Review of "Seasonal extrema of sea surface temperature in CMIP6 models" by Wang et al.

This is an account of seasonal sea surface temperature (SST) biases in CMIP6 models with respect to the observed World Ocean Atlas climatology for the period 1981-2010. The authors document and discuss discrepancies between the multi-model mean as well as individual model seasonal cycles, and observations. They find model diversity in cold and warm season SST biases, as well as annual mean biases and seasonal cycle amplitude biases. Model bias is shown to increase with decreasing vertical ocean model resolution, highlighting promising future model development activities. These findings have important implications for the model specific assessment of regional climate projections. As such, the present study will a valuable reference in the climate model evaluation literature.

I find the manuscript to be interesting, relevant, and well-written. The presented results are novel and important within the ocean modeling community. The paper is therefore a good fit for Ocean Science. After some mostly minor concerns of mine have been addressed, I think this manuscript will make a welcome addition to the scientific literature.

We thank the reviewer for the helpful and constructive comments that helped us in improving our manuscript.

Main concern:

A somewhat substantial concern of mine is the use of the first realisation of each model for drawing fundamental conclusions about the model's biases. Individual ensemble members are subject to substantial chaotic climate variability, which might as well influence the seasonal cycle of SST. To address this issue, I suggest the authors compare r1 to the ensemble mean of individual models (or r2) to get an idea about how strongly SST seasonal cycles depend on the realisation of the models. This would make their analysis more rigorous and give their reported results more weight.

Thank you for your suggestion. We expect differences within each model ensemble to be small as we are examining thirty-year means. We have compared r1 with r2 for the SST seasonal cycles. The differences of Tmax, Tmin, month of Tmax and month of Tmin between r1i1p1f1 and r2i1p1f1 in each model are shown below (Figs 1-4). r1i1p1f1 and r2i1p1f1 are compared when available; r1i1p1f3 and r2i1p1f3 are compared for HadGEM3-GC31-MM and HadGEM3-GC31-LL; r1i1p1f2 and r2i1p1f2 are compared for UKESM1-0-LL. There are no results for SAM0-UNICON and GFDL-CM4 as they have only one ensemble member.

The differences of Tmax and Tmin between two ensemble members are within 0.5^{o}C over most of the global ocean, which are very small compared with the model biases of Tmax and Tmin (Figs. 1-2) for the ensemble member considered in our paper. As for the months of Tmax and Tmin, there are no differences between the two ensemble members over most of the global ocean, while one month differences exist in some specific regions (Figs. 3-4). The small differences between ensemble members demonstrate that SST seasonal cycles do not depend on the realisation of the models, and thus the model biases we report are robust.

The following sentence has been added to line 74 in the manuscript.

"To test the dependence of the biases found on the realisation of models, we compared the first and second ensemble members (except for SAM0-UNICON and GFDL-CM4 as they have only one ensemble member). The differences between ensemble members are very small compared with the model biases (supplementary Figs. S1-4), and thus the model biases we report are robust."

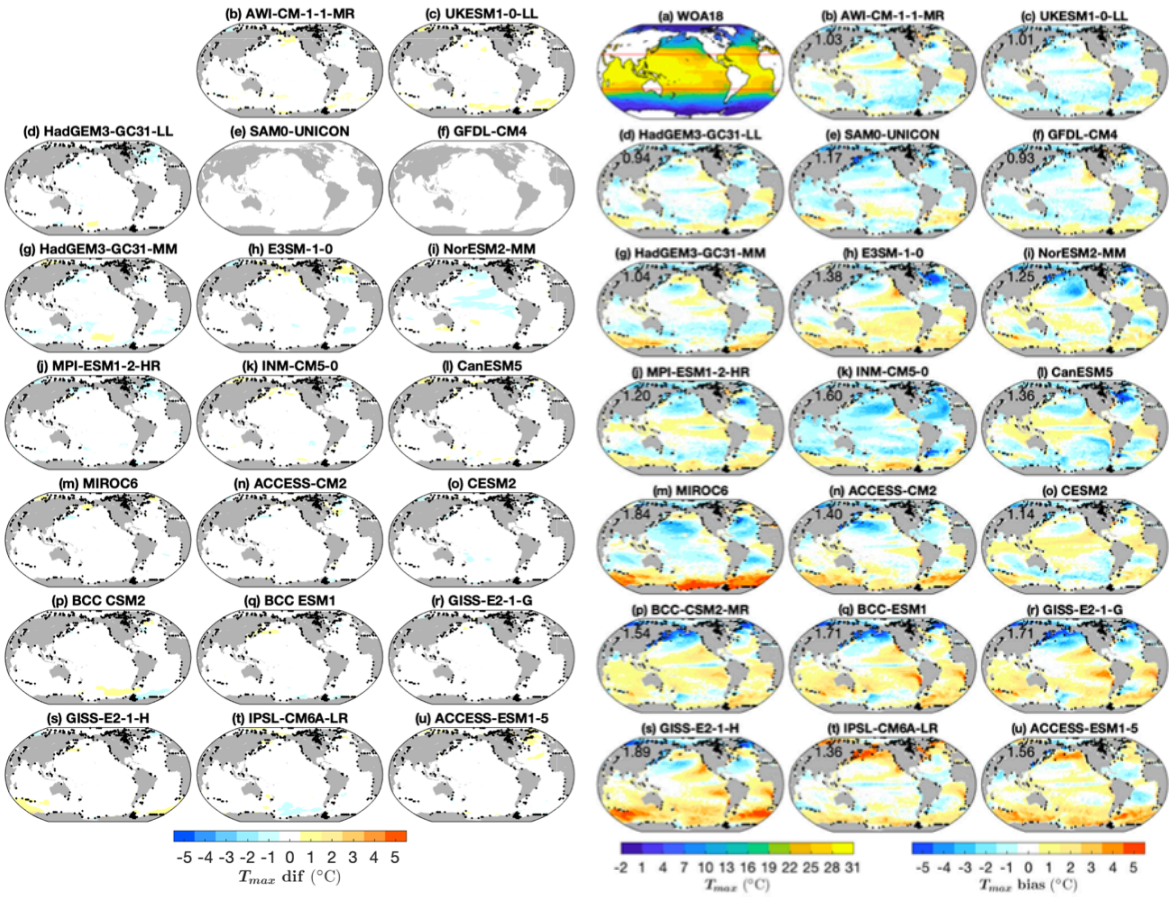

Figure 1. (Left) Differences between two ensemble members for Tmax. (Right) Tmax model biases for the ensemble member considered in our paper (Figure 2 in the paper, reproduced here for ease of comparison). Black dots mark grid points excluded from our analysis.

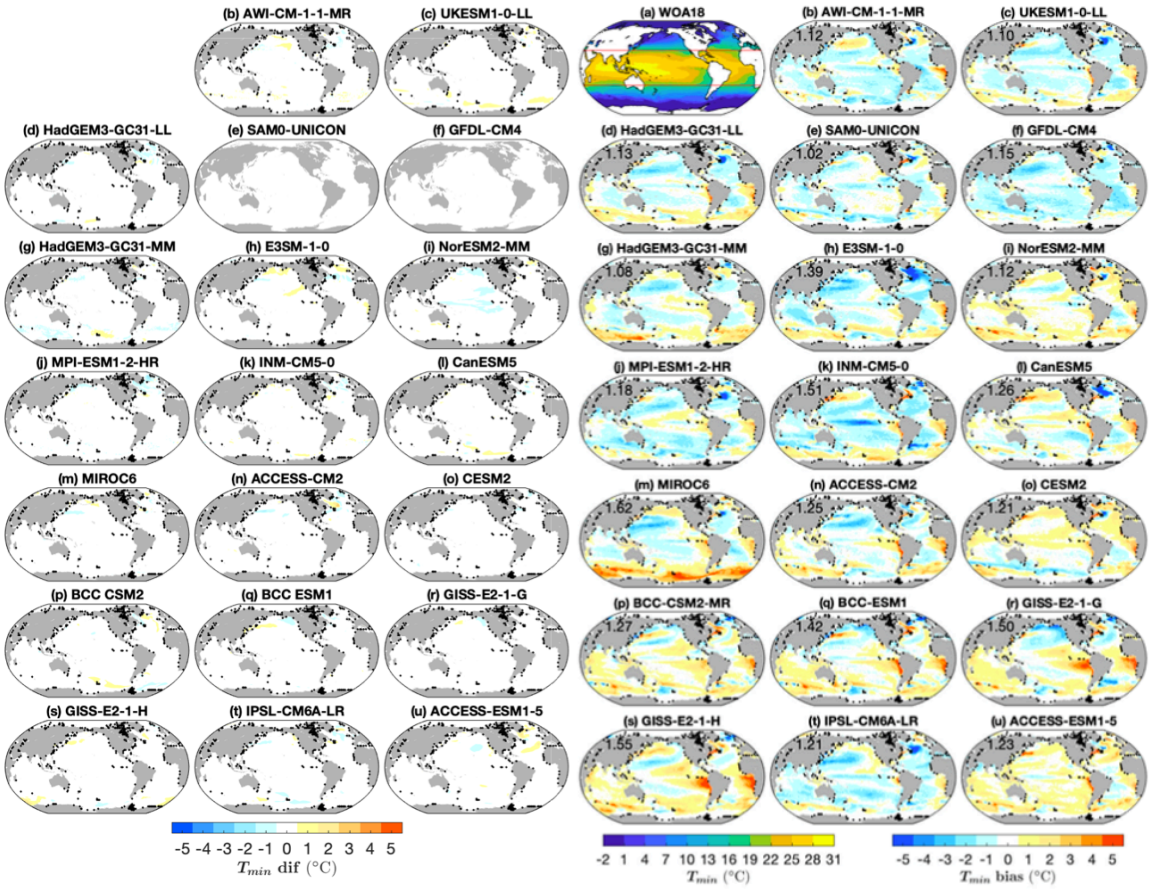

Figure 2. As in Fig.1, but for Tmin.

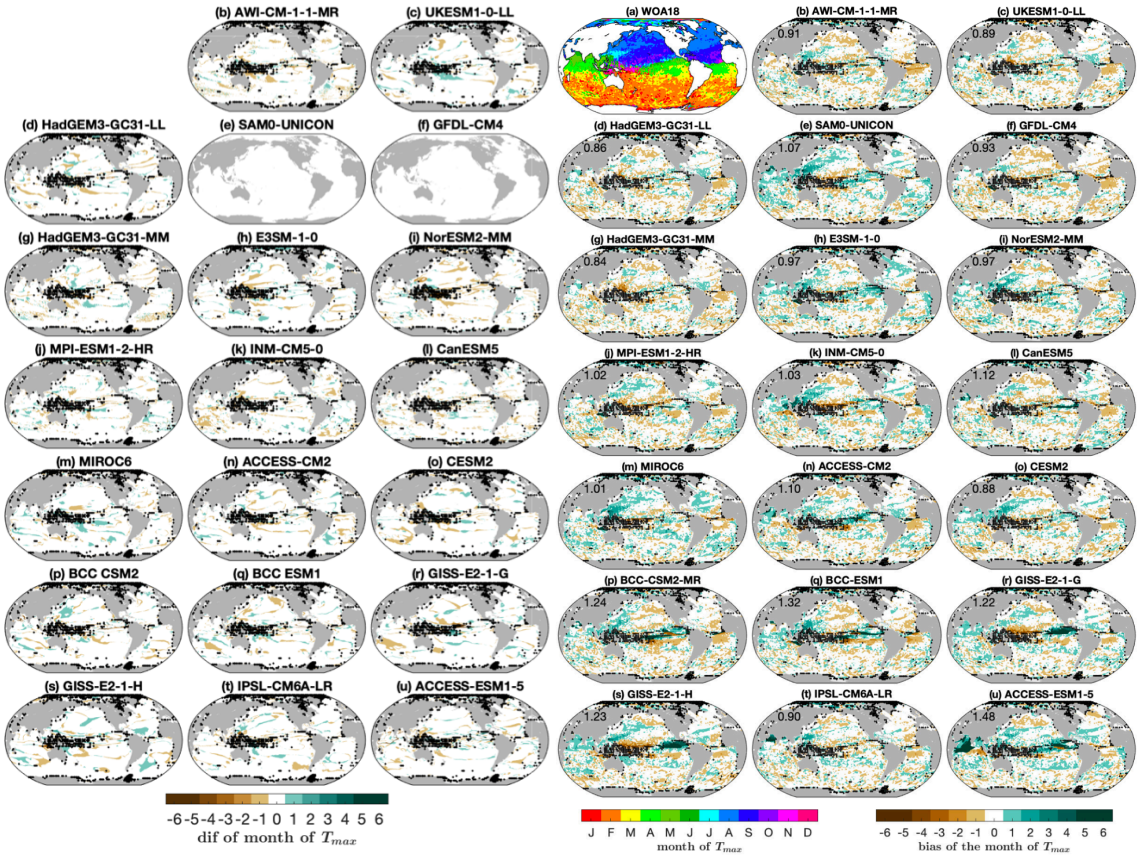

Figure 3. As in Fig. 1, but for the month of Tmax.

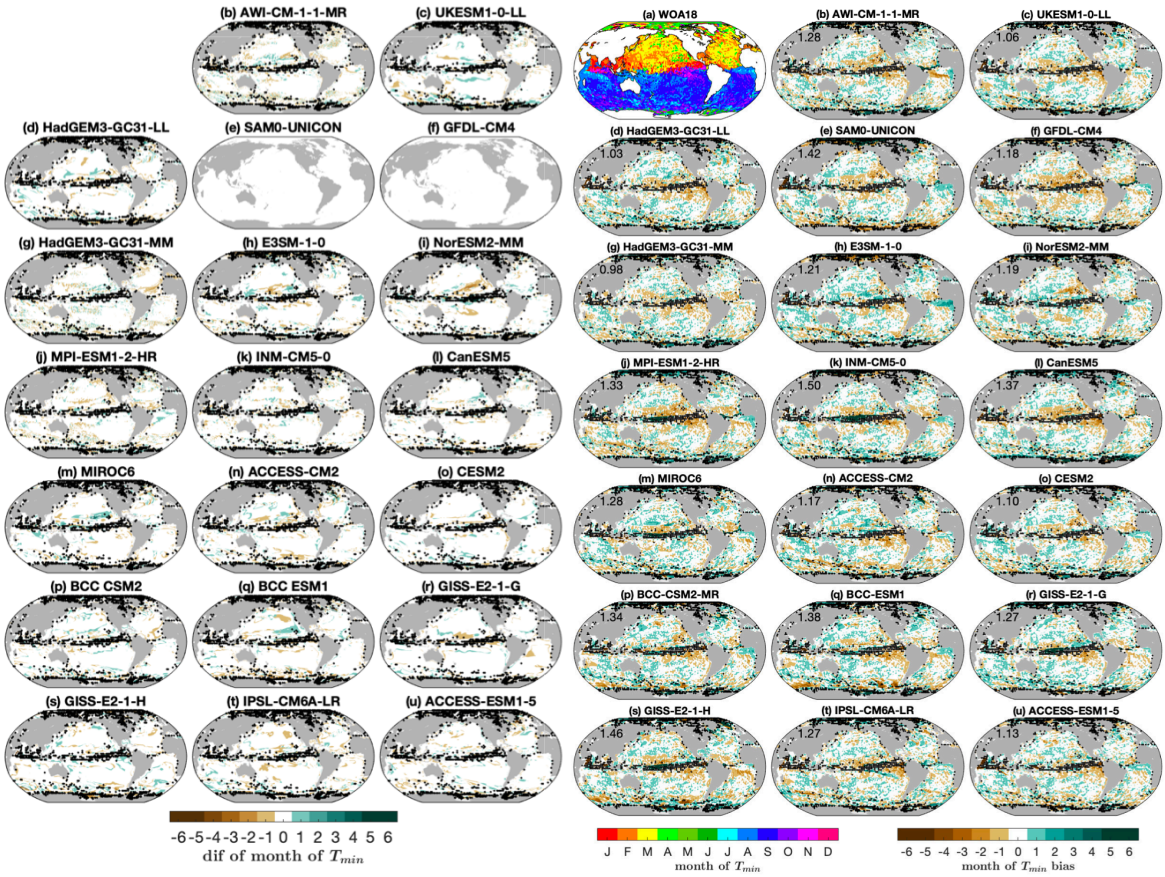

Figure 4. As in Fig. 1, but for the month of Tmin.

Specific comments:

l. 3 Should these "commonly used climatologies" be defined here?

Thank you for the comment. We have rewritten "commonly used climatologies" as "commonly used climatologies (WOA18, WAGHC and HadISST)".

l. 9 I suspect that the authors report "no significant relationship" of SST extrema bias "with horizontal ocean model resolution", but suggest the authors be specific about that.

Thank you for the suggestion. We have rewritten the sentence as "No significant relationship of SST seasonal extrema with horizontal ocean model resolution is found."

l. 21 Suggestions: move the abbreviation "CMIP5" into the brackets and the long form out of the brackets.

Thank you for the suggestion. We have updated the manuscript accordingly.

ll. 31-32 Drawing conclusions about projected seasonal cycles from the analysis presented in this paper assumes i) stationarity of the biases, and ii) that the biases are consistent between r1 (analysed here) and the ensemble mean (used for projections) of the individual models. While i) is difficult to test and could (should?) be discussed as a caveat, I think that ii) requires some attention in the revision of the manuscript (see my main concern above).

Thank you for the helpful suggestions.

For i), we have added the following sentences on lines 269.

"If there is a substantial change in the climate, it should be considered that the pattern of biases in Tmax and Tmin may change."

For ii), we have compared r1 with r2 for the SST seasonal cycles (details are in the response to the main concern above), and the following sentence has been added to line 74 in the manuscript.

"To test the dependence of the biases found on the realisation of models, we compared the first and second ensemble members (except for SAM0-UNICON and GFDL-CM4 as they have only one ensemble member). The differences between ensemble members are very small compared with the model biases (supplementary Figs. S10-14), and thus the model biases we report are robust."

ll. 32-34 I do not see the value of a table of contents here and thus suggest deleting it.

Thank you for the comment. We have removed the sentences.

ll. 41-42 The extra information about INM-CM5-0 (Volodin, pers. comm.) could be placed more appropriately in Table 1. Please consider moving it accordingly.

Thank you for the comment. We have moved the sentences to the footnote of Table 1.

ll. 51-52 I wonder if the averaging of T values from different months in case of a shifted seasonal cycle would be problematic. Could the authors please comment?

Thank you for your suggestion. To calculate multi-model mean, we averaged SST seasonal extrema (Tmax and Tmin) in 20 models and indeed in some regions seasonal extrema occur in different months in different models. However, we don't see it as a serious problem as the timing of seasonal extrema is very similar in most of the global ocean in most models, which is suggested by the small bias (within one month in most of the global ocean in most models) in the timing of seasonal extrema (see Figs 5-6 in the manuscript). Although in some cases we calculated multi-model mean by averaging SST in different months, we cannot pick some specific months for averaging to avoid this problem as SST seasonal extrema occur in different months in different models. In fact, picking specific months would also assume stationarity and it is probably worse. If Tmax and Tmin were to get earlier or later in future climate projections, people could get erroneous results if they look at particular months.

l. 53 To avoid potential confusion, I suggest being specific here that the RMSE of the four different T quantities is calculated against observations for global SST.

Thank you for your suggestion. We have rewritten the sentence "we calculated the area-weighted root mean square error of the model against WOA18 (henceforth RMSE) for global SST." as "we calculated the area-weighted root mean square error of Tmax, Tmin, Tmean and Tcycle of the model against WOA18 (henceforth RMSE) for global SST."

ll. 60-61 Where are the excluded grid points typically located? Maybe add a sentence about that here.

Thank you for the suggestion. The following sentence has been added on line 59.

"The excluded grid points are mostly located in coastal areas, a few regions in the Arctic, and around the ACC, Agulhas Current and Benguela Current."

l. 77 I cannot make out to what the "it" at the start of this sentence refers. Please be specific.

Thank you for the comment. We have rewritten the sentence "It demonstrates that seasonal cycles in CMIP6 models are out of phase with observations." as "The bias in the timing of Tmax and Tmin demonstrates that the seasonal cycles in CMIP6 models are out of phase with observations.".

l. 80 Explicitly stating once in this paragraph that high latitudes show a larger bias than low latitudes would make the entire story easier to follow. Please consider this addition.

Thank you for the suggestion. We have added the following sentence on line 94.

"High latitudes show larger biases than low latitudes."

l. 97 I think it should be "...salinity biases in the Arctic."

Thank you for pointing this out. We have corrected it in the revised version.

Figure 7 The different x axes of a-f compared to h-i are somewhat confusing to me. The message of this figure might be easier to grasp if the figure was split in two separate figures.

Thank you for the suggestion. We have split Figure 7 in two separate figures as you suggested.

ll. 104-105 The two occurrences of "is larger" in this sentence lack clarity without a reference (larger than what?). I recommend using absolute language such as "shows a maximum".

Thank you for the comment. We have rewritten the sentence as "The latent heat loss shows a maximum in summer (Yu, 2007), while the ocean heat advection shows a maximum in winter when meridional SST gradients are greatest."

ll. 131-132 Which part of the cloud is underestimated? From the sentence itself, it is unclear if it is cloud cover, formation, ...

Thank you for the comment. We have rewritten the sentence as "In most models there is a warm Tmean bias in the Southern Ocean, commonly attributed to excessive short wave radiation linked to cloud process representation deficiencies."

l. 150 Separate "Namibia" and "as"

Thank you for pointing this out. We have corrected it in the revised version.

ll. 229-230 What is the significance level at which this correlation is significant? Which significance test was used? Without this information, the statement of significance is not worth much. Similarly, I suggest adding p-values to figures 8 and 9.

Thank you for the suggestion. We have added p-values to figures 8 and 9. When $p<0.05$, we can say that the relationship between global RMSE and ocean vertical resolution is significant. The sentence "The correlations are significant for Tmax, Tmin, and Tmean, with the largest correlation of -0.648 for Tmax." has been rewritten as "The relationship between SST biases and total number of vertical levels is significant for Tmax, Tmin and Tmean (p-values<0.05), with the largest correlation of -0.648 for Tmax."