We thank the reviewer for their careful reading and for their very useful comments on the manuscript.

1 Reviewer Summary:

This study aimed to test the hypothesis that mesoscale resolving climate models can better simulate multidecadal climate variability of the climate system. To do so, the authors compared two simulations with constant 2000 GHG condition from the Community Earth System Model (CESM) – one with 1 degree and one with 0.1 degree ocean grid – with observational records. Specifically, the authors compared 3 climate indices between these datasets: Atlantic Multidecadal Variability (AMV), Pacific Decadal Oscillation (PDO), and Southern Ocean Mode (SOM). The authors also compared surface heat fluxes (SHF), ocean heat content (OHC), and global mean surface temperature (GMST) between these two simulations. The authors found an improvement in simulating these climate modes using the high resolution model. The low frequency of SHF and OHC were also generally larger in the high resolution model. The difference in GMST magnitude between the two simulations were timescale dependent. Based on these results, the authors asserted that using mesoscale resolved climate model could improve the representation of multidecadal climate variability in climate models.

Overall, I thought this study was very interesting and significant. In particular, this study tried to resolve issues raised regarding climate models’ ability to capture ‘low frequency’ climate variability by a series of studies (e.g. Laepple and Huybers (2014b), Laepple and Huybers (2014a), Frankcombe et al. (2015), Cheung et al. (2017), and Parsons et al. (2017)). The approach used in this study also moved past from using idealized model to quantify the effects of small scale processes on large scale circulation. To my knowledge, the analyses done also seemed valid. Though, I believe this study would benefit from further improving the analyses done in this study and the presentation style in certain parts of the manuscript. Detail comments are below.

2 Major Comments:

1. It is a bit unclear what specific timescales the authors are focusing on. Throughout the article, the authors used ‘multidecadal’ variability without specifying the timescales. This becomes confusing when the authors discussed about specific frequencies. For instance, if high resolution model does a better job in simulating a specific frequency but not some other frequency (where both can be classified as ‘multidecadal’ timescale), does that mean the high resolution model is doing a better job or does it not?

We will make the time scale explicit and define it as 10-50 years.
2. One major aspect of this study is the comparison of data in the frequency domain. In general, I find a little bit difficult to convince myself that the high resolution model is indeed better (or significantly different from) than the low resolution model in simulating multidecadal variability by looking at the power spectrum.

(a) For instance, the authors argued that the high resolution model does a better job in simulating the power spectra of all the climate modes by suggesting consistency between spectral peaks shown in observations and high resolution model. However, there are two problems.

• First, this assumes that having the correct spectral peak is what we should aim for. However, as suggested in the text, getting the correct spectral peak that is internal to the climate system from observation is difficult. Spectral peaks can arise from incorrectly removing forced signal from observation. So, having similar spectral peaks between observation and model do not necessarily mean the model simulation is better.

• Second, both high resolution and low resolution simulations seem to have similar spectral power. The main difference between these two simulations is related to the ‘null hypothesis’, where the low resolution climate model has a higher null threshold. As such, I wonder if comparing spectral peaks is a good way to quantify whether the high resolution simulation is better than the low resolution model.

(b) Another example is in SHF and GMST. The effects of reduced variability on ENSO band in the high resolution simulation is undoubtedly pretty clear. However, the difference between high and low resolution simulation becomes a lot less clear on longer timescales (i.e. multidecadal timescales). This again makes comparison on the frequency spectrum less powerful and makes the result less clear.

I suggest the authors can try to come up with some metrics to compare the power spectrum more quantitatively. The simplest way to compare agreements between power spectra would be to compute the spectral coherence. Alternatively, multiple studies have tried to compare the spectral slopes over a range of timescale to argue for/against of underestimation of multidecadal variability (e.g. Dee et al. (2017) and Parsons et al. (2017)). I think these examples can serve as a framework to develop more quantitative comparison between power spectrum.

We will discuss the spectral peaks and their comparison between the simulations and the observations with more nuance. We will also quantify the multidecadal spectral power through the spectral slope, or the mean power in the period range [10,50] years.

3. SHF and OHC are undoubtedly related to SST. However, SHF and OHC were not mentioned explicitly throughout the introduction section and did not come up until the first paragraph of results. This introduction of SHF and OHC is rather abrupt, and it is not totally clear why a comparison of SHF and OHC between high and low resolution model is needed in the context of this study. Having more explicit description on the motivation to analyze SHF and OHC in the context of this study in the introduction part would be helpful.

We will add an introductory paragraph about the connection of OHC to SHF and SST in the context of multidecadal variability.

4. In addition, I am also having a little bit trouble to tie the results in the SHF and OHC sections back to the SST patterns (i.e. AMV, PDO, SOM). Aside from SOM, I don’t think there is enough discussion about how the more variable SHF and OHC are related to the more accurate SST patterns for AMV and PDO (or vice versa). I understand that it is outside of the scope of this study to identify the underlying mechanism that leads to an improvement in simulating multidecadal
variability, but since this study analyzed SHF and OHC in tandem with these climate modes, I think it’s reasonable to at least make more explicit connections between these variables in the discussion section.

We will clarify these connections between the OHC and SHF results to the modes of multidecadal variability.

5. A large portion of results that described how the calculations were done should be moved to the method section. This includes Lines 173 – 184, Lines 249 – 256, Lines 270 – 271. There should also be discussion on how GMST is defined here.

We will move l173-180 to the methods section, but keep Figure 2 and its description (l.180-184) at the beginning of results section 3.1.

We will move the Ocean Heat Content definitions (originally l.249-256 & 270-271) to the methods section.

6. I suggest adding a sentence or two that describes what the implications are based on the results obtained from this study. That way, the readers can understand the major takeaways of this study from the abstract.

We will add a sentence in the abstract and two in the discussion to make this point more explicit.

3 Minor Comments:

1. Line 23, Zhang et al. (2019a) is probably not the best reference for this sentence. There are many other papers that are more relevant and explicit in discussing about the importance to disentangle modes of internal variability from forced changes for detection and attribution studies. Examples include: Hegerl and Zwiers (2011),Bindoff et al. (2013), and Deser et al. (2020).

We will chang the references as suggested.

2. Lines 26 – 27, I believe there are other relevant studies that the author should cite, examples include: temperature extremes (e.g. Ruprich-Robert et al. (2018), droughts (e.g. McCabe and Palecki (2006)and Delworth et al. (2015)), hurricanes (e.g. Zhang and Delworth (2006)). For Atlantic, I think Zhang et al. (2019b) provides a good overview of societally relevant impacts induced by AMV.

We will add the suggested references.

3. Lines 31 – 32, I’m not sure if Atlantic Multidecadal Oscillation was named in Kushnir (1994). To my knowledge, it is more explicitly mentioned and defined in Kerr (2000), although many studies in the 1990s have already identified multidecadal variations in the North Atlantic.

We will adjust the text to include both references.

4. Lines 122 – 123, it is unclear how the data is de-seasonalized. My understanding of how de-seasonalize is commonly done is by first calculating the monthly climatology (Jan – Dec) and then subtracting it from the monthly data (e.g. October 2020 SST – averaged October SST). I think the first part of the sentence described this process, but the ‘i.e. …’ describes something else – annual SST of each year is subtracted from monthly SST of that year. It would be great if this could be clarified.

We will clarify that the canonical deseasonalization procedure was performed.

5. Figure 2d. It should be noted that the standard deviation comparison between observation and models isn’t quite ‘apples’ to ‘apples’ because their temporal lengths are different. Even though I
don’t expect the results would change significantly, I would try to bootstrap the results to account for sampling uncertainty.
We will perform a block bootstrap estimate of the standard deviation.

6. Line 179, please specify which order of Butterworth filter was used.
We will specify that it is a second order Butterworth filter.

7. Lines 198 – 199, it’s worthy to point out Steinman et al. (2015) only focused on North Atlantic and North Pacific, and they did not look at the relationship between Atlantic, Pacific and Indian together.
We will specify this.

8. Lines 218 – 220, please double check the sentence, something is missing. Right now, I have trouble understanding this sentence.
We will simplify the sentence to make it easier to understand.

9. Lines 224 – 226, there are also studies that tried to reconstruct decadal variability in the Pacific. Since the authors mentioned paleoclimate studies in the Atlantic, I think it won’t be complete without a discussion (or at least mentioning) on paleoclimate studies in the Pacific. Examples include: D’Arrigo et al. (2001), MacDonald (2005), Felis et al. (2010), and O’Mara et al. (2019).
We will add these references.

10. Figure 5a&b, the unit on the y-axis is wrong – it should be [heat/time] since it’s flux data.
We will change the y-axis label.

11. Lines 239 – 240, which frequencies of the spectral power for Atlantic and Pacific were integrated over?
We will clarify this (see also major comment 2).

12. Lines 244 – 245, citation is needed. There are studies that showed ocean dynamics (i.e. horizontal divergence) plays a significant role in driving OHC change (e.g. Ruprich-Robert et al. (2018) and Small et al. (2020)).
We will mention the role of OHC divergence.

13. Lines 245 – 246, I think this statement requires clarification. In your previous subsection, you showed that Southern Ocean (and it seems like the global ocean also) SHF exhibit a white noise whereas the Atlantic and Pacific exhibit a blue noise behavior in low frequencies. These results, to me, don’t suggest a particularly strong multidecadal surface heat flux variability, at least relative to high frequency variability.
We will clarify this sentence.

14. Line 249, please explain why and how an interpolation was done here.
We will add a sentence discussing this.

15. Figure 6, I’m curious as to why Indian Ocean is included here. Throughout the manuscript, this is the only place where data from the Indian Ocean was analyzed. Even though I think it’s interesting to look at it, I’m not sure if it is very relevant to this manuscript, where the target is more on global, Pacific, Atlantic, and the Southern Oceans.
We will remove the Indian Ocean panels.

16. Lines 330 – 331, please double check the figure that is referred to. I think it’s supposed to be Fig 6i instead of 5i. If so, Fig 6i represents depth and zonally integrated OHC but not SHF.
We will correct the reference to Fig. 6i and adjust the sentence accordingly.
17. Line 375, note that Mann et al. (2020) showed that CMIP5 do not show multidecadal oscillations (they defined it as significant spectral peak against a null hypothesis) but not absence of multidecadal variability.

We will rephrase the sentence to reflect this distinction.

18. Line 377, ‘multidecadal’ should be multidecadal.

We will fix this typo.

References


