Reviewer comments, Anonymous referee #2, 25th September 2024

I appreciate the authors' thoughtful responses to the second round of reviewer comments. However, although I am happy to defer to the editor and/or other reviewers, my own assessment is that I remain unconvinced that this work is publishable in its current form.

As in my previous two reviews of this manuscript, my main concern regards how the authors use linear trends to compare the models with the observations. In the current version of the manuscript, the authors focus on the possibility that the observations could indicate multidecadal internal variability superimposed on a forced linear ice decline trend (as discussed on line 214* and elsewhere). In this scenario, as the authors correctly note, "evaluating linear trends on increasingly long timescales would capture more of the underlying forced trend" (line 218). In other words, models would exhibit more skill on longer timescales, which the authors claim in the abstract (line 16) and elsewhere (e.g., line 61) is indeed the case for their analysis. Specifically, the authors claim that on timescales as long as 45 years, the models agree with the observations (e.g., line 190), implying that the models may similarly perform well on centennial timescales (line 196). From this they conclude that "one interpretation may be that we can have some level of greater confidence in projections of substantial centennial decline" from the models (lines 194-195).

However, I do not find this to be an accurate assessment of the data that the authors are evaluating. As noted in my previous reviews, the observations roughly resemble a linear ice extent increase during approximately 1979-2015, followed by a dramatic abrupt decrease during 2017-2023 (as shown in Fig B1). It is certainly plausible that the observed behavior really is a superposition of a forced linear ice decline plus internal variability that causes a ~37-year advance and then an abrupt ~7-year decline. But it needs to be emphasized that even if this is the case, including the highly anomalous period 2017-2023 would *not* typically be expected to give rise to a more accurate estimate of the underlying forced linear trend. Rather, one would need to use a period that is long enough to not be substantially influenced by anomalous periods of advance or decline. And to this end, it appears that whether the models and observations agree in this analysis is not an issue of how long a period they evaluate (as they conclude), but rather that it is just an issue of whether the evaluated time interval includes the 2017-2023 period of abrupt decline. For example, Fig 1 shows a greater level of agreement for the 35-year period that ends in 2023 than for the 45-year period that ends in 2023.

*All line numbers in this review refer to the "Author's tracked changes" file (egusphere-2023-2881-ATC2.pdf)

Author Response:

We thank the reviewer again for their thoughtful and carefully considered insights. We respond in detail below, and the edits made to the final text are detailed at the end.

We agree with the reviewer’s first point that the 1979-2023 period is still not long enough to decipher forced trends and variability. Our text at L206-207 only states ‘evaluating linear trends on increasingly long timescales would capture more of the underlying forced trend’, which is correct. However, our following statement could be read to imply that the period to 2023 is long enough to decipher forced trends. We have therefore added to the text here (L214).

The reviewer’s second, related, point is that whether the models and observations agree is an issue of whether the 2017-2023 period is included rather than how long the period is; i.e. the agreement is a function of time period not time scale. As the reviewer points out, the different results for different

35-year periods do point to the role of the specific time period and not just the timescale. Again, we agree with this. We think this is clear in the existing text, since the question we pose is (L44-45) whether the recent rapid decline, i.e. the specific observed sea ice evolution, changes our assessment of model skill. Again, at L53 in the existing text, when framing the paper, we state "these data now include the recent years of observed rapid decline of sea ice". We agree that the inclusion of the period 2017-2023 is critical, and this is stated in the text clearly (e.g. Results, L133). We have also strengthened this point in the new Conclusion section.

However, we do not think this focus on time periods negates our argument about time scales. One interpretation of the changing results on 35-year timescales is that on this timescale, the results are heavily affected by the evolution of multidecadal variability, but considering longer timescales may reduce this. As argued in the existing text, there is some evidence that the satellite era prior to the recent decline was actually unusual in the context of the 20th century (Fogt et al, 2022, Meier et al., 2013; Gallaher et al., 2013, Cavalieri et al., 2003). If there is variability on 40-50 year timescales in the observations, evaluating longer timescales will be less susceptible to the exact period examined. We recognise this is speculative, but it is our motivation for discussing results in the context of timescales.

Since we recognise that the Discussion has become quite involved to cover many nuances of our analysis, we have added a brief 'Conclusions' section. This highlights our key results and the dependence on both timescale and time period, and leaves the reader with a clear summary of our conclusions. On reflection, we realised one key point that needs to come across in the text in light of the reviewer's comments is an acknowledgement that our assessment on the 45-year timescale may still be susceptible to the specific manifestation of variability that has occurred. This is now reflected in the new conclusions section at lines 239-240.

[On one specific note, the higher agreement for 1989-2023 (35 year) than 1979-2023 (45 year) mentioned by the reviewer is a result of both the stronger negative trend over 1989-2023 and the larger ensemble standard deviation for 35-year periods so we don't think this is the key evidence to be considered (the relative role of the standard deviation of modelled trends and the value of the observed trend is discussed in the Results section).]

Edits to text in response to reviewer (line numbers refer to final version with no tracked changes)

- Abstract L14 'timescales' -> 'timescales and periods' (and deleted 'IPCC' initialisation to meet word count)
- L26 'role of internal variability' -> 'relative role of internal variability in the models'
- L57 'timescale' -> 'timescale and period'
- L149 Discussion and Conclusions-> Discussion
- L155 Added 'Over the full satellite era'
- L157 "Trends for the 35-year timescale also fall within the model ensemble for the five most recent 35-year periods" added "showing that the evaluation depends on the exact period analysed, at least over this shorter timecale."
- L184 "However, the existing discrepancy on shorter time scales points to fundamental issues remaining" edited to "However, the existing discrepancy **for earlier, shorter time periods** points to fundamental issues remaining either on shorter timescales in general or specifically for the earlier time period."
- L206 we have split the paragraph for clarity given the additional text.
- L211 Added 'Under this 'multidecadal variability' interpretation…

- L214 Added “However, capturing the underlying forced trends would require a period much longer than the timescale of multidecadal variability”
- L216 ‘However’ -> ‘Nevertheless’
- L234-245: Added ‘Conclusions’ section

Minor typographical corrections:

- L163: removed ‘with’
- L182 “step” -> “step-“