropogenic trends, before the clarifying 'discussion' section about the relative roles of variability and forcing in any discrepancy. Therefore we will remove these references from the results section to clarify that the role of internal variability is accounted for in our methodology. ***We have removed reference to forced trends at lines 125-128.***

Aside from internal variability, quite a few models have ice-free summers for the period of interest (Roach et al 2020) which is obviously going to impact their trends (no ice = no trend)

The influence of mean state biases is investigated in Figure R4, which plots model sea ice area (SIA) trends against model SIA climatology. Figure R4a confirms that there is a weak but highly statistically significant relationship (r^2=0.24, p-value=3E-7, slope= -0.08 decade^{-1}) between summer SIA climatology and annual-mean SIA trend. Therefore, consistent with the reviewer's comments, the models with low summer climatologies are more likely to have less SIA loss. However, this effect is weak on average (the regression has a shallow slope) and the relationship explains only a small proportion of the variance in trends.

We also note that there is no reason in principle why a model with a low summer SIA could not have high annual-mean SIA and therefore a strong trend of annual-mean SIA loss. Since we consider annual-mean trends, we believe that assessing the annual-mean SIA climatology is more pertinent than assessing the summer SIA climatology.

Considering annual-mean SIA, there is again a statistically significant relationship (r^2=0.32,p-value=1E-9, slope=-0.04 decade^{-1}) between climatology and trend (Fig R4b). However, this relationship is even less influential than in summer (the slope is shallower). We observe that a cluster of simulations of the MIROC model variants (MIROC6 and MIROC-E2SL) are a clear outlier, and in fact have less annual mean climatology than the observed summer climatology. There is precedent for excluding these models in the literature (Shu et al, 2020) so we conduct the sensitivity test of removing this data and repeating our analysis. Doing so does not change our conclusion (Fig R4c) that the recent observed rapid reductions bring modelled and observed trends into line for a 2023 end date.

A stricter cutoff might change our results, but it is not clear to us how such a cutoff would be robustly selected. After the MIROC simulations are removed, the model annual-mean climatologies are evenly spread around the observed value in Fig R4b. It is also hard to select a robust cutoff based on summer climatology. Unlike in the Arctic, the observed summer SIA is so low (~2 Mkm^2) that selecting a definition of 'ice free' would require us to exclude models based on a very small absolute error in their mean state. Finally, the underlying philosophy of this paper was to consider whether an analysis of trends alone, incorporating recent observations, leads to different conclusions, and so placing too many conditions on the analysis would draw us away from this goal.

INTENDED ADJUSTMENTS:

- Add Fig R4 to existing Appendix ‘Sensitivity Tests’. ***We have added a new Figure; it is the new Figure C2 panels a) to c).***
- Add to the discussion that there is a weak relationship between summer, or annual-mean, climatology and trends, which is to be expected as very low sea ice constrains trends (though this may not be the only mechanism for the relationship; Holmes et al, 2022). Therefore the models with trends close to observations tend to be biased low in climatology (Fig R4a, R4b). However it is not clear how to robustly choose a cutoff for excluding biased models. Excluding MIROC models which are clear outliers and biased low year-round, as in Shu et al (2020), does not change our conclusion that trends are consistent for an end date of 2023. ***We have added a section on the relationship between climatology and trends to our results section*****, L143-151.**

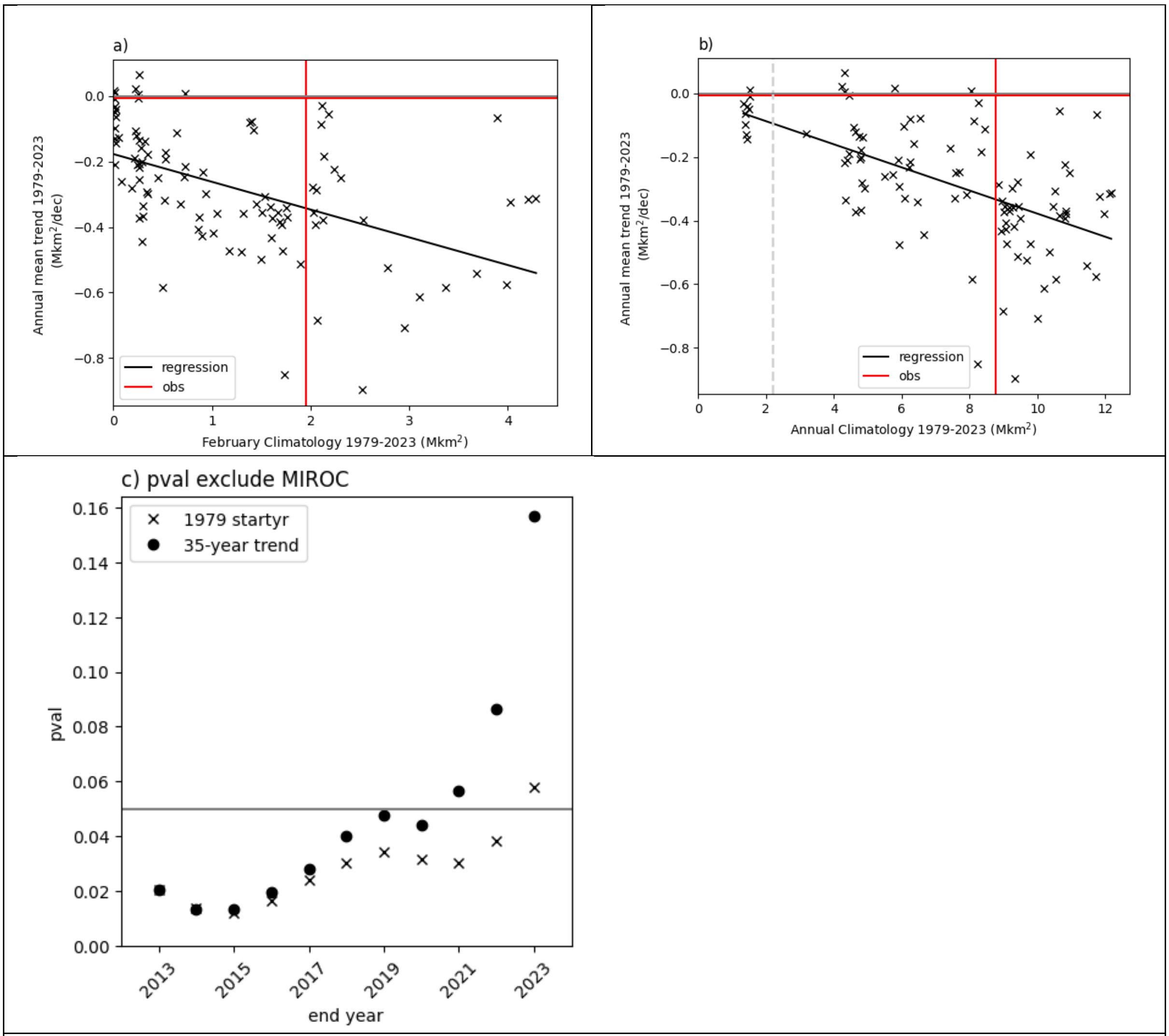


Figure R4: The role of ice-free conditions in explaining inter-model trend spread. a) Scatter plot of summer (February) sea ice climatology for 1979-2023 against the annual-mean trend over 1979-2023. Maximum 6 ensemble members per model shown. b) as a) but for annual mean climatology against trend, with cutoff to exclude MIROC models indicated c) p-value analysis from original manuscript repeated but excluding MIROC models (equivalently, cutoff of annual mean observed climatology/4.)

3) obs - I assume that the 'synthetic' extension of the obs record to end of 2023 is just a placeholder, and that the actual data will be used before publication?

This has now been updated. The observed annual mean is the same to 2 decimal places as the predicted, so this does not affect our results.

Updating Figure 1b to end in 2023 instead of 2022 demonstrates that the annual-mean trend observed is now very weakly negative.

Using observed data visibly alters the monthly trends for October and December in Figure B1 but does not affect the text conclusions based on this figure.

Predicted areas: Oct: 12.7 Mkm2, Nov: 10.3 Mkm2, Dec: 5.6 Mkm2. Observed: 12.5 (less than predicted), 10.3, 5.8 (more than predicted).

INTENDED ADJUSTMENTS:

- Figure 1b (trend histogram) will be replaced with version for 1979-2023. Observed trend now weakly negative. ***This has been done, in what is now Figure 1c), and is mentioned in the text at line 110.***
- Figures 1c, C1a, C1d will be updated by replacing predicted value (grey) with observed value (black). (Results are unchanged). ***This has been done.***
- Text references to extension will be removed. ***This has been done.***
- Figure B1 will be updated by replacing 2022 value for OND (faded) with 2023 value (same formatting as rest of plot). ***This has been done.***
- ***We have also updated the comment in the Introduction about record months in 2023 'so far' to be a statement that reflects all of 2023 (L37)***

References suggested by reviewer:

Hobbs, W. R., N. L. Bindoff, and M. N. Raphael, 2015: New Perspectives on Observed and Simulated Antarctic Sea Ice Extent Trends Using Optimal Fingerprinting Techniques. J Climate, **28,** 1543-1560, 10.1175/JCLI-D-14-00367.1.

Polvani, L. M., and K. L. Smith, 2013: Can natural variability explain observed Antarctic sea ice trends? New modeling evidence from CMIP5Geophysical Research Letters, **40,** 3195-3199, 10.1002/grl.50578.

Zhang, L., T. Delworth, W. Cooke, and X. Yang, 2018: Natural variability of Southern Ocean convection as a driver of observed climate trends. Nature Clim. Change, 10.1038/s41558-018-0350-3.

Zunz, V., H. Goosse, and F. Massonnet, 2013: How does internal variability influence the ability of CMIP5 models to reproduce the recent trend in Southern Ocean sea ice extent? The Cryosphere, **7,** 451-468, 10.5194/tc-7-451-2013.

Citations to Hobbs et al (2015), Polvani and Smith (2013) and Zhang et al (2018) have been added, and additional reference to Zunz et al (2013) has been made.

Anonymous Reviewer 2

Holmes et al. compare observed Antarctic sea ice area trend with the trends in CMIP6 simulations. The authors find that during 1979-2018, the models are not consistent with the observations, as has been widely noted previously. This is because the Antarctic sea ice area tends to steadily decline in response to anthropogenic forcing in GCMs, whereas the observed Antarctic sea ice area increased during the first 35 years or so of the satellite record (roughly 1979-2015). However, strikingly low Antarctic sea ice areas were observed in 2022 and 2023. Due to this, the authors find that when the time period over which the trends are computed is extended to 1979-2023, the models and observations are fairly consistent. The authors conclude that this gives a greater level of confidence in the CMIP6 Antarctic sea ice simulations than previously though.

The manuscript is clearly and concisely written, and the presentation is polished.

We thank the reviewer for their comments and we are glad they found the manuscript concise and readable.

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. This can be readily seen by looking at any time series of observed Antarctic sea ice area or extent (e.g., https://zacklabe.files.wordpress.com/2023/12/nsidc_sie_timeseries_ant_anomalies-5.png). I do not see any meaningful similarity between the steady long-term decline of Antarctic sea ice area in typical CMIP6 simulations (e.g., Historical and then SSP5-8.5 simulations plotted in Fig 4c,d of Roach et al. 2020) and the observed gradual expansion followed by a short-lived abrupt loss during the past two years, even if both time series have similar OLS linear trends.

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.

To this end, the authors do not include any plots of actual time series, only reporting the OLS linear trends, so a reader unfamiliar with the observations and simulations may be substantially misled by this manuscript. (I don't mean to imply that the authors are being intentionally misleading in any way.)

We agree with the reviewer that “the Antarctic sea ice area evolution looks strikingly unlike a linear trend plus any straightforward type of noise” and they make a valid point that we did not present any of the modelled timeseries in the paper. However, on examining the model data we disagree that the typical modelled behaviour is ‘steady decline’. The steady decline that is most obvious in the Roach et al (2020) figure cited is the multi-model mean, which averages out the internal variability and so reflects the gradual forced anthropogenic trend. If we examine all the ensemble members individually (Figure R5), a linear decline is not in general the most prominent feature of variability. Instead the individual realisations have a wide variety of different variability, including some with limited trend followed by rapid decline (e.g. CNRM_CM6_4, HadGEM3_GC31_LL_1), as seen in observations.

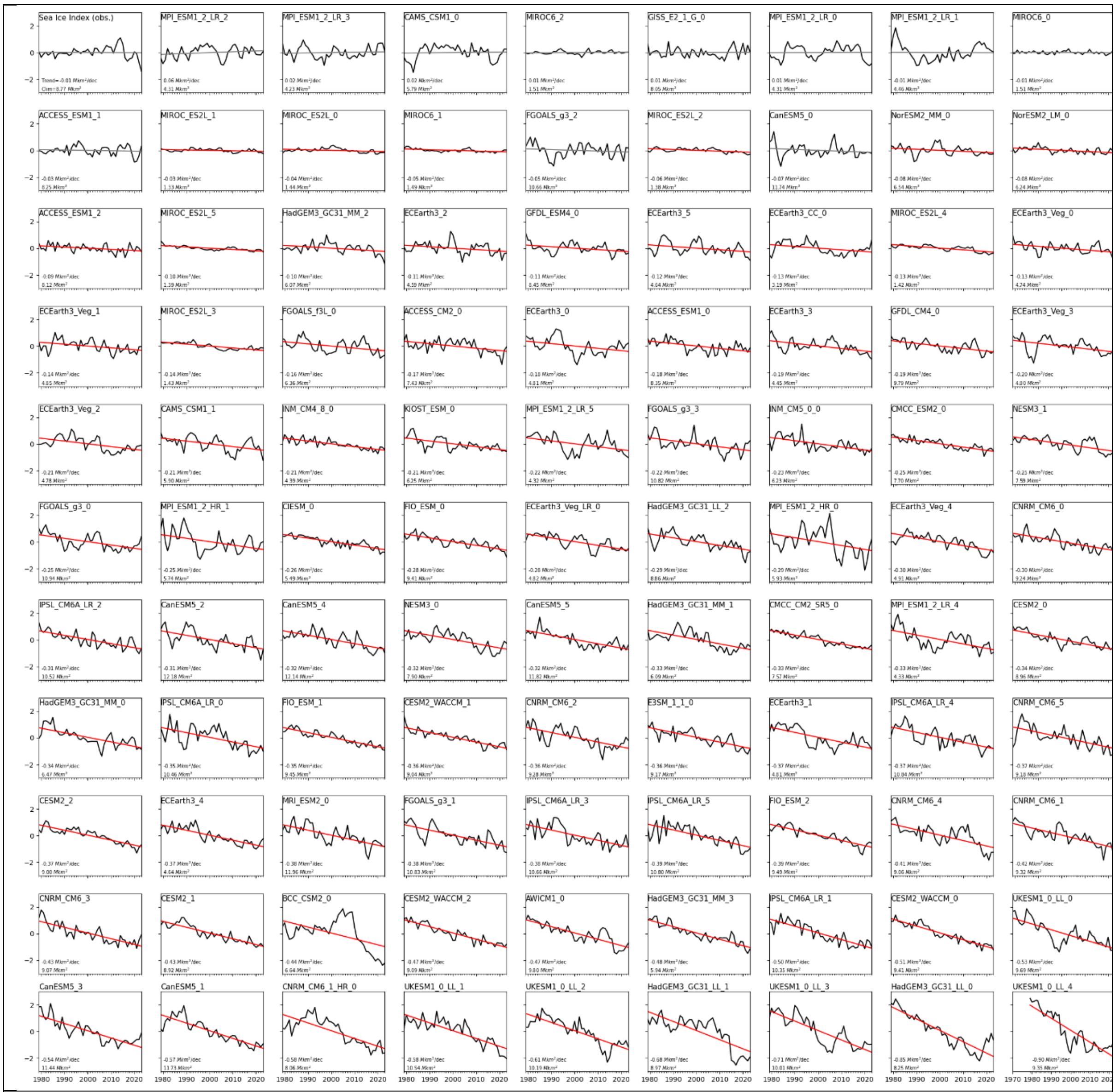


Figure R5: 1979-2023 annual mean sea ice area in observations (Sea Ice Index v3, top left) and in all CMIP6 model ensemble members considered in the analysis, sorted by their linear trend over this period. Linear trends are shown and indicated in red (significant at $p<0.05$) or grey (insignificant). Each panel includes annotation showing the simulation's 1979-2023 climatology and trend.

This wide range of behaviours 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. We use it openly as a simple 'top level' metric, which is often used, and our over-arching purpose is to demonstrate that model-observation consistency 'cannot be ruled out on trends alone'. Our argument is that IF we only consider linear trends of annual-mean ice area THEN observations recently came into line with models. We think that with appropriate discussion this is a valid and noteworthy conclusion.

INTENDED ADJUSTMENTS:

- Add a note to the discussion describing that an OLS linear trend is a limited parametric assessment of a timeseries, and the observed time series does not particularly resemble a linear trend with noise imposed. However, while the multimodel mean resembles smooth decline (e.g. Roach et al 2020), individual members display complex behaviours which more clearly show the interplay of trends and variability. Therefore, while the linear metric gives a first indicator of model performance, this further highlights the need to probe higher-order or non-parametric characteristics of modelled sea ice variability to investigate the models' fidelity. ***Two sentences have been added to the discussion at lines 192-196 as part of the discussion about different ways in which models could be assessed.***
- Add an Appendix 'Timeseries' which will contain Figure R5. ***The timeseries have been added to the paper (Appendix) as the new Figure B1, and are referenced both in the Methods section when the Data is described, and in the discussion about the use of linear trends.***

Furthermore, the authors mention that the observations are consistent with "the 'two-timescale' response to stratospheric ozone forcing whereby increasing westerlies cause a sea ice increase on 'short' timescales and decline on 'long' timescales (Ferreira et al., 2015; Kostov et al., 2017)." It should be emphasized that a later study that included some of the same authors (Seviour et al. 2019, doi:10.1175/JCLI-D-19-0109.1) concluded that this ozone forcing mechanism is unlikely to be a primary driver of the mismatch between the Southern Ocean surface cooling (and Antarctic sea ice expansion) in observations and the Southern Ocean surface warming (and Antarctic sea ice retreat) simulated by GCMs.

We agree that we need to edit the text here. We have re-visited the Seviour et al. (2019) manuscript, which indeed shows that (based on relationships between model climatology and the cooling and warming phases of the SST response), the two-timescale response to ozone depletion is unlikely to be a primary driver of the model-observations mismatch. We do not think this precludes our mentioning this as a possible hypothesis, but it merits a more nuanced discussion.

INTENDED ADJUSTMENTS: We will add a note to the manuscript to confirm that this is a possible hypothesis and referencing the Seviour et al. (2019) paper, and clarify 'models under-estimating the 'two timescale' response… " to 'models under-estimating the timescale or magnitude of the cooling phase of the two timescale response'***. We have done this, with a sentence stating that Seviour et al argue against this mechanism as a primary driver (L183).***

Anonymous Referee 3:

Hobbs et al. compared trends of observed Antarctic sea ice with those of CMIP6 simulations. This investigation sheds light on the model's performance and its ability to capture and reflect the dynamic changes in Antarctic sea ice in the face of recent, dramatic declines.

The authors find that between 1979-2018, the model did not compare well with the sea-ice observations, since the SIE/SIA increased over the first 35-years and the model concurrently showed a decline. However, when the analysis was extended to include the 2022 and 2023 lows (encompassing the full satellite record) a more favorable alignment between the model and observations emerged. They concluded that this was potentially due to being able to assess the trend over a longer time period, which gives a greater confidence to CMIP6 simulations.

Overall, I am not 100 % convinced with the authors' conclusions. The CMIP6 model has consistently underestimated the SIE and SIA in the Antarctic, and thus did not reproduce the 1979-2014 sea-ice trends. Although the model seems to align better with recent data showing a decline in sea ice trends, especially in 2022 and 2033, I find the authors' argument insufficient to establish the model's reliability and instill a higher level of confidence in its capabilities.

Based on these comments and the one below, we believe it is possible that the reviewer may have misinterpreted the data and our aims in the paper. This highlights that we may need to be clearer in the manuscript. CMIP6 is not 'a model' but is instead an ensemble of coupled climate models. In this study, we are considering the results of 98 model simulations from 39 models.

INTENDED ADJUSTMENTS: We will add to the text at line 58 to make the nature of CMIP6 more explicit (mentioning our analysis constitutes 39 models from multiple modelling centres with multiple different model components and resolutions).

We have done this.

It would be beneficial for the authors to provide a more detailed account of the enhancements made to CMIP6 that contribute to its improved simulation of sea-ice trends, particularly when considering the extended satellite record, and recent sea-ice lows. Given the well-documented temporal and regional variations in Antarctic sea ice, where trends and regions often exhibit stark differences, a more comprehensive explanation of the model modifications would help readers better understand the factors leading to its purported improvement in capturing these complexities.

We are not arguing that the CMIP6 models have improved over previous generations of models. Rather, we are arguing that the recent changes in the observations affect our assessment of the models, and that if we use annual-mean ice area trends as a metric, the CMIP6 models now do not disagree with observations. Therefore, a discussion of model enhancements is not relevant.

Of course we still agree that there are many other metrics for which CMIP6 sea ice area does disagree with observations. The manuscript already discusses in detail the disagreement of trends over the 1979-2014 period, as well as other variables on which modelled sea ice may be assessed and found to disagree (penultimate paragraph of the discussion in the original submission).

INTENDED ADJUSTMENTS: We will add text to line 46 to clarify we are not examining a changed set of models but rather our interpretation in the light of changing observations.

We have added to the last line of the Introduction the text ‘while using a consistent set of model data (such that changes are not attributable to changes in model components or resolution)’.

Specific comments:

Line 54: Figure B1 is referenced here and in two other places, but there is only a Table B2 in the Appendix B.

We thank the reviewer for spotting this. Table B2 is labelled in error in Appendix B; we will correct this to ‘Figure B1’. ***We have now reordered the appendices and introduced new figures, but have carefully checked that all new references are correct.***

Lines 81-82: This sentence is a quite confusing and convoluted.

We will rephrase this sentence.

We have changed ‘To estimate the probability that a trend at least as large as observed would occur in the climate model population, the p-value for a one-tailed test is 1-F(x) where x is the observed trend.” to ‘… we calculate the p-value for a one-tailed test as 1-F(x), where x is the observed trend’

Line 105: I would suggest changing the sentence to “This makes it less likely that the observed...”

We will change this as suggested.

We have added the word ‘that’ as suggested.