Response to Reviewer #1

(*Note: Reviewer comments in black and our point-to-point replies in blue*)

Review of "Arctic Ocean Simulations in the CMIP6 Ocean Model Intercomparison Project (OMIP)"

This study compares the simulated Arctic Ocean across models in the latest OMIP experiments (OMIP-1 and OMIP-2 as designated in the manuscript) to those of previous CORE-II experiments by looking at several diagnostics including mean hydrography, liquid freshwater content, and transports through the major Arctic gateways. As a standard model intercomparison paper, the results are straightforward, and I recommend the manuscript for publication after the following concerns are addressed:

Reply: We are very grateful to the reviewer's comments and thoughtful suggestions. We made the revision accordingly, and we believe that the manuscript is much improved by taking into account these comments.

1. Methodology: The inter-model spread is used as a measure of the differences amongst the models and is defined as 1 standard deviation of a given value across the models. However, several of the models used in the OMIP have the same sea ice-ocean components (Table 1). How does this impact the spread and the multi-model mean when effectively some models are being double counted (i.e., those that use NEMO3.6 as the ocean model, or those that are MOM-based models under the hood)? It may not matter a ton here as I think the results are internally consistent by comparing across OMIPs, as long as the standard deviations aren't being used to make statistical inferences, but it might be good to clarify this a bit in the text if it's something that the authors thought about.

Reply: We totally agree with the reviewer that the number of models in different model families affects the results of spread and multi-model mean. For the comparison of OMIP-1 and OMIP-2 in Figures 3, 4, 6, 7, 8, 9, 10, 11, and 14, multi-model mean and inter-model spread are calculated based on the eight models with bold model ID in Table 1, and, coincidentally or not, they use different ocean components. But for the gateway transports in Tables 2-7, we used all model to calculate multi-model mean and inter-model spread, so this may affect the gateway transport results. We added some sentences in the revised manuscript to clarify this, including:

In Section 2.5.1: "Eight models (CanESM5, CanESM5-CanOE, CMCC-CM2-HR4, CMCC-CM2-SR5, CMCC-ESM2, CNRM-CM6-1, EC-Earth3, and IPSL-CM6A-LR) based on NEMO reproduce relatively large net volume transport through the BSO compared with other models. As a result, the ensemble mean may be biased toward NEMO-family models".

In Section 2.5.2: "The eight NEMO-family models simulate relatively large ocean heat transport through the BSO compared with other models".

In Section 2.5.3: "The eight NEMO-family models simulate relatively large freshwater export through the Davis Strait compared with other models".

2. In a couple of locations in the manuscript it is noted that the AWI model performs best. Presumably, the 'best' model is defined is the one that is closest to observations? However, what exactly are the requirements for defining that especially when spatial variation is involved? The text could use a little more clarification regarding this. Another option is to add some difference maps between the models and observations in the supplementary material if that helps make the comparisons more obvious (I will leave that to the discretion of the authors though).

Reply: We found AWI-CM-1-1-LR in OMIP-1 performs well in the simulation of potential temperature profiles, and AWI-CM-1-1-LR in OMIP-2 performs well in the simulation of Arctic Ocean stratification (mixed layer depth and cold halocline base depth).

For the potential temperature profile and winter mixed layer depth simulations, Figures 2a, 2b, S7 and S8 show both simulations and observations together, so it should be easy to compare.

For the cold halocline base depth simulations, we added the root-mean-square error (RMSE) in the figure to make the comparisons more obvious.

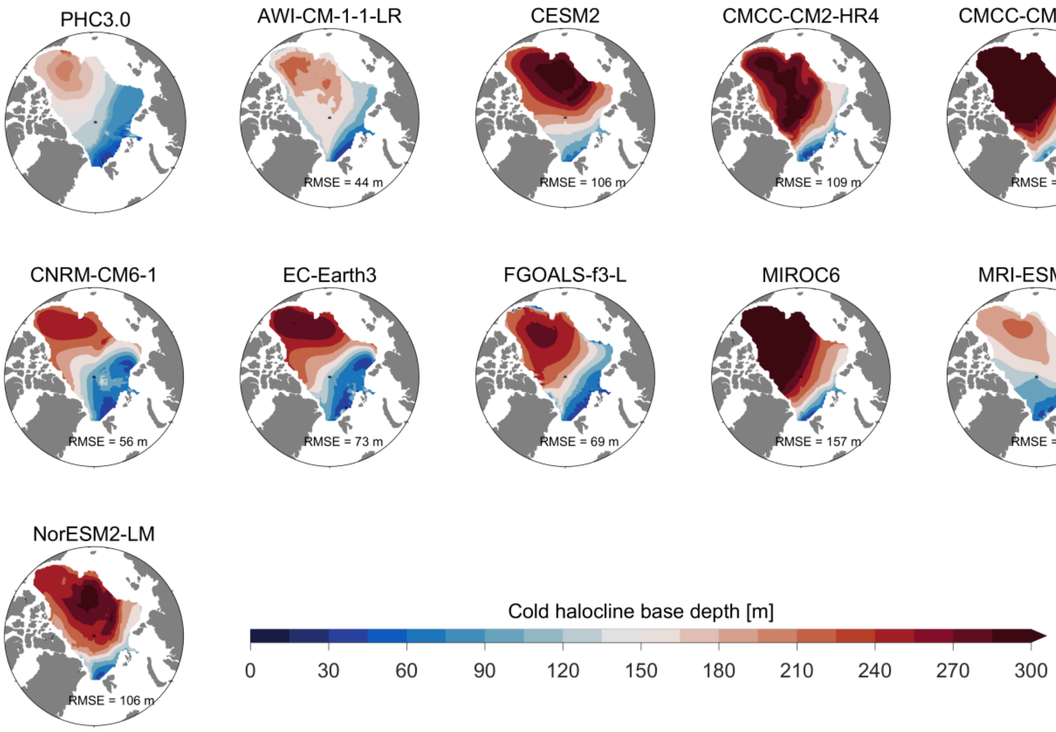

Figure S11. Cold halocline base depth (unit: m) from PHC3.0 climatology and OMIP-2 models average over 1971 to 2000. The root-mean-square error (RMSE) averaged over the Arctic Ocean is labeled in each panel.

3. Citations: there are three other Arctic Ocean CMIP6 studies that could be cited in this manuscript in much the same way that the Wang et al. 2022b and Khosravi et al 2022 papers are invoked to connect the results of this study to its CMIP6 counterparts. Specifically, these studies warrant mention in the introduction and summary/conclusions (e.g., line 390) as well as in specific areas such as Sections 2.3 and 2.5.2 (Heuzé et al., in review), Section 2.4 (Muilwijk et al 2022), Sections 2.2, 2.5.1, and 2.5.3 (Zanowski et al., 2021, e.g., line 267—Zanowski noted the same volume transport trends in the Bering Strait as discussed here but in CMIP6 coupled models).

References below:

Heuzé, C., H. Zanowski, S. Karam and M. Muilwijk: The deep Arctic Ocean and Fram Strait in CMIP6 models (J. Climate, in review). Preprint: https://eartharxiv.org/repository/view/3233/

Muilwijk, M., A. Nummelin, C. Heuzé, I. V. Polyakov, H. Zanowski, and L. H. Smedsrud: Divergence in climate model projections of future Arctic Ocean stratification and hydrography (J. Climate; https://doi.org/10.1175/JCLI-D-22-0349.1)

Zanowski, H., A. Jahn, and M.M. Holland, 2021: Arctic Ocean freshwater in CMIP6 ensembles: Declining sea ice, increasing ocean storage and export, JGR: Oceans, 126, doi.org/10.1029/2020JC016930

Reply: These recent studies are indeed quite relevant to our manuscript. Thanks for the information. They are now cited in appropriate places in the revised manuscript.

4. Summary/Conclusions: The manuscript could benefit from further commentary about what it is we learned about the Arctic Ocean simulations from the OMIP comparisons that has not already been concluded in previous CMIP6 studies such as Khosravi et al 2022, Wang et al 2022b, Muilwijk et al 2022, Zanowski et al. 2021, and Heuzé et al. All of these studies note that simulation of the Arctic Ocean has not improved since CMIP5 (based on the diagnostics in those papers, of course). I am not doubting the validity or usefulness of the OMIPs or CORE simulations, but rather it would be helpful to place this study in the context of the other literature that has come out in the last year or two, and there is a nice opportunity to do that here by expanding this part of the manuscript.

Reply: We added some sentences in this section to connect the simulations from OMIP and CMIP6, including:

“The biases and inter-model spreads in OMIP are relatively small compared with those of CMIP6 fully-coupled models reported by Khosravi et al., 2022”.

“Therefore, it is not surprising that the large biases and inter-model spread are found in the Arctic Ocean temperature and salinity simulations in CMIP6 fully-coupled models, and no significant improvements are found in the Arctic Ocean simulations from CMIP5 to CMIP6 (Khosravi et al., 2022; Wang et al., 2022b; Zanowski et al., 2021; Muilwijk et al., 2022; Heuzé et al., 2023)”.

“So the potential roles of future shoaling of the halocline base in fully-coupled models (Shu et al., 2022) in the atmosphere-ocean-sea ice interactions may be biased low due to the deep biases of the halocline base”.

“OMIP models also have large inter-model spreads in the simulation of the Arctic Ocean stratification, which is consistent with the performance of the CMIP6 fully-coupled models (Muilwijk et al., 2022)”.

5. Lines 265-266 “The reasons for the discrepancy between observations and simulations are unknown and should be further investigated.” Regarding the issues with the models being unable to reproduce the correct sign of the Bering Strait volume transport trend, didn’t Wang et al 2022b (figure 8) suggest that this might be due to changes in the sea surface height gradient between the Arctic and Pacific? That may have been analysis for the future forcing scenarios and not the present-day trends as is the case in this manuscript, but it may also be worth commenting on. I have been wondering about the negative model volume transport trends vs. the positive trends observed by Woodgate et al. as well.

Reply: We agree with the reviewer that the changes in the sea surface height gradient between the Arctic and Pacific can cause the negative model volume transport. We compared OMIP-2 simulated sea surface height difference between the period of 2009–2014 and the period of 2003–2008 with the satellite observations. Figure S16 shows that simulated changes in the sea surface height gradient between the Arctic and Pacific is much larger than the satellite observations, so it may lead to the discrepancy in the trends of volume transport through the Bering Strait between simulations and observations.

In the revised manuscript, this sentence was changed to “The reason for the discrepancy between observations and simulations may be caused by an unrealistic reduction of sea surface height in the Bering in the 2010s in the OMIP-2 simulations (Fig. S16)”.

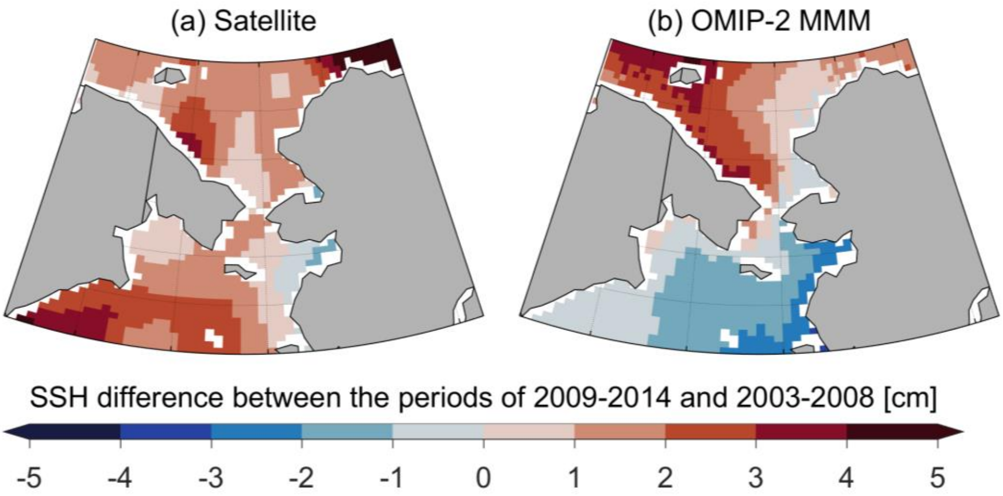

Figure S16. (a) Observations and (b) OMIP-2 multi-model mean (MMM) sea surface height (SSH) difference between the period of 2009–2014 and the period of 2003–2008.

6. A note on rainbow colormaps: It is best to avoid using a rainbow colormap where possible (there’s a lot of information out there about why it’s problematic). Please consider changing the colormaps for the figures to something non-rainbow. If any of the authors use python, cmocean has a nice set of pre-defined colormaps: https://matplotlib.org/cmocean/

Reply: Following the reviewer’s suggestion, the colormaps from cmocean were used in the revised manuscript.