Oct. 10, 2020.

Manuscript Number: os-2020-16.

We would like to thank the editor and reviewer for reviewing our manuscript and giving us good suggestions. According to these suggestions, we have corrected some mistakes in expression and added some discussion.

Responses to the reviewer' comments point by point:

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

Reply: We cut the phrase and modified this sentence into "The identification of the CMIP5 Earth System Models bias is important for the improvement of these models and development of climate projection." **(P4, L61-62)**

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.)

Reply: Both of the two sentences on **L63-66** and **L187-189** in the revision1 are to say that CMIP5 models underestimate the annual mean SST in the North Atlantic. In this revision, we change this sentence to "Meanwhile, Wang et al. (2014b) evaluated the global annual mean SST simulated by the CMIP5 models and found that the SST in the Northern Hemisphere, especially in the NA, is underestimated." **(P4, L65-68)** In addition, we use "a, b" to distinguish the two references from Wang et al., and we have added "s" in the word "biase" in the article title by Wang et al. (2014b). Thank the reviewer for the reminder.

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.

Reply: What we would like to express is that the external forcing will not cause differences between the NAO patterns simulated in these models, because these models are forced by the same external-forcing data. We've changed the sentence to "The differences between the NAO patterns simulated by these models with the same external-forcing data are probably induced by their different model structures and values of parameters." **(P11, L182-184)**

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.

Reply: Taking the reviewer's advice, we have added some discussion: "With a climate system model, Siqueira and Kirtman (2016) found that the change of ocean component model resolution can change the simulated SST variabilities, locations of atmospheric circulation anomalies, and air-sea interactions in the North Atlantic. The change is induced by the impact of the resolution on the ocean dynamics, such as ocean fronts and eddies in the Gulf Stream which can be well resolved in the high resolution model with the horizontal resolution of 0.1°×0.1°. Nevertheless, the highest horizontal resolution of these ocean component models used in this study is 0.4°×0.4 ° (MPI-ESM-MR), and the comparison of MPI-ESM-LR and MPI-ESM-MR, both of which are from the same institution and with different ocean component model resolutions, shows that the SST variability in the Gulf Stream is not significantly different. This indicates that the resolution of these models is still not enough to investigate the SST variability in the Gulf Stream and may induce the deviation between the simulated SST variability and the observed one." **(P11-12, L193-203)**

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

Reply: We've corrected the sentence to “In HadGEM2-ES, the low-pressure action centers of the NAO are slightly further south than observations, and the negative response center of the SST to the NAO is also further south than observations” **(P15, L250-252)**

L255: “abnormal” -> “anomalous”

Reply: Done.

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

Reply: We've changed the sentence to “The distributions of the RCs are similar to those of the SST anomalies against NAO-driven SHF anomalies in a large area of NA. The main difference between the response of the SST to the SHF and to the LHF is that the observed and modeled positive RCs of the SST anomalies against NAO-driven SHF anomalies in the eastern NA around 20°N do not occur in the regression of the SST anomalies against NAO-driven LHF anomalies. It indicates that the influence of the LHF on the SST probably controls the RCs of the SST anomalies against the NAO in this region.” **(P22, L370-374)**

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

Reply: We've changed the sentence to “We also did regression analysis of unfiltered winter average SST anomalies and NAO indices (Fig. S7). It is found that except for the models of IPSL-CM5A-MR and MPI-ESM-L / MR, there is no obvious difference in the distribution of standardized RCs of the SST and NAO between the filtered and unfiltered results, and the main difference is that the RCs from the unfiltered data are slightly smaller than those from the filtered data in the subtropical NA (Fig. 4) of both the observation-based results and most of the modeled results.” **(P26, L450-454)**

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.

Reply: Thank for the reviewer's reminder. The difference between unfiltered and filtered results is more obvious in the subtropical NA, where the regression coefficients between the NAO and SST are positive. The change of the degree of freedom can’t explain this phenomenon, and there are also many problems that can’t be explained by degrees of freedom, such as the inconsistency between the observed data and the model results due to the influence of filtering in the tropical and subpolar regions, so we have deleted these sentences about the degree of freedom. Taking the reviewer’s suggestion, we analyzed the effect of autocorrelation: we removed the autocorrelation from the unfiltered NAO and SST with the Cochrane-Orcutt method before the regression analysis, and found that the magnitude of regression coefficients (Fig. R1) are very consistent with that from the original data (Fig. S7 in this revision). Therefore, we think the autocorrelation is not an important factor that causes the difference between the unfiltered and filtered results. At present, we have not thought of any explanation for this phenomenon. In our follow-up research, we will pay more attention to explore this problem.

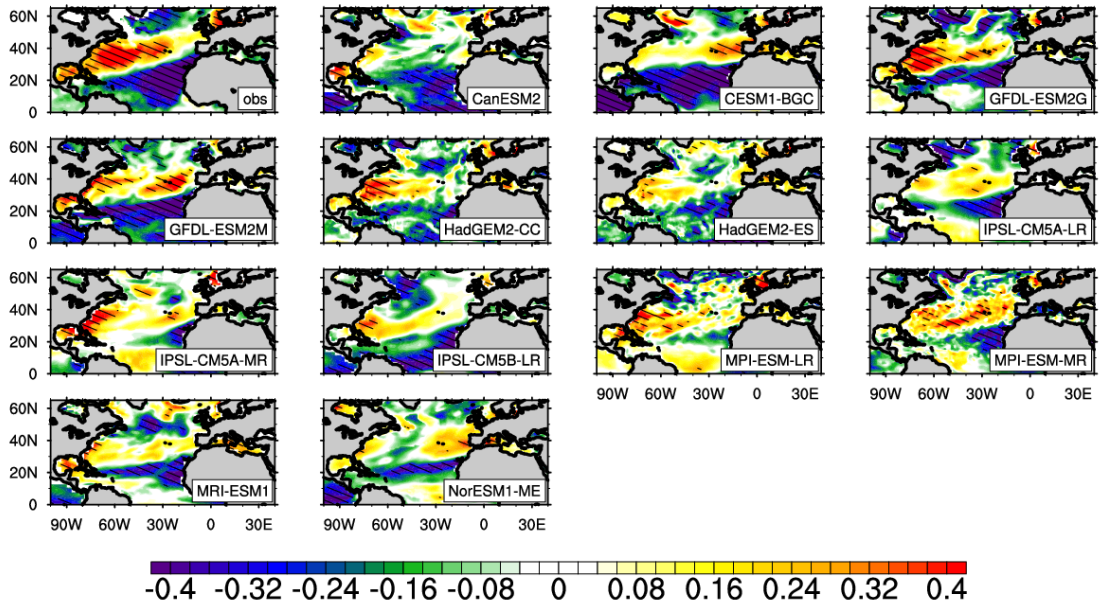

Figure R1 Standardized regression coefficients (removing the autocorrelation of the NAO and SST) of the winter-averaged SST anomalies against the NAO indexes (without data filtering). Shaded areas indicate that RCs are statistically significant at the 95% confidence level of the Student's t-test. The obs is the RCs of observed SST to the NAO indexes provided by NCAR. The time periods for the observation and models range from 1965 to 2015 and 1955 to 2005, respectively. The simulated results are based on historical experiment of CMIP5 (r1i1p1).

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

Reply: Done.

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.

Reply: Yes, we agree with the reviewer. We have modified the paragraph into: "Kay et al. (2015) did ensemble experiments by adding different minute perturbations to the atmosphere as initial conditions to study the internal variability. There are also some ensemble historical experiments in CMIP5 which are initialized with different initial conditions in 1850. The initial conditions of these ensemble members are from the

different integrated time of the piControl experiments, so these initial conditions represent the different time histories of internal variability. The relationship of the NAO and SST simulated by the models with above mentioned different initial fields (r1i1p1 and r3i1p1) are compared (Fig. S11)." **(P29, L493-498)**

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

Reply: After the initial field of historical experiment is provided, the piControl experiment will continue to integrate for 500 years, so the result of the piControl experiment may be more stable than that of historical experiment in terms of the internal variability. The result of the piControl experiments in MPI-ESM-MR is very similar with the historical experiments (r3i1p1), but is different from the historical experiments (r1i1p1). We infer that the initial fields of the historical experiments (r1i1p1) of this model may come from an early integrated time of the piControl experiment. Because we can't find the material to support our conclusion, we delete this sentence in this revision.

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 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).

Reply: Thanks for the reviewer's suggestion. We have read these two references (Scaife

and Smith, 2018; Scaife et al., 2020), and learned a lot from them. They mentioned that the current climate models can predict observed climate variability, although the predictable signal of the climate variability is small, especially in the Atlantic Ocean, and the small predictable signal may arise from an underestimate in the strength of the response to external forcing (such as volcanic forcing, solar variability, and ozone depletion). In our manuscript, based on the comparison of the piControl and historical experiments, we also found that when the external forces are changed, in most models, the NAO-SST is not changed obviously. Scaife et al. (2020) also did a lot of work to overcome the question of low signal-to-noise ratios, which also has a lot of inspiration for us. Unfortunately, since we do not have enough work basis to utilize the results from different models, initial fields and forced fields, we can't carry out the research work in this field within the short term. Thus, we have only added a short discussion in this paper. "Some studies have shown that in the climate models, the amplitude of the response to the external forcing (such as volcanic forcing, solar variability, and ozone depletion) is weak, which leads to weak predictable signals in these models although these models can predict observed climate variability (Scaife and Smith et al. 2018). The weak predictable signals inhibit the estimation of forced climate variability in the Atlantic sector (Scaife and Smith et al. 2018). The weak influence of the external forcing on NAO-SST relationship was also found in the CMIP5 models in this work. Scaife et al. (2020) have argued that a large number of ensemble can overcome the signal-to-noise paradox, which probably provide a reference for the future application of CMIP models in the predications." **(P30-31, L522-528)**