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.

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 A_1 SSS to MLD and A_2 SSS to MLD. Could you elaborate on how these are different and how the calculation is done more 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?

- 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?

- 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)…").

- 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.

- 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).

- 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?

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.

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

- Figure 1: I found labels A_2 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?

- 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?

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

- 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.

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

- Section 3: "The latter is a feature…" can you clarify if the latter refers to mechanism A_2 or the lag-1 memory effect? Maybe "The latter mechanism" for example will make this clearer.

- 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.

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