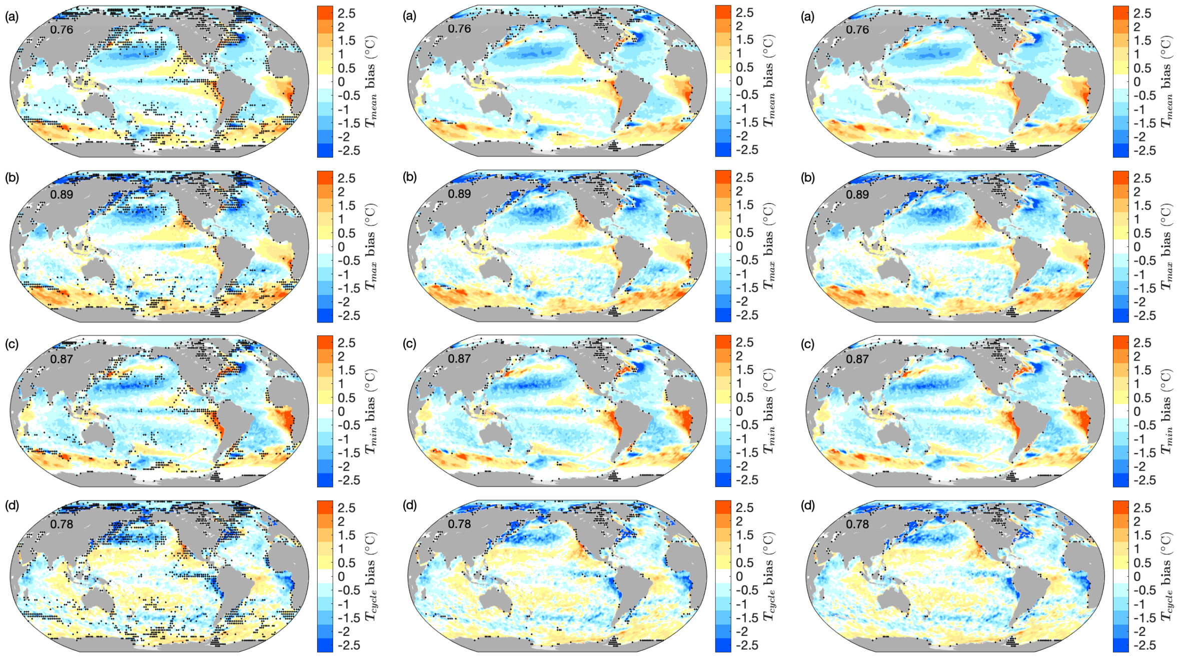
I find that the paper is a valid contribution and I find that the methodology applied is right. I have few minor comments that the authors will hopefully polish for next version of the paper.

We thank the reviewer for the helpful comments that help to improve the manuscript.

1. Page 2, line 58. The authors mention that they assign a 2°C threshold to the biases in Tmax and Tmin to exclude points from the analysis. I wonder whether they have previously performed some sensitivity analysis to analyze the impact of this threshold. I guess that such an error might be more or less important at points with higher or lower variances of SST. I think this should be checked.

Thank you for the comment. We had performed some sensitivity analysis before. The following figures are versions of Figure 1 showing biases of multi-model mean in (a) Tmean (b) Tmax (c) Tmin (d) Tcycle. Black dots mark grid points excluded from our analysis. Figures from left to right used 1°C, 2°C and 3°C thresholds to the maximum differences between the three climatologies, respectively.



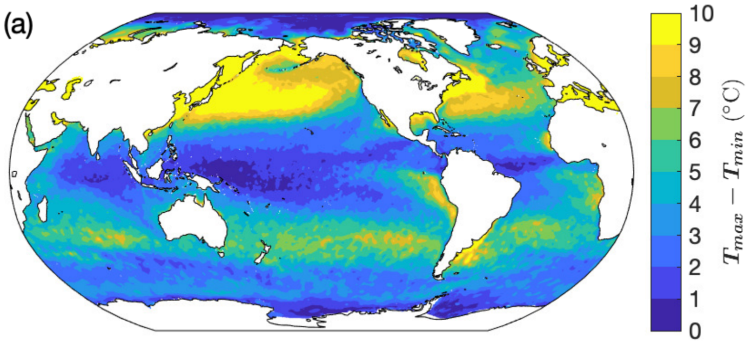
We see that using 1°C threshold will exclude too large areas and many regions with large SST biases will be missing. The 3°C threshold mask covers a few coastal regions, however some points with high uncertainty (up to about 3°C) in the open ocean and coastal regions are not excluded from calculating the model bias. The mask using 2°C threshold mostly covers coastal areas, and it also includes a few regions in the Arctic, around ACC, Agulhas Current and Benguela Current. (Please notice that when plotting the figure, we interpolate the mask from 0.25°*0.25° grid into 2.5°*2.5° grid to make the black dots less dense.)

Threshold=1°C, percentage of the mask: 9% for Tmax, 6% for Tmin, 12% for Tcycle and Tmean

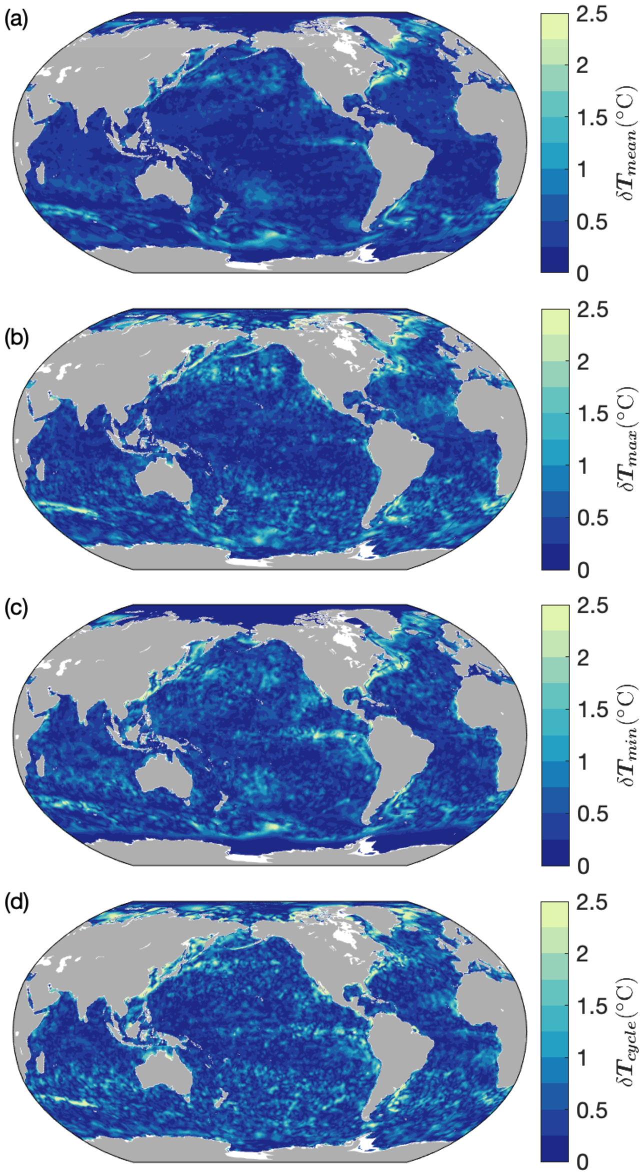
Threshold=2°C, percentage of the mask: 4% for Tmax, 3% for Tmin, 4% for Tcycle and Tmean

Threshold=3°C, percentage of the mask: 2% for Tmax, 1% for Tmin, 2% for Tcycle and Tmean

The seasonal cycle of monthly climatological SST from WOA18 is shown below, suggesting that most of the areas (except polar regions) covered by the masks with 2^{o}C threshold have high variance of SST.



The maximum difference between any two of the three climatologies (WOA18, HadISST and WAGHC) for (a) Tmean (b) Tmax (c) Tmin and (d) Tcycle is shown in the figure below.



2. Page 4, line 70. "a small bias in a multi-model mean...individual models" seems like a little tautology to me.

Thank you for the comment. We will remove this sentence in the revised version.

3. I find that the RMSE values in figures 2, 3, 5 and 6 might be joined in a single table, so that a better analysis of systematic behaviors (if any) could be performed.

Thank you for the insightful suggestion. We plan to add a figure (Figure 7) comparing RMSE values in all models and the following text to the revised paper.

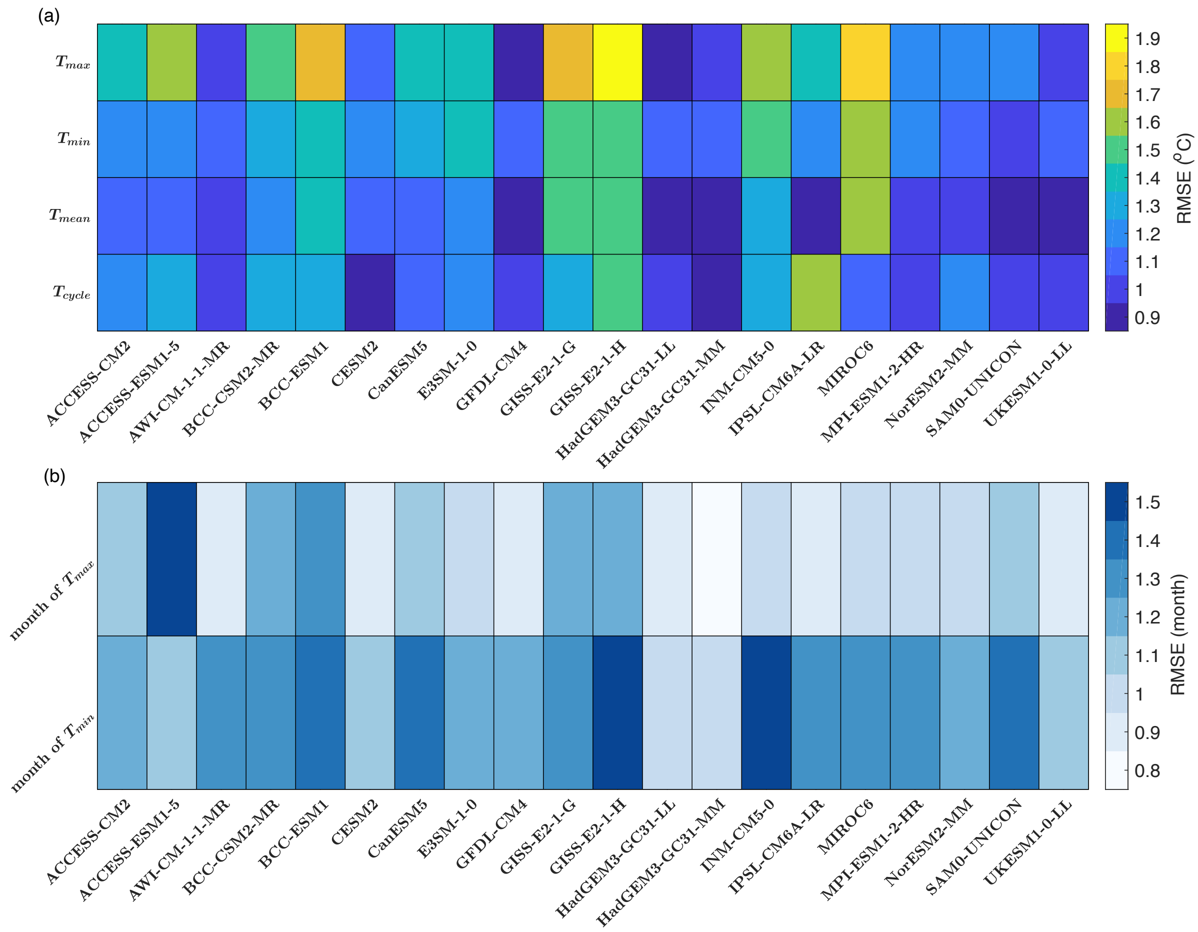


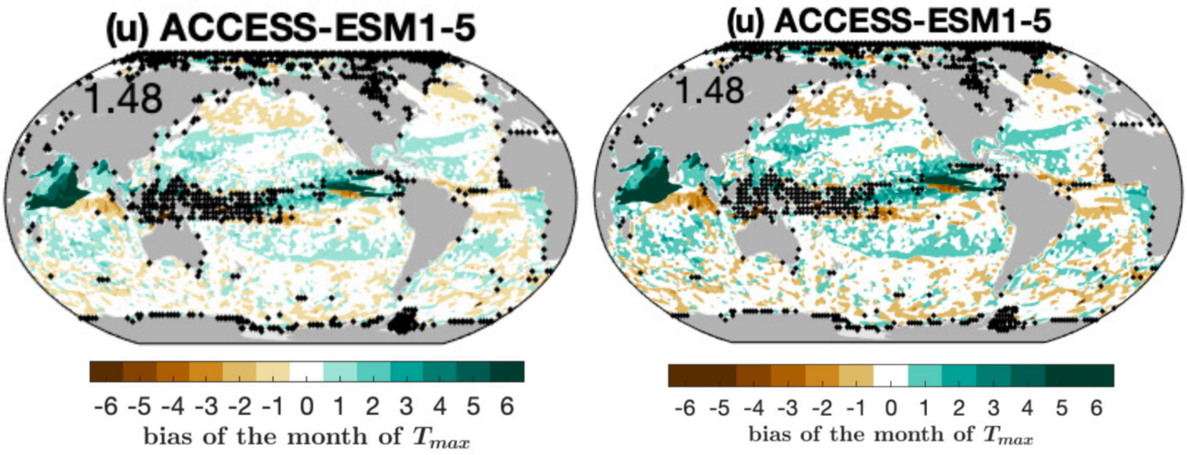
Figure 7. The global area-weighted RMSE of the biases in (a) Tmax, Tmin, Tmean and Tcycle (b) timing of Tmax and Tmin.

"In most of the models the global RMSE is larger in Tmax than in Tmin (Fig. 7a). As the bias in Tmax and Tmin is largely consistent with Tmean bias, Tcycle RMSE is small compared to Tmax and Tmin RMSEs in most models. Different biases in Tmax, Tmin, Tcycle and Tmean suggest that models have different performance in simulating SST seasonal variation and annual mean. The "best" and "worst" models depend on whether you choose SST seasonal extrema or annual mean as your metric. For example, GFDL-CM4 and HadGEM-GC31-MM have the smallest RMSE in Tmax and thus they are best for simulating tropical cyclones and heatwaves; SAM0-UNICON has the smallest RMSE in Tmin and thus it is best for simulating the properties of intermediate and deep waters."

“Models have different performance in simulating the timing of Tmax and the timing of Tmin. All the models except ACCESS-ESM1-5 have smaller global RMSE in the timing of Tmax than in the timing of Tmin (Fig. 7b). HadGEM3-GC31-MM has the smallest global RMSE in the timing of Tmax, whereas HadGEM3-GC31-LL and HadGEM3-GC31-MM have the smallest global RMSE in the timing of Tmin.”

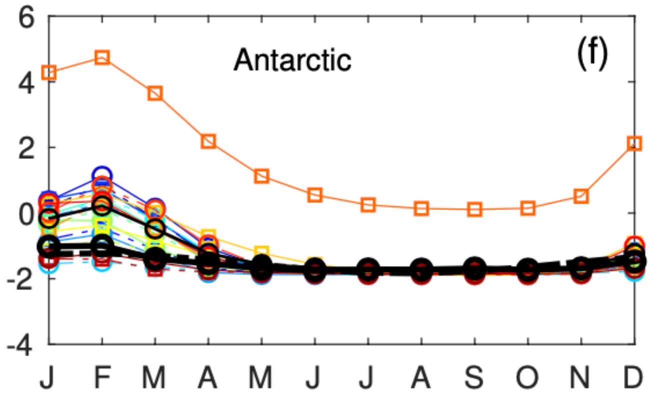
4. I find the palettes in Figures 5 and 6 difficult to grasp. Probably the fact that some isolated points are high in absolute value require a large scale. However, since most points are of a low absolute value, it is difficult to grasp the differences from one map to another. I suggest that perhaps the authors should use a non-equally spaced palette to increase resolution in lower values.

Thank you for the helpful suggestion. We have updated the color bars in Figures 5 and 6 to emphasize the low values on the maps. The following figure shows one of the panels in Figure 5 with the old color bar (left) and new color bar (right).



5. In Figure 7, I suggest some labels referring to regions are added to the individual panels to improve readability “WestEqPac” in panel “a)”, “NWIndOc” in “b)” and so on.

Thank you for the suggestion. We have added labels on the panels in Figure 7. An example panel is shown below.



6. Page 12, line 120. I am not sure that the increase in storminess can be assigned to heat fluxes into the storms but separated from increased atmospheric baroclinicity (Kushnir, 2002), not explicitly mentioned by the authors. I think this point must be revised.

Thank you for the comment. Our original statement seems wrong for the extratropics. Brayshaw et al. (2008) and Nakamura et al. (2004) show that a change in SST gradient in certain key regions impacts the storm track. That is, an SST bias may impact the SST gradient and thus impact storm tracks. Given the uncertainty that the change of SST gradient could be in any direction for any given bias depending upon the background state, here we will remove "Models with a warm bias in Tmin are likely to generate overly intense winter storms, as warm SSTs will increase the storm energy source. Greeves et al. (2007) demonstrated that there was a clear link in the Hadley Centre models between winter SST warm bias to the east of Japan and increased storm intensity in the region."

Brayshaw, David James, Brian Hoskins, and Michael Blackburn. "The storm-track response to idealized SST perturbations in an aquaplanet GCM." *Journal of the Atmospheric Sciences* 65.9 (2008): 2842-2860.

Nakamura, Hisashi, et al. "Observed associations among storm tracks, jet streams and midlatitude oceanic fronts." *Earth's Climate: The Ocean–Atmosphere Interaction, Geophys. Monogr* 147 (2004): 329-345.

7. Page 12, lines 126-129. I guess that Myers et al. (2021) is a good reference to support the authors' hypothesis here.

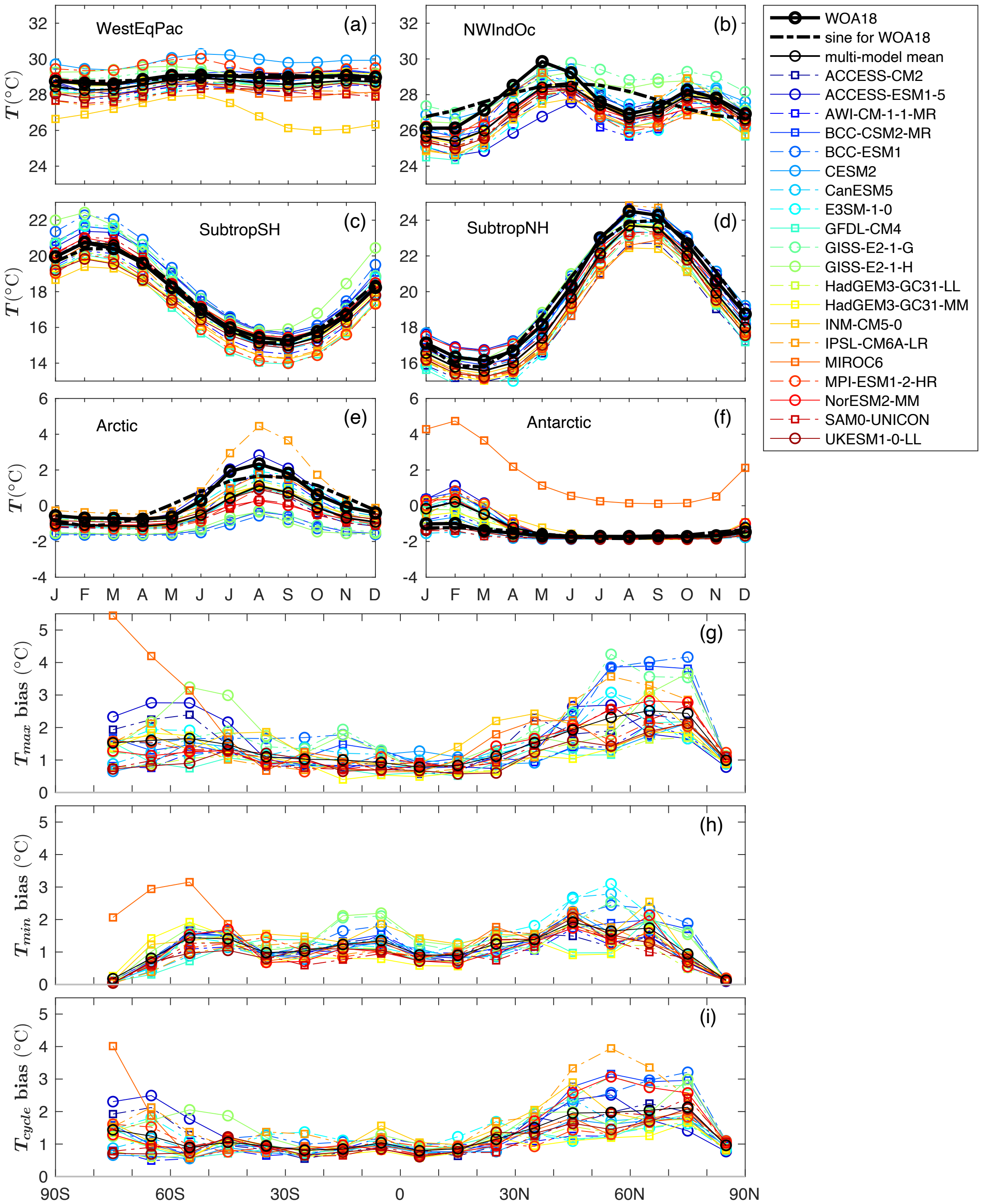
Thank you for your recommendation. We have cited Myers et al. (2021) as "The large cold biases at northern hemisphere high latitudes in BCC-CSM2-MR, BCC-ESM1, GISS-E2-1-G and GISS-E2-1-H, are typically 2-5◦C smaller in Tmin than in Tmax (Figs. 2, 3, 7g-h). These cold biases are likely to be linked to cloud biases due to the cooling radiative effect of low cloud (Myers et al., 2021). The negative cloud radiative forcing is excessive in BCC-CSM2-MR (Wu et al., 2019) and BCC-ESM1 (cloud simulation likely to be similar to BCC-CSM2-MR), while overestimated low-cloud cover in GISS-E2-1-G and GISS-E2-1-H (Kelley et al., 2020) blocks more of the incoming solar radiation."

8. Page 14, lines 179-181. I suggest the authors to fit a simple sinusoidal signal here (period T=12 months) and the fraction of variance explained would allow the authors to show which areas respond to one or the other case.

Thank you for the suggestion. We have fitted sinusoids to the time series in Figure 7 (see below) and the following discussion has been added in the text.

"In regions with fairly sinusoidal SST annual cycles such as the subtropics (sinusoidal signal explains 87% of the observed variances in subtropical Northern Hemisphere and 89% of the observed variances in subtropical Southern Hemisphere), models have realistic SST seasonal cycles with well simulated amplitude and phase of the annual cycle (Fig. 7c-d)."

"In regions with non-sinusoidal SST seasonal cycles such as the western equatorial Pacific, northwestern Indian Ocean, the Arctic and the Antarctic (sinusoidal signal explains 33%, 23%, 58% and 46% of the observed variances), models tend to have biases in amplitudes or phases of their SST seasonal cycles (Figs. 4,5,6,7a-b,e-f)."



Kushnir, Y., Robinson, W. A., Bladé, I., Hall, N. M. J., Peng, S., & Sutton, R. (2002). Atmospheric GCM Response to Extratropical SST Anomalies: Synthesis and Evaluation, Journal of Climate, 15(16), 2233-2256. https://journals.ametsoc.org/view/journals/clim/1 5/16/1520-0442_2002_015_2233_agrtes_2.0.co_2.xml

Myers, T.A., Scott, R.C., Zelinka, M.D. et al. Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity. Nat. Clim. Chang. 11, 501–507 (2021). https://doi.org/10.1038/s41558-021-01039-0