Response to Reviewer #1

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

I would like to thank the authors for their thorough responses to my comments and I have no further suggestions at this time. There are some minor grammatical errors throughout the text that should be corrected moving forward, but I will leave it to the authors to take care of those in the final review.

Reply: We appreciate your positive feedbacks. Thanks for your help.

Response to Reviewer #2

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

Thank you for addressing my comments for the first version of the manuscript. I am pleased to see that you have picked out specific models that stand out and provided some interpretation of the model results. I find the 2nd version of the manuscript satisfactory and recommend it be published after a few minor adjustments. These are specified below.

Reply: We appreciate your positive feedbacks and valuable suggestions. We made the revision according to your comments and suggestions.

Comments

OMIP versus "standard" CMIP6 models

For the reader who is unfamiliar with the OMIP initiative, it is perhaps not so clear what the difference is compared to the CMIP6. Perhaps you could clarify; are the OMIP models global or regional domain? How do they differ from the fully coupled models used in CMIP6? Just one-two sentences before the Methods.

Reply: Following the reviewer's suggestion, the sentence of "Ocean-sea ice models participating in the CMIP6 OMIP are driven by the specified common atmosphere forcing data sets" at the beginning of the Methods was changed to "Most of OMIP models were used as the ocean components of the CMIP6 fully-coupled models; different from the latter, global ocean-sea ice models participating in the CMIP6 OMIP are driven by the specified common atmosphere forcing data sets".

L75: I presume you mean "have not been evaluated in a systematic/collective manner?" I would assume that some of the models have been evaluated individually before?

Reply: "However, Arctic Ocean simulations by the OMIP ocean models, most of which were used as the ocean components of fully-coupled models in CMIP6, have not been evaluated" was changed to "However, Arctic Ocean simulations by the OMIP ocean models, most of which were used as the ocean components of fully-coupled models in CMIP6, have not been evaluated systematically in model intercomparison studies."

L115: Perhaps you can add a sentence mentioning that observations are also not without problems/biases and should not be taken as the complete truth.

Reply: "One has to keep in mind that although the datasets used to evaluate models are mostly based on observations, they also have biases and uncertainties" was added in the revised manuscript.

L130: Using the potential temperature at 400m depth to evaluate the AW makes sense when comparing to previous studies. I would still consider looking at the maximum temperature within a depth interval (or 0-degree isotherm), which could account for differences in the depth of the AW layer between models. By choosing a fixed depth you may bias the results. At least, it is something you should check.

Reply: Following the reviewer's suggestion, we checked the simulations of the Atlantic Water core temperature (the maximum temperature of the Atlantic Water at each grid location) by OMIP models. Figures R1 and 3 indicate that the model performances and patterns in the simulations of Atlantic Water core temperature and 400 m depth temperature are quite similar, and that the conclusions are not changed by the evaluation of the simulated Atlantic Water core temperature.

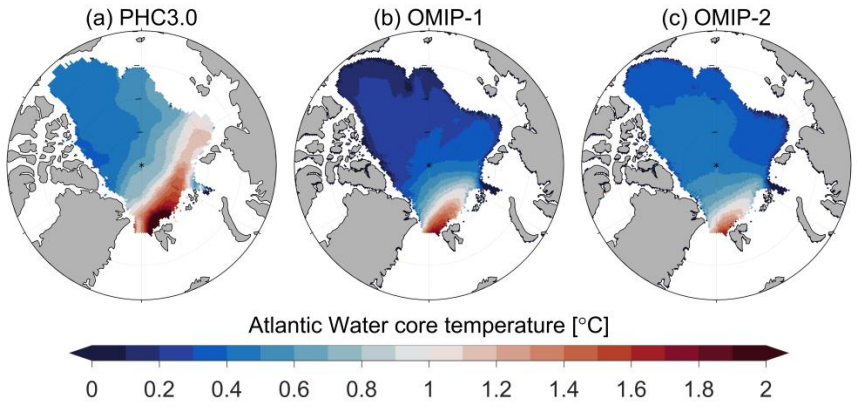

Figure R1. Atlantic Water core temperature (m) in (a) PHC3.0, and (b) OMIP-1 and (c) OMIP-2 multi-model mean results.

L173-175: You state that there are no improvements in the mean state in the MMM. This is somewhat depressing – so I wonder if there are any improvements in any individual model?

Reply: If we use a single metric to evaluate the model, we may find improvements in certain individual models – which are indeed addressed in the paper where relevant. However, no significant improvements are found in individual models for most metrics used in this study.

L183: "meter" —> "meters"

Reply: Done.

L212: It would be nice if you could include a small paragraph about the model's ability to represent stratification in the Beaufort Sea in particular; i.e., how well do the simulate Pacific Summer Water (PSW). This would be interesting to report on since the models seem to capture the volume through BS quite well.

Reply: We added "The discrepancy of mixed layer depth between simulations and observations in the

Chukchi and Beaufort Seas is relatively small. This may be partially attributed to the good performance in the representation of Pacific Water volume transport through the Bering Strait in OMIP models (see the results in Section 3.6.1)" in the revised manuscript.

L311-313: Are the biases in the BG due to incorrect atmospheric forcing or model physics? Please clarify.

Reply: At present, it is difficult to disentangle the two factors and pinpoint the exact reason for the bias in BG. This topic is beyond the scope of the present paper and warrants to be investigated in the future.

L317: Can you please clarify if the freshwater transport include both liquid and solid freshwater contributions? What is the relative importance of solid/liquid transport?

Reply: We only evaluate the liquid freshwater transport in this study. To clarify that, "and freshwater transport FWT through each Arctic Ocean gateway are calculated as follows" was changed to "and liquid freshwater transport FWT through each Arctic Ocean gateway are calculated as follows". Arctic solid freshwater transport mainly outflows through the Fram Strait. The estimation by Haine et al. (2015) indicates that the solid (2300 km^3/year) and liquid (2700 km^3/year) components of freshwater transport through the Fram Strait are comparable.

Refference:

Haine, T. W. N., Curry, B., Gerdes, R., Hansen, E., Karcher, M., Lee, C., Rudels, B., Spreen, G., de Steur, L., Stewart, K. D., and Woodgate, R.: Arctic freshwater export: Status, mechanisms, and prospects, Glob. Planet. Change, 125, 13–35, 2015.

L132: What about vertical resolution?

Reply: We also did not find relation to vertical resolution. "…horizontal or vertical resolutions…"

L354: "… to 2018 and the trend …" —>"…although the trend"

Reply: Done.

L358: Why is the spread in volume transport for the models so low compared to the observations?

Reply: The spreads of models and observations are calculated differently. The spread for simulations is one standard deviation of all simulations by OMIP models. The spread for observations rather indicates observation uncertainty, which is large due to very coarse mooring resolution.

L374: "The OMIP models obtained upward trends …" —> "The OMIP models also obtained positive trends"

Reply: Done.

L380: It's interesting that the models using NEMO seem to have some common biases. Can you comment on why this might be the case? Perhaps mixing schemes? Too strong AMOC preconditioning a stronger inflow across the Greenland-Scotland Ridge? Or representation of bathymetry in the inflow region (e.g Heuzé and Årthun 2019 10.1525/elementa.354)?

Reply: Thank you for the useful information. Tsujino et al. (2020) shows that the AMOC in NEMO-family models (such as CMCC-NEMO and Kiel-NEMO) is relatively weaker in OMIP simulations. We added "It might be related to the different vertical mixing scheme (TKE turbulent closure scheme) or the representation of bathymetry in NEMO-family models" in the revised manuscript. This is a topic that needs further research.

Reference:

Tsujino, H., Urakawa, L. S., Griffies, S. M., Danabasoglu, G., Adcroft, A. J., Amaral, A. E., Arsouze, T., Bentsen, M., Bernardello, R., Böning, C. W., Bozec, A., Chassignet, E. P., Danilov, S., Dussin, R., Exarchou, E., Fogli, P. G., Fox-Kemper, B., Guo, C., Ilicak, M., Iovino, D., Kim, W. M., Koldunov, N., Lapin, V., Li, Y., Lin, P., Lindsay, K., Liu, H., Long, M. C., Komuro, Y., Marsland, S. J., Masina, S., Nummelin, A., Rieck, J. K., Ruprich-Robert, Y., Scheinert, M., Sicardi, V., Sidorenko, D., Suzuki, T., Tatebe, H., Wang, Q., Yeager, S. G., and Yu, Z.: Evaluation of global ocean–sea-ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2), 3643–3708 pp., https://doi.org/10.5194/gmd-13-3643-2020, 2020.

L465: "… implying that changes in the upper Arctic Ocean are captured in the models": In terms of what? Please specify. It's possible for the model to simulate the SSH correctly but still not capture the water mass structure accurately.

Reply: We agree with the reviewer. "implying that changes in the upper Arctic Ocean are captured in the models" was changed to "implying that changes in the upper circulation of Arctic Ocean are captured in the models".

L495: Great! Thanks for suggesting this – I agree.

Reply: Thank you for your valuable suggestion in the first round of review.

L502: I wonder if you really resolve eddies at 4.5 km outside the deep Arctic basin, e.g., on the shelf regions in the Barents and Kara Sea where you have Atlantic Water? Consider changing/adding the reference to Wang et al 2020 (https://doi.org/ 10.1029/2020GL088550) which uses a horizontal resolution of 1 km.

Reply: We agree with the review that 4.5 km resolution cannot fully resolve eddies in the Arctic Basin. "Wang et al. (2018) show that a model with 4.5-km resolution in the Arctic Ocean performs much better than a 24-km resolution model" was changed to "Wang et al. (2018) show that a model

with 4.5-km (marginally eddy-permitting) resolution in the Arctic Ocean performs much better than a 24-km resolution model". We stay with the reference to Wang et al. (2018) here as detailed model evaluation was done in this paper.

Figures

Fig 3+4: If possible I would change the colormap for a,b, c and f, g h to one that is not divergent. For example "Thermal" for a,b, c and "Haline" for f, g, h. The divergent colormap implies differences/anomalies.

Figure 9, 10, 11, 13: Same as above. There are better colormaps available for displaying these metrics (see cmocean) and will help the communicate the figures to the reader more effectively. Note that people associate certain colormaps with specific things.

The comments on the figures are minor details, but I would still encourage the authors to consider updating the figures with more appropriate colormaps. I will leave it to the authors and the editor to make that decision.

Reply: Following the reviewer's suggestion, we tried the colormaps of "Thermal" and "Haline", and we find that the original colormap "balance" is better to show the differences between different panels, such as Figures 3, 9 and 13. So we didn't change the colormaps in the revised manuscript.

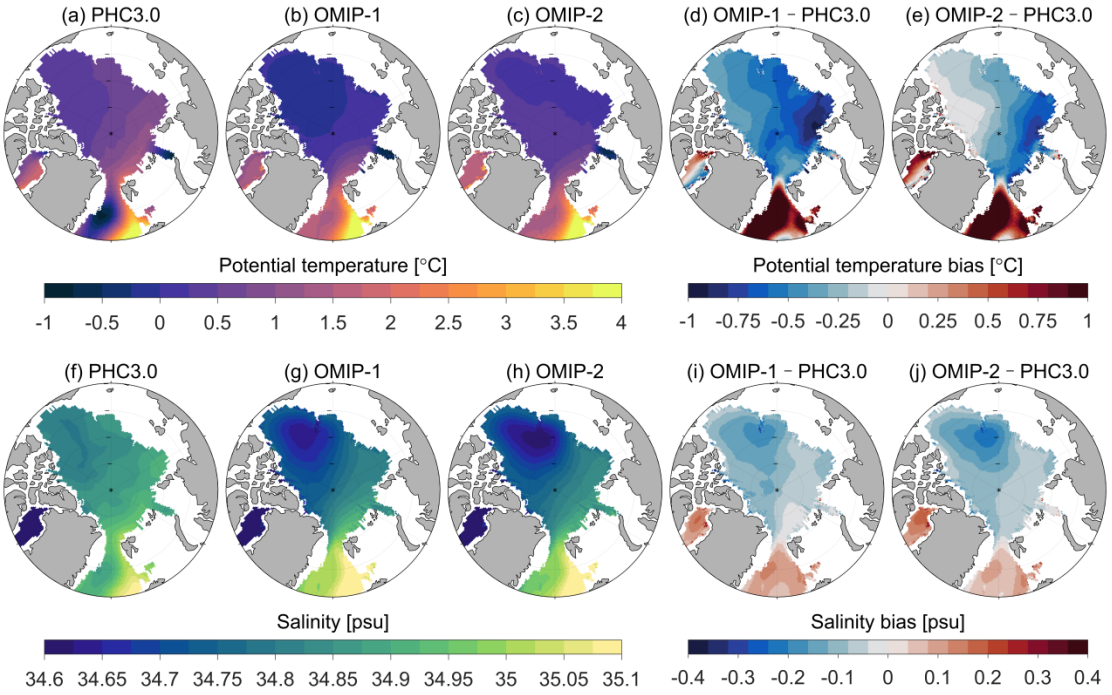

Figure 3. (upper) Potential temperature and (bottom) salinity at 400 m from (a,f) PHC3.0, (b,g) OMIP-1, and (c,h) OMIP-2 multi-model mean results, and the biases of (d,e) potential temperature and (i,j) salinity of (d,i) OMIP-1 and (e,j) OMIP-2.

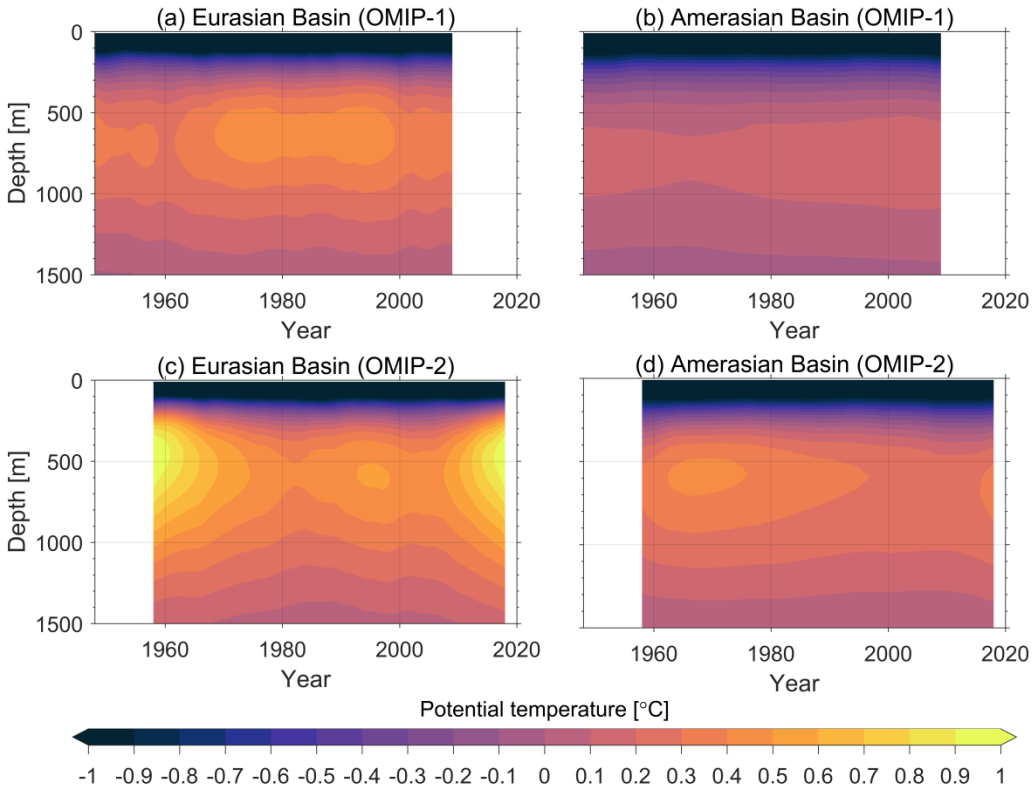

Figure 9. Hovmöller diagram of basin-mean potential temperature (unit: °C) for the Eurasian Basin (a, c) and Amerasian Basin (b,d) in OMIP-1 (a,b) and OMIP-2 (c,d).

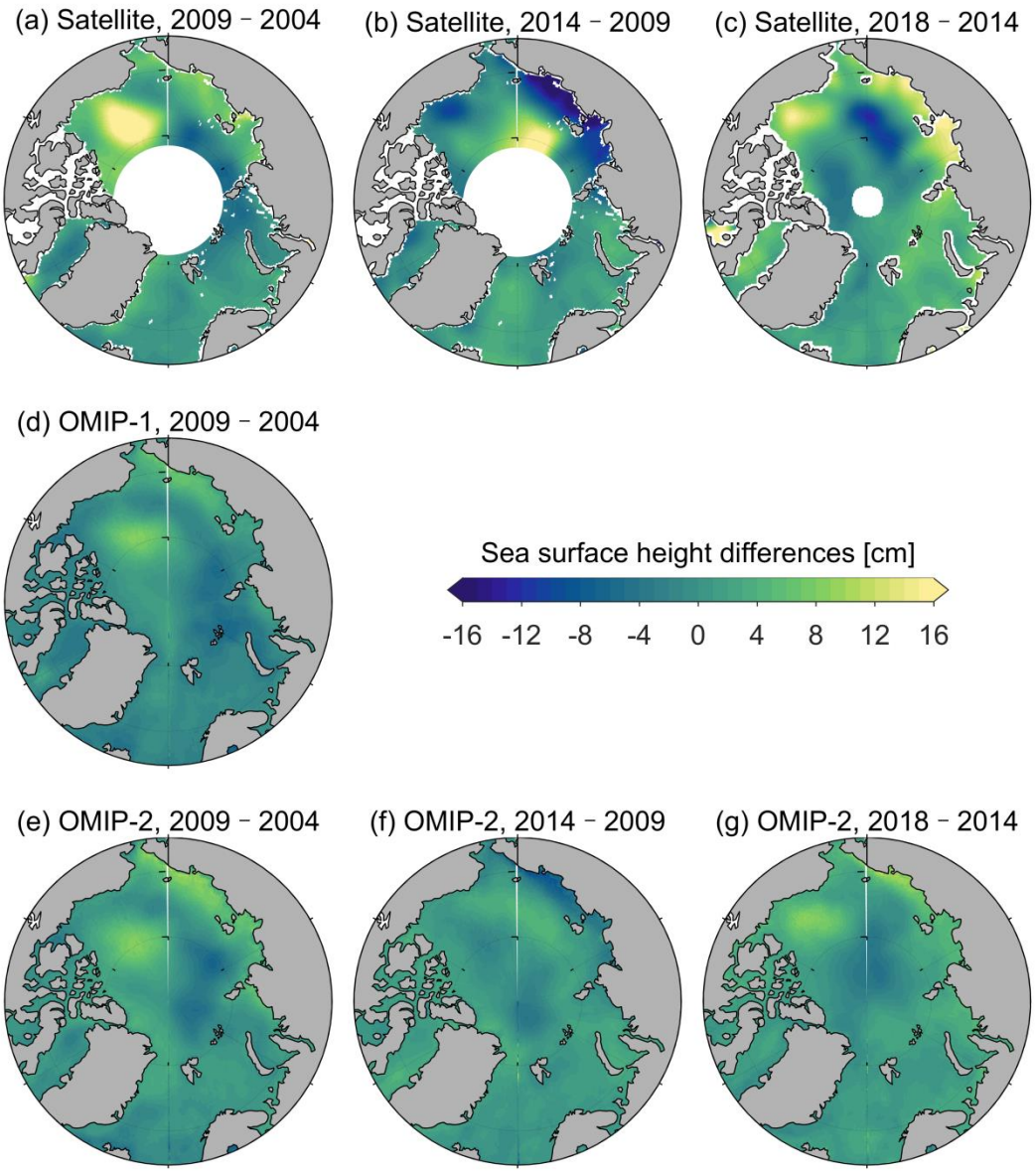

Figure 13. Sea surface height differences between (a) 2009 and 2004, (b) 2014 and 2009, and (c) 2018 and 2014 from satellite observations. (d) The same as (a), but for the OMIP-1 multi-model mean. (e)(f)(g) The same as (a)(b)(c), but for the OMIP-2 multi-model mean.