Response to Reviewer 2

Review of "Causal mechanisms of SPG variability in CMIP6 models"
This study aimed at assessing whether the hypothesized physical mechanisms that could lead to a tipping of SPG convection are effectively represented in modern climate models. To this end, the authors analyzed unforced piControl simulations from CMIP6, and evaluated those interactions that could give rise to SPG convection variability and/or bistability. They applied causal inference and causal effect estimation techniques to test the presence, strength, and direction of the commonly proposed causal links among the key variables in the SPG system. The results show that while interactions between salinity, temperature, and mixed layer depth are relatively well represented in many CMIP6 models, the feedbacks involving density and gyre strength are either inconsistently captured or missing entirely. Only one model, i.e. CESM2, exhibits full consistency with the proposed feedback mechanisms underlying SPG bistability.

1. General assessment
The study is highly relevant and valuable, as it tackles a crucial issue for understanding and predicting future climate change: to what extent climate models can capture the key mechanisms governing SPG dynamics. The SPG convection, commonly proposed as an independent tipping element of the climate system, has already been projected to collapse in some future projections, with severe impacts on the North Atlantic. However, the models exhibiting such a collapse do not appear to share common patterns and underlying processes. Furthermore, the SPG convection is closely linked to the AMOC, another recognized major tipping element, whose disruption could lead to even broader climatic consequences.

The comprehensive coverage of CMIP6 models allows the authors to draw robust conclusions about model spread and biases when reproducing mechanisms in the SPG. In this context, the results for CESM2 are particularly noteworthy: it is the only model that shows both a convection collapse in future projections and a consistency with the physical mechanisms underpinning SPG bistability. This finding supports the plausibility of the potential bistability of the SPG convection, thus reinforcing its potential role as a climate tipping element. Conversely, the absence of such mechanisms leading to bistability found in the other models suggests that SPG tipping, as shown in other studies focusing on future projections, might rather reflect a nonlinear and irreversible response to external forcing. Nonetheless, the lack of detected bistability mechanisms in these models may also stem from methodological limitations in the current analysis, as discussed by the authors in the text and also detailed below in the major comments. I think that these limitations do not undermine the overall quality of this study, but rather point to valuable directions for future research.

A key strength of the paper lies in its innovative application of causal inference and effect techniques, i.e. the PCMCI+ framework, to disentangle the causal structure and feedbacks within the SPG system. This represents a notable methodological advancement when investigating SPG dynamics, which traditionally relies on

correlation-based diagnostics. However, the experimental design used to analyze mechanisms A1 and A2 could possibly benefit from further refinement to match the overall rigor and quality of the paper, as detailed below in the major comments.
The figures in the manuscript are well-designed and effectively informative, while a couple of figures referenced in the SI appear to have been inadvertently omitted. The manuscript is grounded in a clear hypothesis framework, and generally follows a logical narrative. However, I find certain parts technically dense and too hard to understand at a first read. This mainly reflects the inherent complexity of both the methodology used and the SPG dynamical system, yet there is space for some simplification or improvement as detailed below.

Overall, I think this is a strong and well-executed paper that provides timely and insightful results through the use of an innovative methodology. The few limitations appear to likely arise from methodological constraints and minor omissions. With a few additional refinements and clarifications, I believe the manuscript would be very suitable for publication.

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

2. Major comments

- *Mechanisms A1 and A2*

I find the analysis of mechanisms A1 and A2 rather weak with respect to the overall study. This is for the following reasons:

(i) Violation of causal sufficiency: as mentioned by the authors, this analysis omits potential common drivers, which conflicts with the otherwise high level of scientific rigor in the study.
(ii) Redundancy with mechanism B: the main outcome of this part of the study, i.e. the strong causal link SSS → MLD in mechanism A1, is already robustly captured in the more comprehensive analysis of mechanism B.
(iii) Lag structure ambiguity: the methodology for estimating causal effects in A1 and A2 is not fully clear. I suppose that links have been evaluated at lag 0 (?), i.e. contemporaneous links, which raises concern for mechanism A2. Indeed, the restratification feedback (MLD → SST) would require a lag of at least one year to be detectable, as deep convection occurs in winter, while restratification starts in the following spring. The effects on SST would be therefore visible only from lag-1 on.

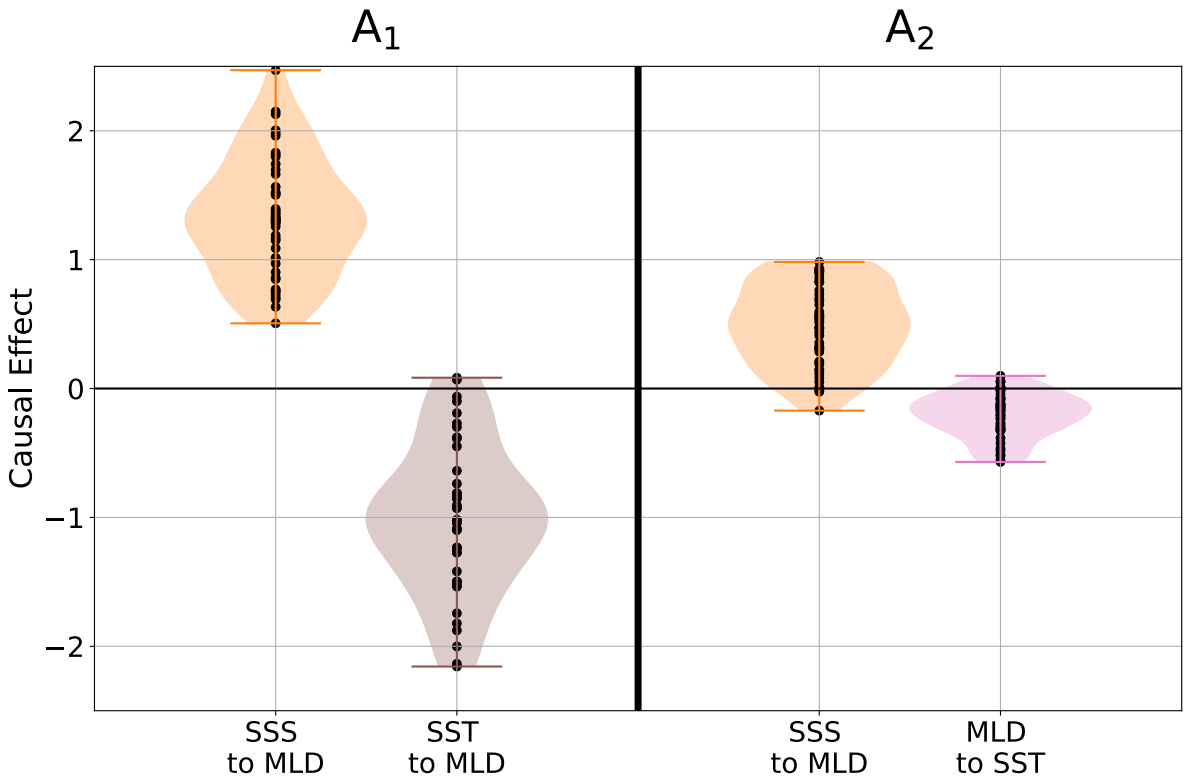
More broadly, I cannot understand why mechanisms A1 and A2 were treated independently. To my understanding, the PCMCI+ algorithm should be able to handle causal networks that include bi-directional links at different lags (i.e., SST → MLD at lag 0, and MLD → SST at lag 1). If the separation between A1 and A2 is due to a limitation in the PCMCI+ algorithm, it would be helpful to clarify this point in the text. Otherwise, applying PCMCI+ to a unified A1+A2 structure may yield a more consistent representation of the SPG convection feedbacks. If this is technically feasible, and in a relatively short time, I would therefore encourage the authors to consider a unified

A1+A2 structure, with also the inclusion of a shared parent for SSS and SST, e.g. by including the SPG gyre strength in this loop (or, said in a different way, by expanding the yellow box to SPG gyre strength in Fig.1). The use of alternative confounders of SST and SSS, such AMOC or NAO, would also offer valuable insights into the mechanisms of SPG convection. However, given the additional effort that such an analysis would require, it is reasonable to present this as a potential direction for future work in the Conclusions section.

The reviewer is correct that the causal inference step of PCMCI+ in principle is able to detect the links between the variables. However, when doing this the strongest link identified is one between SST and SSS at different lag times. Common drivers lead to a strong (lagged) correlation between the two, meaning the outcome of causal inference is hard to interpret. Therefore, we focussed our analysis for mechanisms A1 and A2 on the causal effect given the hypothesis of these mechanisms.

The difficulty with expanding the current analysis to include the SPG strength to have a causally sufficient mechanism is that the analysis in mechanisms A1 and A2 is on the local processes, with the location being different between all models (whereas the analysis for mechanism B is for the Labrador sea). This means that the inclusion of the SPG strength as a common driver would give results that are hard to interpret given the different locations that are considered (see Fig. 2) and therefore we refrain from doing so. The different locations used to study the different mechanisms also means that the A1 and A2 analysis is not redundant, as it allows for better identifying the local convection dynamics. For example, the signal of the SSS->MLD link is significantly stronger when considering the model-specific convection locations.

We will clarify the above reasoning on the distinction between local and Labrador sea processes in the text. In addition we will better describe the lag structure used for both mechanisms. The reviewer is correct that all links here are contemporaneous. We will add a discussion of lags for mechanism A2 and thank the reviewer for noting this likely is more relevant for the MLD->SST link at lag-1. We have now computed these, showing an improvement on the identified restratification feedback as seen in the plot below, and will add the results to the revised manuscript.



- *The sensitivity of causal links to the model-specific sites of deep convection*

Concerning the mechanism B, results show a generally weak causal connection between SubT, rho, and SPG (gyre strength) across most CMIP6 models. This is an important and intriguing finding, in apparent contradiction with proposed mechanisms of bistability of the SPG convection. The authors explained these outcomes as associated with the use of the barotropic flow rather than baroclinic as a measure of SPG (gyre strength), and with the role of other confounding factors not considered here, e.g. the atmospheric circulation. While these interpretations are physically plausible, I am concerned that the result may more merely reflect limitations of the experimental design. Specifically, for the analysis of mechanism B, the authors considered a fixed spatial box over the Labrador Sea. However, in at least 10 out of 32 models, convection may be substantially weaker or even absent in this region. This spatial mismatch, i.e. the different location of the major convection sites among models, could severely impact the estimation of causal relationships in the fixed box, and introduce large inter-model spread, as indeed seen in Figure 5, notably for SubT→rho and rho→SPG. I believe this caveat warrants explicit discussion in the manuscript.

In this context, it would also be helpful to clarify which depth levels of potential temperature (thetao) were used to compute SubT. Given the varying convection depths across models, this information is essential for assessing the robustness and comparability of the results. Also, it is not clear to me if the “density at the centre of the SPG” is the surface density averaged over the entire Labrador box or it has been defined in a different way.

The reviewer is correct in noting that the choice of the Labrador sea box is one of the limitations of this study. For example, the CanESM-1 model, in which no significant

links are identified, does rarely show deep convection occurring in the Labrador sea (instead it occurs in the Nordic seas). We will add some discussion on this in the manuscript. In addition we will clarify the depth levels considered (50-1000m depth). The reviewer is correct that the density is computed over the entire Labrador sea box.

- *General readability*

The manuscript is very well written, yet a few minor revisions could further improve the clarity and flow of the text. First, references to figures in the SI should be more precise, ideally citing the specific figure numbers rather than using a too generic "see SI." Note that several supplementary figures referenced in the main text appear actually to be missing from the SI. Please fix it!

We apologise for part of the SI missing and will correct this. We will also add references to the exact figures and tables in the SI.

Also, I recommend reconsidering the use of the term "direct link" as a synonym for "contemporaneous link" (i.e., causal connections at lag-0, which can be "directed" or "unoriented"). This terminology may lead to confusion, as "directed link" in causal graphs refers more broadly to the presence of a causal direction, regardless of the lag. This apparently subtle issue led me to some initial misinterpretation of the text, which can be easily avoided by adopting a less ambiguous term.

We thank the reviewer for noting this and will consistently use contemporaneous link throughout.

3. Specific minor comments and technical corrections

Page 3 – Fig. 1 → why the potential link "SPG strength" to "SST in SPG" has not been considered for the analysis of mechanism B? Also, I suggest to clarify in the caption the reason for framing SST with dashed lines. Without instructions, I assume this is intended to highlight the dual role of SST within the mechanisms, as it acts as a parent in A1 and a child in A2. Is it correct?

Mechanism B is a hypothesis is based on the study of Born, in which SST was not found to be of strong relevance for the mechanism. Therefore, we do not include it here.

The reason for the dashed lines around SST is to indicate it is only included in mechanisms A1 and A2, and not in mechanism B. We will clarify this in the caption

Page 5 – Fig. 6 → are the causal effects between variables in mechanisms A1 and A2 estimated at lag 0 (i.e., contemporaneous), while accounting for persistence through lag-1 autocorrelation? Please provide more details on this point.

The reviewer is correct that they are estimated at lag-0 with lag-1 autocorrelation. We will clarify this in the manuscript.

Page 5 – “Secondly, the time series is assumed to be stationary, in the sense that the presence and strength of the links does not change in time.” → instead of referring to “stationary time series” I would rather express this as a “stationary set of causal links”.

We thank the reviewer for the suggestion and will change this phrasing.

Page 6 – “The number of models for which each of the five links is significant are shown in Figure 4 for lags up to 10 years (individual model results are shown in the SI).” → this is not shown in SI.

We will correct the SI and ensure all figures are included.

Page 7 – “All but one model capture at least one step of the process (B) (only CanESM5-1 does not).” → it would be interesting to know more about the MLD pattern in this model, particularly the location of its maximum relative to the Labrador Sea box. Unfortunately, Fig. 2 does not allow to identify it, likely because “not all markers are visible due to overlapping maxima.” Could the inability of this model in capturing any step of the mechanism B be related to the specific MLD pattern? Clarifying this point could possibly help interpreting the inter-model spread.

The location of maximum MLD in CanESM5-1 is located in exactly the same spot as that in CanESM5, south of Iceland. Both CanESM5 models are among those that show the least convection in the Labrador sea, instead their convection happens outside the SPG region in the Nordic seas. We will add a note of this in the manuscript.

Page 7 – “The links to and from SPG are found in only 10 and 17 models respectively.” → I think that it is not “SPG” here but rather “rho”.

We thank the reviewer for noting this mistake and will double check the numbers.

Page 8 – “In Figure 5 the causal effect of each link is shown for models with and without that link being significant following the causal discovery step (values for individual models are given in the SI).” → this figure in SI is actually missing.

We will correct the SI and ensure all figures are included.

Page 8 – “At lag-1 both these links are found with the opposite sign in most models where they are significant, contrary to the theoretical model.” → here “both these links” means SSS to MLD and MLD to SubT? Please specify it.

We will specify this in the manuscript.

Page 9 – “Studying this system (see SI) shows large disagreement between models on the sign of the SubT→SPG link at lags of one and two years.” → this is actually not shown in SI.

We will correct the SI and ensure all figures are included.

Page 10 – “The absence of the interaction between density and gyre strength means that we cannot be certain of the existence of the bistability mechanism in many CMIP6 models, and thus whether they are able to exhibit tipping behaviour in SPG convection.” → this phrasing seems to suggest that tipping behaviour necessarily depends on the presence of bistability, which is not always the case. While bistability does allow for tipping through bifurcation, a tipping point more generally refers to a critical threshold at which a small change in forcing leads to a rapid shift in the state of a dynamical system. Such behaviour can also occur without bistability, e.g., through strongly nonlinear but irreversible responses. I suggest to rephrase this part.

We thank the reviewer in noting that we implicitely use a definition of tipping that is linked to bistability. We will clarify the link between tipping and SPG bistability based on the mechanistic understanding in the introduction. Furthermore, we will rephrase this part to be in line with the added clarification.

Page 10 → (Gou et al., 2024) should be (Gu et al., 2024).

We will correct this.

Page 10 → “grye strength” should be “gyre strength”.

We will correct this.

Page 14 → “Tippint” should be “Tipping”; “Jonathen” should be “Jonathan”.

We will correct this.