Article: AMOC weakening across latitude and time in CMIP6 future scenarios

Review

The manuscript, "AMOC weakening across latitude and time in CMIP6 future scenarios," investigates the latitude-time evolution of AMOC weakening in CMIP6 historical and ScenarioMIP simulations, with emphasis on differences between depth-space and density-space overturning and on a threshold-based framework intended to quantify the southward propagation of weakening. The manuscript is clearly written, visually coherent, and ambitious in scope. I also think the attempt to move beyond a single-latitude AMOC index is a genuine strength. The figures are coherent and well constructed with a clear physical narrative in mind.

The authors conclude that density space may offer a more suitable early warning signal. They further argue that the propagation speed associated with a given weakening threshold decreases as the threshold increases.

That said, I have several major concerns regarding the methodology and the strength of the conclusions. In my view, the manuscript, in its current form, does not yet provide a sufficiently robust basis for inter-model comparisons and for some of the stronger physical interpretations, for reasons discussed below. Unless these issues are addressed thoroughly, I do not believe the manuscript is suitable for publication in Ocean Science.

We thank the reviewer for taking the time to assess our manuscript and provide thoughtful and thorough feedback. Their comments have helped us clarify and substantially improve the manuscript.

Major Comments

1. The study uses 15 CMIP6 models, with one ensemble member per model, chosen on the basis that they either provide overturning streamfunctions directly or provide the variables needed to reconstruct them. Table 1 shows that the model set is a mix of archived depth-space streamfunctions (msftmz or msftyz), archived density-space streamfunctions (msftmrho), and offline reconstructions from vmo, or from vo, so, thetao, and sometimes thkcello. Only eight models provide depth-space streamfunctions directly, and only four modelling centres provide density-space streamfunctions directly. The ensemble is diagnostically heterogeneous, and this is not adequately validated.

   This is, in my view, the central methodological concern. Archived overturning diagnostics generated by modelling centres are not automatically interchangeable with offline reconstruc- tions generated by the authors. The manuscript explains the offline calculations in some de- tail, including interpolation of salinity and temperature to the V-grid, calculation of sigma-2 with GSW/TEOS-10, density binning, and reconstruction on native-grid V-points. However, nowhere do the authors seem to show that, for models where archived streamfunctions are available, the offline method reproduces those archived products.

Such validation is necessary before the mixed use of archived and reconstructed products can be considered robust. Without it, part of the inter-model spread may reflect diagnostic-method differences rather than model physics. This is particularly important because the paper later uses these fields to derive threshold-crossing years and "weakening gradients," which are more sensitive than the raw streamfunctions themselves.

I strongly recommend that the authors add a validation section in which archived and offline- reconstructed overturning are compared directly for the overlapping subset of models. At minimum, the comparison should quantify differences in: mean AMOC strength, latitude of maximum overturning, latitude of onset of weakening, threshold-crossing year, weakening- gradient estimates.

Without such a calibration step, I do not think the mixed use of archived and reconstructed products is sufficiently justified.

We recognise the importance of a consistent set of variables for comparison and thank you for raising this.

For the depth-space streamfunction, we have changed to use the vo variable for all models except ACCESS-CM2 and ACCESS-ESM1-5. Those two models lack available vo data for ssp126 and ssp585 therefore we use the online msftmz version.

For the density-space streamfunction, we have added a verification in Appendix B comparing online and offline streamfunctions. We find that the offline calculated streamfunctions are weaker than their online counterpart and the AMOC strength is ~1-3Sv less. As our method focuses on percentage change, we find the threshold lines and onset largely agree between the online and offline versions. We argue the mixed use of online and offline density-space streamfunctions do not have a significant impact on our analysis.

For the sake of completeness, we include in Appendix A the comparison of online, vmo and vo streamfunctions to illustrate that using vmo to calculate the AMOC streamfunction near-perfectly reproduces the online msft- streamfunction. We also show a comparison of the Hovmöller plot calculated using vo and vmo/online streamfunctions and they are very similar. As we use vo in the density-space streamfunction calculations, we opt to use the vo streamfunctions for consistency.

2. The manuscript argues that weakening originates further north in density-space than in depth- space, and concludes that density-space is therefore better suited for early warning. It then presents the remainder of the analysis in density-space, arguing that this captures a more complete representation of weakening.

   This may be physically plausible. However, given the heterogeneity described above, I do not think the paper has cleanly isolated a coordinate effect from a method effect. Density-space overturning is more often reconstructed offline than depth-space overturning in this model set. Therefore, the comparison is not purely "density-space versus depth-space." It is partly "offline-reconstructed density-space versus a mixed set of archived and reconstructed depth-space products."

That distinction matters because the paper uses this comparison to support one of its headline conclusions. At present, I would regard the result as suggestive rather than robustly demon- strated. The authors should either validate that the coordinate-space conclusion survives when only diagnostically comparable models are used, or else temper the language.

We thank you for highlighting this. As discussed in the previous point, we have recalculated most of the depth-space streamfunctions offline using vo. This should make them more directly comparable to the offline calculated density streamfunctions. The new depth-space streamfunctions show very similar result to those using the online streamfunctions; density-space weakening still occurs sooner and at higher latitudes than depth-space weakening. We therefore are confident these results are not due to the method of calculation used but that it is the density-space representations itself that means it is better for early warning than depth-space representations.

3. The paper defines AMOC weakening as percentage reduction relative to 1850–1899, computes 10-year running means, and then identifies the first year each weakening threshold is exceeded at each latitude. These threshold contours are then fit linearly between 40°N and 30°S to produce a "weakening gradient," interpreted as a propagation speed. This is helpful. However, it rests on several strong assumptions: that threshold lines are sufficiently smooth and physically interpretable, that the first exceedance year is a robust measure in the presence of internal variability, that a linear fit over such a large latitude range is meaningful, that the resulting slope can be interpreted as a propagation speed rather than a property of a derived contour.

   The manuscript does acknowledge important weaknesses of the method later. It states that threshold lines become noisy in models with substantial internal variability or non-smooth weakening, that confidence in the gradient calculation is reduced in such cases, and that with one ensemble member per model it is not possible to disentangle forced weakening from internal variability for all models.

   However, in my view, its implications are not fully carried through. If the threshold-gradient diagnostic is unstable in variable models, then results based on exact threshold timing, gradient curves, and especially model categorization should be treated much more cautiously. At present, I think the manuscript treats the threshold-line method as more robust than the paper's own caveats justify.

   Following comments from Reviewer 1, we have introduced multiple ensemble members for 5 of the 15 models. The ensemble mean of these models are used and threshold lines calculated from these are treated with more confidence than the single ensemble realisations. We have re-focused the analysis in Section 3.3 to these 5 models. The ensemble analysis highlights that some models (e.g. CanESM5) have very similar behaviour across the ensemble set, whereas other models (e.g. CNRM-CM6-1) have very varied behaviour across their ensemble set. As such, single ensemble members may be representative of the model (like CanESM5) or may fluctuate as seen in CNRM-CM6-1. Due to this, we agree that the threshold lines for models with single ensemble members should be interpreted cautiously and caveat analysis that uses these models.

4. The classification of models into “gradient behaviour” categories is not yet convincing. The manuscript classifies models into categories such as constant gradient, continuous decline, decline to plateau, decline to increase, and inconclusive. The classification is based on differences between early, mid, and late parts of the model’s threshold-gradient curve under ssp585.

   I do not find this categorization sufficiently robust in the current form. First, it depends on the threshold-gradient diagnostic, whose fragility in variable models is already acknowledged. Second, the paper itself notes that increasing gradients in some models may be more a result of internal variability and methodological limitation than of any physically meaningful increase in propagation speed. Once that is admitted, the classification scheme begins to look less like a robust physical taxonomy and more like an exploratory grouping.

   I suggest either substantially softening this section and presenting these groupings as tentative, or supporting them with stronger robustness tests, ideally using multiple ensemble members where possible.

   We agree the categorisation of gradient behaviour is not robust and have reduced the scope and complexity of categorisation to whether the gradient is decreasing, increasing or varying. This is done in Sect 3.3.

   We have expanded the analysis to multiple ensemble members and reworked Sect. 3.3 to focus on the models with multiple ensemble members. One model, CNRM-CM6-1, had a wide range of weakening gradients in its ensemble set with as much spread as seen in the entire model set. Other models have much more similar weakening gradient behaviour between their ensemble members. Because of the example of CNRM-CM6-1, we realise any definitive categorisation of model behaviour for a model with a single ensemble member cannot be considered robust. By focusing purely on whether the gradient is increasing, decreasing or varying, we give an overview of the general trends in the model set but agree these groups are purely tentative and would require further work to definitively categorise.

5. The manuscript distinguishes between the explicit multi-model mean (EMMM), derived from threshold lines of the MMM field itself, and the statistical multi-model mean (SMMM), derived from averaging model-wise weakening gradients at each threshold. The authors state that their analysis will primarily focus on the EMMM. In a heterogeneous ensemble, averaging first and diagnosing later can create smooth structures that are not representative of individual-model behaviour. This is especially relevant here because the manuscript later shows that the number of models contributing to the SMMM declines at high thresholds and that the EMMM and SMMM diverge there. That is a warning sign. Since the main scientific question concerns AMOC weakening behaviour across models, model-by-model results should be primary, and the EMMM should be treated as an illustrative summary rather than the central physical object.

   We agree the EMMM is not a physical object but we argue the EMMM Hovmöller diagrams are useful objects for providing insight. The comparison between depth-space and density-space and of the different scenarios could be made on a model-by-model basis but it is much more straightforward to illustrate these large structural differences with the EMMM.
   We believe the divergence between the EMMM and SMMM gradients is an argument in favour of the use of EMMM over the SMMM. Taking the mean Hovmöller of all the models to

then analyse provides physically plausible weakening gradients (see MIROC6) and produces Hovmöller diagrams that are middle of the road for this model set. The SMMM however is skewed at high thresholds by the models which experience the greatest degree of AMOC weakening.
We do acknowledge your concern with the EMMM and avoid reference to it when discussing plausible mechanisms, focusing instead on interpreting the models with multiple ensemble members.

6. Some conclusions are stronger than the evidence currently supports. The abstract and discussion make fairly firm claims: (a) the weakening originates at high latitudes and propagates southward, (b) density-space is better suited for early warning, (c) propagation speeds are similar in depth-space and density-space south of 40°N, and (d) models can be categorized according to propagation-speed behaviour.

   These claims may well be directionally reasonable, but given the heterogeneous diagnostic inputs, the lack of archived-vs-offline validation, one member per model, acknowledged instability under internal variability, and the strongly derived nature of the threshold-gradient metric, I think the conclusions should be softened. The paper is presently stronger as an exploratory methodological study than as a firm characterization of forced AMOC propagation across CMIP6.

   Given the conversion to generally offline vo calculated depth-space streamfunctions, as well as the verification of density-space streamfunctions, and the introduction of multiple ensemble members for 5 models which are thus analysed, we argue that there is sufficient evidence in support of the points you've listed as a, b and c. However, we agree the language used is stronger than perhaps is justified and soften our language accordingly. We agree with point d that model categorisation is not robust and have removed the emphasis of model categorisation from the paper.

Minor Comments

- The manuscript normalizes AMOC change as percentage weakening to facilitate comparison across models with different mean AMOC strengths. But the authors should acknowledge more explicitly that percentage normalization does not remove method-dependent structural differences in the underlying AMOC diagnostics.
  - We have added this comment to the methods (Lines 164-165).
- The choice of a 10-year running mean is acceptable for reducing interannual noise, but the authors should say more about the sensitivity of threshold timing and gradient estimates to this smoothing choice.
  - We have added a comment in the methodology about the effect interannual variability would have on the confidence of the gradient calculations (Lines 189-192).
- The use of a single ensemble member per model is a major limitation for any threshold-based "first exceedance" framework. This is acknowledged, but the limitation should be stated more prominently in the abstract or conclusions as well.
  - The first exceedance threshold method has been added to the abstract and is mentioned in the discussion (Lines 548-549).

- The manuscript states that there is no relationship between propagation speed or propagation category and mean AMOC strength. Given the relatively small sample and the methodological uncertainties, this statement should be phrased carefully as a result for this model set rather than as a general conclusion.
    - Given the changes to the dataset with the introduction of multiple ensemble members, a statistically significant positive trend ($p < 0.05$) has been shown between reference AMOC strength and low threshold weakening gradients. A plot showing this has been added to the discussion (Fig 12), but given uncertainties with the weakening gradient method for models with one ensemble member, we are cautious to infer too much from this relationship (Lines 527-535).
- Since the paper argues that density-space is more informative north of 40°N, it would be useful to show at least one targeted robustness figure restricted to the subset of models with directly archived density-space products, even if the sample is smaller.
    - Added an analysis as an Appendix B comparing online and offline density-space products.
- The manuscript states that so and thetao are interpolated from T-grid to V-grid points before sigma-2 is calculated, but it does not describe the interpolation procedure. Please specify the interpolation method, how masks and partial cells are treated, whether sigma-2 calculations were done on monthly values, and whether this is handled explicitly by the authors' code or by the NEMO Cookbook routines. This step is important for reproducibility and for assessing the robustness of the offline density-space AMOC reconstruction.
    - Added these details to the methods section (Lines 136-140).
- Line 142: "AMOC weakening refers to the reduction AMOC strength relative to the reference period." This should presumably read "the reduction in AMOC strength."
    - Corrected.
- Line 161: "a mean absolute r-value," not "an mean absolute r-value."
    - Corrected.
- Line 417: "variation to the speed and extent of southward propagation dependant on future forcing..." should be "variation in the speed and extent..." and "dependant" should likely be "dependent."
    - Corrected.
- Line 428-429: "As such when considering latitudinal connectivity AMOC decline the use of density-space is likely more informative..." is missing a preposition. Something like "when considering the latitudinal connectivity of AMOC decline" would be clearer.
    - Corrected.
- Line 431: "for individual models while AMOC weakening often is first onset in the SPG..." appear incorrect. It should be "while AMOC weakening often first begins in the SPG..." or "often has its onset in the SPG..."
    - Corrected.
- Line 469: "there are still models that still exhibit noticeable variability" repeats "still" unnecessarily.
    - In rewriting the discussion, the sentence has been removed.
- Line 473: "there is the question to how representative the threshold gradient is..." should be "the question of how representative..."
    - Improved the grammar of this sentence.
- Line 477: "It is therefore not achievable to disentangle the weakening signal from the internal variability for all models" is understandable but can be improved. "It is therefore not possible..." would be better suited.
    - In rewriting the discussion, this sentence has been removed.

- A general style point: the manuscript sometimes shifts between “AMOC decline,” “AMOC weakening,” “weakening signal,” and “propagation speed of weakening” in ways that are understandable but not always fully controlled. Some terminological tightening would help.
    - We have now adopted a preference to use “AMOC weakening” to refer to the decrease in AMOC strength and the use of “Weakening gradient” when relating to changes measured from the threshold lines. Reference to “Propagation” or a “Weakening signal” are only mentioned as an interpretation of the weakening gradients.