Jing et al. 2020 Second Review

This manuscript shows that in several CMIP5 models the wintertime sea-surface temperature response to the NAO in the North Atlantic subpolar gyre is not consistent with observations on interannual timescales. They attribute this model bias to and incorrect (or inconsistent) turbulent heat flux response to the NAO induced SST tripole.

Overall, I'm satisfied with the scientific content of the manuscript and recommend acceptance pending minor revisions. Further, I would highly recommend the authors seek out additional (perhaps professional) copy editing. I think this is a useful and timely paper, so I encourage this copy editing to make sure the authors' work is understood by the community at-large.

Minor Revisions

L59 – 61: I recommend cutting the phrase "In recent years, more and more people have realized" (and adjusting the rest of the sentence).

L63 – 66: The wording is slightly ambiguous here. As you point out on L187-189, Wang et al. (2014) found that *mean* SSTs were too cold. (Also, there are two references for Wang et al. 2014. Can you please differentiate in some way? In the second Wang et al. (2014) reference, I believe that there is a missing "s" in the word "biases" in the article title.)

L180-183. I think that both of these sentences are correct individually – but I am not following the logic between them. I would recommend cutting or moving the first sentence.

L189 – 192 and Figure 2: This reminds me of Siqueira and Kirtman 2016 who show a change in ocean resolution can change the location of atmospheric circulation anomalies (their Figure 3).

 Siqueira, L., and B. P. Kirtman (2016), Atlantic near-term climate variability and the role of a resolved Gulf Stream, Geophys. Res. Lett., 43, 3964–3972,doi:10.1002/2016GL068694.

L239: "are slightly [further] south than observations" or "are slightly south [of] observations" (and again on L240 – 241).

L255: "abnormal" -> "anomalous"

L358 – 363: I recommend breaking this up into multiple sentences.

L438 – 443: I also recommend breaking this up into multiple sentences.

L443: It is true that the unfiltered timeseries should have more degrees of freedom, but I'm not sure why we would expect that to influence the magnitude of the regression coefficients in one direction (negative) over another (positive). Perhaps this is just removing the autocorrelation induced by the comparing two filtered timeseries?

L466: see previous comment.

L475 and L476: "indexes" -> "indices"

L481: I believe the CESM Large Ensemble is initialized with minute perturbations in atmospheric temperature. Given these are free running models, I suspect that the differences the authors find are a result of the different time histories of internal variability that result from different initial conditions – and not from ocean initialization (as in a forecast model). I encourage the authors to make this distinction, if they agree.

L505 – 506: I'm not sure how you draw this conclusion.

Optional suggestion:

Section 5.4:

First, I found this additional section very helpful. I know it was a lot of work, but I hope it will make your paper more impactful. One reason I suggested this analysis was related to Scaife and Smith (2018)'s "signal-to-noise paradox", wherein models produce NAOs that are more like observations than themselves. In other words, the signal-to-noise ratio in the NAO in climate models is unrealistically low. I suspect that the authors are finding something similar (through a very welcome mechanistic approach). It may be useful for the impact of this work to tie these ideas together here. I'll note that since the first version of this manuscript, Smith et al. (2020) have published a high-impact paper that claims to overcome the signal-to-noise paradox through a very large ensemble (169 members).