REFEREE REPORT(S):
Referee: 1

General Comments:

This manuscript presents an assessment of the HighResMIP models regarding four well-known, long-standing biases. Authors analyse simulations of 5 climate models in at least two different resolutions. Four of the five models have increased resolution in the atmospheric and oceanic components simultaneously, whereas 1 model has only increased resolution in the atmospheric component. The set of experiments consists of historical runs from 1950 to 2014 in two versions: AMIP and coupled simulations. The author's results suggest some improvements of the model performance regarding the analysed biases with increased atmospheric resolution. Yet, those improvements are not consistently found in all models. No systematic improvement is shown by increasing the resolution from eddy-parametrized into eddy-permitting ocean models. However, the only eddy-resolving ocean model presents improvements in reproducing the North Atlantic temperatures and the path of the Gulf Stream. Overall this study shows limited benefits for reducing the long-standing biases from the high resolution based on the ensemble mean of HighResMIP. This result yields a recommendation for the modelling community: in addition to model resolution, future efforts should be oriented to improve model physics.

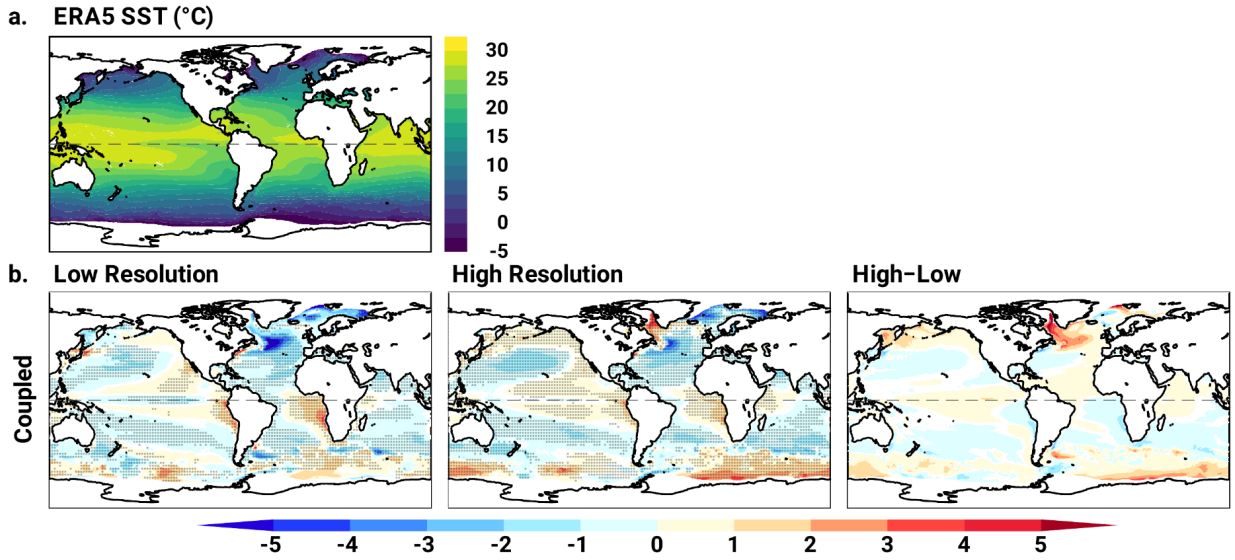
The manuscript is well written and structured, the methodology is correctly explained, and the results are relevant. All in all, I believe that this manuscript is relevant for the climate modelling community, and it is worth to be published in GMD. In some places, the results described in the text are difficult to see in the figures, and I think that the original manuscript could be improved with a moderate revision. I report in what follows some comments and suggestions.

We thank the Reviewer for the thoughtful comments. In the following we answer each specific point (in blue).

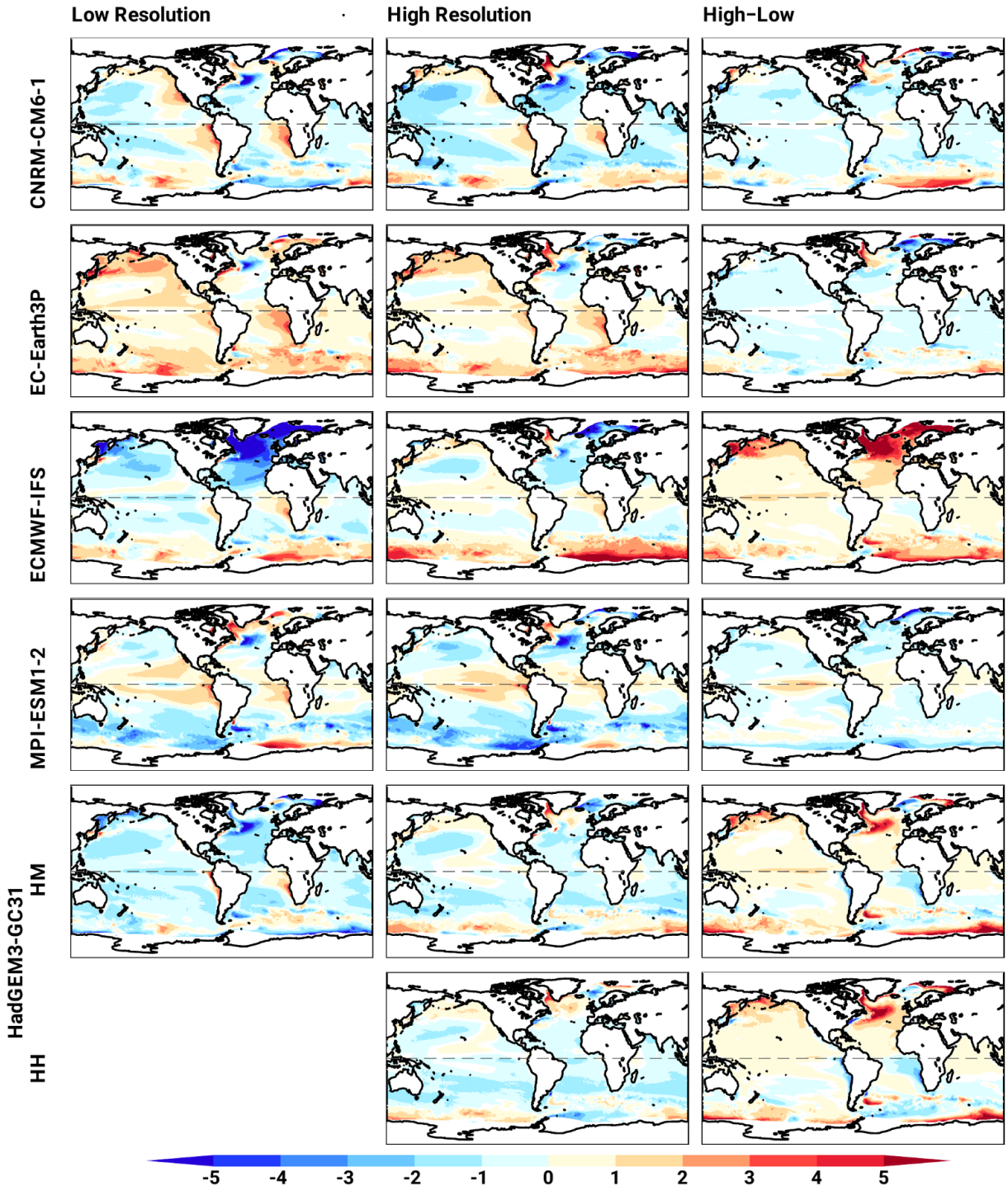
Specific Comments:

1) All the variables assessed in this manuscript come from the atmospheric model component. I think that analysing the model bias of at least sea surface temperature (tos) could help complete the study. For example, the known Southern Ocean warm bias or North Atlantic cold bias in climate models are reported for the upper-ocean temperatures. The authors assess the 2m-air temperature (tas), which is fair, but it is not the same value as tos in the open ocean. Analysing the tos bias in climate models has an only sense for the coupled runs. But still, a different dataset (from ERA-interim) as observations should be used, and you could get new results. I would recommend looking at sea surface temperatures too.

Biases in sea surface temperature (SST) in the coupled models are nearly identical to those in the near-surface air temperature (SAT; compare the two Figures below with Figs. 1 and S2 in the submitted manuscript). Given that SSTs add no new information about temperature biases in the models (and are limited to the coupled models only, in contrast to the SAT), we will not include these figures, although we mention the results in Section 2 of the revised manuscript:
"Biases in SST (not shown) are very similar to those in SAT, which suggests SAT biases are dominated by the SST ones over the ocean."



Responses Fig. 1. (a) ERA5 sea surface temperature (SST; in °C), which are derived from the HadISST2 and OSTIA SST/SIC products (Hersbach et al., 2020). (b) Left and middle: Multi-model ensemble mean bias in SST (in °C) in the coupled models at low (left) and high (middle) resolutions. Biases are with respect to ERA5 (shown in a.). Stippling masks where at least four out of the five models agree on the anomaly sign. Right: Difference between the two resolutions. In all panels non-significant anomalies at the 5 % level (based on a two-tailed Student's t test) are masked white. The Equator is a dashed line in all the panels.



Responses Fig. 2. Mean bias in SST (in °C) in each individual coupled model at low (left) and high (middle) resolutions, and as the difference between the two (right). Biases are with respect to ERA5. Non-significant anomalies at the 5 % level are masked white. The Equator is a dashed line.

2) When discussing model biases regionally (in particular in sections 4.1, 5.1, 5.2), I think that a table similar to Table 2 (which I found very clear and helpful) could help to summarise the results. I suggest including additional tables in which the metrics RMSD and Bias are regionally computed.

Tables with regional biases have been included in the Supplementary Material (Tables S1–S4 in the revised manuscript) and the results are discussed throughout the revised manuscript. For example in Section 5.1:

"Increased resolution has a mixed effect on the SO warm bias and, although it seems to increase in the ensemble mean (Fig. 1), this varies substantially across models (Fig. S3 and Table S3): the CNRM-CM6-1 model experiences a reduction of a cold bias over the Weddell Sea up to about 4 °C; the EC-Earth3P warms along the Antarctic coast and its widespread SO warm bias persists at HR; the ECMWF-IFS model shows an increase of its temperature bias by about 1.5 °C on average and very strongly locally in the Weddell Sea by over 5 °C; the MPI-ESM1-2 shows a mean cooling over the SO of about 0.5 °C and becomes cold biased especially to the west of the Antarctic Peninsula; and the HadGEM3-GC31 model shows a reduction of its coastal cold bias, developing instead a more widespread warm bias with local values up to about 1–2 °C—although the cold bias over the Weddell Sea persists in the HadGEM3-GC31 eddy-rich model."

3) I may be wrong since I am not a native English speaker, but I find the concept 'improving/worsening the bias' (used several times along the manuscript) a little bit confusing. I would say that the model is improved/worsened by a reduction/increase in bias.

This has been rephrased throughout the text. Two examples from the revised manuscript:

"A reduction in the temperature and cloud biases in the eastern tropical oceans might reduce current uncertainty about climate sensitivity (Andrews et al., 2019)"

"This paper examines whether increased horizontal resolution reduces four well-known, long-standing climate biases in five global models developed within the PRIMAVERA project"

Minor Comments:

L182. Please re-order (Kato et al., 2018; Loeb et al., 2018).

Done.

L189-190. 'The statistical significance of the anomalies between models and observations is calculated based on a two-tailed Student's t-test at the 5 % level.' I do not understand this very well. For example, in figure 1, the stippling mask indicates areas where four out of five models agree on the sign of the bias. In the same figure, however, non-significant anomalies at the 5% level are masked white. What do you mean by 'anomalies' here? Are you referring to the bias itself? How did you consider the degrees of freedom? How many independent observations do you consider here? Please, specify.

Figures 1,2,4,5,6 include two sorts of significance analysis. The first one (stippling) masks areas where four out of the five models agree in the anoamy sign (where anomaly means the difference between a model, or the ensemble mean, and an observation). The second test (white mask) is a standard two-tailed Student's t test of whether the two samples (the model and the observation) have the same mean (null hypothesis), assuming they have different variances (Section 6.6: Test of the Mean; in Von Storch and Zwiers, 1999. *Statistical Analysis in Climate Research*. Cambridge Univ. Press). The sample size is the number of the years over which the mean is computed, indicated in Section 2.2. This point has been clarified in the revised manuscript:
"The statistical significance of the differences between models or the ensemble means and the observations is calculated for each variable based on a two-tailed Student's t test at the 5 % level, in which the null hypothesis is that the two samples (model and observations) have the same mean over the above-mentioned periods, assuming the two samples have different variances (von Storch and Zweirs, 1999). The associated non-significant values are masked in white in Figs. 1,2,4,5,6 and all the Supplementary Figures. An additional test is applied in Figs. 1,2,4,5,6 (shows as stippling) to measure the agreement in the difference's sign of the ensemble members with respect to observations.

L222. 'the bias is especially persistent': do you mean persistent across the models? Isn't the term 'persistent' related to time?
Rephrased: "At LR, the bias extends over the eastern tropical South Atlantic and South Pacific from the coast equatorward."

L240. In order of appearance, I would change the figure numbering and name Fig. 2 to current Fig. 4. Similarly, Fig. 5 would be Fig. 3.
Figures 2 and 3 are referred to in the Introduction, in the double ITCZ subseccion (1.1.2). Therefore, the figure order has not been changed.

L308. I guess that it is Fig. S6 instead of S8.
Corrected.

L317-318. Is the dry bias at SH mid-latitudes associated with the warm bias in the SO? I see the dry bias for each LR and HR model (Fig. S4) and even in the AMIP runs (Fig. S3).
This has been rephrased.
"The LR coupled models also present a dry bias at mid-latitudes (Fig. 2)."

L323. What do you mean by: 'an improvement of the cold bias in the Weddell Sea'?
It means the cold bias over the area gets reduced with increased resolution in the CNRM-CM6 model. This has been clarified as: "The CNRM-CM6-1 model experiences a reduction of a cold bias over the Weddell Sea up to about 4 °C".

L330-333. A table as table 2 with RMSD and mean bias computed regionally in the SO would help in this kind of statement.

Tables with regional biases have been included in the Supplementary Material (Tables S1–S4) and the results are discussed throughout the revised manuscript. For these particular lines, the revised text now is: "Increased resolution has a mixed effect on the SO warm bias and, although it seems to increase in the ensemble mean (Fig. 1), this varies substantially across models (Fig. S3 and Table S3): the CNRM-CM6-1 model experiences a reduction of a cold bias over the Weddell Sea up to about 4 °C; the EC-Earth3P warms along the Antarctic coast and its widespread SO warm bias persists at HR; the ECMWF-IFS model shows an increase of its temperature bias by about 1.5 °C on average and very strongly locally in the Weddell Sea by over 5 °C; the MPI-ESM1-2 shows a mean cooling over the SO of about 0.5 °C and becomes cold biased especially to the west of the Antarctic Peninsula; and the HadGEM3-GC31 model shows a reduction of its coastal cold bias, developing instead a more widespread warm bias with local values up to about 1–2 °C—although the cold bias over the Weddell Sea persists in the HadGEM3-GC31 eddy-rich model. In contrast to temperature, biases in cloud cover and net cloud radiative effect remain relatively unchanged between LR and HR (Figs. 4 and 5). The CNRM-CM6-1 shows a 1 % reduction in its mean cloud cover bias over the SO, while the ECMWF-IFS and MPI-ESM1-2 models show a 1–3 % increase over the SO (Table S3). Similarly, the ECMWF-IFS model shows a 1.5 Wm–2 mean reduction while the MPI-ESM1-2 model shows a 4 Wm–2 mean increase in their net cloud radiative effect biases over the SO (Figs. S6–S9)."

L333-337. Couldn't it also be the ocean model resolution or physics?

We have expanded this line to include the possibility that ocean resolution/physics have played a role in the Souther Ocean biases both in Section 5.1 (Southern Ocean bias) and Section 6 (Discussion and Conclusions):

"Given the small reduction in the cloud cover and net cloud radiative effect biases with increased resolution, the change in the temperature bias over the SO might be related to a change in the sensitivity of the HR coupled models to the similar cloud and radiation biases, or to development of further biases, for example, in the sea ice, mixed layer depth, air–sea heat fluxes, or the strength of the Antarctic Circumpolar Current (e.g., Roberts C.D. et al., 2018b). Some of these biases might, in turn, be linked to the disabling or not of the mesoscale eddy mixing at higher resolution (Roberts C.D. et al., 2018b), as discussed in Section 6."

"When additional model configurations are available, the benefit of bias reduction from increasing ocean resolution alone can be assessed. For the ECMWF-IFS model, increased ocean resolution from 1° to 0.25° reduces the North Atlantic, Arctic, and equatorial Pacific temperature biases but increases the SO warm biases (Roberts C.D., et al., 2018b). For the HadGEM3-GC31 model, increased ocean resolution up to an eddy-rich one (0.08°) improves the Gulf Stream

separation (Roberts M.J. et al., 2019) and representation (Moreno-Chamarro et al., 2021), although the eddy-rich resolution by itself has a modest impact on reducing surface temperature biases compared to the eddy-present (0.25°; Fig. S3 and Roberts M.J. et al., 2019). For the MPI-ESM1-2 model, the North Atlantic temperature and the Gulf Stream separation are also more realistic in an eddy-rich ocean (~0.1°) compared to the LR and HR versions used in our study (Gutjahr et al., 2019). These results thus suggest that an eddy-rich ocean resolution might be key to reducing North Atlantic and Southern Ocean temperature biases, which is consistent with previous studies (e.g., Mertens et al., 2014; Xu et al., 2014b). Particularly important for such biases might be the treatment of the mesoscale eddy mixing at the eddy-present resolution because mesoscale eddies become smaller at higher latitudes and are therefore not fully resolved at the eddy-present resolution (0.25°). Thus, for example, while the CNRM-CM6-1 (Voldoire et al., 2019b), EC-Earth3P (Haarsma et al., 2020), and MPI-ESM1-2 (Gutjahr et al., 2019) HR models respectively use a Smagorinsky scheme, and the Gent and Mcwilliams (1990) and K-profile parameterizations (KPP), the ECMWF-IFS (Roberts C.D., et al., 2018b) and HadGEM3-GC31 (Roberts M.J. et al., 2019) HR models switched off the Gent and Mcwilliams (1990) parametrization. Subtle differences in the model physics due to increased resolution might therefore exert a strong influence on model biases."

L365. 'All models but MPI-ESM1-2', not sure if also EC-Earth3P and CNRM-CM6-1. Again, here it would be helpful to have a table with the regional metrics.
This has been clarified: "Increased model resolution reduces the magnitude of the cold bias by about 1 °C on average (Table S4) and locally up to 2–3 °C in the ensemble mean (Fig. 1). There are, however, important differences across the ensemble members (Fig. S2). The EC-Earth and CNRM-CM6-1 HR models show relatively small local reductions of the cold bias by about 0.5–1 °C over the central subpolar North Atlantic. The lack of a clear improvement in these two HR models might be related to the unchanged ocean physics between the low and high resolutions (Section 2). The MPI-ESM1-2 shows no changes in the biases between resolutions over the subpolar North Atlantic but a strong cooling up to about 4 °C over the Nordic Seas, likely related to misrepresented local sea ice."

L373. 'In all the HR models, the cold bias over the subpolar North Atlantic is replaced by a warm bias'? I don't see it. Are you referring to the warm bias west of Greenland? If it is the case, it is not valid for MPI-ESM1-2.
This has been clarified: "On average at HR, the cold bias over the subpolar North Atlantic is replaced by a warm bias up to about 2–3 °C over the Labrador Sea (Fig. 1). The warming of the entire subpolar North Atlantic is, in fact, one of the most remarkable differences at increased resolution in the ensemble mean. The warming is especially prevalent in the NEMO models at the 0.25° resolution, in which the warm bias is likely related to a stronger oceanic heat transport in the North Atlantic and a reduced sea ice (Roberts M.J. et al., 2020b) than at lower resolutions, linked to a too strong ocean deep mixing in the Labrador Sea (Koenigk et al., 2021). In the

MPI-ESM1-2 models, by contrast, a warm bias is already present at LR and, together with the cold bias in the central North Atlantic bias, remains unchanged at HR (Fig. S3)."

L379. 'the cloud cover bias remains relatively unchanged' isn't easy to quantify from the figures.
In the ensemble mean and in most of the individual models, changes in cloud cover between resolutions are of about ±5 % over the entire North Atlantic, with no evident changes in the pattern (Figs. 3, S5, and S6). This has been rephrased in the revised manuscript as "The change in the cloud cover bias in the ensemble means is relatively small, of about ±5 % over the entire North Atlantic, with no clear changes in the pattern (Figs. 3)." In addition we have included tables for the mean biases over each region to help quantify the changes (Tables S1–S4).

L398. 'tropics' instead of 'tropis'.
Corrected.

L416. How do you see that 'the Gulf Stream separation improves'? In HH, it appears a cold bias close to the coast.
This has been clarified. "For the HadGEM3-GC31 model, increased ocean resolution up to an eddy-rich one (0.08°) improves the Gulf Stream separation (Roberts M.J. et al., 2019) and representation (Moreno-Chamarro et al., 2021), although the eddy-rich resolution by itself has a modest impact on reducing surface temperature biases compared to the eddy-present (0.25°; Fig. S3 and Roberts M.J. et al., 2019)."

L432. In general, I would prefer the terms' eddy-parametrised/permitting/resolving' for the ocean resolution. But you are using 'eddy-parametrised/present/rich', so eddy-permitting here is not consistent. Besides, I think that you are referring to 'eddy-rich' instead of eddy-present' here.
We would like to keep the current eddy-parametrized, -present, -rich naming because it describes the models more accurately. For example, an "eddy-resolving" resolution of 1/12° does not actually resolve the typical ocean mesoscale eddies at high latitudes; a finer ocean resolution (1/20° or so) would be needed. The eddy-rich naming has become more frequent in recent literature (e.g., Malcolm M.J. et al., 2020). Nonetheless, the sentence the Reviewer points out has been corrected to follow this convention "Even though we acknowledge that our conclusions might be both model and region dependent, taken together, our analysis suggests that to remove model biases i) a refinement of the atmosphere resolution up to ~50-km alone might not always be sufficient, and ii) reaching eddy-rich ocean resolutions (1/12° or fine) might be needed. The increase in ocean resolution from eddy-parametrized (~100 km) to eddy-rich (~10 km) allows models to represent the first baroclinic Rossby radius and might therefore improve the representation of small-scale dynamical processes and then biases."

L455. I guess that is LL instead of LR.

The HadGEM3-GC31-LL configuration is referred to as LR (low-resolution) throughout the manuscript to simplify model naming. This is indicated in Section 2.1 and Table 1.

Figures and tables:
Table 2. Adding the global tos bias in the coupled runs could help determine the possible added value of increased ocean resolution.
As described before, biases in SST and SAT are very similar.

Figure 6. The stippling mask is missing in this figure.
The stippling mask has been modified so it does not look like horizontal lines.