Review of the manuscript entitled “The role of buoyancy forcing for northern North Atlantic SST variability across multiple time scales”

The authors describe in the introduction that they aim to:
1) “document existing SST anomalies and phase relations in observational data, CMIP6 Shared Socioeconomic Pathways (SSP)126 experiments and Pliocene alkenone SST reconstruction from the North Atlantic, Norwegian and Iceland Seas.”
2) “Address why different SST phase relations may emerge and exist across different climate states, time scales and forcing scenarios.”
3) “Investigate impacts of changes in buoyancy forcing on the phase relationship between SSTs in the North Atlantic, Norwegian Sea, and Iceland Sea.”

Combining information from past, present and future climates, and paleoclimate data, observational data and idealized ocean model simulations; and use all of that that to improve our understanding of spatial patterns of climate change across a wide range of temporal scales is a courageous undertaking with potentially large impacts. However, in my view the authors currently don’t fully succeeded to logically combine all this information or derive underlying mechanisms and explanations for the observed patterns.
In the following I will detail all my major, minor and technical comments. I hope my comments will help the authors to improve the manuscript.

Major comments:

<u>Underlying hypothesis and research questions:</u>
The starting point of this study seems to be that on time scales longer than the advective time scale of SST anomalies between the three regions under consideration, the SST changes in the three regions should always be in-phase. I suppose this is theoretically true if all else remains equal. If on longer time-scales there are, however, changes in e.g. sea-ice, ocean currents, vertical mixing, boundary conditions etc, then this does not need to be true. Given that there is internal climate and Earth system variability on all possible time-scales plus forced changes, it seems to me that this theoretical starting point will very often not be met. I think this is also what the authors conclude on lines 687-688 “The in-phase situation is the norm relative to the mean background climate state under weak forcing”. (Where ‘forcing’ in this context is to be understood as any change that is external to the coupled system of the northern North Atlantic, the Norwegian Sea and the Iceland Sea; in contrary to forcings external to the climate system in general.) To me it seems that this should in fact be the starting point of this study (for reasons explained above), not a conclusion. This could potentially lead to an overall manuscript structure that is much easer to read and follw. I would like to ask the authors to reflect on this.

Then there are all the cases in which the ‘forcing is not weak’ (again ‘forcing’ would be any change that is external to the coupled system of the northern North Atlantic, the Norwegian Sea and the Iceland Sea). Given the wide range of time scales that are considered here, it seems to me that there are potentially many, very different, causes of ‘strong forcings’ and thus reasons why SST’s in the three regions under consideration would not vary in concert. The authors mention a number of them throughout the manuscript (changes in the Bering Strait, Arctic sea-ice export, strength of the different boundary currents in the region, gyre circulation, just to name a few). Is it the aim of this study to pin-point all these different causes for the time-scales that are considered in this study? Or to identify

commonalities between these causes? I think it should be more clearly described what the authors want to learn from the assessment they are presenting.

The third aim of this manuscript, according to the aims as I list them in the top of this review, is to use the MITgcm to study mechanisms underlying the observed 'phase relationship between SSTs'. Sensitivity studies allow one to test different hypothesis. The experiments presented here only focus on surface buoyancy changes. Why is that? Are there indications that on all the time scales under consideration here the drivers of the SST patterns were buoyancy changes? Is that the hypothesis of this study? Most of the explanations that the authors discuss for the observed patterns (changes in AMOC strength, changes in strength of the East Greenland Current, changes in salinity and/or sea-ice export via the East Greenland Current or a change in the connection between the Northern North Atlantic with the Nordic Sea, from the main connection being with the Norwegian Sea to the main connection being with the Iceland Sea) are not captured in the set of sensitivity experiments so I wonder if these experiments are the right ones to test the mechanisms underlying the observed, reconstructed and simulated patterns.

<u>The usage of the terms 'in-phase' and 'out-of-phase':</u>
The terms 'in-phase' and 'out-of-phase' for me are linked to temporal behavior. In the context of 'equilibrium' results (Pliocene time-slices, CMIP simulations and MITgcm results) I find these term rather confusing. If for instance a CMIP model shows cooling for the North Atlantic domain for 2070-2100 and warming in the Norwegian Sea they indeed have a different sensitivity to the forcing, but would you call that 'out-of-phase'? For the description of the observational record, I think usage of the terms 'in-phase' and 'out-of-phase' is appropriate. It is perhaps 'only a wording issue', but this for me confuses the whole concept behind the study. It raises questions like whether the authors mean to say that the in-phase Pliocene SST changes are part of an internal mode of variability? (like the described observational changes). Perhaps use words like 'spatially homogeneous versus spatially heterogeneous'? Coherent versus different?

<u>Data sources:</u>
Aims number 2 and 3, as described at the beginning of this review, are interesting research questions. However, aim number one is not a research question by itself. Describe more clearly why one would like to document SST anomalies and 'phase-relationships" in this particular combination of data-sets. What is the rationale of combining these data-sets?

Line 101: It seems to me that the easiest starting point to test the main hypothesis of this study (on time scales longer than the advective time scale of SST anomalies between the three regions under consideration, the SST changes in the three regions should always be in-phase) would be to look at CMIP/PMIP multi-millennial pre-industrial (control) simulations and past-2k simulations. Instead, future runs are used which introduce all kinds of complexities. Please explain the underlying reasoning.

The comparison with the CMIP6 results at the end of the century with the other data sources is difficult because, as the authors say, they show transient climate change, not equilibrium climate change. So why not include for instance experiments forced with a doubling of CO2, that would provide you with an equilibrium response that is more comparable. What is the added value of investigating the transient climate change at the end of this century? Please explain the underlying reasoning.

Line 101: why only look at Pliocene SST observations while Pliocene model simulations also exist.

Paleoclimate reconstructions:
Are the results for proxies local while the model results are large-scale averages? Does this impact the results? Did the authors test how the model results would look like if only data is used from the grid cells in which the proxy data is found?

Table 2: The information provided here for the Pliocene comes from many different sources. How good are the age constraints? Are they sufficient to assume that all of the presented climatic indicators happened during a single interval as defined by the authors?

MITgcm simulations:
The setup of the MITgcm is focusing on the connection between the North Atlantic, Norwegian Sea and Iceland Sea, but not the connection with the Arctic (sea-ice changes or opening of the Bering Strait). Nonetheless, those mechanisms are found to be important by the authors, questioning whether the setup of the idealized simulations is appropriate for the questions that are asked in this manuscript. Please explain why this approach is taken.

Line 206: why force buoyancy changes on the north of the ridge while the starting point of the manuscript is on the impact of advecting buoyancy anomalies from the region south of the ridge to the region north of the ridge?

Figures 1 and 5: Are the geographical regions used for the Iceland Sea and the Norwegian Sea in figure 1 (and so for the observational results and the CMIP results) comparable to the definition of the Iceland Sea and the Norwegian Sea in figure 5 (and so for the MITgcm results)? In the latter it seems really a comparison between a boundary current (Norwegian Sea) and the ocean interior (Iceland Sea), but is that a good representation of the other observational, modeling and proxy-based reconstruction results?

Minor comments:
Line 66: I understand what is being said here, but that is only because of the lines that follow. Please try to describe more clearly what is meant here. ‘Relationships’ between what? And why does the continuous northward transport of heat imply this?

Lines 208-215: I find the description of the modeling setup very confusing. You start by describing three different reference experiments. Why do we need three different reference experiments and how are they different? And then at some point in this text you continue describing the actual perturbation experiments. Why does the second set of experiments (G1-P1, G1-P2 etc) use G1 and not G0 as SAT forcing? Refer to table 1 at this point as the different experiments are nicely summarized in there, or include a new table to show this information?

Line 304: ‘enhances the SST in the Iceland Sea’, what does that mean? Please clarify.

Line 354: why would you call an insignificant response ‘in phase’ and why not simply an insignificant response? I guess there could also be a significant in phase response?

Lines 515-517: The opening of the Bering Strait could indeed play a role the Pliocene cooling over the whole North Atlantic, but what about other potential drivers? Higher CO2 levels for instance, how well are those constrained?

Lines 541-349: Many GCM experiments exist in which sea-ice melt increases and/or NADW is reduced. What do such experiments show? Is that in line with the findings described here for these specific Pliocene intervals? Please discuss.

Lines 576-577: I don't agree that the time-scales are comparable. The changes in the observational record play out on decadal time-scales while in the future runs, we are comparing roughly the period 2000-2050 with the period 2050-2100. Indeed figure 8 shows that on top of this 'long-term' variability, there is also decadal variability similar to the observational record for which the relationships between SST's in the three different regions are again different from what is described in this section (and perhaps more similar to the variability in the observational record). Please clarify.

Line 609: I don't quite understand this part. Do the authors mean that there is an indirect link between the radiative forcing and the 'out-of-phase' relationship between the North Atlantic and the Nordic Seas? And if so, what kind of link would this be?

Lines 632-633: Increased influence of the East Greenland Current can explain a cooling of the Iceland Sea, but how does it explain the corresponding cooling of the North Atlantic?

Lines 635-639: Wouldn't you still need to weaken advection between the Norwegian and Iceland Seas? Otherwise wouldn't warm North Atlantic water enter the Norwegian Sea via the Iceland Sea and still result in a warming in both regions?

Lines 688-690: what is meant here with a 'weakened ocean circulation'? Please clarify.

Lines 705-706: Isn't this what one would expect? That the amplitude of SST changes depends on the radiative forcing?

Figure 1: How are the domains depicted in figure 1 determined and what is their influence on the presented results?

Figure 7: clarify in the figure caption that these are all multi-decadal variations. What does it mean that only the bandpass filtered data are significant and the running mean data are not?

Figure 10: The SST response seems very small in many cases. How significant are these results and how does the magnitude of these responses compare to the magnitude found in the observational results, CMIP models and proxy-based reconstructions?
And on line 241 you mention that the North Atlantic is set to constant, so why are anomalies simulated for that region in figure 10? Clarify in a little more detail how the experimental forcing is defined.

Table 2: Why is information from the observational period, one of the three main periods discussed in this manuscript, not included in this table? Please explain.

Technical comments:
Figure 1: include letters a,b and c in the figure. Also I don't see the described blue and red boxes in figure 1. Please clarify.

Figure 6: What do the grey and white vertical bands mean in this figure? The ‘multi-decadal’ periods over which the anomalies in the inserted maps are calculated?
In general I find figure 6 not a very clear representation of the different in-phase and anti-phase relationships on interannual-to-decadal and on multi-decadal time-scales as is described on lines 258-263.

Figure 8B: why no inserted maps for the middle two models?

Table 1: The experiment name that is given in this table (exp) seems different from what is used in the main text and in figure 10. Consider changing for clarity.