Author responses to reviewer comments

1 Responses to Reviewer #1 (Frank Bryan)

General Comments

- **Reviewer comment:**

  This manuscript is the logical successor to Griffies et al (2016) defining the OMIP protocol and Tsujino et al (2018) describing the construction of the JRA55-do forcing dataset. The collective efforts of the world's leading ocean modeling groups both in preparing for this study (preparing the forcing data, developing and agreeing on the protocol, running the experiments) and in collecting and collating the results is monumental.

  The manuscript documents that the massive effort took place, but to be perfectly honest, it is as dull as dirt. Paragraph after paragraph begins "Figure N presents ...", "Figure N+1 presents ...", going through the catalog of standard metrics. The authors set a low bar by declaring they will offer only a "glimpse rather than an in depth view of the many elements of ocean model performance" (line 93). With 150 pages of material and several hundred figures, I believe there is more than can be described as a "glimpse", but agree that an in depth view is not offered. I guess this is the consequence of CMIP-ification of climate science. I am sure many groups will use the figures at some point to calibrate their efforts going forward, and more in depth studies will follow, but I would have hoped that we might have found a little more introspection on the successes and shortcomings of the protocol as well as more on the impacts of the structural changes in the forcing data. For example:

  - **Does each metric considered add value to the assessment, e.g., Do we need 0-700m heat content and SSH metrics or would one or the other be sufficient to discriminate among the included models?**
  - **Will these metrics be relevant as resolution (and resolved variability) increase? There is already some indication that certain of these metrics become misleading.**
  - **Does a change in ordering among models in various metrics in OMIP-1 vs OMIP-2 suggest the importance or not of different aspects of the forcing? What does the change in spread across the ensemble imply about the forcing?**
  - **Are there any obvious groupings of models (e.g. the NEMO models or the hybrid coordinate models) in model skill metrics or not?**
  - **Did variance in the solutions during in the pre-satellite era change more or less as compared to the later years between OMIP-1 and OMIP-2?**
  - **The Tsujino et al (2018) manuscript calls out several "notable differences between CORE and JRA55-do" (pg 106, first pp). Are these apparent in the solutions?**
  - **How did the additional variability in runoff included in JRA55-do forcing impact the solutions?**

  In short, what are the high-level conclusions that we can draw about the value of this exercise? I doubt that this question will be addressed in subsequent studies, so this is the obvious place to address it.

- **Author's response:**

  Firstly, we would like to thank reviewers for their time and effort to review this paper and to provide constructive comments. We acknowledge that the discussion paper is not clearly summarizing the outcome of the overall effort and is failing to convey some important messages to the reader. The following are the main conclusions based on the analysis originally conducted and the additional analysis conducted for this revision:

- Both OMIP-1 and OMIP-2 ensembles capture observations, while the multi-model spread greatly exceeds the difference caused by the change in forcing datasets.
- Many ocean climate indices are very similar between OMIP-1 and OMIP-2 simulations, and yet we could also identify key qualitative improvements in transitioning from OMIP-1 to OMIP-2, which represents a new capability of the OMIP2 framework for evaluating process-level responses.
- A clear distinction is found between the metrics that are directly forced and those that require complex model adjustments, causing well-ordered and potentially less-organized responses among models to a change in forcing, respectively.

In the revised version, the following modifications will be incorporated:

- One of our key findings, that models tend to disagree with each other more than the forcing products do, or more specifically, the multi-model spread greatly exceeds the difference between the two datasets, will be more highlighted in the revised version. To reinforce this conclusion, we explicitly quantify as many metrics as possible. For example, the figure of SST bias assessment (Figure 5 of the discussion paper) will be revised and look like Figure 1. The mean ensemble standard deviation exceeds the root-mean-square-difference between OMIP-1 and OMIP-2 simulations. Also, as shown in the middle panels, the regions where the observation is outside the 90% confidence range of the model spread ($\pm 2\sigma$) are generally less than 15% of the global ocean, implying that both OMIP-1 and OMIP-2 ensembles capture the observation. The discussion will become clearer with such quantification. Also, we will extend the quantitative assessment of model biases and spreads to MLD and zonal mean temperature and salinity (Figures 9 through 12 of the discussion paper).

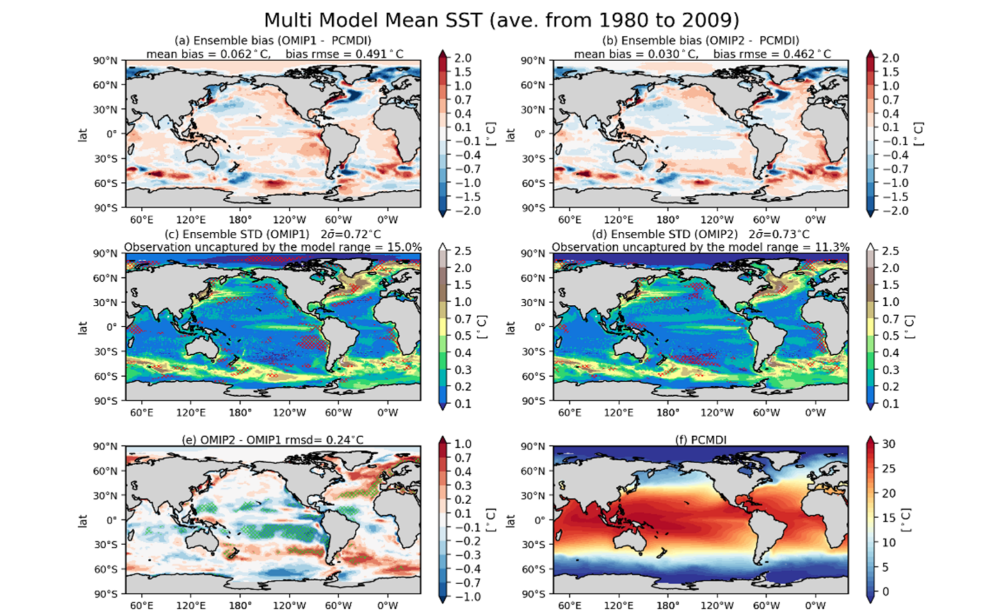

Figure 1 (to replace Figure 5 of the discussion paper): Evaluation of the simulated mean sea surface temperature (°C). Upper two panels show the bias of the multi-model mean, 30-year (1980-2009) mean SST relative to an observational estimate provided and updated by Program

for Climate Model Diagnosis and Intercomprison (PCMDI) following a procedure described by Hurrell et al. (2008) (hereafter referred to as PCMDI-SST). (a) OMIP-1 and (b) OMIP-2. The middle two panels show the standard deviation of the ensemble, with the regions where the observation is outside the 90 % confidence range of the model spread (±2σ) are hatched with red. (c) OMIP-1 and (d) OMIP-2. (e) Difference between OMIP-1 and OMIP-2 (OMIP-2 minus OMIP-1), with the regions where the difference is significant at 95% confidence level hatched with green. The uncertainty of multi-model mean is computed based on the method proposed by Wakamatsu et al. (2017). (f) 30 year (1980-2009) mean SST of PCMDI-SST.

- In doing the above analysis, we also found some features of potential importance about the ordering among models in the metrics. These quantifications will be summarized in new Tables and the outcome will be highlighted throughout the paper. The specific features are as follows.
  - For SST (rmse and mean), SSS (rmse and mean), SSH (rmse), sea ice extent (mean), MLD (rmse and mean), zonal mean temperature (rmse) in the Indian and the Pacific Oceans, and Indonesian Through Flow (mean), the change in ordering among models is small between OMIP-1 and OMIP-2. This may indicate that the behaviors of these metrics are largely determined by settings used by each model.
  - On the other hand, for some circulation metrics such as AMOC and GMOC (bottom water circulation), ACC, zonal mean temperature (rmse) in the Southern Ocean and Atlantic Ocean, zonal mean salinity (rmse), and the drift of vertically averaged temperature, the ordering among models is less consistent between OMIP-1 and OMIP-2. This may indicate that those metrics that involve thermohaline adjustment in models are sensitive to the differences in the forcing dataset.

  Here we list some examples. Figure 2 shows scatter diagrams of rmse and mean of SST and SSS bias of the OMIP-1 and OMIP-2 simulations. The linear fitting and its r^2-score are also depicted. It would be notable that these metrics correlate well between OMIP-1 and OMIP-2. In other words, the ordering among models does not change significantly between OMIP-1 and OMIP-2. The implication would be that the behaviors of these metrics are largely determined by the settings used by each model.

  Figure 3 shows the similar diagrams for metrics related to large scale circulations. Correlation coefficients are generally low except for the Indonesian Through Flow, which is thought to be determined by the model topography by the first order approximation. This implies that the metrics that involves thermohaline adjustments could show significantly different behaviors to different forcings.

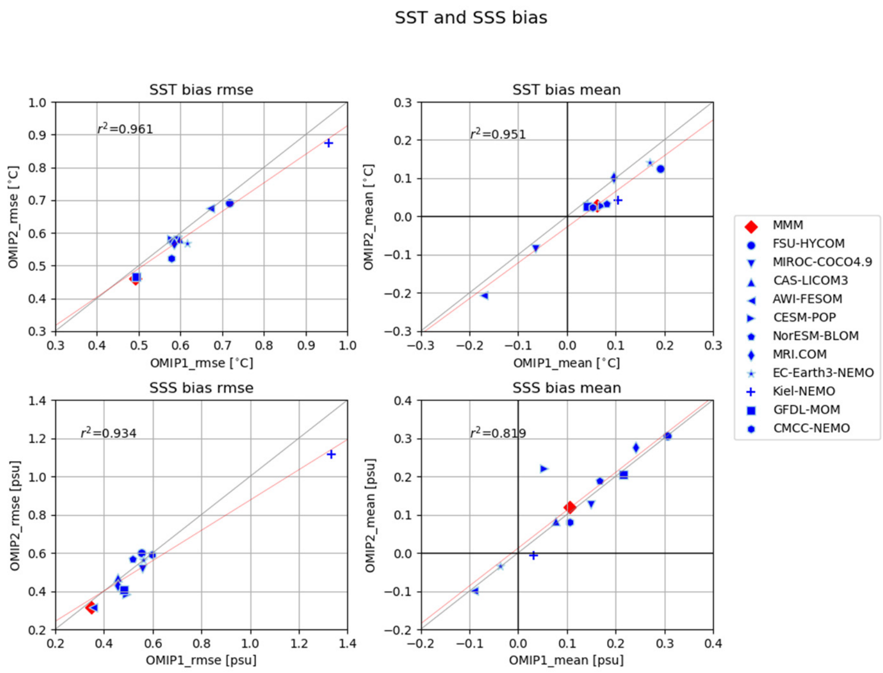

Figure 2: Scatter diagram with linear fitting (red line) and its score (r^2) comparing SST bias rmse (upper left), SST bias mean (upper right), SSS bias rmse (lower left), and SSS bias mean (lower right) from OMIP-1 (abscissa) and OMIP-2 (ordinate). Note that this figure will not be used in the revised version, only tables that list specific values will be included.

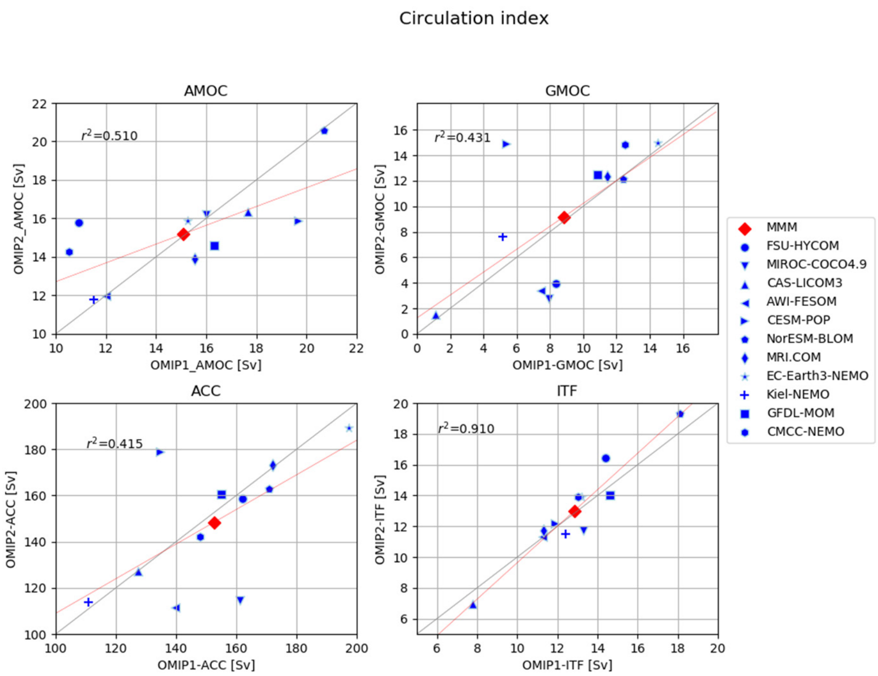

Figure 3: Scatter diagram with linear fitting (red line) and its score (r^2) comparing AMOC (upper left), bottom water circulation (upper right), ACC (lower left), and ITF (lower right) from OMIP-1 (abscissa) and OMIP-2 (ordinate). Note that this figure will not be used in the revised version, only tables that list specific values will be included.

It may not be necessary to respond to all of the points suggested by reviewer #1 as examples for more careful consideration toward the improvement of the manuscript, we list our responses to them, whether positive or negative, in the following:

- **Comment: Does each metric considered add value to the assessment, e.g., Do we need 0-700m heat content and SSH metrics or would one or the other be sufficient to discriminate among the included models?**
  Response: In the revision, with an intention to more streamline the description of the main text, we will add a sentence or two to discuss about the meaning and usefulness of the chosen metrics when each metric is assessed.

- **Comment: Will these metrics be relevant as resolution (and resolved variability) increase? There is already some indication that certain of these metrics become misleading.**
  Response: It has been shown that it will not be appropriate to apply some common metrics to both eddying and non-eddying models (e.g., interannual variability of sea surface height). This point will be highlighted more in the revised version.

- **Comment: Does a change in ordering among models in various metrics in OMIP-1 vs OMIP-2 suggest the importance or not of different aspects of the forcing? What does the change in spread across the ensemble imply about the forcing?**
  Response: The revised version will be more quantitative and take care of the ordering among the models as described above. Regarding the change in spread across the ensemble, we did not observe particularly notable changes in spread due to the change in forcing datasets, except perhaps for the larger spread in OMIP-2 for the metrics involving thermohaline adjustments such as vertically averaged temperatures. We do not have a clear conclusion about this relatively larger spread in OMIP-2. It might be due to the lack of experiences with the OMIP-2 forcing dataset of modelling groups, which will be mentioned in the text.

- **Comment: Are there any obvious groupings of models (e.g. the NEMO models or the hybrid coordinate models) in model skill metrics or not?**
  Response: In this assessment, we did not notice any obvious grouping of models in model skill metrics in terms of model formulation and model code. This will be mentioned in the conclusion.

- **Comment: Did variance in the solutions during in the pre-satellite era change more or less as compared to the later years between OMIP-1 and OMIP-2?**
  Response: In this assessment, we did not notice major change in the variance in the solutions between the pre-satellite and the satellite era (e.g., Figs. 17 through 20).

- **Comment: The Tsujino et al (2018) manuscript calls out several "notable differences between CORE and JRA55-do" (pg 106, first pp). Are these apparent in the solutions?**
  Response: The positive heat flux anomaly during 1980s in the CORE forcing dataset (Fig. 22e of Tsujino et al. 2018) may explain the failure of OMIP-1 simulation to reproduce the gradual increase of SST during 1980s. This will be explicitly mentioned in the paragraph that discusses Figure 19.

- **Comment:** **How did the additional variability in runoff included in JRA55-do forcing impact the solutions?**

  Response: More fresh water discharge from Greenland in the JRA55-do forcing may have at least partly impacted the initial decline of AMOC in the OMIP-2 simulations. Our internal assessment implies that the recent increase in the runoff from Greenland does not have major impact on the AMOC variability and trend. But this is worth investigating further in the future studies. These will be stated in the text.

- **Author's changes in manuscript**

  The above features/findings will be more clearly stated in the abstract and conclusion. The quantitative assessment will be explicitly listed on Tables and relevant discussions will be included in the main text.

Specific comments and author responses

- **Comment:** **Figures: I find the color bar used for positive definite quantities (e.g. 2b,d,f,h) very difficult to interpret. More contrast would be helpful.**

  Response: A more contrasting color sequence will be used. Figure 2 of the discussion paper will look like Figure 4 of this document.

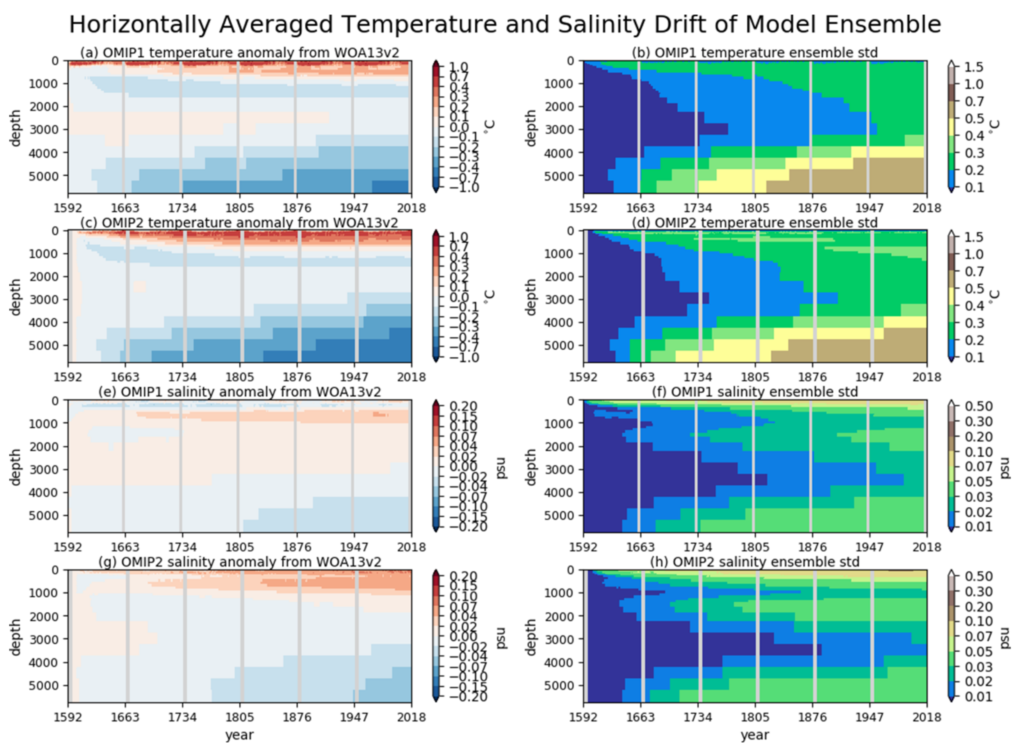

Figure 4 (to replace Figure 2 of the discussion paper): Globally averaged drift of multi-model mean horizontal mean (a, c) temperature (°C) and (e, g) salinity (practical salinity units (psu)) as a function of depth and time. The drift is defined as the deviation from WOA13v2 (Locarnini et al. (2013) and Zweng et al. (2013) for temperature and salinity, respectively). For each, time evolution of the standard deviation of the model ensemble is depicted to the right. (a, b) OMIP-1 temperature, (c, d) OMIP-2 temperature, (e, f) OMIP-1 salinity, and (g, h) OMIP-2 salinity.

- **Comment:** **Figure 1 and similar: Some explanation of what accounts for the nearly instantaneous development of the ensemble spread in upper ocean heat content, SST etc would be helpful. Perhaps maps of the year 1 bias in each model and how it compares to the longer term mean bias. What structures are responding this rapidly? What can we learn from experiments integrated for a few years vs 360?**

  Response: Regarding the apparently instantaneous development of the ensemble spread in some metrics, in particular the upper ocean heat content, the reason is that the models have somewhat distinct initial conditions. There are many details about model initialization that can create differences across models, most notably the methods each group uses to interpolate/extrapolate WOA to their grid/topography and how they initialize sea ice. In particular, the choices in how the bottom topography is constructed for a given model can result in significant differences in such volume average fields. And these differences could affect the initial adjustment processes in models as well. This issue was encountered by the earlier CORE studies such as Griffies et al (2009) and Griffies et al (2014). We continue to perform the initialization using distinct methods across groups for CMIP6-OMIP. This relaxed protocol for initialization is partly because we are not here focused on prediction (an initial value problem) but instead are most concerned with variations and trends after the initial adjustment phase. However, we should think about this issue more carefully in the next phase of this comparison effort. Note that, we will add a new figure showing spin-up behavior of SST and SSS in the simulations (as Figure 1 of a revised manuscript).

  Regarding the implications of the first years of integration for later model biases, the spatial pattern of biases in later years is indeed discernible in the initial years of SST and SSS as shown in Figure 5. This may not necessarily apply to other metrics, but we think that this would be worth mentioning and are considering to include Figure 5 in the revised version.

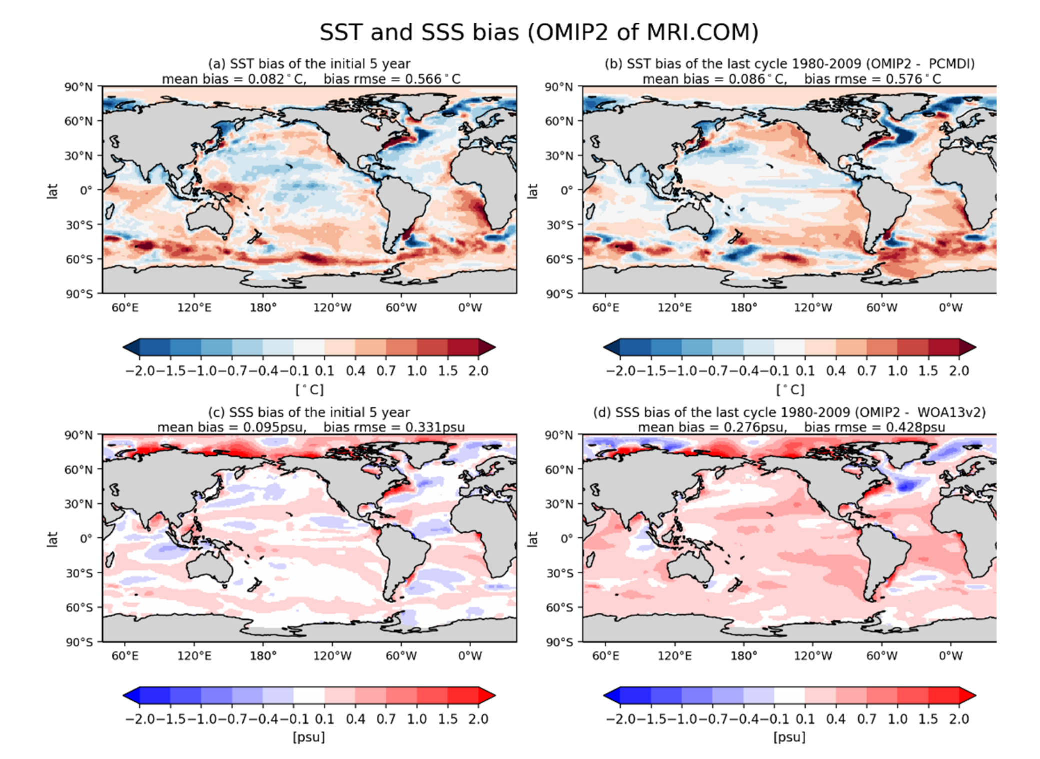

Figure 5 (to be inserted after Figure 6 of the discussion paper): Comparison of SST (a,b) and SSS

(c,d) biases relative to observation for the initial 5-year mean (left panels) and the long-term mean (1980-2009) in the last cycle (right panels) of the OMIP-2 simulation using MRI.COM. Pattern correlation of biases between the initial 5 year mean and the last cycle climatology is 0.75 for SST and 0.85 for SSS.

- **Comment: Line 303 and following: A comparison at a subsurface (maybe 50m) depth would be more enlightening to factor out the influence of salinity restoring.**
  Response: We compared salinity distributions at 0 m, 50 m, and 100 m depths but they look qualitatively similar (not shown). Instead, we will show the difference between salinity used for restoring sea surface salinity in OMIP-1 and OMIP-2 (Figure 6f). Figure 6f indicates that the difference in salinity used for restoring is having nontrivial effect on the simulated difference in salinity of the Arctic Ocean (Figure 6e), although a more dedicated analysis would be necessary to thoroughly understand the simulated difference considering the many other processes contributing to determining the salinity fields in the Arctic Ocean.

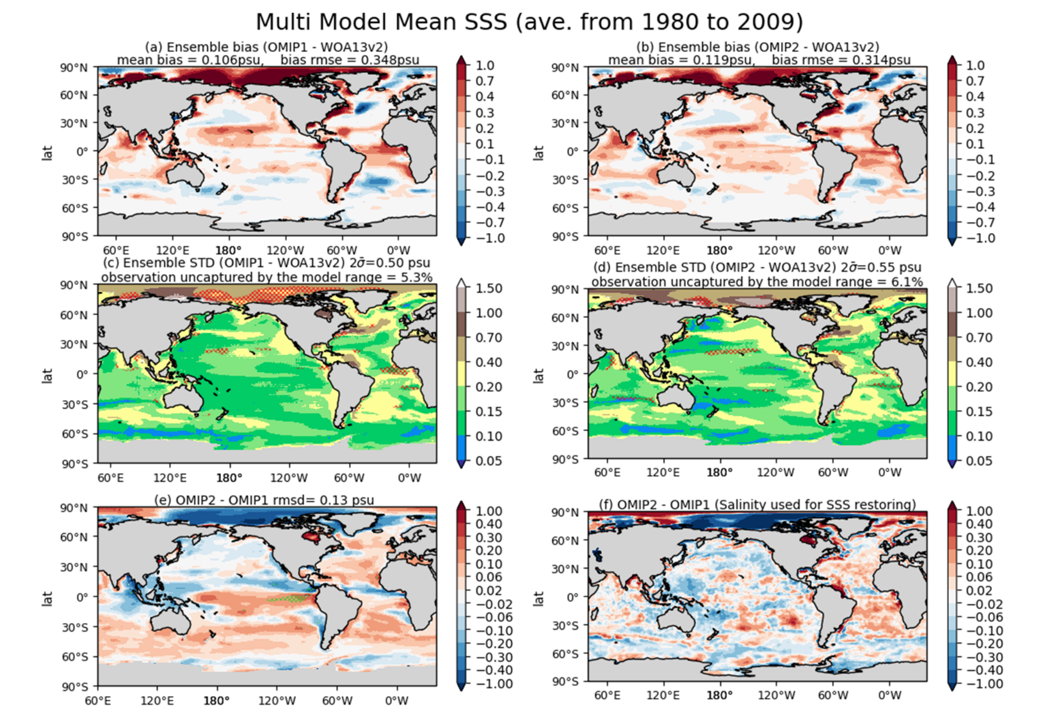

Figure 6 (to replace Figure 6 of the discussion paper): Evaluation of simulated sea surface salinity (psu). Upper two panels show the bias of the multi-model mean 30-year (1980-2009) mean SSS relative to WOA13v2 (Zweng et al. 2013). (a) OMIP-1 and (b) OMIP-2. The middle two panels show the standard deviation of the ensemble, with the regions where the observation is outside the 90 % confidence range of the model spread (±2σ) are hatched with red. (c) OMIP-1 and (d) OMIP-2. (e) Difference between OMIP-1 and OMIP-2 (OMIP-2 minus OMIP-1), with the regions where the difference is significant at 95% confidence level hatched with green. (f) Difference of salinity used for restoring for OMIP-1 and OMIP-2 (OMIP-2 minus OMIP-1).

- **Comment: Line 330: Would not a simple broadening of the front (irrespective of the occurrence of recirculation gyres) result in such a dipolar structure?**
  Response: As shown in Figure 7f, the observation (CMEMS: red) show a pair of positive and negative bumps relative to the multi-model mean (blue), which seems essential for the sharpening of the front along 35°N. It would also be notable that the observed sea surface

height shows a peak just to south of the front (~33°N), implying the existence of a recirculation gyre. We would like to keep the text unchanged in the revised version.

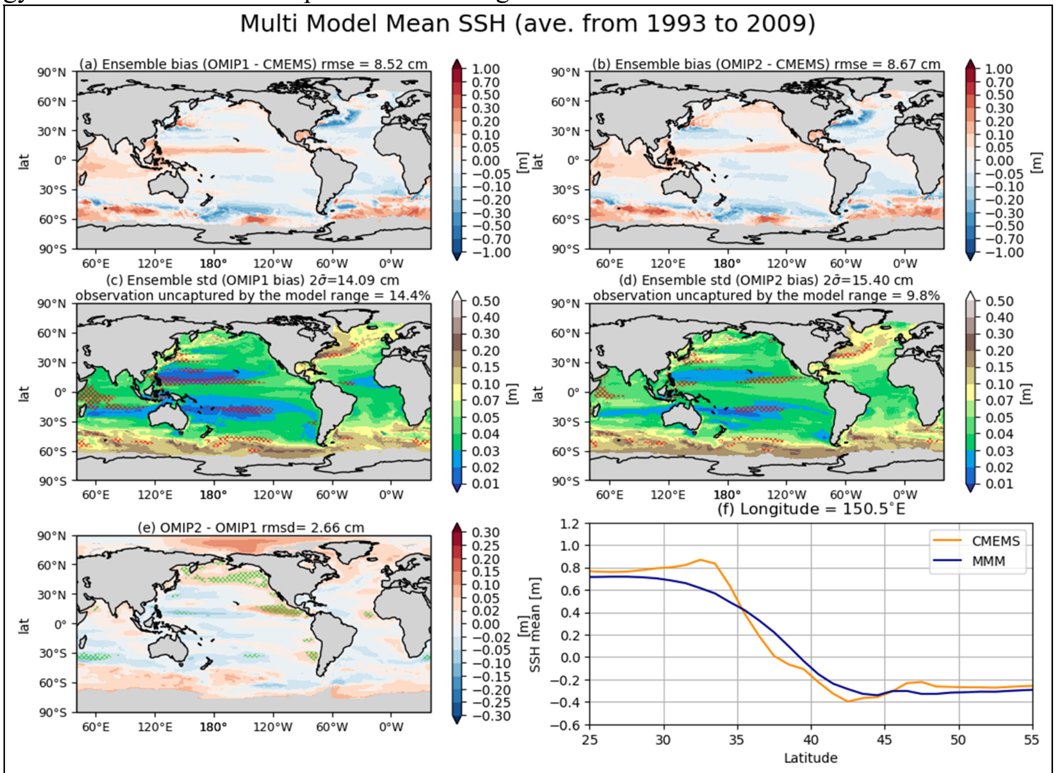

Figure 7 (to replace Figure 8 of the discussion paper except for (f), which will be kept unchanged (climatology of CMEMS)): Evaluation of simulated sea surface height (m). Upper two panels show the bias of the multi-model mean, 17-year (1993-2009) mean SSH relative to CMEMS. (a) OMIP-1 and (b) OMIP-2. The middle two panels show the standard deviation of the ensemble, with the regions where the observation is outside the 90 % confidence range of the model spread (±2σ) are hatched with red. (c) OMIP-1 and (d) OMIP-2. (e) Difference between OMIP-1 and OMIP-2 (OMIP-2 minus OMIP-1), with the regions where the difference is significant at 95% confidence level hatched with green. (f) Annual mean SSH of CMEMS (red) and OMIP-1 multi-model mean (blue) along 150.5°E in the northwest Pacific (cutting the Kuroshio Extension). Note that all SSH fields are offset by subtracting their respective quasi-global mean values before evaluation as described in Appendix C.

- **Comment: Figure 9a,b: A nonlinear color scale would be helpful to bring out more than the deep water formation sites.**
  Response: In the revised version, we will show biases of the simulated mixed layer depths (Figure 8a and 8b), but a nonlinear color scale will be used to show observational distribution of mixed layer depth (Figure 8f), which will certainly clarify the detailed distribution in the relatively shallower mixed layer depth region.

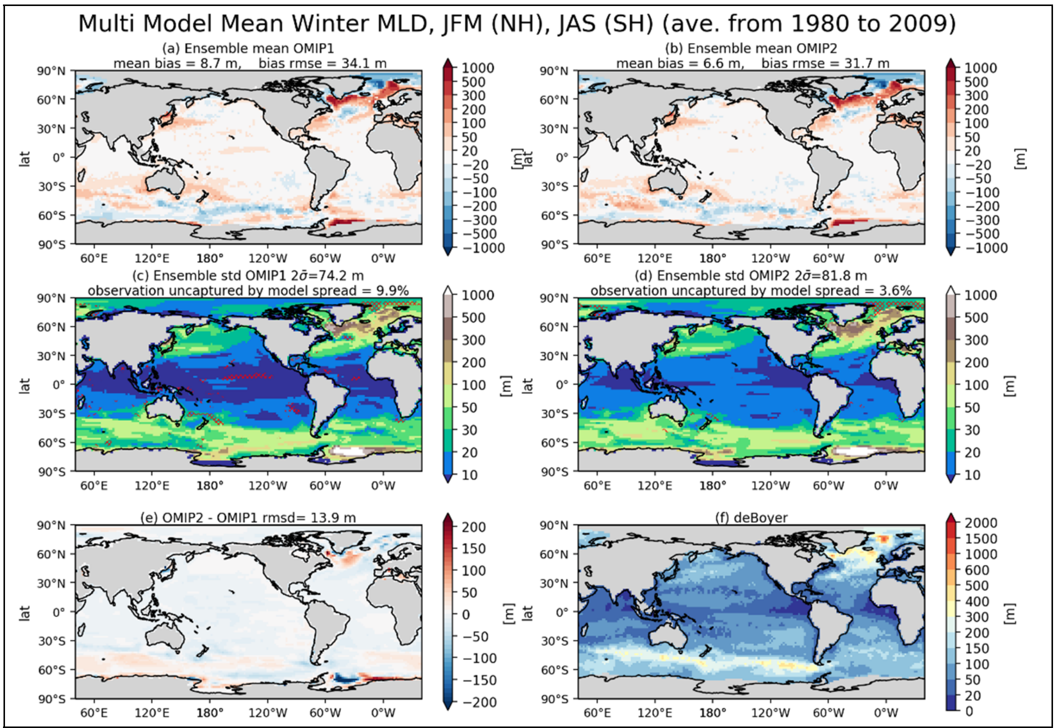

Figure 8 (to replace Figure 9 of the discussion paper): Evaluation of simulated mixed layer depth (m). Upper two panels show the multi-model mean 30-year (1980-2009) mean winter mixed layer depth bias in both hemispheres. January-February-March mean for the northern hemisphere and July-August-September mean for the southern hemisphere. (a) OMIP-1 and (b) OMIP-2. The middle two panels show the standard deviation of the ensemble, with the regions where the observation is outside the 90 % confidence range of the model spread (±2σ) are hatched with red. (c) OMIP-1 and (d) OMIP-2. (e) Difference between OMIP-1 and OMIP-2 (OMIP-2 minus OMIP-1). (f) Observationally derived mixed layer depth data from de Boyer Montégut et al. (2004).

- **Comment: Line 448 and following: It is notable that the SH mean bias improves more because the worst models get better.**
  Response: The text will be revised according to this suggestion, which will read:
  > "The overall reduction of the mean bias in the southern hemisphere in OMIP-2 in both seasons is due to the improvement of outliers."

- **Comment: Line 455 and following, Figure 22: I found this to be perhaps the most important figure when considering the limitations of the wash-rinse-repeat OMIP cycling. We really do not capture 60 years of variability with a 60 year cycle. Worth emphasizing more strongly.**
  Response: This limitation will be more emphasized throughout the paper in the revised version. For example, the following description will be added to this part:
  > "Overall, the OMIP simulations under the protocol of repeating many cycles of the entire period of the atmospheric forcing dataset do not capture variability of heat content and thermosteric sea level in the entire atmospheric dataset period. Only recent (after 1990s) upper layer heat content variability is reproduced. This limitation should be taken into account in analysing the results of the OMIP simulations."

- **Comment: Appendix B2: Figure B4 a bit of over kill to make the point (did we really**

think Drake passage transport might depend on small differences in the properties of moist air?), but oh well, only four more panels among 400!
Response: Figure B4 will be removed.

Minor typos etc and author responses:

- **Comment: Line 100: The four ... or All four ...**
  Response: This will be corrected accordingly.

- **Comment: Line 224: smaller drift**
  Response: This will be corrected accordingly.

- **Comment: Line 227: "subsurface" not clear what depth range is being described**
  Response: The depth range of 100-500 m has been intended, which will be reflected in the revised text.

- **Comment: Line 609: piston velocity**
  Response: This will be corrected accordingly.

- **Comment: Line 610: 6 cycles (to be constant with rest of text)**
  Response: This will be corrected accordingly.

- **Comment: Line 662: (CESM)**
  Response: This will be corrected accordingly.

- **Comment: Line 730: with OM4 configured**
  Response: This will be corrected accordingly.

2 Responses to Reviewer #2

General Comments

- **Reviewer comment:**

The manuscript describes overall results of ocean model intercomparision organized in the framework of OMIP-2. After the development of new surface boundary forcing dataset (JRA55-do; Tsujino et al. 2018), the performance of various ocean model simulations forced by this new dataset is now reported here.
Under the same protocol proposed by the authors, eleven state-of-the-art global ocean models are forced by not only newly developed JRA55-do atmospheric dataset but also previously referred CORE forcing. This design makes it possible for the authors to clearly evaluate what stems from the difference from the surface forcing and what is from inter-model differences.
In previous OMIP-1 comparisons, the CORE forcing by Large and Yeager (2009) was developed for surface forcing dataset. This dataset has been widely used for ocean model community but not updated after 2009, therefore, its replacement by newly developed JRA55-do is awaited. The results reported here provide us with the solid evidence that new JRA55-do dataset is good enough to replace CORE forcing as a new forcing dataset for

global ocean simulations. The manuscript also presents timely and valuable assessment about the overall performance of the state-of-the-art global ocean models.

Although the manuscript demonstrates the overall performance of global ocean simulations rather than detail analysis about specific topics, such documentation fits the scope of GMD and the ocean model and related communities will benefit from the results reported in this manuscript very much. Therefore, I can recommend the publication of this manuscript in GMD after minor revision. I have several comments which I hope will be useful for the authors to revise the manuscript before its publication.

- **Author's response**

Firstly, we would like to thank reviewers for their time and effort to review this paper and to provide constructive comments. Please read the following for how we have responded to your specific comments/suggestions.

Specific Comments and author responses

- **Comment: Line144: "absolute wind vector"–> "wind vector"**
  Response: This will be corrected accordingly.

- **Comment: Line157-163: It was difficult for me to understand the content of this paragraph. The authors appear to point out the possibility of weak bias of wind in JRA55, but its reasoning provided here is not clear. Is this related to the adjustment method of wind discussed in Sun et al. (2019)?**
  Response: There are two issues (relative versus absolute wind and with versus without surface ocean current imprints on winds) involved. This paragraph will be revised by adding a few sentences including referencing to relevant papers to complement the explanation. The paragraph will read as follows:

  > "There also remains ambiguity as to what is represented by the prescribed winds ($\overrightarrow{U_a}$) depending on the way they are constructed from the satellite-based and reanalysis atmospheric wind products. This ambiguity becomes an issue with the OMIP-2 dataset. First, its wind field is based on the JRA-55 reanalysis, which assimilates scatterometer winds yet not necessarily reproduces winds identical to scatterometer winds depending on the level of assimilation constraints. Since scatterometer winds represent wind relative to the surface current (e.g., Plagge et al., 2012) and contain imprints of surface currents (Renault et al., 2017, 2019b), assimilating scatterometer winds directly, yet not identically, to the absolute surface winds would make the feature of surface winds of the JRA-55 reanalysis somewhat ambiguous. Second, only the long-term mean JRA-55 winds are adjusted with respect to the satellite-based winds in constructing the OMIP-2 dataset (JRA55-do). As a result, the long-term mean winds of the OMIP-2 (JRA55-do) dataset could be regarded to be replicating the scatterometer winds, but ocean current imprints on them have not been clarified yet. Thus, it is not possible to accurately correct them to reconstruct absolute winds using the wind speed correction approach described above. On the other hand, in short time scales, ocean current imprints on winds are shown to be small, if not negligible, in the OMIP-2 (JRA55-do) forcing dataset (Abel, 2018), which would make them possible to be treated as absolute winds at least in short time scales. A future version of the OMIP-2 dataset should resolve this ambiguity. Readers are referred to Renault et al. (2020) for more discussion on the issues of using satellite derived winds to force uncoupled ocean models.".

- **Comment: Line240-248: I think that the content of this paragraph appears to focus merely on a technical issue of the model and is not very useful.**
  Response: The paragraph is intended to explain the reason why we do not adopt global mean salinity, which would be virtually constant, as metrics. In the revision, we will more explicitly state this point. Specifically,

  "Note that in contrast to heat content, the total salt content in the ocean–sea-ice system is essentially constant in nature. In most participating models, the global salt content in the ocean–sea-ice system is explicitly conserved, which is achieved by removing the globally integrated salt flux arising from salinity restoring at each time step (salinity normalization). The same adjustment is applied to surface freshwater flux in most participating models, resulting in conservation of total mass of water in the ocean–sea-ice system. Thus, in such models, variation of global mean salinity only occurs due to variation of sea-ice volume and the global mean salinity would not be normally employed as a metric for the purpose of model intercomparison. Figure 3 implies that global mean salinity increases for the first 10 to 15 years of each forcing cycle and then decreases for the rest of the cycle in both OMIP-1 and OMIP-2 simulations. It also implies that a long-term drift of global mean salinity does not occur in those models that have applied both salinity and freshwater normalization."

- **Comment: Line295-296: In Figure 5, improvement from OMIP1 to OMIP2 can be found generally around the Eastern boundary regions of both Pacific and Atlantic basins. Therefore, rather specifically referring to Benguela region, the sentence here could be modified such as "It is also the case for the Eastern boundary region in the Atlantic basin, but the warm bias is somewhat exacerbated offshore in OMIP-2".**
  Response: Thank you for the suggestion. The text will be corrected accordingly.

- **Comment: Line323-325. This sentence is not clear. Do the authors just describe slight difference between OMIP-1 and OMIP-2 in (northern) equatorial Pacific area?**
  Response: Yes, both OMIP-1 and OMIP-2 ensemble spreads fail to capture the observation there and we thought that this is worth mentioning. This part will be revised as follows (see also Figure 7 of this document):

  "A zonally elongated pattern of positive bias occurs from the western to central basin in OMIP-1 and from the central to eastern basin in OMIP-2. Both OMIP-1 and OMIP-2 ensemble spreads fail to capture the observation there (Figs. 8c and d (Figs.7c and d of this document))."

- **Comment: Line335-336: How about mentioning about the largest difference in the Arctic Ocean? (This seems related to salinity difference there)**
  Response: The largest SSH difference in the Arctic Ocean will be mentioned along with the salinity difference that could possibly explains this difference, which will read as follows:

  "A large difference in sea surface height is found in the eastern Arctic Ocean, with OMIP-2 higher than OMIP-1. This difference is presumably related to the fresher upper ocean salinity (and thus less dense) found in OMIP-2 (Figure 6e)."

- **Comment: Line444-445: It would be better to replace the word "hiatus" by "slowdown".**

Response: The word “hiatus” will be replaced by “slowdown” throughout the manuscript.

- **Comment:** **Section 6 (Line492-525): Many figures are prepared for this section (Figs. 25-31) with very short description provided. It is nice to see improvement from OMIP-1 to OMIP-2 in some statistics here but it appears better that the authors focus on the key result in the main text and most of the figures will be moved to Appendix.**
  Response: Following the suggestion, section 6 will be moved to Appendix D.

- **Comment:** **Line573-574: “will be therefore become”–>”will therefore become”**
  Response: This will be corrected accordingly.