<u>Reviewer 1</u>

Review of "Global ocean and sea ice variability simulated in eddy-permitting climate models", by Yushi Morioka, Eric Maisonnave, Sébastien Masson, Clement Rousset, Luis Kornblueh, Marco Giorgetta, Masami Nonaka, Swadhin K. Behera, submitted to *Geoscientific Model Development*.

Oceanic mesoscale eddies account for 70% of kinetic energy in the global ocean and play a significant role in the transport of mass, heat, and nutrients. The accurate representation of oceanic mesoscale eddies in numerical models is thus essential. By developing an oceanic eddy-permitting (~25 km) coupled climate model SINTEX-F3 and comparing with the coarse-resolution counterparts, this study assesses the model performance at various resolution in simulating global ocean and sea ice variability as well as the climate variability over different regions including the tropical ocean, western/eastern boundary current regions, and polar regions. I recommend major revision prior to the publication of this manuscript and my concerns are listed below.

We thank the reviewer very much for providing constructive comments on the original manuscript. Those are very helpful for further improving the manuscript. In particular, we have examined physical processes underlying the improvement of ocean and sea ice variability in the high-resolution models and discussed the differences in the ocean and sea ice variability between the eddy-present and eddy-rich models. With this revision, we believe we have addressed concerns raised by the reviewer. Our point-by-point responses to the reviewer's comments are provided below with a blue front.

Major comments

1. In this study, the authors attributed most improvement in the western/eastern boundary current regions and sea ice variability to increased oceanic resolution. However, the findings are primarily descriptive and lack sufficient mechanism analysis or plausible mechanism conjecture on how the increased oceanic resolution, and thus, the better presented mesoscale eddies contribute to these improvements through various oceanic dynamic processes or air-sea interaction processes. In addition, models with finer oceanic resolution are typically accompanied by higher atmospheric resolution, and the simple comparison across LR and HR ensembles is not sufficient to highlight the relative importance of refining oceanic and atmospheric resolution to benefit the model simulation. The resolution hierarchy including solely increasing oceanic and atmospheric resolution for ECMWF-IFS and HadGEM-GC31 may provide a testable approach to address this issue.

This study has clearly demonstrated the importance of increased model resolution particularly in the form of ocean mesoscale eddies that play an important role in the SST and sea ice variability in the western/eastern boundary current and polar regions. To understand the impact of ocean mesoscale eddies, we have calculated the mixed-layer depth and sea surface height anomalies regressed onto the climate indices (Figs. 10-11, 16-17, 22-23). In the western/eastern boundary current regions, SINTEX-F3 and CMIP6-HR tend to have insignificant variations in the mixed-layer depth but significantly higher SSH in the core area of warmer SST, while SINTEX-F2 and CMIP6-LR show significantly shallower mixed-layer and higher SSH in broad areas. These results suggest that the warmer SST in the low-resolution models may be more related to shallower mixed-layer that enhances the warming of mixed-layer by shortwave radiation (e.g., Morioka et al. 2010), while in the high-resolution models, warm mesoscale eddies tend to advect more warm water from the subtropical region and offset the shallower mixed-layer through anomalous downwelling.

In the Antarctic seas, the observation shows that sea ice increase in the open water is associated with the shallower mixed-layer and lower SSH that prevents upwelling of warm subsurface water due to stronger stratification, whereas in the coastal region, it is accompanied by the thicker mixed-layer and lower SSH that represents weaker stratification as a result of brine rejection. SINTEX-F3 and CMIP6-HR can simulate the shallower mixed-layer in the open water but fail to simulate the thicker mixed-layer in the coastal region probably due to insufficient model resolutions. Therefore, the improvement of Antarctic sea ice variability requires not only an improvement of sea ice model physics but also an increase in the model resolutions.

Second, to clarify the impact of the increased ocean resolutions on the SST and sea ice simulations, we have discussed the difference in the resolution hierarchy of ECMWF-IFS and HadGEM-GC31 in the main text. For example, the increased ocean resolution from the eddy-free ECMWF-IFS-LR to the eddy-present ECMWF-IFS-MR improves the global SST bias (Table S1) and Arctic SIC bias (Table S2), NINO3.4 SST (Table S3). On the other hand, the increased ocean resolution from the eddy-present HadGEM-GC31-HM to the eddy-rich HadGEM-GC31-HH leads to stronger NINO3.4 SST (Table S3), IOD SST (Table S4), larger Antarctic SIC (Table S8), and smaller Arctic SIC (Table S9), but the last three variables get worse than the observation. Given these results, it is clear that the improvement of the SST and sea ice variability is not only related to the increased ocean resolution but also to the increased atmospheric resolution. We have added this discussion in the relevant paragraphs.

2. The authors categorized 13 CMIP6 HighResMIP models into 3 low-resolution (CMIP6-LR) models and 10 high-resolution (CMIP6-HR) models for comparison with the SINTEX-F2 and SINTEX-F3 models. In fact, following Hewitt et al. (2020), the simulations can be classified into three regimes according to the resolution of their ocean components under the guidance of Rossby deformation radius: eddy-free (≥50 km), eddy-present (~25 km), and eddy-rich (~10 km). They suggested that whether the ocean mesoscale is explicitly represented in eddy-rich regime or parameterised in eddy-free/present regimes affects not only the mean state of the ocean but also the climate variability and the future climate response. Given such difference between eddy-rich and eddy-present models, is it reasonable to simply classify them together into CMIP6-HR? In addition, I noticed a considerable intermodal dispersion indicated by the large standard deviation among the CMIP6-HR ensemble. Is it partly attributable to disparities in different resolution regimes? Since the metric calculated from CMIP6-LR is included in the one standard deviation interval of CMIP6-HR, taking Figure 2b as reference, is the difference between CMIP6-LR and HR considered to be statistically significant?

We categorized the eddy-present and eddy-rich models as the HR models, because only two models (HadGEM3-GC31-HH and CESM1-CAM5-SE-HR) were available for the eddy-rich models and there were small differences in statistical metrics (Tables S1-9). For example, the eddy-rich HadGEM3-GC31-HH shows almost similar values with the eddy-present HadGEM3-GC31-HM for global SST bias (Table S1) and Antarctic SIE bias (Table S2), ATL3 SST (Table S5), Agulhas SST (Table S6), and Dakar SST (Table S7) regression values. On the other hand, HadGEM3-GC31-HH shows stronger NINO3.4 SST (Table S3), stronger IOD SST (Table S4), larger Antarctic SIC (Table S8), and smaller Arctic SIC (Table S9) than HadGEM3-GC31-HM, but the last three variables get worse than the observation. Given these results and a limited availability of the eddy-rich models, it is difficult to discuss statistical differences in the metrics between the eddy-present and eddy-rich models. Therefore, we have combined the eddy-present and eddy-rich models together in this study, and compared the eddy-rich HadGEM3-GC31-HH against the eddy-present HadGEM3-GC31-HM as an example to discuss the impact of the increased ocean resolution in the main text.

Furthermore, to examine the large intermodal differences in the time series of the climate indices, we have performed Welch's test to the time series of the sea ice extent between CMIP6-LR and CMIP6-HR models. For example, the climatological difference in the Antarctic sea ice extent is not statistically significant (Fig. 2b), while that in the Arctic sea ice extent is significant (see open circles in Fig. 3b). On the other hand, we have confirmed significant increase in the standard deviation of the Antarctic sea ice extent anomalies (Fig. 19a) but we have not seen significant differences for the Arctic sea ice extent anomalies (Fig. 19b). These results indicate that the increased model resolutions have different impacts on the sea ice mean state and variability in the polar regions. We have added these discussions in the relevant paragraphs.

3. In this study, the authors defined a series of climate indices within the same region and made comparisons of the simulation performance across multi-models at various resolutions. Could this introduce additional errors into the mode evaluation? Taking Figure 9 as an example, it seems that the defined region is not focused on where the salient center of ocean heat release is located. In addition, it has been widely recognized that the simulated western boundary currents might present meridional displacement bias compared to observation. Should it be more suitable to adjust the reanalysis regions according to the model behavior?

We defined the box regions for climate indices following previous studies (e.g., NINO3.4) or the standard deviation of the observed SST (e.g., Fig. S8). These climate indices, defined in earlier studies to characterize basin-scale ocean-atmosphere interactions, are evaluated in models to determine the fidelity of their representation. For other indices, for example, in the Agulhas region, we have found that the rectangular box in the Agulhas Retroflection region was wrongly displayed in the figures, inconsistent with the description of the box region (15-30° E, 40-42° S) in the figure caption. So, we have corrected the rectangular box in the figures. Furthermore, we have tested different boxes in the Agulhas Retroflection region, but we have found that the results are almost similar to the case when using the original box. For example, the corresponding figures using a larger box (20-35° E, 40-44° S) are shown in Figs. S9-13. The spatial patterns of the warmer SST, larger upward surface heat flux, shallower mixed-layer depth, higher SSH, and higher

SLP regression values do not differ from those using the original box. Therefore, we have remained using the original box to discuss the physical process in the section.

4. This study demonstrated limited improvement of simulation for the tropical climate variability such as ENSO in most of the eddy-permitting models, and found a remarkably reduced warm bias over the tropical ocean in SINTEX-F3 model compared to the coarse-resolution SINTEX-F2. This is different from the finding of Liu et al. (2022), which highlights that eddy-present models improve ENSO patterns because of the realistic mean state and associated SST-net heat flux feedback and underscores the reduced model cold biases of the equatorial SST mean state in higher-resolution models, which is attributed to increased oceanic resolution, and thus, better resolved eddy-driven heat transport. What factors contribute to these divergent conclusions?

We tend to believe the limited improvement of ENSO amplitude in high-resolution models is consistent with the results of Liu et al. (2022). This is because not all high-resolution models in their paper show significant improvement of several ENSO characteristics compared to low-resolution models (see their Fig. 2b). Only PRIMAVERA models excluding CESM1-3 model show significant improvement of ENSO pattern only (see their Fig. 2c). In particular, for ENSO amplitude, three out of five high-resolution models (i.e., CESM1-3, CNRM-CM6-1, and HadGEM3-GC31) show degradation and simulate weaker ENSO than the observation, similar with SINTEX-F3. Liu et al. (2022) clearly mentioned that the uncertainty of modeled ENSO amplitude tends to be heightened in high-resolution models, because the inter-model spread becomes larger in high-resolution models compared to low-resolution models. Therefore, our SINTEX-F3 results support their argument on the limited improvement of ENSO characteristics in high-resolution models. We have added this in the Discussion section.

Reference:

Hewitt, H. T., and Coauthors, 2020: Resolving and parameterizing the ocean mesoscale in Earth system models. Curr. Climate Change Rep., 6, 137–152, https://doi.org/10.1007/s40641-020-00164-w.

Liu, B., and Coauthors, 2022: Will increasing climate model resolution be beneficial for ENSO simulation? Geophys. Res. Lett., 49, e2021GL096932, https://doi.org/10.1029/2021GL096932.

Minor comments

1. In the introduction section, the authors mentioned the role of ocean mesoscale eddies as well as their impacts on the overlying atmosphere in the climate system. In fact, recent studies (Gan et al. 2023; Yu et al. 2024) reveal that such mesoscale and frontal-scale ocean–atmosphere coupling could significantly increase the formation of subtropical mode water, which is likely exerting far-reaching impact on the regional and global climate. It should be explicitly addressed in the introduction.

We have cited those papers in the last paragraph of the Introduction.

Reference:

Yu, J., B. Gan, H. Yang, Z. Chen, L. Xu, and L. Wu, 2024: Mesoscale Ocean–Atmosphere Coupling Effects on the North Pacific Subtropical Mode Water. J. Phys. Oceanogr., 54, 1467–1488, https://doi.org/10.1175/JPO-D-23-0148.1.

Gan, B., and Coauthors, 2023: North Atlantic subtropical mode water formation controlled by Gulf Stream fronts. Natl. Sci. Rev., 10, nwad133, https://doi.org/10.1093/nsr/nwad133.

1. Table 1: Spatial resolution of “Ocean-Sea Ice” should be specified for the SINTEX-F2 and SINTEX-F3 coupled models, which represents one of the most significant differences between the two versions of the SINTEX-F model.

We have specified.

2. Line 216: Replace "Is" with "is".

We have corrected.

3. Line 222: Replace "^{5t}h" with "5th".

We have corrected.

4. Line 245: You mentioned “the atmospheric component of the SINTEX-F2 is based on the ECHAM6”. This disagrees with the model information listed in Table 1. Which is correct? Please check it carefully.

This is a typo. We have replaced it with SINTEX-F3.

5. Section 2.3: As for the CMIP6 HighResMIP models, which variant labels did you use? Please clarify. Looking closer to Table 2, it is better to express oceanic resolution in kilometer units to keep correspondence with the atmospheric component as well as the information provided on the official website. In addition, it is wrong that the HadGEM3-GC31-HH model is categorized into LR.

We used the model experiments with the variant name of r1i1p1f1. Following the suggestion, we have expressed the oceanic resolution in km and corrected the HadGEM3-GC31-HH in Table 2.

6. Lines 256–258 and Lines 273–275: A linear trend of the variables in SINTEX-F model was removed to avoid the influence from model drift. Given that the 30-50 years spin-up period in CMIP6 may be insufficient especially for HR models, was identical data processing applied to the CMIP6 model outputs?

Yes, we removed a linear trend from the variables of CMIP6 models over the analysis period considering the model drifts. We have added this data processing in the Methodology.

7. The shaded regions denoting one standard deviation of the model spreads for LR and HR ensemble overlap with each other in most of the pictures in the manuscript. Maybe try changing the color or increasing the transparency to make it clear.

Following this and second reviewers' comments, we have modified the colors of all figures and increased the transparency of the color shade to make them clearer.

8. It is mentioned many times in the figure caption that because of the limited number of the CMIP6-LR models, we did not apply statistical test to the regression values for the CMIP6-LR models. How do you apply the statistical test for multi-model ensemble? Should you conduct the regression analysis before calculating the multi-model ensemble mean, or directly conduct regression in the multimodel-averaged time series? Why do the model numbers limit the statistical test? Please be specific.

We conducted the regression analysis for each model and calculated the multi-model ensemble mean of the regression values. Then, we performed the statistical test to the multi-model ensemble mean values, so the degree of freedom for the LR models is too small (i.e., 3-1=2) to evaluate the statistical significance. We have clarified the details on the statistical testing in the caption of Fig. 5.

9. Remove line numbers from blank lines.

We have removed.

10. The Results section should provide a concise synthesis rather than exhaustive phenomenological descriptions without interpretative conclusions.

We have modified the Results section to make it succinct with more solid evidence (e.g., statistical metrics and significance testing in Tables S1-9) rather than narratively vague descriptions.

11. The manuscript needs to be further polished and have more professional phrasing. I recommend having the text checked for English style.

We utilized a professional English editing service for scientific journals and had the manuscript proofread by native English speakers to further improve the readability.