Response to Reviewer 1

Review of "Causal Mechanisms of Subpolar Gyre Variability in CMIP6 Models"

Summary of paper

This paper explores the representation of hypothesised mechanisms that drive subpolar gyre variability in the CMIP6 piControl ensemble. The motivation is to understand how well represented the processes believed to be important for the subpolar gyre tipping point and collapse of deep convection are in models. The authors apply causal inference and causal effects approaches to show that most models only capture some elements of the important processes, and just one features all of them. This has implications for the reliability of CMIP6 models in modelling the subpolar gyre tipping point.

Overview thoughts and recommendation

I believe this paper makes an important contribution to understanding both the mechanisms of subpolar gyre variability, and how well models capture them. Since we are highly reliant on models to understand the risk of climate tipping points such as that of the subpolar gyre, this paper gives significant insight into how reliable these models are. Alternatively, the results could also hint at deficiencies in the popularly proposed mechanisms of subpolar gyre variability/collapse that might warrant further study. As subpolar gyre convection collapse is highly topical at present, this paper lays important groundwork for additional questions that can be addressed in future research.

I found the paper overall well written and easy to read. In particular, the use of colour coordination between the mechanisms schematic and later plots showing results was intuitive and aided understanding.

I recommend this paper for acceptance but with the following comments and questions addressed.

We thank the reviewer for their positive evaluation of the paper.

Major comments

- I am confused about the difference between the SSS to MLD pathway between the A1 and A2 mechanisms. From the schematic, it appears that that element is shared between them. However, Figure 3 makes it clear that there is a difference in the calculation for A1 SSS to MLD and A2 SSS to MLD. Could you elaborate on how these are different and how the calculation is done more in the text.

The reviewer is correct that the SSS-MLD pathway is shared between the two mechanisms. The difference that arises in Figure 3 is because of the difference in the variables on which one conditions in determining the causal effect. That is, to determine the causal effect of the link in A1 one conditions on SST, since it is also a driver of MLD, while for the causal effect in A2 this is not done as SST is no driver of MLD. We will clarify this in the text.

- Section 3: Is there any way to reformulate the hypothesised mechanisms A1 and A2 to meet the causal sufficiency condition needed for your method? For example, later you mention that atmospheric interaction has an important role – could including variables such as wind stress, E-P, latent and sensible heat help resolve this? Related to this, I wondered if there is any possible benefit to using ocean-only simulations (e.g. OMIP) to inspect these mechanisms. Does a freely evolving atmosphere obscure some of the important processes?

To fulfill the causal sufficiency condition all relevant variables need to be included. As the reviewer mentions, variables like wind stress and heat fluxes likely are drivers of both SSS and SST. The difficulty with including them in a causally sufficient network is that e.g. the wind stress can affect the heat flux and vice versa, as well as both being driven by e.g. the Jetstream, shifting the causal sufficiency condition one level up. For that reason, a "closed" hypothesis like the Born mechanism, is one that can readily be tested, but others tend to keep expanding making them unfeasible.

An ocean-only model would indeed allow studying the mechanisms in more detail, where one could use idealized atmospheric noise patterns or observations. However, to connect to the Swingedouw et al. (2021) results, the model would have to be forced by very specific atmospheric noise patterns as these are important for setting the locations where convection occurs. We will make a remark on this in the revised discussion.

- Section 4.2: "Physically these links could be related to the atmospheric dampening…". Related to my previous question, if you tested this in ocean-only experiments, would you then expect the sign to stay the same?

We do not expect these links to depend on the freely evolving atmosphere and no changes in the text needed.

- Page 9: I believe there is a figure or two missing from the SI that you refer to on this page. I cannot see one relating to omitting rho and studying just SubT -> SPG, or one relating to SPG to SPG at lag-2 ("Studying this system (see SI)…" and "…circulation strength on itself in most significant models (shown in SI)…").

We apologise for missing some figures in the SI. We will include those in the revised version.

- You mention that CESM2 is one of the CMIP6 models that Swingedouw et al. 2021 show features an SPG convection collapse. How do the other 3 models that they find show this tipping point perform in your analysis? Are they better on average than other models at representing these mechanisms? If not, how do you reconcile this with the fact they feature a collapse of convection under future scenarios.

The other models in which a collapse of SPG convection is found are CESM2-WACCM, MRI-ESM2-0 and NorESM2-LM. It could be the case that the mechanism is present in

these models does exist, but is not found to be significant for one or more links. Looking at the sign of the causal effect, representing the direction of the interaction, there is a small set of models which have the same sign as CESM2 for all links (with the constraint of the strength being at least 0.01, i.e. different from zero): CESM-FV2, **CESM2-WACCM**, **MRI-ESM2-0**, **NorESM-LM**. This is exactly the set of models in which Swingedouw et al. identified abrupt shifts (CESM-FV2 was not considered), indicating the interaction is present in these models, but obscured by internal variability. The links that are most relevant for distinguishing this set of models are those that include the SPG strength (SPG->SSS and Rho->SPG at both lags). If only constraining on these three links we find one additional model; ACCESS-CM2. This model has a different sign for the SubT->Rho links. Since it is not included in Swingedouw et al. (2021), we cannot determine how relevant this is. We will add a mention of the above in the paper and thank the reviewer for the suggestion to look into this.

- You mention that mechanism B is in part driven by the role of eddies. However most CMIP6 models that you look at are too low resolution to be eddy permitting or eddy resolving. I am intrigued that you mention there was no discernible impact from model resolution on the strength of the link. What was the range of model resolutions? As far as I can tell, none of the models you look at go above 0.25□ resolution – which is not enough to resolve eddies well. I think to really know if resolution has an important role you would need an eddy-resolving model (e.g. ~ 1/10□ resolution or better).

The reviewer is correct that none of the models have a high enough resolution to resolve eddies, hence the presence of the mechanism likely depends on how well they are parameterized. It would be valuable to test a model that is eddy resolving for this mechanism. We make an additional remark on this in the revised text.

- I think there are multiple promising ways to build on this work, for example looking at CESM2 in more detail to understand why it performs better, looking at high resolution models like I mention above, or including atmospheric variables in the mechanism etc. Could you add something to the discussion summarising what you think might be important next steps?

We thank the reviewer for the suggestion and will extend the discussion to include possible next steps.

Minor comments

- Page 2: “Convection also feeds back to the surface as water from depth is mixed with the surface water, lowering the surface density.” This statement may not be immediately intuitive. Could you add some additional descriptive words (e.g. “warmer water from depth is mixed…”) to aid understanding.

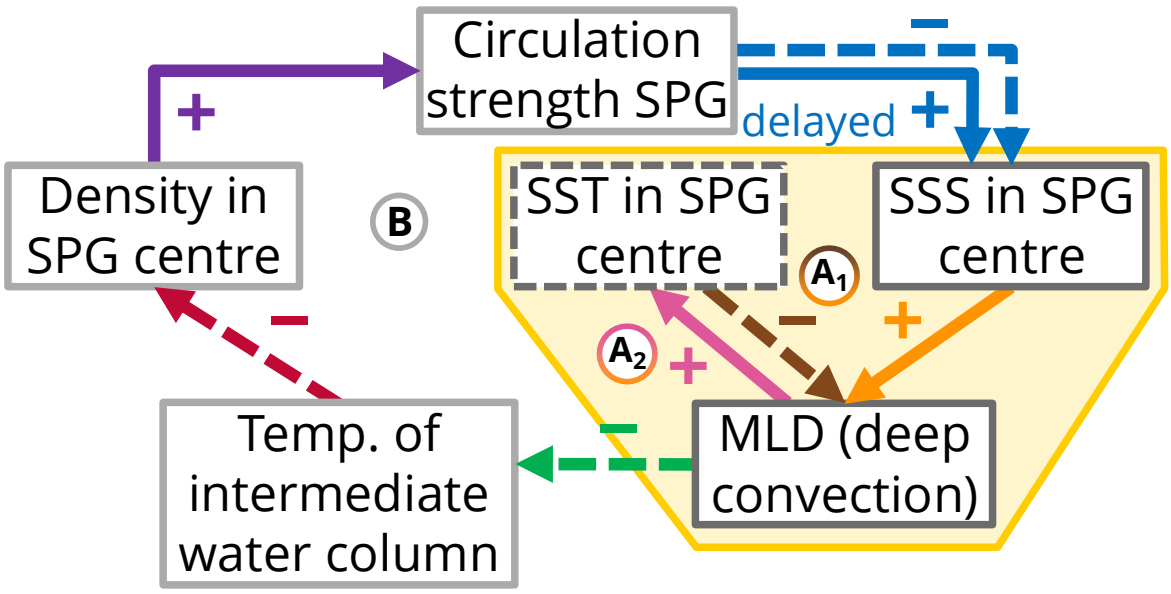
We will clarify this sentence to “Convection also feeds back to the surface as warmer water from depth is mixed with the surface water, lowering the surface density.”.

- Page 2, Born and Mignot 2012 – citation needs brackets.

We will change this.

- Figure 1: I found labels A2 and B being the same colour as one arrow in the full mechanism is a bit confusing. It implies B refers to just the green arrow for example.
- Maybe move label B to a more central location and change its colour to more obviously refer to the full mechanism?

We have updated this figure (below), changing the colours of the labels and moving label B to a different location.



- For models where it is not output, is it possible to derive barotropic streamfunction using velocities from the model? Is there difficulty in doing this that prevented it for this study?

Computing the barotropic streamfunction from the velocities in the model in principle is possible. However, this requires downloading the 3D velocity fields and then integrating over those, which requires a lot of memory and computational power. For one model this would be feasible, however for doing it for all models would take a lot of computational resources and hence we refrained from it in this study.

- Section 2.2: Python not capitalised; too instead of to ("...the longer it takes too compute..."

We will correct this.

- Section 2.2: Can you explain in more detail what an "intervention" refers to. I did not follow why this cannot be done with time-series data.

With an intervention we refer to interfering in the system, like can be done when studying waves in a wave tank by changing the e.g. forcing frequency. Such an intervention allows for cleanly separating cause and effect. However, in the climate system such an intervention, for example changing the salinity in the Labrador sea is unfeasible and undesirable. In climate models such interventions are possible, but computationally very expensive and therefore unfeasible. We will clarify this point in the text.

- Section 3 (and hereafter): when you refer to the SI, can you label the particular figure you're referring to.

We will do this.

- Section 3: "The latter is a feature…" can you clarify if the latter refers to mechanism A2 or the lag-1 memory effect? Maybe "The latter mechanism" for example will make this clearer.

This refers to the lag-1 memory effect. We will clarify this.

- Section 4.1: "SPG -> SSS link is absent in 22 out of the 32 models". From Figure 4, I cannot see where 10 models show this link. Is that meant to be the sum of purple and orange? This does not seem to add to 10 though.

The number of 22 refers to the absence of the link for all lags. With 8 models showing the link at lag-1, and two more at lag-0 (next to some that have both as significant). We will add a note that this is for all lags in the text.

- Section 4.2: Do you have an alternative suggestion for gyre strength metric that could better capture the baroclinic nature of the gyre?

We also considered sea surface height as a measure of gyre strength. We found both were strongly linked, but that the links with the barotropic streamfunction were stronger. We have thought about a better metric to capture the baroclinic contribution of the gyre circulation. However, any definition would depend on the density in the gyre one way or another, meaning such a gyre metric would automatically be linked to the density (and thus temperature and salinity). This way you likely "build in" a link between e.g. rho and the SPG strength, which would complicate a proper verification of the mechanism.