Response to Reviewer 1 (Will Hobbs)- Reviewer comments in black, author comments in blue

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.

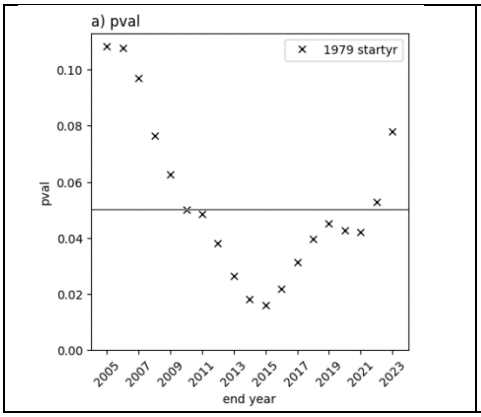


Fig R1: Analysis as in paper but with end dates as early as 2005 (for fixed start date trends only, since we cannot calculate a 35-year trend in observations ending before 2008).

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. '
- 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 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.
- 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."
- 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.

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

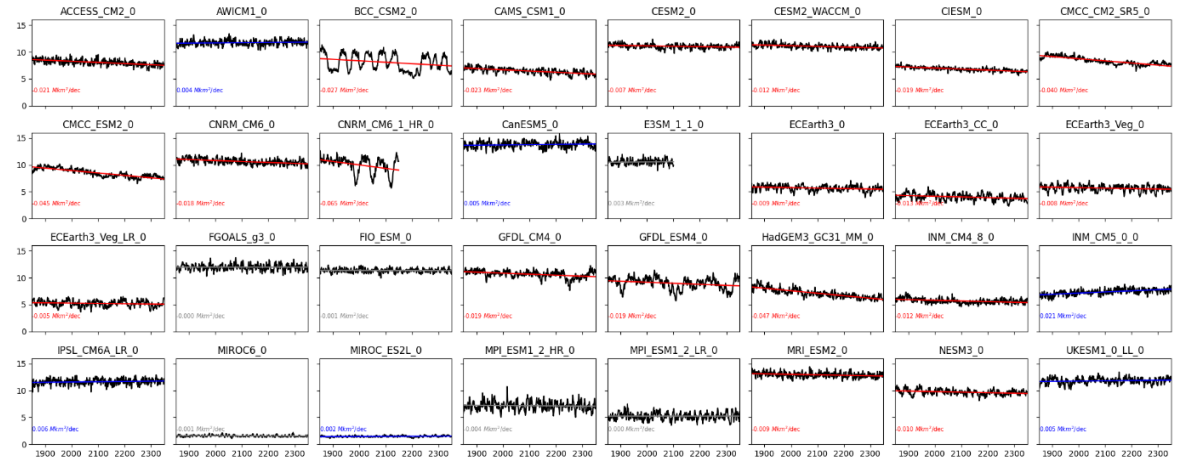
2a) Firstly there is no drift correction which is essential for dealing the historical simulation trends. Most models will have little drift in SIA but some on the list (e.g. MIROC) are known to have quite large drifts. I think at the very least proof from the piControl experiments that spurious trends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

a)



b)

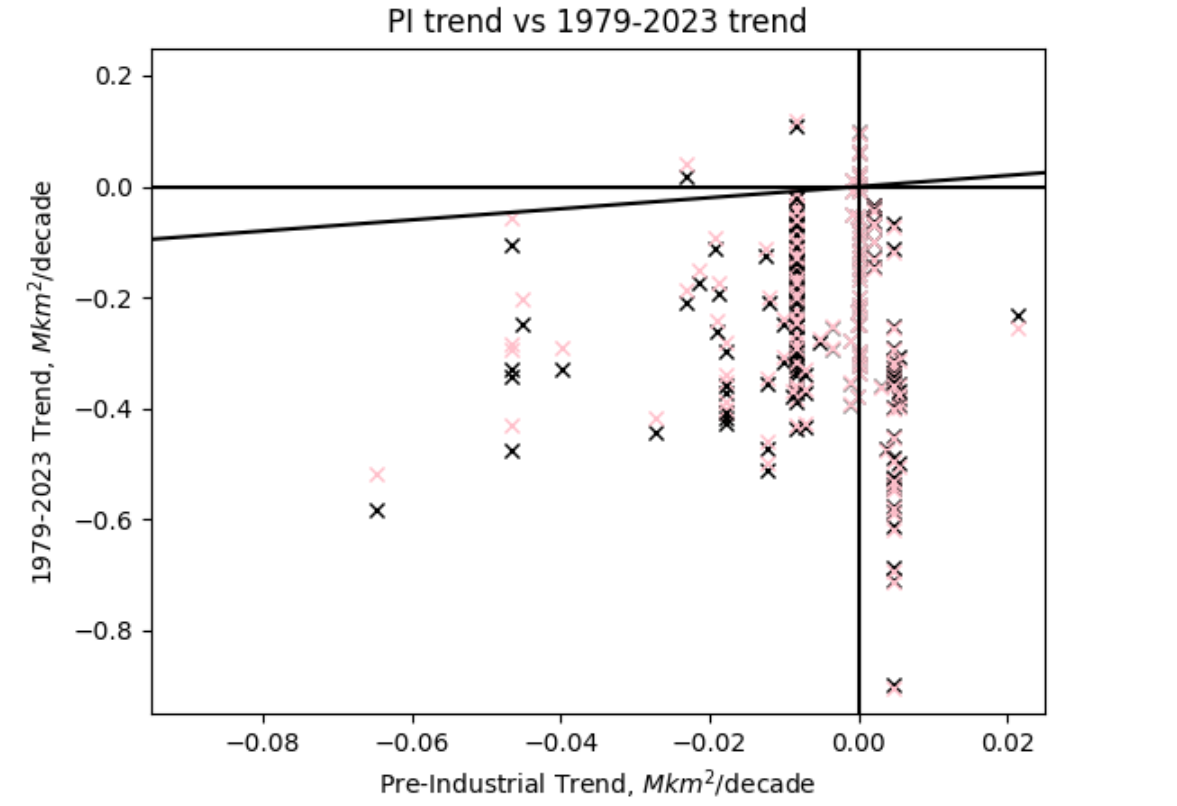


Fig R2: Pre-industrial trends and their relationship to historical era trends. a): Pre-industrial time series for each model analysed, with trend indicated in grey (not statistically significant), red (statistically significant reduction), or blue (statistically significant increase). Models are sorted 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. 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. We will add Fig R3c to the existing appendix 'Sensitivity Tests'.
- 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.

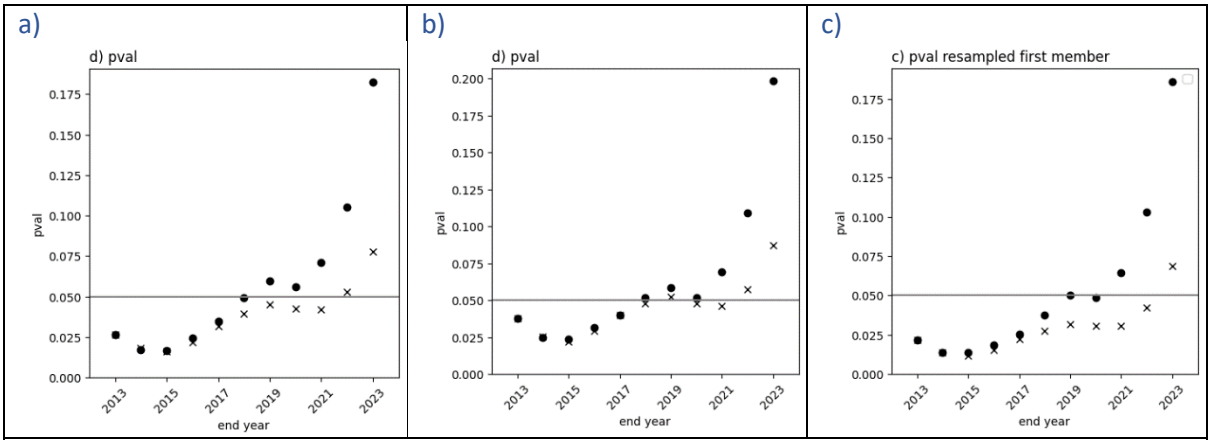


Figure R3: Sensitivity to truncating contribution of each model. LH panel: First 6 ensemble members from each model (version in original submission). Middle panel: First ensemble member from each model. Right panel: 1 random ensemble member from each model, resampled 10000 times; mean of pvals.

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.
- 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.

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'
- 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.

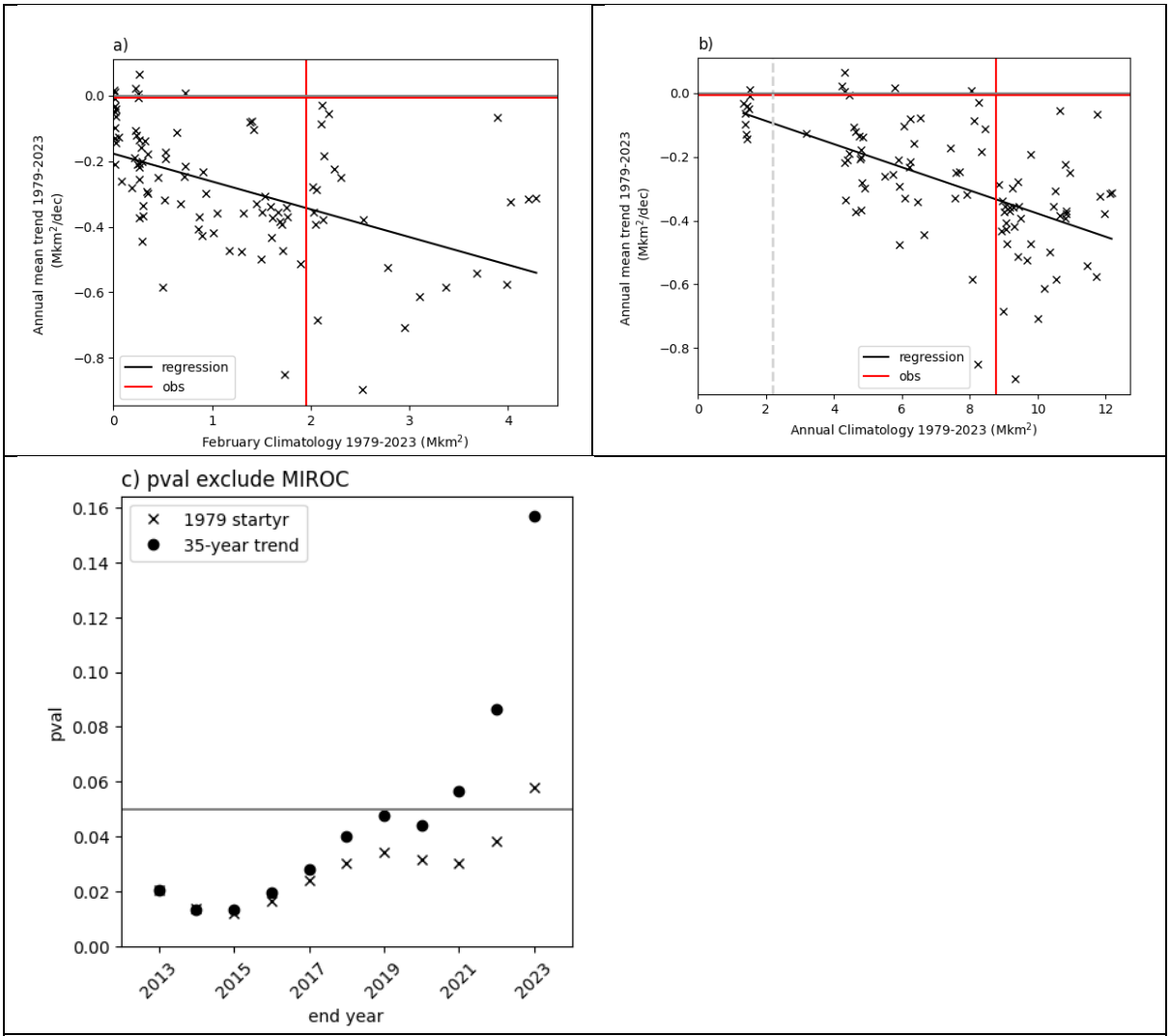


Figure R4: The role of ice-free conditions in explaining inter-model trend spread. a) Scatter plot of summer (February) sea ice climatology for 1979-2023 against the annual-mean trend over 1979-2023. Maximum 6 ensemble members per model shown. b) as a) but for annual mean climatology against trend, with 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.
- Figures 1c, C1a, C1d will be updated by replacing predicted value (grey) with observed value (black). (Results are unchanged).
- Text references to extension will be removed.
- Figure B1 will be updated by replacing 2022 value for OND (faded) with 2023 value (same formatting as rest of plot)