Responses to Reviewer 1 – Pepijn Bakker

Below you will find our responses to Pepijn Bakker's review of our manuscript "The role of buoyancy forcing for northern North Atlantic SST variability across multiple time scales", egusphere-2022-959. His comments are provided in blue while our responses are shown in black, directly following the individual reviewer comments.

We appreciate his positive view on the overall goal of the paper and his very thorough, constructive, and useful feedback that clearly will help improve the manuscript. We acknowledge the main points about coherence, terminology etc. of the manuscript and will put major efforts into clarifying these.

Best regards,
Bjørg Risebrobakken, on behalf of all co-authors

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:

Underlying hypothesis and research questions:
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 easier to read and follow. 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 pinpoint 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.
We thank the reviewer for making us aware of the need to clarify the aim of the study, the hypothesis to be tested and the need to look carefully at the phrasings to avoid misunderstandings. We will put major efforts into addressing these issues.

We have considered the suggestion to change the starting point of the study from expected coherency to expected non coherency. We will, however, like to keep the observational record as our starting point and to set the expectation based on how oceanography of the focus region is known from observations. To do so, we want to focus and tighten the introduction to express more clearly what we want to do and why, including the reasoning for the choices made.

We agree that on long time scales, many factors, e.g. changes in continental configuration, may impact the large scale oceanographic features. Keeping that in mind, we see that how we have used the background knowledge from literature (Table 2) might be confusing. We will go carefully through the text to clarify how we use the information given. While many of the factors mentioned in Table 2 can be seen as forcing in themselves, they are more often used to infer how buoyancy may have been different under given conditions. E.g. sea ice may work as a force directly, however, more sea ice can also be seen as an indicator for more freshwater in the system, and hence a buoyancy change.

Furthermore, we will go carefully through the text with the aim to pick up other potential unclear statements. One example of how the text will benefit from clarifications is the statement mentioned above in original lines 687-688; here we should refer to buoyancy forcing specifically (as the sentence refers to the MITgcm results), not any change to the earth system. This has now been specified.

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?
Yes, we hypothesise that buoyancy plays a key role at all the time scales considered. We will make a clear statement in the introduction about why we focus on buoyancy.

Two key factors impact the inflow of Atlantic Water from the North Atlantic, over the Greenland Scotland Ridge and into the Nordic Seas; wind and buoyancy. From a present day physical oceanography point of view, buoyancy is considered most important at multidecadal and longer time scales, while wind is more important at interannual to decadal time scales. We acknowledge that wind may still be important, however, in this study we have chosen to focus on buoyancy. We expected buoyancy changes to be of major importance and wanted to see to what degree the observed SST patterns could be explained solely through changing buoyancy. While all cases may not be explained through these experiments, most cases can, emphasising the importance of buoyancy.

In the revised manuscript we will stress that wind may also impact the inflow, even though it is not a focus of this study to investigate the effects of wind changes. It could be an interesting future study to test the potential effects of changes in wind forcing on SST patterns. We may add a sentence on this in the outlook section.

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.
Thanks for making us aware of the need to clarify the purpose of Table 2 and how we use the information given there. We do not intend to investigate all possible factors that may have an impact. Rather, we wanted to provide background information on factors that may have had an impact on buoyancy over the region at a given time, or factors that may compare to responses seen in the idealised experiments to changes in buoyancy. We will consider the phrasings used throughout, to avoid misunderstandings on how we consider the relevant information at any point in the manuscript. We acknowledge that some of these factors may work as forces themselves, and this will be clarified.

The usage of the terms 'in-phase' and 'out-of-phase':
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?

We see that using the terms in and out of phase may cause confusion and will change the wording throughout the manuscript to spatially coherent and dissimilar or noncoherent.

Data sources:
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?
We totally agree that to document the anomalies is not a research question. This sentence was included to describe the different steps of the paper and is in that sense just a description of what we do to address what type of SST anomaly patterns that exist in the selected data sets. We will delete this sentence to avoid further confusion.

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 reason for looking into the future projections is twofold:

1. The spatially incoherent SST pattern, with a cold North Atlantic/warm Nordic Seas was detected in the RCP 8.5 CMIP5 runs but is not fully understood. As specified in the introduction, we wanted to see if this still holds in the lower emission scenario CMIP6 runs. Our results show that this is the case for the most sensitive model.
2. At the end of the century the CO2 concentration of the ssp126 scenario is in the range of the Pliocene CO2 level. Therefore, we wanted to look at the long-term SST pattern of the Pliocene, to see how long-term equilibrium SST patterns compares to the short-term transient response to a comparable atmospheric CO2 concentration.

When the work started out, we did consider looking at the longer historical runs in addition to the future runs, however, at that time the available runs were few and from different models than the ones we had selected for the analysing the projections. Since it would not have been possible to keep consistency between analysed models' consistency and to avoid adding even further complexity to an already complex study we decided not to continue in this direction.

Part of our motivation came from how the high emission scenarios seemed to break the expectation of spatially coherent SST anomalies over the North Atlantic, Norwegian and Iceland Seas. The question that came up then was if this signal was restricted to the high emission scenarios or if similar responses could be seen under lower emission scenarios, or in a past comparable climate state.

We see that a logical continuation of this study might be to look at the CMIP/PMIP multi-millennial pre-industrial (control) simulations and past-2k simulations and will add this as a potential future outlook in the summary section of the paper.

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.
We appreciate the suggestion to look at a 2*CO2 equilibrium response rather than the CMIP6 scenario experiments. As mentioned above, part of our motivation came from how the high emission scenarios seemed to break the expectation of spatially coherent SST anomalies over the North Atlantic, Norwegian and Iceland Seas, as set by the observational record. The question that came up then was if this signal was restricted to the high emission scenarios or if

similar responses could be seen under lower emission scenarios. The next question then is if the response relates to the transient nature of the ssp experiments, or if similar spatially noncoherent SST anomalies also existed during a past climate state with comparable CO_2 but in equilibrium. We could not have been able to address these questions by a doubling of CO_2 experiment.

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

When we started out, we did consider the possibility to look at PlioMIP experiments as well. However, in the end we decided not to investigate the Pliocene experiments. The reason why we did not analyse PlioMIP simulations is that they are time slice experiments. Analysing those could have given us the SST anomalies relative to pre-industrial, while the SST reconstructions (and the SST anomalies of the other time intervals) are all looked at as the SST anomalies relative to the mean of the respective investigated time series. Hence, the results would not have been comparable to our other results.

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?

The proxy records provide information from the individual sites, however, the available sites are from locations considered representative for the area that they represent. If any impact on the results when comparing with the model results that are averaged over a larger area, one may expect the proxy results to show a somewhat larger amplitude of change. We have by purpose avoided looking at the model results from individual grid points. It is not given that the grid points surrounding a specific site provides the best representation of that site, or that the representation would be constant between the different models. To address this question from the reviewer we will check the impact of using single grid points versus averaged domain.

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?

The shortest interval we look at lasts for 200 000 years and we only look at the mean state of the individual intervals. Looking at the mean conditions over such long time intervals, we do consider the age constraints to be sufficiently good for the use. Pliocene chronologies are mostly constrained by tuning to LR04 and/or using tie points from magnetic reversals. The tuning error is generally considered to be no more than a few thousand years but may exceed 10 ka prior to 4.3 Ma due to a less certain obliquity variance (Lisiecki and Raymo, 2005). If we had looked at shorter orbital time scales the 10 kyr uncertainty would have been an issue. We will add somewhat more details about this in Section 2.3. When feeding information to Table 2, we are consistent in treating the data in the same way as we have treated the North Atlantic/Norwegian Sea/Iceland Sea records.

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.

The model set-up is indeed idealized (but with full physical equations), including sea-ice and a possble key gateway missing, However, we have some experience in using the model in similar set-ups before – jointly to disentangle cause-and-effect and, broadly, regionalities (e.g., Våge et al. 2011; Jensen et el. 2018). And this we have found useful.

The approach has in general been spearheaded by Mike Spall, using it in numerous applications/process studies for the North Atlantic, Nordic Seas, and, in some cases, including the Arctic, on the nature of flow, exchanges, and external forcing (buoyancy, wind). Including making regional inference (with care).

We will clarify better in the revised manuscript how comparison is or can be made between the idealized – ("case study" – model, and both the other data and research questions of our mansusctipt. In particular, we will acknowledge the missing part of sea-ice and a northern gateway, and how that may reflect on our results.

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?
This goes to the nature of the approach/the model set-up. The restoring boundary condition in the south is also similarly a forcing – it is both the representation of the (infinite) source of Atlantic Water and the experiment's energy source (heat and buoyancy input). This water mass then experience the surface forcing applied, water mass transformation takes place, and a consistent ocean circulation is set up, incl setting the hydrography of the different regions. (The southern boundary energy input and northern surface heat loss balance when the model have reached (quasi-)equilibrium; they are accordingly equally important for the experiment.)

We will explain this better in the revised version of the manuscript.

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?
The reviewer is right that the Norwegian and Iceland Seas domains are slightly different when looking at the output from the MITgcm experiments compared to the CMIP output. Thanks for making us aware that we have missed adding information about this and the reasoning behind. In the revised manuscript we will add information on the reasoning for this choice in Section 2.4. We will also correct the caption for Figure 1. The MITgcm setup is idealized, therefore it is not possible to set exactly the same domains and expect comparability to the observations and CIMIP results.

The reviewer is also correct that MITgcm output is for the Norwegian Sea representative of the boundary current and the Iceland Sea the interior of Iceland basis. This is done as we consider these areas to be best representative for what the proxy data from the core sites represent. Site 642 in the Norwegian Seas is under direct influence of the Norwegian Atlantic Current represented by the eastern boundary current, while Site 907 is from the interior part of the Iceland basin, and hence not within the core of the western boundary current.

We believe regional inference broadly can be made from these experiments, but we will take greater care in outlining the basis/caveats for interpretation and comparison (see also response above). The reviewer's point regarding highly advective regions (Norwegian Sea) versus more "stagnant" interior basins (Iceland Sea), is a good one.

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?

We see that this could be unclear and will rephrase the start of this paragraph. Thanks for pointing it out.

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?

We thank the reviewer here and in general to identify where our overall presentation and lines-of-argument are in need of improvement, incl making specific and constructive suggestions. We will clarify and in particular use Table 1 more actively to structure the presentation of experiments/results as suggested by the reviewer.

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

We will rephrase this sentence to clarify the content.

In the model version with a lower horizontal resolution in the atmosphere the SST of the Iceland Sea is higher at the end of the century than in the model version with a higher horizontal resolution in the atmosphere. Such an effect is not seen in the Norwegian Sea or the North Atlantic.

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?

From the MITgcm experiments, all in phase SST changes are less than 2*std of the reference experiment, hence the notation insignificant in-phase. We will consider possible ways to clarify the message, e.g. by marking the inserts differently when the response is considered insignificant.

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?

Increasing or decreasing atmospheric CO2 concentrations will have an impact on the atmospheric temperatures over the region. For example, one might imagine that with arctic amplification a potentially increased atmospheric CO2 concentration may entail a larger temperature change over the Nordic Seas than over the North Atlantic, hence, comparable to idealized experiments relative to REF1 and REF3 where we change the buoyancy forcing by increasing the SAT over the Nordic Seas.

The CO2 level is unfortunately not well enough constrained before ca. 3.5 Ma to evaluate what role CO2 changes might have been for the individual intervals.

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.
Thanks for the suggestion. We will look at the literature and consider how to implement information from such studies before submitting the revised manuscript.

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.
We will clarify this point in the revised manuscript.

The SST anomalies of the observational record take place over multiple decades, hence multi decadal time scales. For the future we focus on the conditions at the end of the century, 2068-2098 relative to the mean of the future runs (2023-2098), hence, a 30 year long period, or a multi decadal signal.

For the end of the century (2068-2098), CNRM-ESM2-1 SSP126 and NorESM2-MM ssp585 show the discussed phase relation, with a cold anomaly in the North Atlantic and a warm anomaly in the Norwegian and Iceland Seas. For MPI-ESM1-2-LR SSP126 you cannot really distinguish the different members from each other. Therefore, we did not include the insert there (however, the same cold anomaly in the North Atlantic and a warm anomaly in the Norwegian and Iceland Seas is seen in the member presented in Figure 8b). NorESM2-MM is the least sensitive of the models, and for that model there are still spatially coherent SST anomalies at the end of the century for the SSP126 experiment, showing how the sensitivity of the model plays a role.

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?
Thanks for making us aware of this unclear statement. We will rephrase this sentence when we revise the manuscript.

The CO_2 forcing of SSP126 is comparable to the high end of the Pliocene CO_2 range (Meinshausen et al., 2020; De La Vega et al., 2020), suggesting that the amplitude of the SST anomalies of the North Atlantic and the Nordic Seas are set by the atmospheric CO2 level and that the response time may be as short as within a century.

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?
The latter is, admittedly, not trivial from how we present the case. A stronger East Greenland Current – more cold water – could do both. We will elaborate this issue better in the revised manuscript.

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?

We see that this is not very well explained. The argument would go something like this: the AW-source region is in a warm state. It is provided to the Iceland Sea more directly by a stronger Irminger Current, and the former warms. For the Norwegian Sea, fed by the Norwegian Atlantic Current (NwAC), the weakened NwAC decreases the heat input more than what is contributed from the AW source being warmer (the less volume transport, the more cooling from a given heat loss). And evetual influence of the Irminger Current on the Norwegian Sea via the Iceland Sea would be margial compared to the larger and more direct impact of NwAC.

We will elaborate this issue better in the revised manuscript.

Lines 688-690: what is meant here with a 'weakened ocean circulation'? Please clarify.
This relates to similar arguments as outlined in our response above. We will elaborate this issue better in the revised manuscript.

Lines 705-706: Isn't this what one would expect? That the amplitude of SST changes depends on the radiative forcing?
We agree with the reviewer that it may be expected that the amplitude of SST change depends on the radiative forcing. However, we still find it interesting that we see the same amplitude at the end of the century, after quite a short time of increasing CO2, as for Pliocene anomalies representative of hundreds of thousands of years. It emphasises how quickly change may take place. We will add a specification here.

Figure 1: How are the domains depicted in figure 1 determined and what is their influence on the presented results?
The domains are selected to be representative of the three sites where we have Pliocene data. We do consider that SST averaged over the domains to better represent the variability than a single grid point (see reply to comment above). More details on the selection of the domains will be added to section 2.1.

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 7 was included to show the anti-phase relations at interannual to decadal scales. We see that it may be confusing the way it was included and we will remove Figure 7 from the revised manuscript. The text will be edited accordingly.

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?
In Table 1 it is shown which of the MITgcm results are significant. The magnitude of the anomalies for each case is also defined in Table 1. In addition, we will add information about significance in the figure caption of Figure 10. All "red" experiments see significant anomalies in both the Norwegian and Iceland Seas. The "blue" experiments see a significant change in the Norwegian Sea, but not in the Iceland Sea. The "grey" experiments see no significant change in either the Norwegian or Iceland Seas.

The significant SST anomalies seen from the MITgcm experiments are in the range of 0.17 to 0.69°C for the Norwegian Sea and 0.51 to 2.42°C for the Iceland Sea; the insignificant MITgcm responses are 0.0-0.1 and 0.0 to 0.08, respectively (Tabel 1). The mean anomalies as seen in HadlSSTs, CMIP6 models and Pliocene reconstructions are within the same range for the

Iceland Sea. In the Norwegian Sea the mean anomaly exceeds the MITgcm range in the most sensitive CMIP6 model (CNRM-ESM2-1 ssp126), and in four of the six Pliocene cases. We will elaborate in the revised manuscript.

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.
We will check this and clarify in the revised manuscript.

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.
Related climate change for warm and cold periods in the observational period have been documented previously in studies about AMV. However, in this study, information from the instrumental period is primarily used to set the expectation, and confirm that the result previously shown by others (e.g. Årthun et al., 2017) for specific stations along the pathway of Atlantic Water transport also is valid when looking at the domains that we focus on in this study. We will add an explanation to the text.

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.
The letters a, b and c will be added to the figure. The reference to blue and red boxes was added to inform on how the boxes/defined domains are expressed in following figures. We see that the phrasing may be confusing and will delete this part of the sentence.

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
The grey bars are inserted to highlight the periods with positive spatially coherent SST anomalies. We will add a specification to the figure caption.

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.
The focus of Fig. 6 is to show in-phase relationships on multidecadal time scales, not the anti-phase relation at shorter time scales. Hence, you should not expect to see the anti-phase reflation. As mentioned above, figure 7 was included for the purpose of showing the anti-phase at interannual to decadal time scales, but to avoid confusion we will delete figure 7 from the revised manuscript and rephrase the text.

Figure 8B: why no inserted maps for the middle two models?
Inserted maps are not shown for the two middle models due to a weak signal and/or that there was not possible to say anything from between the different model members. We will consider adding inserts here as well, with specifications in the caption.

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.
Thanks for making us aware that this may be a source of confusion. We will consider restructuring the information about the exp names in Table 1.

References

de la Vega, E., Chalk, T. B., Wilson, P. A., Bysani, R. P., & Foster, G. L. (2020). Atmospheric CO2 during the Mid-Piacenzian Warm Period and the M2 glaciation. *Scientific Reports*, *10*(1), 11002, https://doi.org/10.1038/s41598-020-67154–67158.
Jensen, M. F., Nisancioglu, K. H., & Spall, M. A. (2018). Large changes in sea ice triggered by small chnages in Atlantic water temperature. *Journal of Climate*, *31*, 4847–4863.
Lisiecki, L. E., & Raymo, M. E. (2005). A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography*, *20*, PA1003, doi:10.1029/2004PA001071.
Meinshausen, M., Nicholls, Z. R. J., Lewis, J., Gidden, M. J., Vogel, E., Freund, M., Beyerle, U., Gessner, C., Nauels, A., Bauer, N., Canadell, J. G., Daniel, J. S., John, A., Krummel, P. B., Luderer, G., Meinshausen, N., Montzka, S. A., Rayner, P. J., Reimann, S., … Wang, R. H. J. (2020). The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500. *Geosci. Model Dev.*, *13*, 3571–3605, https://doi.org/10.5194/gmd-13-3571–202.
Våge, K., Pickart, R. S., Spall, M. A., Valdimarsson, H., Jónsson, S., Torres, D. J., Østerhus, S., & Eldevik, T. (2011). Significant role of the North Icelandic Jet in the formation of Denmark Strait overflow water. *Nature Geoscience*, *4*, 723–727, doi:10.1038/NGEO1234.
Årthun, M., Eldevik, T., Viste, E., Drange, H., Furevik, T., Johnson, H. L., & Keenlyside, N. (2017). Skillful prediction of northern climate provided by the ocean. *8*, 15875, doi:10.1038/ncomms15875.